5th INTERNATIONAL INDUSTRIAL SIMULATION CONFERENCE

2007

ISC'2007

EDITED BY

Jaap Ottjes

and

Hans Veeke

JUNE 11-13, 2007

DELFT, THE NETHERLANDS

A Publication of EUROSIS-ETI

Cover Picture: The Geographer (1669) by Johannes Vermeer

5th Industrial Simulation Conference 2007

DELFT, THE NETHERLANDS

JUNE 11-13, 2007

Organised by

ETI- The European Technology Institute

Sponsored by

EUROSIS - The European Simulation Society

Co-Sponsored by

AEKI

CREAX

ENSAIT

Ghent University

Fern Universität Hagen

IPK Fraunhofer

UPV

Hosted by

Delft University of Technology

Delft, The Netherlands

EXECUTIVE EDITOR

PHILIPPE GERIL
(BELGIUM)

EDITORS

General Conference Chair

Professor Lodewijks, Delft University of Technology, Delft, The Netherlands

General Program Chairs

Hans Veeke, Delft University of Technology, Delft, The Netherlands
Jaap Ottjes, Delft University of Technology, Delft, The Netherlands

Track Chair Simulation in Electronics Manufacturing

Lars Moench
Fernuniversität Hagen
Hagen, Germany

Track Chair Workshop on Modelling and Simulation in the Textile Industry

Vladan Koncar
GEMTEX-ENSAIT
Roubaix, France

INTERNATIONAL PROGRAMME COMMITTEE

Simulation Methodology
Carmen Bobeanu, Ghent University, Ghent, Belgium
Olivier Dalle, INRIA, Sophia Antipolis, France
Juan de Lara, Universidad Autonoma de Madrid, Madrid, Spain
Helge Hagenauer, Universitaet Salzburg, Salzburg, Austria
Mauro Iacono, SUN II, Universita di Napoli, Naples, Italy
Bjorn Johansson, Chalmers University of Technology, Gothenburg, Sweden
Panagiotis Katsaros, Aristotle University of Thessaloniki, Thessaloniki, Greece
Pierre l'Ecuyer, Universite de Montreal, Montreal, Canada
Jan Lemeire, VUB, Brussels, Belgium
Lenin Lemus, UPV, Valencia, Spain
Aziz Naamane, ISIS-EPUM, Marseille, France
Jose Neves, Universidade do Minho, Braga, Portugal
Werner Pohlmann, Universitaet Salzburg, Salzburg, Austria
Jiri Safarik, University of West Bohemia, Plzen, Czech Republic
Frederic Serin, IUT du Havre, Le Havre, France
Jaroslav Sklenar, University of Malta, Msida, Malta

INTERNATIONAL PROGRAMME COMMITTEE

Simulation in Manufacturing
Andre Astolfi, Ecole Navale, Brest, France
Pascal Berruet, Universite Bretagne Sud, Lorient, France
Ana Camacho, UNED, Madrid, Spain
Luis Camarinha-Matos, Universidade Nova de Lisboa, Lisbon, Portugal
Remy Dupas, Universite d'Artois, Bethune, France
Alexander Felfernig, University of Klagensfurt, Klagensfurt, Austria
Michel Gourgand, Universite Blaise Pascal, Clermont-Ferrand, France
Bernard Grabot, LGP-ENIT, Tarbes, France
Brian Hollocks, Bournemouth University, Bournemouth, United Kingdom
Peter Lawrence, Monash University, Clayton, Australia
Habtom Mebrahtu, APU, Chelmsford, United Kingdom
Pascal Meyer, Forschungszentrum Karlsruhe, Karlsruhe, Germany
Fabrizio Micari, University of Palermo, Palermo, Italy
Henri Pierreval, IFMA, Aubiere Cedex, France
Marcos A.Rodrigues, Sheffield Hallam University, Sheffield, United Kingdom
Flavio Tonelli, University of Genova, Genova, Italy
Markus Vorderwinkler, PROFACTOR GmbH, Steyr, Austria

Simulation in Automotive Systems
Naoufel Cheichrouhou, EPFL, Lausanne, Switzerland
Jesus Felez, UPM, Madrid, Spain
Morteza Montazeri, Iran University of Science and Technology, Tehran, Iran
Rob Walker, Anglia Ruskin University, Chelmsford, United Kingdom

Simulation in Robotics
Leopoldo Arnesto Angel, UPV, Valencia, Spain
A. Chatzinikolaou, Athens, Greece
Andrzej Dzielinski, Warsaw University of Technollogy, Warsaw, Poland
George L.Kovacs, Hungarian Academy of Sciences, Budapest, Hungary
Terje Kristensen, NTNU, Trondheim, Norway
Martin Mellado, UPV, Valencia, Spain
Bogdan Raducanu, Computer Vision Centre, UAB, Barcelona, Spain
Angel Valera, UPV, Valencia, Spain

Simulation in Electronics, Computer and Telecommunications
Andrea Aquaviva, University of Urbino, Urbino, Italy
Kamel Barkaoui, CEDRIC-CNAM, Paris, France
Christos Bouras, University of Patras, Patras, Greece
Boguslav Butrylo, Bialystok Technical University, Bialystok, Poland
Christophe Claramunt, Ecole Navale, Brest, France
Tom Dhaene, University of Antwerp, Antwerp, Belgium
Domenico Giunta, ESA-ESTEC, Noordwijk, The Netherlands
Clemens Heitzinger, Purdue University, West Lafayette, USA
Paulo Maciel, University of Pernambuco, Recife, Brazil
Ana Pajares, UPV, Valencia, Spain
Marco Roccetti, University of Bologna, Bologna, Italy
Fernando Boronat Segui, UPV, Valencia, Spain
Renate Sitte, Griffith University, Gold Coast, Australia

Simulation in Electronics Manufacturing
Track Chair: Lars Moench, Fernuniversitat Hagen, Hagen, Germany
Theresa Roeder, San Francisco State University, USA
Oliver Rose, Dresden University of Technology, Germany
Gerald Weigert, Dresden University of Technology, Germany

INTERNATIONAL PROGRAMME COMMITTEE

Simulation in Logistics, Transport and Harbour Simulation
El-Houssaine Aghezzaf, Ghent University, Ghent, Belgium
Isabel Garcia Guttierez, Univ. Carlos III de Madrid, Madrid, Spain
Martin Grunow, Berlin University of Technology, Berlin, Germany
Daniel Krajzewicz, German Aerospace Centre, Berlin, Germany
Johannes Krauth, SDZ GmbH, Dortmund, Germany
Marie-Ange Manier, UTBM, Belfort, France
Jaap Ottjes, TU Delft, Delft, The Netherlands
Selwyn Piramithu, University of Florida, Gainesville, USA
Ahmad A.Sharieh, University of Jordan, Amman, Jordan
Rik van Landeghem, Ghent University, Ghent, Belgium
Hans Veeke, TU Delft, Delft, The Netherlands

Hospital Logistics Simulation
Track Chair: Joel Colloc, MODYS, University Lumière Lyon 2, Lyon, France
Co-Chair: Giorgio Romanin-Jacur, University of Padova, Vicenza, Italy
Francois Chapuis, SIS, University Lyon 1, Lyon, France
Claude Dussart, CESH, University Lyon 1, Lyon, France
Jose Machado, University of Minho, Braga, Portugal
Claude Petit, CESH, University Lyon 1, Lyon, France
Peter Summons, University of Newcastle, Australia

Complex Systems Modelling
Ivan Bruha, McMaster University, Canada
Frantisek Capkovic, Slovak Academy of Sciences, Bratislava, Slovak Republic
Werner Dubitzky, University of Ulster, Coleraine, United Kingdom
Alessandro Genco, Universita di Palermo, Palermo, Italy
Krzysztof Fujarewicz, Silesian Technical University, Gliwice, Poland
Eugene Kindler, Ostrava University, Ostrava, Czech Republic
Marielle Mokhtari, Defence R&D Canada, Valcantier, Canada
Alexandre Nketsa, Universite Paul Sabatier, Toulouse, France
Marc Parenthoen, CERV, Plouzane, France
Francesco Quaglia, University of Rome I, Rome, Italy
Filippo Queirolo, Hua Yi Plastic Products, Longhua Town, P.R.China
Miguel Rocha, University do Minho, Braga, Portugal

Simulation in Aerospace
Martin Adelantado, ONERA-CT, Toulouse, France
Wolfgang Kuehn, University of Wuppertal, Wuppertal, Germany
Martin Spieck, DLR, Goettingen, Germany
Zdravko Terze, University of Zagreb, Zagreb, Croatia
Tao Wan, University of Bradford, Bradford, United Kingdom

Simulation in Industrial and Product Design
Chiara Catalano, IMATI-CNR, Genoa, Italy v
Yan Luo, NIST, Gaithersburg, USA
Jorge Perez Mateos, Industria de Turbopropulsores SA, Madrid, Spain

Simulation in Engineering Processes
Chrissanti Angeli, Technological Institute of Piraeus, Athens, Greece
Blazej Balasz, Technical University of Koszalin, Koszalin, Poland
Jose Barata, Universidade Nova de Lisboa, Lisbon, Portugal
Fouad Riane, CREGI-FUCAM, Mons, Belgium
Christopher Schlick, RWTH Aachen, Aachen, Germany
Jan Studzinski, Polish Academy of Sciences, Warsaw, Poland
Joao Tavares, University of Porto, Porto, Portugal

Simulation in Energy and Power Systems
Saleh Al-Jufout, Tafila Technical University, Tafila, Jordan
Janos-Sebestyen Janosy, KFKI Atomic Energy Research Institute, Budapest, Hungary
Mohammad Montakhab, University of West England, Bristol, United Kingdom

INTERNATIONAL PROGRAMME COMMITTEE

Simulation in Multibody Systems
Track Chair: Bernd Schaefer, DLR, Wessling, Germany
Jose Diaz Lopez, Dynasim AB, Lund, Sweden

Simulation in Chemical, Petroleum and Mining Engineering
Diganta Das, University of Oxford, Oxford, United Kingdom
Klaus Hellgardt, Imperial College, London, United Kingdom
Vahid Nassehi, Loughborough University, Loughborough, United Kingdom
Mohamad R. Riazi, Kuwait University, Kuwait

Simulation in Military and Defense
Track Chair: Mike Letherwood, US Army RDECOM, Warren, USA
Co-Chair: Aleksander Kurec, US Army TACOM, Warren, USA
Carlos Palau, UPV, Valencia, Spain
Matthias Reuter, CUTEC Gmbh, TU-Clausthal, Clausthal, Germany

Virtual Reality and Graphical Simulation in Industrial Simulation
Track Chair: Guodong Shao, NIST, Gaithersburg, USA
Co-Chair: Kiran Mahajan, Heinz Nixdorf Institute, Paderborn, Germany
Anders Hast, University of Gavle, Gavle, Sweden
Fabrizio Lamberti, Politecnico di Torino, Turin, Italy
Sudhir Mudur, Concordia University, Montreal, Canada
Mike Sips, University of Konstanz, Konstanz, Germany

Verification, Validation and Accreditation
Track Chair: Alexei Lisounkin, Fraunhofer IPK, Berlin, Germany
Roberto Revetria, University of Genoa, Genoa, Italy
Dmitri Zaitsev, Odessa National Telecom University, Odessa, Ukraine

Simulation and Training
Manuel Alfonseca, Universidad Autonoma de Madrid, Madrid, Spain
Dorin-Mircea Popovici, Ovidius University, Constanta, Romania
Gerhard Schreck, Fraunhofer IPK, Berlin, Germany
Ahmed Bin Subaih, University of Sheffield, Sheffield, United Kingdom
Agustin Yague, Technical University of Madrid, Madrid, Spain

Workshops

Workshop on Modelling and Simulation in the Textile Industry
Track Chair: Vladan Koncar, ENSAIT, Roubaix, France
Peter Farber, Hochschule Niederrhein, Krefeld, Germany
Jiri Militky, Technical University Liberec, Liberec, Czech Republic
Catarina Rizzi, University of Bergamo, Bergamo, Italy
Ali Shams-Nateri, Guilan University, Rasht, Iran

Workshop on Intelligent Transport Systems
Paul Davidsson, Blekinge Institute of Technology, Ronneby, Sweden
Jairo Montoya Torres, Universidad de la Sabana, Chia, Columbia

NANOSIM
Javier Marin, ETSI, University of Malaga, Malaga, Spain

Workshop on Simulation in Lean Manufacturing
El-Houssaine Aghezzaf, Ghent University, Ghent, Belgium
Hendrik Van Landeghem, Ghent University, Ghent, Belgium

All articles published in these Proceedings have been peer reviewed

EUROSIS-ETI Publications are ISI-Thomson and INSPEC referenced

Selected papers of this conference are published in scientific journals.

For permission to publish a complete paper write EUROSIS, c/o Philippe Geril, ETI Executive Director, Ghent University, Faculty of Engineering, Dept. of Industrial Management, Technologiepark 903, Campus Ardoyen, B-9052 Ghent-Zwijnaarde, Belgium.

EUROSIS is a Division of ETI Bvba, The European Technology Institute, Torhoutsesteenweg 162, Box 4, B-8400 Ostend, Belgium

Printed in Belgium by Reproduct NV, Ghent, Belgium
Cover Design by Grafisch Bedrijf Lammaing, Ostend, Belgium

EUROSIS-ETI Publication
ISBN: 978-90-77381-34-2
EAN: 978-90-77381-34-2

INDUSTRIAL SIMULATION 2007

PREFACE

The benefits of modeling and simulation technology are nowadays well-known and documented. Forecasting the behavior of complex systems yields tangible benefits in industry, research and human and social services. The design of new systems is not complete without a detailed simulation of the new system.

This tendency puts great responsibility on the simulation society. Errors and mistakes in simulation may lead to failure of the system, to be built. Simulation is not only a technique, but requires a sound modeling theory in the first place. Especially the growing complexity of systems requires more than a nice visual interface. Distributed simulation and hardware-in-the-loop applications ask for a full cooperation between experts of different domains.

We at the Delft University of Technology and the department of Marine and Transport Technology in particular, emphasize this issue in our research. Therefore, we are excited to welcome scientists and experts from all parts of the world to Delft. The Industrial Simulation Conference 2007 (ISC-2007) is the fifth annual International Simulation Conference, which aims to give a status report on advances to the present day. It is a meeting place for researchers and practitioners in industry.

Delft is a wonderful city for science. The city is located between The Hague (the seat of government of The Netherlands) and Rotterdam, one of the biggest harbours of the world.

The university itself, is the oldest technical university of The Netherlands. In 1842, King Willem II founded the 'Royal Academy for the education of civilian engineers, for serving both nation and industry, and of apprentices for trade'. The Academy also educated civil servants for the colonies and revenue officers of the Dutch East Indies. An Act in 1863 imposed regulations on technical education as well as bringing it under the influence of the rules applying to secondary education. Then, in 1864, a Royal Decree was issued, ordering that the Royal Academy in Delft be disbanded in order to make way for a new 'Polytechnic School'. The School went on to educate architects, and engineers in the fields of civil works, shipbuilding, mechanical engineering and mining. Since 1905 the academic level of the School's technical education was acknowledged as it became a 'Technische Hogeschool', or an 'Institute of Technology'. The Institute was granted corporate rights by an Act in 1956. It was an Act which took effect in 1986, that officially transformed the Institute of Technology into Delft University of Technology, also known as 'TU Delft'.

On behalf of the International Programme Committee of ISC-2007, we would like to welcome you to the TU Delft. We are certain that this conference will be an event of inspiration, new ideas and contacts for you.

Gabriël Lodewijks, Jaap Ottjes, Hans Veeke
Conference chairs

CONTENTS

CONTENTS

COMPLEX SYSTEMS SIMULATION

COMPLEX SYSTEMS SIMULATION METHODOLOGY

SOFTWARE MODEL DEVELOPMENT

HUMAN BEHAVIOURAL MODELLING

MANUFACTURING SIMULATION

STEEL MANUFACTURING SIMULATION

CONTENTS

CONTENTS

MILITARY OPTIMIZATION AND RISK ASSESSMENT

TEXTILE SIMULATION

MODELLING OF TEXTILE STRUCTURES

SMART TEXTILES MODELLING AND DEVELOPMENT

CONTENTS

MODELLING FOR EVALUATION AND PREDICTION

LATE PAPER

SCIENTIFIC PROGRAMME

TUTORIAL

scMod/Sim.exe
For Supply Chain Modelling and Simulation

J. Manuel Feliz-Teixeira

March 2007

GEIN – Section of Management and Industrial Engineering
Faculty of Engineering of the University of Porto, PORTUGAL
Url: http://www.fe.up.pt/~feliz
Email: feliz@fe.up.pt

KEYWORDS: supply chain, modelling, simulation, flexibility, event scheduling, computer application.

ABSTRACT

In this paper we present an application dedicated to modelling and simulating Supply Chain systems (*scMod/Sim.exe*), developed with C++, with which a wide number of structures and policies can be analysed. The model construction is based on primitives resembling the real objects involved in these systems, like trucks, transport paths, retailers, factories, warehouses, stocks, etc. The result is a very simple environment for representing the Supply Chain in a wide range of situations. The simulator application, with its various tools for automatically representing the relevant metrics, gives the analyst an excellent opportunity for studying the behaviour of his/her Supply Chain, from the operational basis to the level of strategic decisions. In addition, the simulator includes a tool for helping to create the *rigidity matrix* of the Supply Chain (Feliz-Teixeira & Brito, 2004), with which the analyst can get an idea about how much his/her system will be *flexible* to demand variations.

1. Introduction

Developing a model commonly implies the representation of a system with the help of general purpose primitives, like the basic blocks of logic used in *Petri nets*, *System Dynamics*, UML, etc., or even the wide spread *state transitions diagrams*. Building a model therefore implies the appropriate ability to construct based on such blocks. Of course, today it is easy to agglomerate a certain number of these blocks and assign them the "self" of an "entity", using the properties of inheritance offered by the modern object-oriented languages, for example, but even so the modeller must have some knowledge about the basis behind these approaches, since it is frequently necessary to readjust such "entities" to the specific contours of a project. A substantial number of companies purchase generalist modelling tools with the intent of addressing specific problems, but later realise that it would also be necessary to contract a specialist on modelling, since the system to model reveals itself to be considerably more complex than expected. The usual destiny for such modelling systems is therefore the wastebasket. Exceptions exist, of course, among companies with a high capability for management and contracting, where groups of engineers are specifically dedicated to modelling and simulating the most relevant processes. This emphasis on the simulation as a matter for specialists seems an interesting policy, and for sure the results emerge accordingly.

2. About *scMod/Sim.exe*

The double application named *scMod/Sim.exe*, comprised of *scMod.exe* (modelling) and *scSim.exe* (simulation), aims to dissolve the need for special skills in building the Supply Chain model, while at the same time maintaining the need for reasonable skills in handling the simulation process. The

entities of the Supply Chain are represented as much as possible as they appear in "reality", thus, no special knowledge is needed to construct a model other than that of creating a replica of the system based on "scaled elements of the reality". This could perhaps be interesting for Supply Chain specialists who have few skills in modelling, but a good background in simulation.

In effect, each object is intended to be seen by the user as an "imitation" of an object existing in reality, with a certain functionality encapsulated, instead of exposed to the modeller. Thus, the level of abstraction required in the process of modelling comes closer to that of a planner. At this level, one does not need to define or represent the dynamics of the internal processes of the Supply Chain elements, since they are pre-defined in the respective objects. The user only has to tailor these generalised objects – named *Customer Supplier Units* (CSU) – in order to represent the particular Supply Chain with the accuracy needed. The next figure shows various CSUs prepared to be linked in a future Supply Chain network.

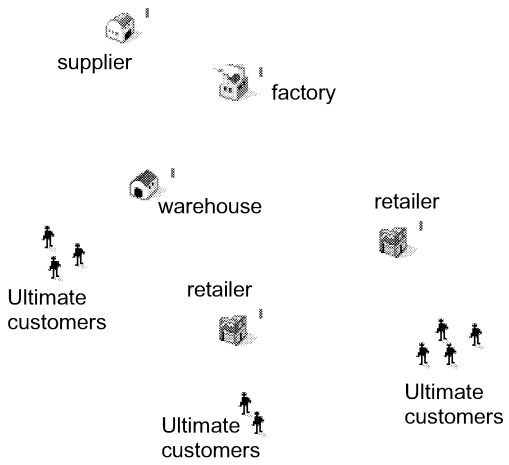


Fig. 1 Elements for creating the Supply Chain model.

This approach may also be considered a "data-driven" approach, since there is no need for programming in order to model or to simulate the system. The proposal also reflects the tendency of providing the industry with simulators for specific areas of knowledge, trying to imitate the natural specialisation observed in the business world. As there are companies specializing in airport projecting, urban traffic design, nuclear central building, etc., our point of view is that it would be advantageous to reproduce the same tendency in industrial simulation tools, instead of systematically developing models from the basis. This would provide the modelling community with an extensive and very useful "data base" of specific simulators, for specialist planning. A huge data base of highly reliable specific simulators is what we modestly intended to contribute to. More detail on the processes involved in this application can be found in Feliz-Teixeira & Brito (2003) and in Feliz-Teixeira (2006). A visit to the author's website could probably be of interest too:the site can be accessed at http://www.fe.up.pt/~feliz.

3. Modelling with *scMod.exe*

Here we will try to give an idea of how easy it is to create a model of a Supply Chain, even when it is to represent a complex network. This will be done with the aid of some images taken from *scMod.exe*, the application for modelling (Fig.2).

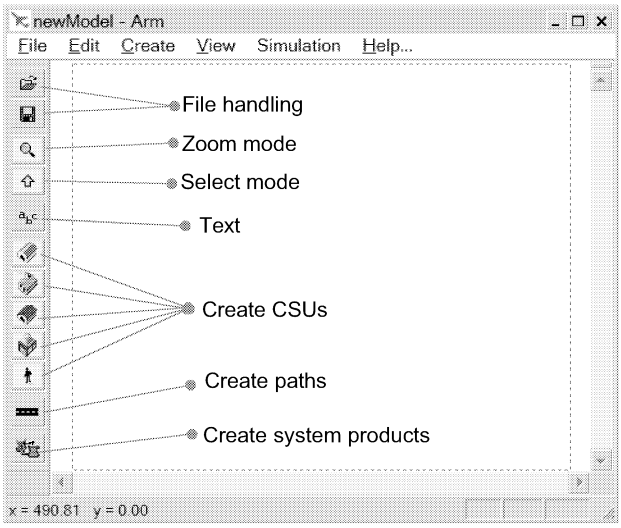


Fig. 2 A general view of the *scMod.exe* application.

The Supply Chain model will be represented in the application's window with the aid of the buttons of the main *toolbar*, which allows not only the creation of diverse Supply Chain facilities (CSUs), but also the products, sub-products and raw materials available to the system, and the transport paths for connecting these CSUs.

One can basically say that to build the Supply Chain network will be as simple as to place into the application's window the various Supply Chain facilities, that is, factories, warehouses, retailers, etc., and then connect them in the appropriate way. Notice that the dimensions of the "*operations theatre*", the space where the system lies, will be represented in scale to the reality. The user establishes these dimensions when a new model is

created. From then on, when the user moves the mouse around this area, he/she will see the dimensions (x, y) in kilometres displayed down at the application's *status bar*.

Next figure (Fig. 3) shows a new Supply Chain represented in the *scMod.exe* window, with 2 suppliers, 2 factories, 2 warehouses and 6 retailers. The retailers supply several customers, and these will be responsible for injecting the demand into the system. Notice that the facilities are connected via *transport paths*, with distances, maximum speeds, road charges, etc., previously configured by means of a property dialog box.

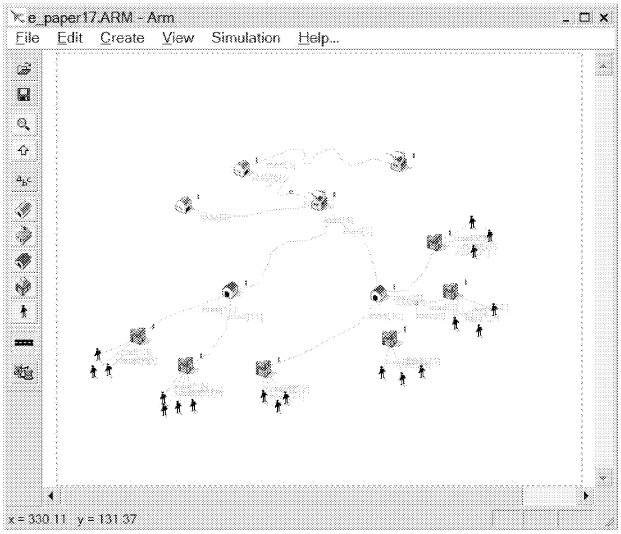


Fig. 3 A new Supply Chain represented in *scMod.exe*.

Once the Supply Chain network has been represented in the application's window, one must prepare for the implementation of the following steps: (1) define the products available to the system; (2) establish the type of transport paths between each pair of facilities; (3) Create the network nodes; (4) configure each facility to handle its particular products; (5) define the demand at each customer. In the next sections, we will try to explain how these steps can be completed.

As we see, after representing the network much more time will be consumed in carefully and properly configuring the various elements. And this task will get harder with the increase of the Supply Chain complexity and the number of products considered. In each facility (CSU), at least the *ordering policy* specification and the *supplier* specification are needed for each product handled in the facility.

Concerning the *transport paths*, they may also be configured for a specific distance, giving the modeller the ability to "force" certain paths to a convenient distance of travel. This is very useful when certain facilities are located close to each other, but others are not, as for instance, in certain complex Supply Chains including regional and trans-national links, as in the case of ZARA fashion clothes, for example, where some suppliers are located in China, while the main factory and a great number of retailers are located in the Iberian Peninsula.

3.1 *Products available to the system*

The products available to the system, or *system products*, can be created and accessed using the button at the bottom of the main toolbar: 

This option, which must be the very first option to use when configuring the Supply Chain, gives access to a *dialog box* (Fig. 4) where the user may name, characterise and link products, sub-products and raw materials, as well as establish among them a kind of *Materials Resource Planning* (MRP) tree, if necessary. The example shown in the figure is purely didactic.

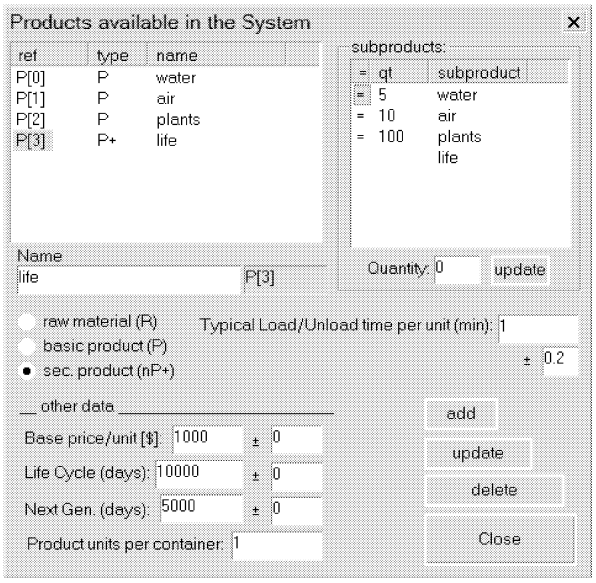


Fig. 4 Dialog box for creating system products.

Notice that these products may or may not be used by the CSUs, depending on if they are handled or not handled by its inventory. Notice also that in the example of the figure, the secondary product named "*life*", is defined as 5 "*water*" + 10 "*air*" + 100 "*plants*". During the simulation process, this information will be considered invaluable to good stock management.

3.2 *Type of transport paths*

After connecting the appropriate CSUs with *transport paths*, these may be configured in order to represent *roadways*, *railways*, *airways*, *seaways*, *information paths* (used to order by catalogue, by telephone, or by Internet), or *xTunnels*, a special kind of path for modelling pipeline-like flows. In addition, some other properties can be configured, like the exact length of the path, the maximum speed allowed, charges, type of charges, etc. Figure 5 shows the *dialog box* where all this can be done.

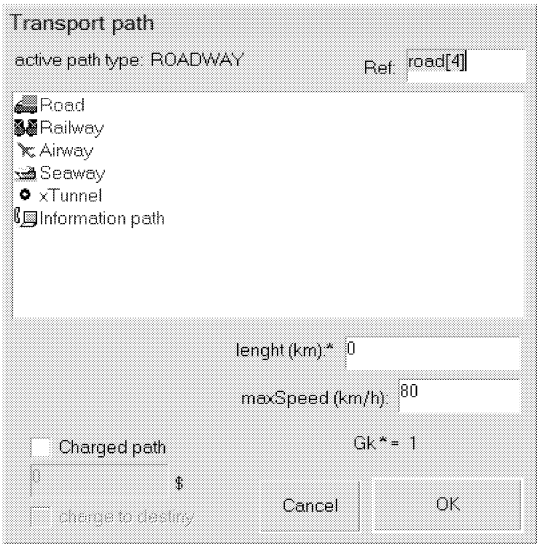


Fig. 5 Dialog box for creating system products.

3.3 *Creating the network nodes*

The network nodes, very important elements for the process of simulation, are automatically created in the application in response to the menu option "*Create->SupplyChainNodes*". This creates the nodes based on the proximity of the endpoints of the various paths, associates paths with nodes and nodes with paths, and establishes links between nodes and CSUs. This ensures that all the elements in the system will be correctly interconnected. Some nodes can then be configured to act as points of pause in the transport network, or even points where the material is transferred from a certain type of vehicle to a vehicle of another kind. This is very useful when modelling interfaces between normal roads and the railway, for example.

3.4 *Configure each facility to handle products*

Figure 6 shows the *dialog box* for configuring the products that the CSU will handle in its stock (or inventory), the maximum available space for those items, and their initial stocks. Notice that on the left list of this *dialog box* all the previously created system products are shown, while on the right side list appear only those particular products the CSU handles.

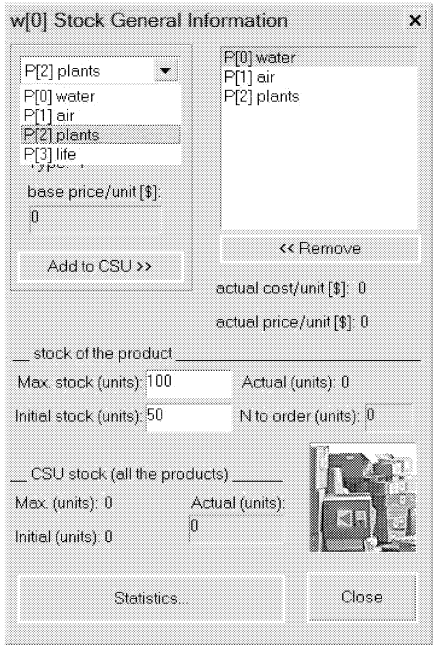


Fig. 6 Dialog box for configuring the CSU's stock.

The second most important task in the configuration of the CSU products has to do with the choice of the *ordering policy* to be used in each product, and the choice of its *supplier*. This is perhaps the part of the configuration process where the modeller must be more cautious, since small deviations from the appropriate representation of such policies will possibly lead the outputs to significantly diverge from reality. In effect, the choice and the configuration of the inventory model will definitely affect not only the internal dynamics of the CSU, but also the traffic on the transport paths and the pressure on suppliers. In a complex structure of CSUs, such mistakes of modelling may well produce chain reactions where from no reliable conclusions can be obtained. So, it is principally through this *dialog box* (Fig. 7) that each CSU is customized in order to behave the best possible in accordance to the reality.

Notice that for each product handled, there are basically six fields of data to be defined: (1) <u>*when to order*</u>, which identifies the method used in the time domain to define the moment for ordering from the supplier; (2) <u>*How much to order*</u>, which identifies the method used in the volume domain to compute the amount to order; (3) <u>*how to order*</u>,

specifying if the order is to be sent to the supplier *now* or if it can go at the end of the day with other orders; (4) *who to order from*, choosing a supplier for the product; (5) *cool time*, the lead time from the supplier that would be considered 100% appropriate; (6) *one order cost*, as the cost of ordering from the supplier.

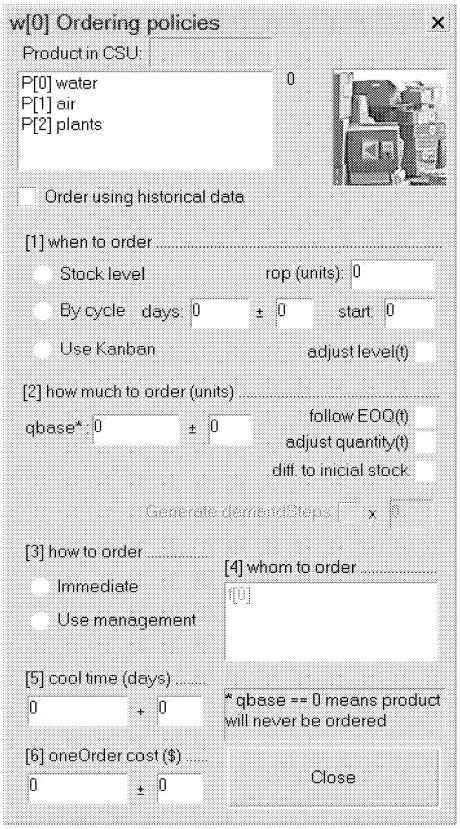


Fig. 7 Configuring the order policies, the supplier, etc.

A similar *dialog box* is used for *orders of production*, if the facility is to include a production process. Notice, however, that most of these parameters will not be taken into account when using historical data, imported from a text file.

3.5 *Defining de demand at each customer*

Since the customer (ultimate customer) is also a CSU, it will be configured in the same manner as any other CSU. Nevertheless, it will not handle any stock resources or production facilities, and it is only allowed either a cyclical ordering of a certain amount of product (see again figure 7, and the *by cycle* option), with the frequency given in days, or, instead, ordering based on historical data. The ultimate customer is, however, an element with which the demand at a certain facility can be made as complex as needed, using several customers to demand different amounts of product at different moments . Each *customer* is also able to generate demand steps, that may be used to help construct the *rigidity matrix* of the Supply Chain, a way for estimating the *flexibility* of the overall structure (see Feliz-Teixeira & Brito, 2004).

4. Simulating with *scSim.exe*

The simulator application (*scSim.exe*) looks very similar to the modeller application, apart from the main *toolbar* which now only gives access to some file handling options and to the configuration of the system products. In the menu of the application, the most important option is the "*Simulation*" option, which leads to the simulation start. Notice that the simulation may be set to be *deterministic*, instead of *stochastic*, if the option "*Simulation->UseRandom*" will not be checked. Notice also that a model must have been created previously with *scMod.exe* in order to be opened and simulated with *scSim.exe,* represented in figure 8.

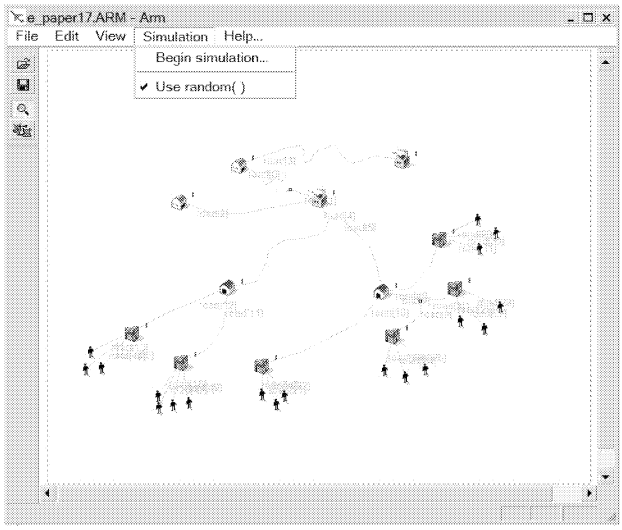


Fig. 8 A model in the *scSim.exe* application.

When the user chooses the option "*Simulation->BeginSimulation*", another toolbar appears at the bottom of the application's window, from where the simulation process will be controlled (Fig. 9). Most of the options in this toolbar require no explanation, since it resembles a TV controller. That is, once the *"start"* button is pressed, a sequence of actions is started, first testing the consistency of the model, and then connecting and preparing its objects for the simulation process. In the end, the process starts and movement can be seen in the application's window.

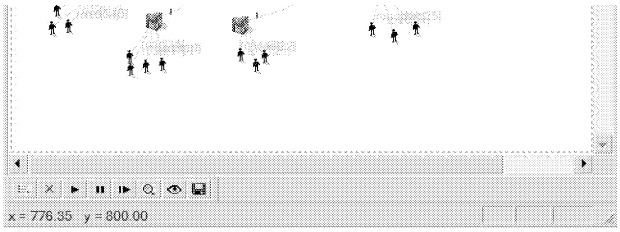


Fig. 9 The simulation toolbar.

During the simulation process, it is possible to inspect how the important Supply Chain indexes are performing, for example, by simply clicking on a CSU with the *right-mouse-button* and choosing the appropriate option from the menu list that would meanwhile appear (Fig. 10).

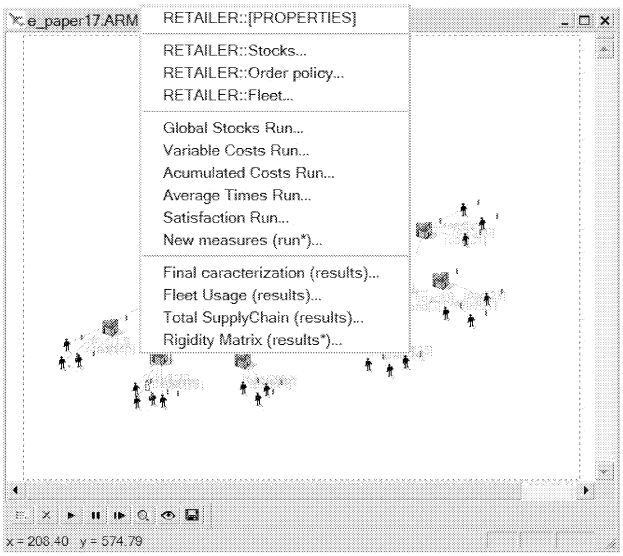


Fig. 10 Diverse parameters can be inspected during the simulation process, and the conditions reconfigured.

The next figure shows as an example on how the global inventory is performing at the facility *w[0]*.

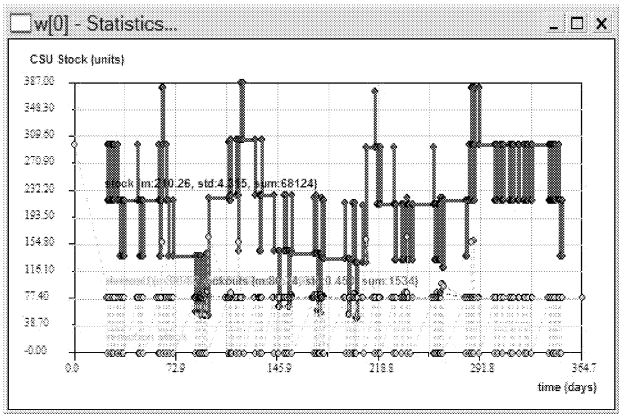


Fig. 11 A dynamic view of the inventory at *w[0]*.

In addition to these, several other graphics can be visualized even while the simulation is running, providing the analyst with a powerful tool for inspecting the behaviour of its system.

5. Metrics and outputs

The outputs of this simulator are firstly focused in a group of relevant operational measures, like the behaviour of the *global inventory* (all items), *variable costs*, *accumulated costs*, *average times*, *satisfaction with suppliers*, etc., with which the analyst may easily estimate the performance of each facility. Then, there is a second group of metrics which are computed from these, raising the analysis to a certain *operational-tactical* level, like the standard measures of *turnover*, *service level*, *stockout ratio*, among others (see Fig. 12). Finally, a tool for helping to measure the *flexibility* of the overall Supply Chain is included, allowing the studying of the network structure as a whole, an important instrument for strategists.

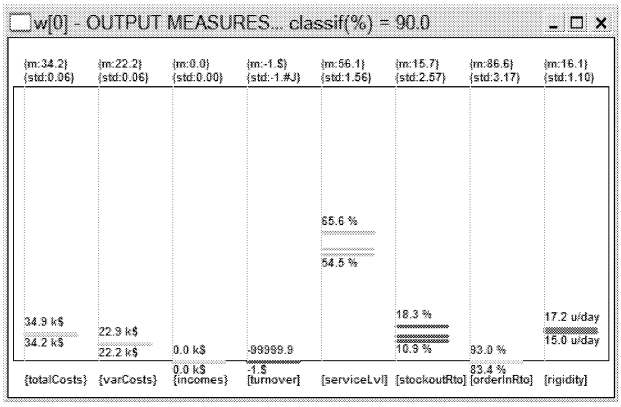


Fig. 12 Some standard measures[1] related to *w[0]*.

6. The rigidity/flexibility matrix

As told before, the simulator includes the ability of helping to compute the *rigidity* of the entire Supply network in the form of a matrix (Fig.13).

	<NO STEPS>	r[0]	r[1]	r[2]	r[3]	r[4]	r[5]
s[0]	10.9 ± 4.3	f 00	f 01	f 02	f 03	f 04	f 05
s[1]	16.9 ± 7.1	f 10	f 11	f 12	f 13	f 14	f 15
f[2]	85.9 ± 29.5	f 20	f 21	f 22	f 23	f 24	f 25
w[1]	45.0 ± 6.6	f 30	f 31	f 32	f 33	f 34	f 35
w[0]	16.1 ± 2.2	f 40	f 41	f 42	f 43	f 44	f 45
r[0]	7.5 ± 0.5	f 50	f 51	f 52	f 53	f 54	f 55
r[1]	9.0 ± 3.0	f 60	f 61	f 62	f 63	f 64	f 65
r[2]	7.2 ± 0.8	f 70	f 71	f 72	f 73	f 74	f 75
r[3]	8.9 ± 1.4	f 80	f 81	f 82	f 83	f 84	f 85
r[4]	7.7 ± 0.4	f 90	f 91	f 92	f 93	f 94	f 95
r[5]	8.9 ± 1.4	f 100	f 101	f 102	f 103	f 104	f 105

Fig. 13 Matrix of the "residual" *rigidity* of the system.

This matrix is computed by means of generating

[1] In this case the *turnover* was not computed, since no base prices were given to the products.

steps of demand at each customer, which introduce a perturbation in the system, and then this perturbation is measured along all the network nodes. At the end of this process, a matrix can be drawn with the results. Notice, however, that only the "residual" rigidity of the system has been measured in the present case, that is, the rigidity exhibited by the facilities without any abnormal perturbation (no-steps). Anyway, from this matrix we may already conclude that the worst rigidity observed was at the factory *f[2]*, exhibiting around 85.9 units/day of stock imbalance.

7. Conclusions

The modelling/simulation application presented in this article, specifically devoted to Supply Chain, resulted in a powerful tool for analysis based on an intuitive ambience of interaction with the user. Several unusual tools are included, like the *demand steps* for helping computing the *flexibility* of the overall structure in terms of a *rigidity matrix*, transport paths for modelling *pipeline* type flow of materials, facility related metrics, etc.

The application, made of a modeller *scMod.exe* and a simulator *scSim.exe*, was developed by the author in C++, and is considered a *beta* version academic software. The simulator *scSim.exe*, however, is available as free software, and can be used for directly running and testing models previously created with *scMod.exe*. We invite interested people to try to simulate their own Supply Chain models in this way.

Authors Biography:

J. Manuel Feliz-Teixeira graduated in Physics in the Faculty of Sciences of University of Porto, Portugal, and received an MSc and PhD from the Faculty of Engineering of the same university. His work has been related to various matters, from optical communications, solar energy and seismology to, more recently, the simulation of complex systems in management science, like warehouse and supply chain. His PhD thesis is on "Flexible Supply Chain Simulation".

References:

Feliz-Teixeira, J. M. (2006). *Flexible Supply Chain Simulation (published thesis)*. Porto, Portugal: Publindustria Publisher.

Feliz-Teixeira, J. M., & Brito, A. E. S. C. (2003). *An Approach for Dynamic Supply Chain Modelling.* Paper presented at the 2003 European Simulation and Modelling Conference, Naples, Italy.

Feliz-Teixeira, J. M., & Brito, A. E. S. C. (2004). *On Measuring the Supply Chain Flexibility.* Paper presented at the 2004 European Simulation and Modelling Conference, UNESCO, Paris, France.

METHODOLOGY

META-KNOWLEDGE SIMULATION

HARD AND SOFT APPROACHES IN A SIMULATION META-METHODOLOGY

László Muka
Elassys Consulting Ltd.
Bég utca 3-5.
H-1026 Budapest, Hungary
e-mail: muka@t-online.hu

Gábor Lencse
Department of Telecommunications
Budapest University of Technology and Economics
Magyar tudósok körútja 2.
H-1117 Budapest, Hungary
e-mail: lencse@hit.bme.hu

KEYWORDS

simulation methodology, simulation meta-methodology, hard method, soft method, collaborative modelling, organisational information and communications systems, business process system

ABSTRACT

A simulation project for Information and Communication Technology (ICT) system design and Business Process (BP) design in an organisation usually starts with an unstructured problem situation. We have already outlined a simulation meta-methodology addressing the situation. Now, we give a detailed description of the proposed simulation meta-methodology dealing with both the hard and the soft aspects of the problem. We describe the detailed process and functioning of a hard six-step simulation methodology which is an integrated part of our simulation meta-methodology.

We define the role of the traditional SSM (Soft Systems Methodology) in the simulation meta-methodology, and also show how our previously proposed Soft Systems Methodology with Modified Conceptual Models (SSM-MCM) can function as a bridge between hard and soft approaches. There is also introduced a new fast simulation approach, based only on the Traffic Flow Analysis and the Entity Flow-phase Analysis.
We also examine the functioning of the simulation meta-methodology in a collaborative environment, which is a frequent situation.

INTRODUCTION

A simulation project for Information and Communication Technology (ICT) system design and Business Process (BP) design in an organisation usually starts with an unstructured problem situation.

In a situation like this, there is a need for using a methodology which is able to deal with both the unstructured, soft aspects and the hard aspects of the problem, thus this methodology has to integrate the soft and hard approaches.

In (Muka and Lencse 2006) we outlined a simulation meta-methodology addressing these requirements: we proposed a simulation meta-methodology supporting problem-structuring and effective goal definition, also increasing efficiency by precise localization of systems to be modelled and by supporting decisions on the use of parallel simulation helping in speeding up the simulation.

In (Muka and Lencse 2006) we introduced new concepts to SSM (Soft Systems Methodology, Checkland 1985, 1989), hardening-up the traditional SSM ((Gregory 1993) and (Jackson and Keys 1984)). Our new concepts, Soft Systems Methodology with Modified Conceptual Models (SSM-MCM), help in the analysis of the organisational information systems and also give support to timing decisions in modelling.

In this paper we give an exact definition of the elements and cycles of the outlined simulation meta-mehodology.

We also describe a new approach of the preliminary modelling step, a method, using only TFA (Traffic Flow Analysis) and EFA (Entity Flow-phase Analysis) for this purpose.

In this paper we also introduce how SSM-MCM functions as a bridge between soft and hard approaches and we exactly define the functions of classic SSM in the process of the simulation meta-methodology.

In (Muka and Lencse 2006) we proposed a six-step hard simulation methodology to the simulation meta-methodology.
Now, we give a detailed description of the proposed hard methodology, introducing the working process and describing the functions of the hard simulation methodology.

Finally, in this paper we also examine the simulation meta-methodology process in a collaborative environment, which can be a usual situation, by comparing meta-methodology to a multi methodology, using the approach described in (Sierhuis and Selvin 1996) and (Sierhuis and Clancey 2002) as a starting point.

DETAILS OF THE SIMULATION META-METHODOLOGY

In this paper we examine how the simulation meta-methodology works in detail.

First, we introduce the hard and soft approaches used in building the meta-methodology.

Hard Approach: Detailed Formulation of the Proposed Six-step Simulation Methodology

Traditionally, *simulation* is considered to be a *hard systems approach* as a typical *predict and control method* (Jackson 1991).

As an element of the *simulation meta-methodology* (MM) we describe in detail a hard *simulation methodology* (SM) comprising six steps.

The six-step process of SM is shown in Figure 1.

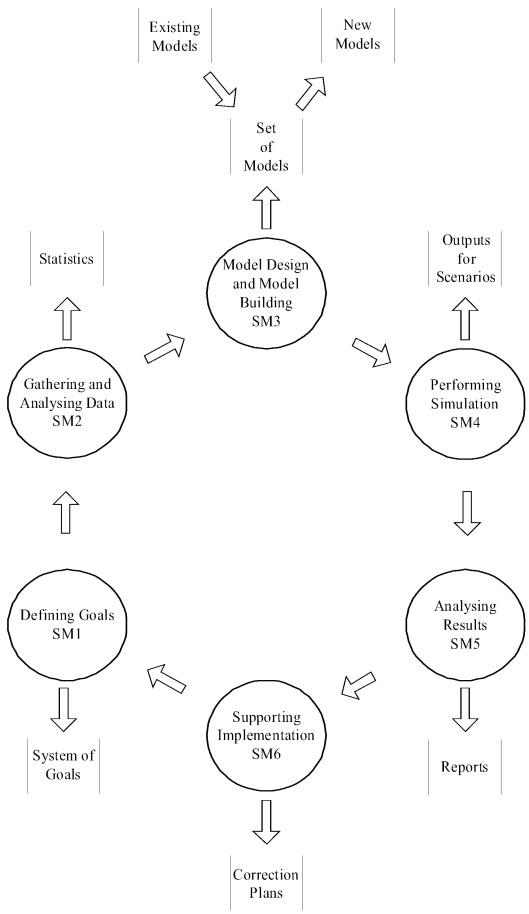


Figure 1.The Six-Step Process of the Proposed Classic Simulation Methodology

This is an iterative approach where an *implementation support step is closing the loop*. SM can be applied to simulate both P and IT elements of an organisational information system (Muka and Lencse 2006).

SM1 Defining Goals: - including preliminary design of models

At this step we define the goals of the simulation: the features to be simulated, the resolution and interval of simulation.

A preliminary model design (which should be a simplified and fast model design and simulation) serves to help successful goal definition by giving a general view about the model and the simulation. The result is a *System of Goals* containing also the relationships of goal elements.

SM2 Gathering and Analyzing Data:

Here all the data necessary for simulation should possibly be collected: data about the "as-is" state and the states *"before-the-as-is"* of the systems to be simulated for a reasonable time interval and scope. It should be a thorough *data mining* activity: we should try to find data about the normal and the extreme modes of operation, and if there is any information about disasters and disaster recovery situations.

The preliminary simulation model mentioned in SM1 may help to find the right data and the correct limits to data gathering.

Here we should find all the input data for simulation, data about the model intended to be built and its loads.

The collected data should be analyzed and the *Statistics* produced to help the necessary understanding of the systems.

SM3 Model Design and Model Building:

This is a tool-specific step: in many simulation environments (like ARIS, System Architect, Brahms, etc.) there can be a separate step of designing a formal, *static* model. In other environments design and building of simulation models is performed in one step. It is the direct *dynamic* approach (see Simprocess, ImiNet, ImiFlow, etc.).

According to the defined goals usually a *Set of Models* is designed and built. To improve efficiency, *Existing Models* can be used but usually it is also necessary to design and build *New Models*.

SM4 Performing simulation:

This is also a tool-specific step and closely related to SM3. At this step, run-analyze-debug cycles are performed and as a result we will have model-runs of validated and verified models.

According to SM1 a set of "as-is" and "what-if" scenarios are examined and the required *Output for Scenarios* is produced.

In a hard approach, the decision about parallel simulation is usually made at this step: the only criterion to use parallel simulation is the software-hardware capacity limit sufficient to perform the simulation with the required features.

SM5 Analyzing Results:

At this step, a *data mining* activity is also required because of the usually large quantity of output data. It may happen that we have got the necessary results but we cannot find them. The other important requirement is the *visualization* of the results (In visualization, the *animation* feature of our tool may have importance.). Using data mining and visualization in the analysis we produce *Reports* about the simulation for the defined users of the results.

SM6 Supporting Implementation:

This is an unusual step in the methodology but it may be useful to consider the implementation aspects immediately. It may reveal the necessity of producing *Correction Plans*, because there can be different alternatives in the steps and timing of implementation. It may be useful to perform SM6 in a preliminary way and include it into *Reports* produced at step SM5.

Soft Approach: Two Aspects of Using SSM in the Simulation Meta-Methodology

The classic SSM is a soft approach. In MM two different aspects of SSM are important:

1. SSM as a learning method
2. SSM as a method for information system analysis and design

1. SSM as a learning method plays multiple roles in MM:
   a. It is a method of facilitation in the team work
   b. It is a method of enquiry
   c. It is a method of consensus building

2. The modified SSM, that is SSM-MCM (Soft Systems Methodology with Modified Conceptual Models (Muka and Lencse 2006)), is a soft method that has been *hardened up* (Jackson and Keys 1984) to support information system analysis and design.

The Simulation Meta-Methodology: Elements and Cycles

In the following, we describe the elements and cycles of the MM using the concept of combined hard and soft approaches introduced in (Muka and Lencse 2006).

(The elements and cycles in working process of MM are shown in Figure 2.)

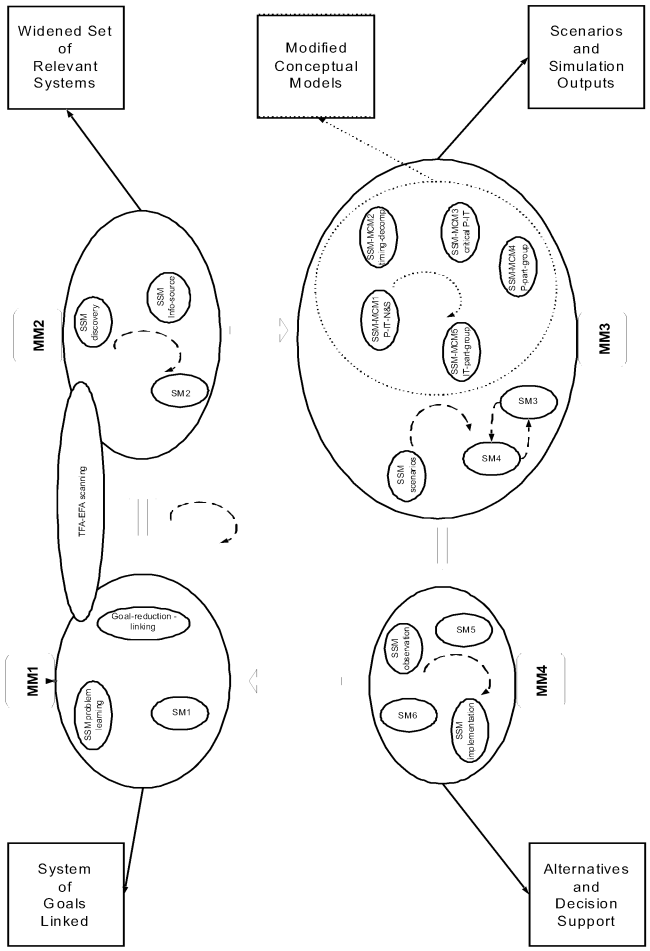


Figure 2. Elements and Cycles of the Meta-Methodology

Main points of MM are as follows:
MM1: Goal Definition
MM2: Identification of a Widened Set of Relevant Systems
MM3: Development of Conceptual Models and Simulation
MM4: Analysis and Support for Implementation

Now let us consider the detailed work of MM:

MM1 consists of the following elements: SSM problem learning, Goal-reduction-linking, SM1 and TFA-EFA scanning.

SSM problem learning: a classic SSM is used for learning the situation, learning the goals and their relationships for the derivation of requirements for simulation models.
Goal-reduction-linking: this is a hard element, a method of getting the simulation project goals from the organisational goals. The basic idea of this method is the goal reduction (Koubarakis and Plexousakis 1999) and linking the goals to the processes to be simulated and to the process of the simulation project itself.
SM1: Defining goals
TFA-EFA scanning:
By performing this methodological element, we execute preliminary simulation. At this step, we do not want to get exact final results but rather to see the quality and scope of the results, to introduce the simulation method, to get information about the necessary inputs, and to get information about the future models and simulation, in order to help goal definition, and to make the objectives clear.

For this purpose *fast full MM cycles* are to be used that is fast cycles including MM1-MM4.

To perform fast simulation in this phase, methods like EFA and TFA may be efficient (Lencse and Muka 2006a, 2006b). (As an additional advantage, models constructed at this step may be used later in the methodology at phase MM3.)
(For further research it may be fruitful to examine the use of the Statistical Synchronization Method (Pongor 1992) (Lencse 1999) as a fast simulation method at this step of MM, because of its minimized communication needs between model segment.)

MM2 contains: SSM discovery, SSM data source, TFA-EFA scanning, SM2.

SSM discovery: This is a classic learning SSM, whose purpose is to identify all the systems possibly influenced by the simulation project, to gather and analyze informal data too, to identify the widened set of relevant systems.
The widened set of relevant systems consists of systems that are sources of data for simulation, systems that can be probable users of simulation results, and systems probably to be simulated.
SSM info-source:
This is a classic learning SSM used in order to identify formal and informal data sources and also to identify typical and critical data configurations.

The method supports identification of typical and critical data configurations for all relevant systems. These data configurations should be identified for the simulated time interval, or for a reasonably longer interval having influence on the simulation.
TFA-EFA scanning: The purpose of using the method is to help data source identification and to find the users and potential users of data in order to maximize effectiveness and efficiency of simulation project.
SM2: Gathering and Analyzing Data – It is a method of gathering and analyzing *formal* data.

MM3 consists of elements: SSM scenarios, SSM-MCM1 P-IT-N&S, SSM-MCM2 timing-decomp, SSM-MCM3 critical P-IT, SSM-MCM4 P-part-group, SSM-MCM5 IT-part-group, SM3, SM4.

SSM scenarios: This is a learning SSM to identify simulated scenarios and also to verify and validate simulation models.
SSM-MCM(1-5): These SSM methods are using MCM (Modified Conceptual Models) introduced in (Muka and Lencse 2006).

SSM-MCM1 P-IT-N&S: to identify P and IT subsystems, and elements to N&S conditions
SSM-MCM2 timing-decomp: to define time relations in models, to synchronize models and to make time decomposition
SSM-MCM3 critical P-IT: to define critical P and IT elements to be simulated
SSM-MCM4 P-part-group: to make decisions on partitioning and grouping of P elements for *parallel simulation*
SSM-MCM5 IT-part-group: to make decisions on partitioning and grouping of IT elements for *parallel simulation*

Considering the application of parallel discrete-event simulation (PDES) we make decisions about grouping or partitioning the elements and also (already having the information about critical elements) we may have considerations about using DES-P DES-IT, TFA and EFA and also about using the combination and interworking approach in parallel simulation (Lencse 2004), (Lencse and Muka 2006a), (Muka and Lencse 2006).
If a subsystem seems to be too large we may try to use expansion for partitioning (Muka and Lencse 2006). After PDES decisions the simulation elements SM3 and SM4 are completed. At performing SM4 it may be necessary to reuse the SSM scenarios method.
SM3: Model Design and Model Building
SM4: Performing Simulation

MM4 has elements: SSM observation, SM5, SM6, SSM implementation.

SSM observation: This is a learning SSM application helping in the evaluation of the simulation results we got from performing SM5. Based on formal information of the analysis in SM5 and on informal assessment of SSM, alternatives for decisions can be made.
SM5: Analyzing Results
SM6: Supporting Implementation
SSM implementation: This is a learning SSM application which is used in the evaluation of implementation features revealed by SM6. To complete a successful simulation project, it is advisable to close the methodological loop and to think about implementation aspects as a decision support element.

The Simulation Meta-Methodology and the Organisational World

In Figure 3. we show the simulation meta-methodology with its elements and their relationship in the Organisational World, in the world where the simulation takes place and where the meta-methodology should work.
The organisational world is divided into two parts: the Hard Thinking World and the Soft Thinking World.

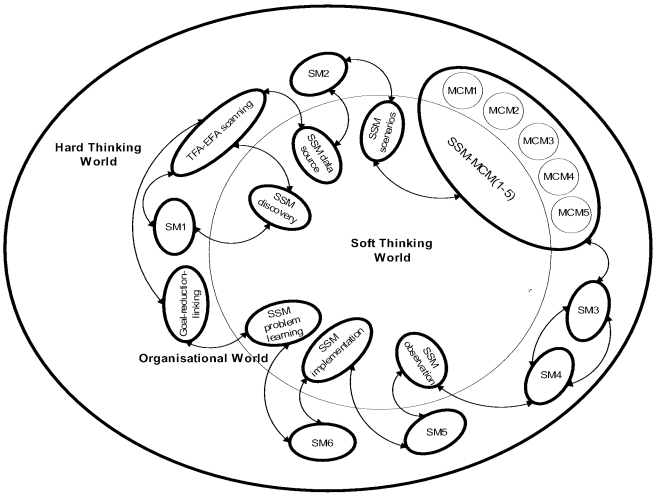


Figure 3. Hard and Soft Thinking in the Process of the Meta-Methodology

The conclusions may be the following:

- the classic SSM works in Soft Thinking World
- the SM woks in the Hard Thinking World
- the SSM-MCM works "between the worlds"
- the MM cycle starts and ends in Soft Thinking World
- usual soft and hard methods have only logical connections: after completing their cycle we go on to use the next method
- in the case of SSM-MCM hard steps are introduced into the soft cycle therefore it operates in a transient way.
- the starting and finishing method (SSM problem learning) is in the Soft Thinking World

ANALYSING SIMULATION META-METHODOLOGY AS A COLLABORATIVE APPROACH

A description of a *framework for collaborative modelling and simulation* was given in (Sierhuis and Selvin 1996) and (Sierhuis and Clancey 2002). The *collaborative modelling and simulation process* is viewed as a *holon* in the sense of

Soft Systems Methodology. (*SSM type holon* is a particular kind of holon, a Human Activity System (HAS), that is a set of activities connected to make a purposeful whole (Checkland and Scholes 1990).) In this approach there are two design teams - a soft thinking team and a hard thinking team – and there is a *multi-methodology* that consists of methodologies M1-M4. (Any of the methodologies are also holons in this framework.)

Now we will summarize the features of the methodologies in the *multi-methodology*:
M1 is a methodology to perceive a system and to construct soft static models. This is a structured methodology to observe and analyze the "as-is" state of the system (or problematic situation) with the goal to gather data describing the system.
M2 is a methodology to design simulation models.
This is a methodology to co-design formal static models, to translate soft models into formal models.
M3 is a methodology to create and to run simulation models. This is a tool-specific methodology to implement, to debug, and to validate dynamic simulation models and to run the simulator with the formal model as input, and the dynamic performance (behaviour) of the simulated system as an output.
M4 is a methodology to observe the simulation output. The purpose of this methodology is to observe and to investigate simulation results by comparing the results with the existing system in order to create shared understanding of the system.

For our analysis, we ***modify*** the *collaborative modelling and simulation process* described in (Sierhuis and Selvin 1996) and (Sierhuis and Clancey 2002).

The modifications (shown in Figure 4.) are summarized in the points below:
1. The analyzed HAS is an organisational information system with P and IT elements in it. (The *Organisational Systems* with P and IT elements that are in the focus of our analysis may also be viewed as holons.)
2. The observed *Simulation Output* shows the dynamic behaviour of the organisational information system.
3. The *Soft Team* may involve *Decision Makers*.
4. The *Hard Team* may contain two collaborating design sub-teams: a *P Team* and an *IT Team*.
5. In the Soft Team and also in the Hard Team, there is a modeller familiar with the meta-methodology and having practice with SSM-MCM.
6. A new methodology *M23* is introduced. M23 is a methodology of controlling the collaboration of P Team and IT Team in the parallel and co-design of the set of P and IT simulation models.

We examine the following questions:
- What is the relationship between meta-methodology and the methodologies of the multi-methodology?
- How does the meta-methodology meet the requirements of the collaborative approach?

For the analysis, we compare the elements of the multi-methodology and the meta-methodology in the modified *collaborative modelling and simulation process*:
Meta-Methodology and M1:
M1 is covered by MM1, MM2 and MM3 phases of our meta-methodology. Here the resolution of the meta-methodology is higher: In the case of MM1 the goal setting is explicit, while in the case of M1 the goal setting is only implicit (the goal is to model and to simulate human activity systems in a way of building useful models). In MM2 the relevant data sources are selected carefully, but in a wider manner: not only about the "as-is" state (for example it may be decided to collect data about state "before-the-as-is", and also about the "to-be" state, etc.) and both formal and informal data may be gathered.
MM3 and its steps MCM1-MCM5 (in the part performed by the soft team) will help to eliminate the methodological gap between M1 and M2, between the soft (informal) and hard (formal) approaches, respectively.
Meta-Methodology and M2:
M2 is covered by SM3 of MM3 and by the part of MCM1-MCM5 performed by both sub-teams (P-team and IT-team) of the hard-team. MM3 will help to increase the efficiency of the co-design performed by the collaborative teams.
Meta-Methodology and M3:
M3 is covered by SM4 of MM3, but in our approach it should be a more compound step, it is about the creation of a set of models by hard-teams: P-models, IT-models, according to the scenarios derived from the system of goals.
Meta-Methodology and M4:
M4 is covered by MM4, with steps SM5 and SM6, but in the meta-methodology the simulation outputs are not only observed, but the results of the observation are to be presented to the decision makers and the question of implementation is also discussed.
Meta-Methodology and M23:
In the meta-methodology, M23 is supported by MCM1-MCM5 of MM3 fulfilling the requirements of the collaborative approach.

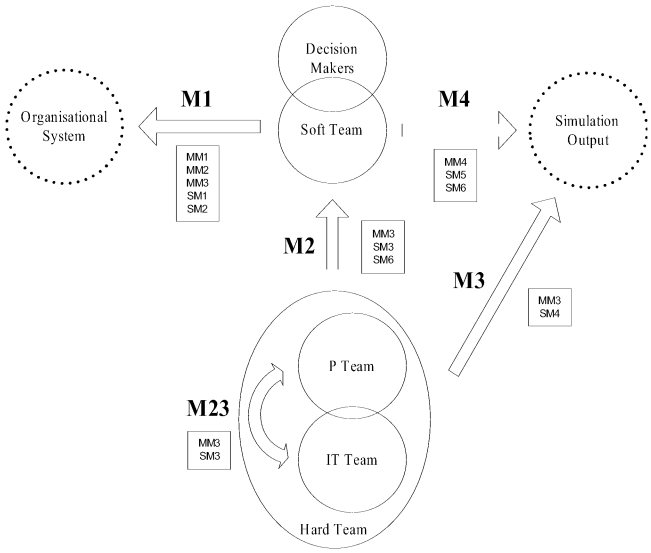


Figure 4. Collaborative Simulation Model Development Using the Meta-Methodology

In Figure 4. we show the methodologies of the multi-methodology and the meta-methodology elements realizing the function of the given methodology.

CONCLUSIONS

In this paper, we have given a detailed description of our previously outlined simulation meta-methodology, defining the full set of methods and the process of the simulation meta-methodology.
We have given an exact description of our hard six-step simulation methodology. We have defined the main output of the methodology at each step.
We have also defined the role of traditional SSM in the meta-methodology.
We have proposed a new approach, to use only TFA-EFA for preliminary modelling in the meta-methodology.
We have examined the relationship of hard and soft methods of the meta-methodology in the organisational world.
We have described the use of the simulation meta-methodology in a collaborative environment.

REFERENCES

Jackson, M.C., Keys, P. 1984 *Towards a System of Systems Methodologies* J. Opl. Res. Soc. Vol. 35, No. 6.

Jackson, M.C., 1991. *Systems Methodology for the Management Sciences* Plenum Press, New York, London

Checkland, P. 1985 *Achieving "Desirable ad Feasible" Change: An Application of Soft Systems Methodology* J. Opl. Res. Soc. Vol. 36, No. 9.

Checkland, P. 1989 *Soft systems methodology* In Rational Analysis for a Problematic World, Edited by J. Rosenhead, John Wiley & Sons Ltd

Checkland, P., Scholes, J., 1990. *Soft Systems Methodology in Action* JohnWiley & Sons Ltd., Chichester UK

Gregory, F. 1993. *Cause, Effect, Efficiency and Soft Systems Models* J. Opl. Res. Soc. Vol. 44, No. 4.

Koubarakis, M., Plexousakis, D. 1999 *Business process modelling and design – a formal model and methodology* BT Technol. J. Vol. 17, No. 4.

Sierhuis, M., Selvin, A.M. 1996. *Towards a Framework for Collaborative Modeling and* Simulation, Workshop on Strategies for Collaborative Modeling & Simulation, CSCW '96, Boston, MA

Sierhuis, M., Clancey, W.J. 2002. *Modeling and Simulating Work Practice: A Method for Work System De*sign IEEE Intelligent Systems, September/October 2002, Vol. 17, No. 5. pp.32-41.

Elassys Consulting Ltd. 2007 *Iminet and Imiflow Systems* http://www.elassys.hu

Lencse, G. 1999. *Applicability Criteria of the Statistical Synchronization Method* Proceedings of Communication Networks and Distributed Systems Conference (CNDS'99), (San Francisco, CA, USA, January 17-20.) SCS, 159-164

Lencse, G.: 2004. *Combination and Interworking of Traffic-Flow Analysis and Event-Driven Discrete Event Simulation* Proceedings of the 2004 European Simulation and Modelling Conference (ESM®'2004) (Paris, France, Oct. 25-27. 2004.) EUROSIS-ETI, 89-93.

Lencse, G. 2005. *Speeding up the Performance Analysis of Communication Systems* Proceedings of the 2005 European Simulation and Modelling Conference (ESM®'2005) (Porto, Portugal, Oct. 24-26.) EUROSIS-ETI, 329-333.

Lencse, G; Muka, L., 2006a *Convergence of the Key Algorithm of Traffic-Flow Analysis* Journal of Computing and Information Technology, Vol. 14, No 2, June 2006, 133-139.

Lencse, G., Muka, L. 2006b *Expanded Scope of Traffic-Flow Analysis: Entity Flow-Phase Analysis for Rapid Performance Evaluation of Enterprise Process Systems* Proceedings of the 2006 European Simulation and Modelling Conference (ESM'2006) (Toulouse, France, 2006. Oct. 23-25.) EUROSIS-ETI, 94-98.

Muka, L., Lencse, G., 2006. *Developing a Meta-Methodology Supporting the Application of Parallel Simulation* Proceedings of the 2006 European Simulation and Modelling Conference (ESM'2006) (Toulouse, France, 2006. Oct. 23-25.) EUROSIS-ETI, 117-121.

Pongor, Gy., 1992. *Statistical Synchronization: a Different Approach of Parallel Discrete Event Simulation* Proceedings of the 1992 European Simulation Symposium (ESS'92) (Nov. 5-8, 1992, The Blockhaus, Dresden, Germany.) SCS Europe, 125-129

BIOGRAPHIES

GÁBOR LENCSE received his M.Sc. in electrical engineering and computer systems from the Technical University of Budapest in 1994 and his Ph.D. in 2000. The area of his research is (parallel) discrete-event simulation methodology. He is interested in the acceleration of the simulation of communication systems. Since 1997, he works for the Széchenyi István University in Győr. He teaches computer networks and networking protocols. Now, he is an Associate Professor. He does R&D in the field of the simulation of communication systems for the Elassys Consulting Ltd since 1998.
He is a member of IEEE Hungary Section, IEEE Computer Society and IEEE Communications Society.

LÁSZLÓ MUKA graduated in electrical engineering at the Technical University of Lvov in 1976. He got his special engineering degree in digital electronics at the Technical University of Budapest in 1981, and became a university level doctor in architectures of CAD systems in 1987. Dr Muka finished an MBA at Brunel University of London in 1996.
Since 1996 he has been working in the area of simulation modelling of telecommunication systems, including human subsystems.
He is a regular invited lecturer in the topics of application of computer simulation for performance analysis of telecommunication systems, at the Széchenyi István University of Győr.

A CLINICAL DECISION METAKNOWLEDGE MODEL

Joël COLLOC , Marie-Hélène CUVELIER
CRIS-LEACM, Université de LYON
Université Lyon2, ISH, 14-16 Avenue Berthelot
69007 Lyon, France
Joel.colloc@ish-lyon.cnrs.fr
mariehelenecuvelier@wanadoo.fr

Peter SUMMONS
The University of Newcastle
University Drive, Callaghan, NSW 2308, Australia
Peter.Summons@newcastle.edu.au

ABSTRACT:
This paper presents a clinical metaknowledge model that can be applied in a wide range of application domains that use Knowledge Base Systems (KBS). It relies on a general description of the clinical decision process. This model can be then specialized to solve similar problems in all domains where processes similar to clinical decisions are applied. We show the relevance of our model in three different domains: medicine; management; and education.

KEYWORDS: metacognition, management, medicine, education, prognosis, living being, case based reasoning, multi-agent systems.

INTRODUCTION

We propose a clinical metaknowledge model to build Decision Support Systems (DSS) and Knowledge Bases (KB) in different scientific and technical areas where the clinical approach is relevant. Section 1 presents the model that provides us with a framework to express the clinical decision process independent to any specific application domain. The model relies on a definition of the main steps of the clinical process and provides operators to specialize and adapt these steps to other situations and domains. Section 2 presents the clinical decision process used in medicine. Section 3 shows how the same process can be applied in management and explains why a company can be considered as a living being. Section 4 looks at the education domain, where one can find similar clinical steps used to cope with the cognitive deficiencies of some pupils. Finally, Section 5 describes methods to implement the metaknowledge model.

1. A METAKNOWLEDGE CLINICAL MODEL

We propose a metaknowledge model, which relies on the decision process similarities found between practitioners' behaviour with other specialists and can be seen in many varied domains such as medicine, management and education. A decision process problem can be defined as a choice between several options in order to achieve a goal as efficiently as possible.
Four main steps define the model: diagnosis; prognosis; treatment; and follow up. These steps make up the clinical decision loop. Each step is a complex task that involves similar means and moreover similar goals.

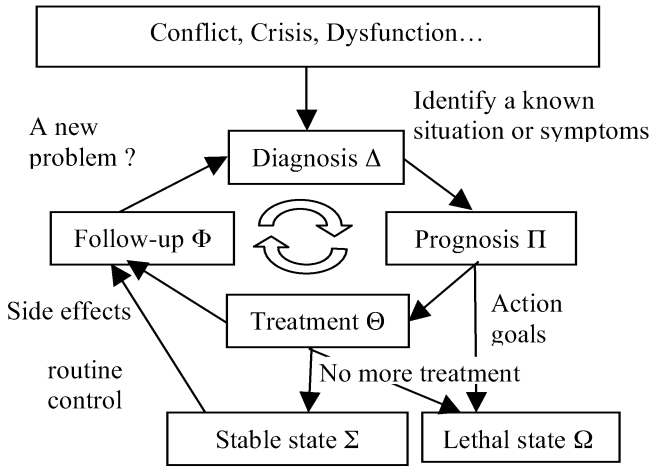


Figure 1. a clinical metaknowledge model

These main goals are: to recognise a situation as a clinical picture; to make a prognosis of the situation, firstly if nothing is done, and secondly if something is done to modify the current evolution.
Thus, the prognosis should predict the probable outcome of the situation. According to the prognosis, treatment goals are established that try to enhance the situation and, if possible, to solve the problem. Most of the time, it is necessary to monitor and control the treatment's efficiency and the possible occurrence of side effects that might cause new problems (that would then need to be diagnosed). The complete clinical decision loop is illustrated in Figure 1.

2. MEDICINE

2.1. the medical context

The medical decision activity could be defined as a decision chain (more exactly a network) that involves the following steps/stages: diagnosis; prognosis; therapeutic and treatment follow-up.
As some of these steps become more and more complex, a decision aid tool would be useful, especially during the prognosis and therapy stages.
Paradoxically, many physicians are interested in information systems and some of them have developed small, or sometimes more sophisticated, Decision Support Systems (DSS). However, these systems are very locally used inside a hospital or a research department by a small group of specialists who have to solve very specific and localised problems.

In France, the system end user and designer is very often the same person who prefers well suited home-made solutions. In such clinical situations the medical activity is often so specialised that there is no generic outlet for the resulting DSS. Only one copy of the software may actually be produced and used. The complexity of the medical domain is another complication in building efficient DSS. To sum up: on one hand, the physicians know very well the interesting domains where a decision tool is required, but as clever they are, their computer science knowledge is seldom sufficient to design and develop such a system. On the other hand, computer scientists must cope with the difficulty of understanding the doctors' needs and the complexity of their medical domains. Thus, they often provide expensive and inappropriate tools. The legal and ethical aspects of using a DSS to cure a patient should be also be taken under consideration. The responsibility of taking inappropriate and harmful actions scares physicians away from using a DSS during their practice. Hence, clinical DSS must be very secure and the transactions should be journalised.

2.2. the clinical process in medicine

We believe that the medical decision activity contains underlying general-purpose knowledge that can be useful for decisions in a large number of situations, provided that the reasoning process is similar.

Therefore, the problem is to detect the decision paths, which are independent from a specific medical field.

Every medical decision needs to correctly diagnose the disease(s) and to evaluate the clinical state of the patient. Before this can be done, the physician must know the previous therapies and their results in order to choose an appropriate treatment to improve the patient's health. The physician must also predict the disease's evolution. These three general tasks are respectively called: diagnosis Δ, therapy Θ and prognosis Π. They interact together and then contribute to provide a relevant decision, which is, most of the time, a treatment strategy to apply. The diagnosis, the therapy and the prognosis need four different types of information (Gabrielli and Buffalo, 1982):

1. Patient-related information such as current complaints, recent clinical course, past medical history and genetic and social background.
2. Related clinical experiences such as experience with *similar* cases.
3. Related formal medical knowledge such as segments of textbooks pertinent to diagnosis, therapy, prognosis, and the clinical course.
4. Theories, concepts, experimental hypotheses concerning all related areas.

But, to this classification above, we must add the very important knowledge related to the decision processes used by the physician.

2.3. Diagnosis Δ

The diagnosis is based on medical and surgical semiology, that is to say on the study of clinical signs and events, which guides the physician to identify the disease. Semiology is a science that studies the signs nature and how their combinations and their evolutions define clinical pictures. One complex classification of the sign concept's different meanings is proposed by Umberto Eco (Eco, 1988). Medical semiology is a method used to teach the different necessary steps required to elaborate a correct diagnosis to medicine students. The experienced physicians have memorized these decision schemes, made them unconscious and, thus, difficult to formalize. There are three different kinds of diagnosis steps involved in the decision process:

Positive diagnosis step (Δ+): During this step the physician considers current complaints, medical history, genetic and social background, and, through the examination of the patient, searches for the clinical signs which are classified into three different types: pathognomonic signs (characterizing a disease), evocative signs (making sense to think about some diseases), and accessory signs (completing the clinical picture). During the Δ+ step, the physician is making hypotheses relevant to the patient's clinical state.

Differential diagnosis step (Δ≠): During this step the physician is searching for the existence of, or the lack of, specific signs in order to eliminate those which are not relevant from the previously elaborated hypotheses.

Etiological diagnosis step (Δet): During this step the physician studies the causes of the disease manifestations, that is, the process of the disease's appearance and evolution.

To sum up, the clinical diagnosis stage includes the positive and differential diagnosis steps. Most of the time, it makes use of static superficial knowledge (heuristic), while the etiological diagnosis is based on «deep knowledge» that is able to explain the causes of the disease's appearance and the subsequent events that are going to take place during its evolution. The latter information will be useful to the prognosis (Π) as well. An object-oriented model is used to represent composite disease objects. Clinical pictures and signs are defined as sub-objects involved in a composition hierarchy. The disease's evolution is handled with dynamical functions, which, according to a time scale, modify the composition of the disease objects, so as to represent the known evolution modes (Colloc and Summons, 2006).

2.4. Prognosis Π

Prognosis is an activity in which the physician tries to predict the patient's clinical state evolution and the probable outcome (healing, stabilization, death). This prediction is very difficult to do because one must take into account not only the disease's classical clinical pictures, but also the patient's individual parameters such as: their genetic and social background, their psychological factors, their physical characteristics, their past medical history, and of course, the probability for the chosen therapy (Θ) to succeed.

The necessary knowledge and facts which help to elaborate the prognosis derives from several sources: the etiological diagnosis (provided by Δ), the patient's general clinical state, the stage of the disease, the knowledge of well known

clinical pictures, scenarios representing (like a movie) the sequence of virtual clinical pictures representing therapy alternatives. Epidemiological studies provide knowledge of the probable disease evolution from statistical data. More relevant knowledge is the physician's past experience in observation of any similar cases, which may suggest that the evolution will be similar.

The latter knowledge can be handled through a case-based reasoning (CBR) approach (Aamodt and Plaza 1994). We propose to represent the prognosis as a decision process that finds the most probable scenario, represented as a sequence **(Δ0 ... Δn),** of usual clinical pictures according to the available patient data, the knowledge of typical disease evolution, and the usual actions of known therapies Θ. This sequence is a path in a graph of clinical states that defines the prognosis.

Each path of the graph should be evaluated on the qualitative aspect as the best path towards healing the patient, or if that is not possible then a path that providea a relief of pain and enhancea comfort.

These paths contain clinical scenarios, represented by clinical picture objects, linked by time arcs and triggered by dynamical functions (Colloc and Summons, 2006).

2.5. Therapy Θ

The physician must know what treatments will enhance the patient's clinical state. To cope with this problem, the practitioner must have different kinds of knowledge concerning the available therapies, such as knowledge of indications for administering drugs, pharmacology data, contraindications, drug interactions, drug toxicity and adverse effects. The physician has previous prescription experiences of past cases that provide knowledge to predict the effects of a therapy strategy in similar cases.

First, a therapy is prescribed to achieve a goal. Therefore, there is a therapy classification:

- **Curative treatments goal** is to cure the patient or consolidate the patient's clinical state (to stop the disease's evolution). For example: an antibiotic prescription to cure tonsillitis. The physician must be provided with the etiological diagnosis (made previously in the Δ step) to be able to choose an appropriate curative treatment.
- **A preventive treatment** is prescribed to prevent a serious disease's occurrence. For example a vaccination, or an antibiotic prescription before a septic surgical operation.
- **Symptomatic treatments** are used to cure disagreeable effects, that is to say functional manifestations of a disease, without knowing the etiological diagnosis. For example the steps to cure an ordinary headache. Note however that the specific type of headache could be a symptom of a more serious disease, such as a cerebrovascular stroke.
- **Palliative treatments** are prescribed when the prognosis Π is hopeless and when we only try to cure the disease's manifestations (for example pain relief in cancer) to provide more comfort for the patient.

Therefore, the physician initially tries to define the treatment goal, according to the diagnosis and the prognosis. Then knowledge concerning the available therapies is required (knowledge of indications, contraindications, toxic and adverse effects, drug addiction...), which is used to build up a therapy strategy composed of drug administration according to an appropriate periodicity. For example multiple therapies to treat tuberculosis.

If the etiological diagnosis is not discovered, the physician must prescribe complementary investigations (provided time is available). The physician must also prescribe a symptomatic therapy to keep the patient alive and to gain time in order to do further investigations to finally find the etiological diagnosis and then set up a curative treatment. Thus, some heuristic strategies are available to help and/or cure the patient in emergency cases where it was, at first glance, difficult to understand what was happening. Therefore to provide completion, the decision schema shown in Figure 1 must also consider a strategic level where a decision is taken according to the available heuristics.

Physicians often use analogical reasoning and epidemiological knowledge to compare similar clinical courses (stories) where different treatments were prescribed and different strategies were used. However, to fight death is always a very tricky task.

3. MANAGEMENT

In the management domain, we propose a clinical method based on the metaknowledge model provided in section 1.

A company or a firm is unique, like a patient in medicine or a pupil at school, and more generally in all the domains in which the clinical reasoning mode is relevant.

The main goal of the method is to be able to measure the relevance of the treatment and the actions that could enhance a situation.

In management, just as in medicine, it is very useful to find similarities between their entities (patients and companies) but it is impossible to find equalities. Semiology must be invented in order to set up a new management nosology. If a company is to be considered as similar to a living being it is necessary to define the company's vital parameters. Thus, the decision-maker must rely on an efficient information system to recursively adapt the advice sources between the available measurement tools and the observed variables in the real world. The decision-maker sets up the stop condition of this recursive loop. This is obviously one of the main abilities of a good manager.

One can find other similarities between a patient and a company. A company, like a person, has an identity: a logo, a trademark, a recognised activity and is mainly composed of human beings. Consequently, a company should be considered as a living being and thus can suffer and die.

Dying companies experience disorganization and various dismantling processes. Therefore, it is also interesting to compare a human being's death with a company's death.

Considering the similarity with the medical anamnesis, it should be necessary to describe the genesis of the dysfunctions (internal side), the environment of the company (external side), communication between these two levels, and the company's way of life. The previous history

and genetic inheritance has the same importance for a company as it does for a human being.

Where is the company coming from? Where is it going? The first question is a diagnosis task **Δ**; the second one is a prognosis task **Π**.

In both domains (management and medicine), the study of the similarities between individuals (company or patient) and their environment is essential. On one hand, it is necessary to study the impact of politics, the state, and local administration on a company's life and its environment; on the other hand it is essential to study the family and the way of life of the patient.

Thus, the management practitioner has a very similar role to that of the physician. A manager is in charge of making an audit of the company that is a living and an organised being. This function is based on the manager's tact and confidentiality, and in a company's confidence in its manager being able to perform the company's diagnosis and to evaluate its health state (on the strategic, human, resources, financial aspects). This is exactly the same type of relationship that a physician has to set up with a patient. A manager has to make the prognoses of the spontaneous evolution, including other different scenarios that are relevant, according to the actual evolution of the environment of the market, the competitors, the available skills, expertise and knowledge. A manager must also provide new recommendations to enhance the company's development, its results and moreover the blooming of women and men who are giving their best to the company success (Kitous, 2004). These recommendations have a therapeutic goal and the management practitioner's expectation is to be able to solve the company's problems towards its healing and then (hence) to avoid its death.

3.1. A clinical approach of companies

How to detect the company's dysfunction and how to define the health of a company? Medical clinics use analogical reasoning practices based on recognition of similar situations. Available operators include analogy, induction, deduction and abduction. The clinical approach conflicts with the logic used in the market, which is based on a search to reach the optimum. The therapy may sometimes destroy the company, as a treatment may cure a condition but still kill a patient. Clinical knowledge relies on the experience, analogy and observation of similar cases, and the skills of the clinical practitioners.

Knowledge should be reused and adapted to solve new, but similar, cases. It is also important to provide an expandable tool to reuse and adapt the knowledge to treat new similar situations. CBR is well suited to the clinical reasoning mode (Aamodt and Plaza, 1994), (Gupta, 1994), however the main difficulty is the lack of adaptation operators.

A living being is protected from external attacks by an immunity system that is tuned to the being's identity (Anspach and Varela, 1992). A plasmic membrane establishes a border between the inside and the outside of a cell of the being. This membrane tunes the interactions and maintains cell homeostasis. For example, a potassium injection will cause death by cardiac arrest because the Na+/K+ equilibrium between the internal and external sides of cells is broken.

Like a living being the company has a communication interface that controls what should be disclosed and what should be kept secret. Most of the time the identity and the existence of a company relies on the know-how and knowledge it possesses, such as its manufacturing secrets. This biological metaphor is essential because the companies are composed of several, or many, human beings who are themselves metazoa, composed of sixty thousand milliards of cells. Therefore, it is relevant to compare the disease notion in both medicine and management. This notion relies on the observation of cases and especially experts who carry on the autopsy of company into liquidation, like a physician will do the autopsy of a corpse. The autopsy is very important to gain knowledge about, to understand and to describe, the death process. A manager can learn about the story and from it elaborate new means and therapeutic methods to avoid the same fatal result in similar cases. These clinical cases make the CBR approach very suitable.

However, in management the strategic level is very important (just like in medicine) because some heuristics are essential in order to take appropriate decisions based on the clinical model. It is therefore necessary to study the relationships between the company and its environment as there are economic events that are not predictable and the involved actors produce perturbations. The company has influences on its environment. Thus, the clinical approach in management relies on the study of both internal and environmental aspects of the company and the exchanges between them. This is one of the main characteristics of a living being. The company's health should not be only defined by quantitative economic parameters but also by quality signs evaluating the human life experiences. It is necessary to unify these two aspects to enhance a better understanding of the social and psychological life in organisations.

The management clinical model relies on a triptyque: the company's finance, that allows a diagnosis Δ and a short-term prognosis Π but it is unable to take into account the human dimension.

The organisation clinic studies the company's internal or external dysfunctions and weaknesses (human resources, information system, finance, marketing and communication).

A nosology that describes a classification of companies' dysfunctions, clinical pictures and signs and their evolution is necessary. In management, as in medicine, time modelling is also essential. It defines the seizure occurrence: cyclic or step-by-step (Colloc and Summons, 2006). These clinical objects could be defined as the ontology of a clinical approach for companies. However most of the time, a company's health is only defined in terms of its competitiveness or its financial market value, forgetting that the health is an equilibrium state that relies on changing situations within a hostile economic environment (Kitous, 2004). In this context clinical ground

experience provides very valuable knowledge that should be stored and reused through a CBR approach (Aamodt and Plaza, 1994). Some predictive indicators are useful to model and foresee the evolution of a company's situation, necessary to allow a prognosis to be made and thus to choose appropriate behaviours and means of enhancing the situation. The psychological aspect is an essential indicator because a company is firstly composed of human beings.

4. TOWARDS A CLINICAL APPROACH OF EDUCATION

According to previous research, a double metacognition loop involves aspects of the clinical metaknowledge model: firstly in the diagnosis and treatment of a pupil's lack of understanding, and secondly in the self detection and a cure for a teacher's inappropriate behaviour that will in turn help to enhance the teacher's teaching skills. We provide an education framework that employs aspects of the clinical decision model. We firstly talk about the teacher's clinical approach intended to allow the teacher to learn about their attitude, to provide a reflective view on their teaching methods and abilities, to build the expertise and clinical tools that can enhance their observation of their pupils. All of these aspects are diagnoses tasks (Δ).

4.1 The education diagnosis Δ

In education the diagnosis task relies on many evaluation methods, such as observations, interviews, intellectual tests, and personality tests. A clinical observation of the pupil consists of detecting all of the significant cognitive, linguistic, behavioural, and emotional symptoms and signs. The clinical reasoning activity is mainly based on the appropriate observation that should be carried out and the appropriate methods and tools that might be employed to correct detected problems.

Therefore, clinical interviews provide a means of forming an accurate analysis of the intellectual structure of the pupil. Piaget provided us with a general knowledge development theory based on the assimilation and the accommodation processes (Piaget, 1967).

The assimilation process relies on an integration of pre-existing structures. In other words, the assimilation process means that pupils are able to solve new problems in new situations with the help of previous knowledge. In assimilation, the pupil's reasoning process uses intellectual operators that they have already acquired. Conversely, the accommodation process is necessary when pupils are obliged to modify their own intellectual structure: more precisely, when the learner is unable to solve a problem with the help of their previous knowledge and experience. This unsteadiness situation leads the learner to modify their action schemas. Following the accommodation process, these newly acquired schemas will allow pupils to solve new, but similar, problems and to do new assimilations. To summarize, according to Piaget, the learner is doing an "equilibration majorante" allowing progression and development of their intellectual capabilities.

The explanation interview allows the teacher, by the way of questioning, to know the pupil's approach to carrying out a task, to solving a problem, to reaching a research goal. This can then allow the teacher to provide appropriate support, to find an educational remedy, and to provide guidance. This both helps the pupil to be informed and improves the teacher's knowledge (Vermersch, 1996). This defines what we term the double metacognitive loop (Colloc and Cuvelier, 2004). Our approach shows that the intellectual activity according to Piaget and Vermersch are complementary.

4.2 Semiology of cognitive dysfunctions

Among pupils who have apprentice difficulties, we distinguish three main groups, pupils who:

- have a well-known visual, auditory, motor or physical disability: such as dysphasia, dyslexia, dyspraxia, autism, and trisomy, etc;
- experience problems caused by factors external to the school: such as tongue mastering, culture, etc;
- have learning problems due to bad educational methods or to difficulties provided by individual characteristics such as language, attention, memory, working strategies, and metacognition concerning the learning goal.

Two commonly quoted difficulties learning problems are dyslexia (that affects reading and writing abilities) and dyscalculia (which concerns the acquisition of mathematical principles resulting in difficulties for a pupil in areas such as their use of numbers; calculation abilities; use of operators and algorithms, lack of logic, lack of organisation, and can result in spatial representation disorders and abstraction disabilities). Most of the time, a physician or a psychologist makes the diagnosis for:

- Pupils who have severe learning difficulties,
- Pupils who have personality troubles

Once diagnosed, specialist teachers are able to choose the relevant tools to provide help according to the pupil's difficulties.

Specialised teacher's work also has a clinical dimension where they must be able to distinguish the cognitive signs of an intellectual difficulty from a sensory or neurological disability. The specialist teacher knows the clinical signs that will require the advice of another practitioner, such as a paediatrician, psychiatrist, psychologist, or a speech therapist. A clinical observation is essential when pupils have language troubles. The observation focusses on infraverbal communication indicators such as body attitude and position, gestulations, eye expressions, and prosody.

In the case of young pupils it is essential to detect any sign of trouble very early to stop a disease's evolution. As an example with oral language capabilities, special tools were developed to detect language troubles of 5-6 years old children (Zorman, 2004) or phonemic capabilities (Kippfer-Piquard, 2007). Detection of language difficulties depends on the type of difficulty, typically oral problems should be detected at the nursery school (before the pupil is 3-4 years old), reading difficulties before 5-6 years old and writing

difficulties before 6-7 years old, to retain the possibility of taking efficient corrective actions. In clinical reasoning of these problems, time is essential and the prognosis changes according to how early or late an appropriate action is taken to prevent or to cure the problem (Colloc and Summons, 2006).

4.3 Prognosis: how to foresee what it is possible to do?

A main difficulty for successful prognosis is that the disease must have been previously diagnosed and an analysis must be done to provide some possible treatment paths.

Each intellectual disabled child will not able to cope with all education stages and thus it is necessary to rely on their individual potentials to choose the most appropriate educational action. In fact, teachers search for the pupil's strengths to build tailored help projects that allow the pupil to succeed. In other words, we prognose what the pupil will be able to do.

The prognosis Δ relies on the diagnosis described in the previous section.

4.4. Therapy

In education, the therapy stage consists of an individual, or a group, project. This project defines the obstacles, the objectives to surmount them, the reasons and the methods of remedial teaching.

5. IMPLEMENTATION

The Multi-Agents Decision Support System (MADSS) uses four agent categories (Figure 2):

- *Task Specialized Agents (TSA)* {**Θ, Δ, Π, Σ, Φ**} commit to tasks that are selected by the supervisor agent to be the current system goal (e.g. Diagnosis Δ).

The TSAs access the domain relevant bases containing knowledge and data represented with the appropriate models. TSA are embedded with a Task Finite State Automaton (TFSA), that manages the execution of a specific cognitive step e.g. diagnosis. They are cognitive agents that are specialized according to the knowledge domain (medicine, education, management…), their capabilities, and the knowledge model. They have access to specific knowledge bases and databases (whose details are not presented in this paper for space reasons). For example a Θ knowledge base must deal with indication, contraindication, and the side-effects of treatments.

- *Ontological agents (**O**)* formalize the terminology, and the entity definitions used during the information transactions between agents (Gruber, 1993), (Falasconi et al., 1997).

- *Case based reasoning agenst (**CBR**)* memorize and retrieve the experience in solving similar previous problems. This capability allows CBR to express previous experience knowledge and thus to use it to enhance a decision in new (similar) situations. The basis of CBR is that, by computing distances between new cases and previously stored (and indexed) cases, and by then retrieving those cases that are similar, we can apply the corresponding decision and actions of the old cases to solve this new problem. A CBR cognitive agent type encapsulates the necessary modules to run the CBR functionalities.

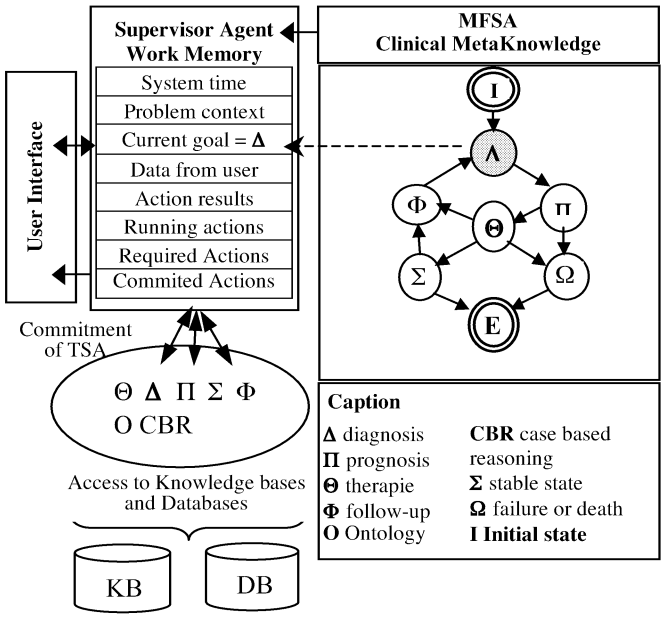


Figure 2. The MAADS architecture

- *A supervisor agent* manages the agent community and insures the interface with the user. It uses job oriented cognitive scenarios represented with a stored Metaknowledge Finite State Automaton (MFSA) which represents the clinical metaknowledge based on the decision model of Figure1. The supervisor's working memory describes the current situation and intermediate results. The MFSA schedules the whole clinical decision process. The supervisor controls task commitment and task execution by available, appropriate, unoccupied Task Specific Agents (TSA's). Each TSA uses its reflexive knowledge to decide whether it is able to do the job or not and if it is useful to cope with the current goal set by the supervisor. The agent's specialization degree, the task type, the available knowledge, and the required reasoning mode are the criteria that decide whether to commit to the current task or not. If the agent accepts the task, the supervisor notes this down in the *committed actions* table and traces the state of the task in the *running actions* table. The TSA gets access to the user data via the supervisor. It may require information from the user through the supervisor interface. The answers are made available to all agents committed to the current task in the *action results* table. When the goal is achieved, the supervisor will set the next goal corresponding to the transition.

A TSA can drop its task execution if the context is modified, such as if a solution was found, the system goal is changef, the allocated delay has expired, etc.

When a goal is completed the supervisor agent will set the next current goal according to the MFSA' next transition and the type of the previous results (Figure 2). Thus, a new cognitive step will then begin (e.g. a Therapy **Θ**). The system stops when the *End* state is reached. All transactions

are journalised by the supervisor agent and provided to the user via the interface.

CONCLUSION

We have proposed a metaknowledge model, which describes the necessary tasks to take for a clinical decision (diagnosis, prognosis, therapy and follow-up). The clinical approach is used in many domains where living beings are involved such as medicine, management and education. We have shown the similarities between these three domains and especially that the same clinical reasoning mode can be applied in each. Moreover, the same clinical decision model can be applied in many other areas.

We have provided an implementation schema for the metaknowledge model capable of carrying out the steps for a clinical decision. The implementation uses a multi-agent system including task specialized agents. These agents are cognitive autonomous agents and have a reflexive knowledge about their abilities that allows them to be efficient in a specific clinical task within a particular domain. These agents have access to specialised knowledge bases and databases providing the appropriate guidelines in situations where they can act with relevance.

In medicine, some applications have already been developed. In the other areas considered, management and education, it is still necessary to enhance the nosology definition and to develop appropriate knowledge bases. We are currently doing investigation into both these domains.

REFERENCES

Aamodt A., Plaza E. 1994, Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, *AI Communications*, IOS Press, 1994, Vol. 7 n.1, 1994

Anspach M. R. and Varela F. 1992, Le système immunitaire : un "soi" cognitif autonome.", in *Introduction aux sciences cognitives*, (ss dir. D. Andler), ed. Gallimard, , pp 489-508., ISBN 2-07-032577

Colloc J and Léry N. 1997 A multi-expert decision support system in medical ethics., *Health and Systemics Journal*, ed. HERMES, vol. 1, no 1, 1997, pp.39-55.

Colloc J. and Cuvelier M.H. 2004, A Multi-Agent Approach of Modelling Metacognition and Emotion in Education, SEG'O4 in the conference *2004 Jordan International Conference on Computer Science and Engineering*, JICCSE 2004, Al-Balqa' Applied University, October 4-7, 2004, Al-Salt, Jordan, 6 p.

Colloc J. and Summons **P.F**. 2003 Modelling the Temporal Information in Medical Decision Support Systems: *IMACS Multiconference Computational Engineering in Systems Applications CESA'2003,* Lille, France, 2003, Electronic Publication ISBN2-9512309-5-8.

Colloc J**.** and Summons P.F. 2006, An Object Oriented Time Model of a Decision Support System for Intoxication Diagnoses *Proceedings of the 4th International Industrial Simulation Conference ISC'2006*, June, 2006, Palmerno, Italy, pp387-392,

Gruber T.R. 1993 , A translation approach to portable ontology specifications, *Knowledge Acquisition*, vol.5. no 2,1993, pp. 199-220

Gupta, U.G. 1994 How Case-based Reasoning Solves new Problems, *Interfaces*, Vol.24, no 6

Kipffer-Piquard A., 2007 Prédiction dès la maternelle de la réussite et de l'échec spécifique à l'apprentissage de la lecture en fin de cycle 2, in *Les troubles du développement chez l'enfant Prévention et prises en charge*, ed by Bordin B., Habin-Gayle M.,Le Driant B., Vandromme L. Université of Picardie Jules Verne, 2007, Paris, l'Harmattan, p 151 à158

Kitous B. 2004 Research Accreditation PhD, *Vers une approche clinique des organisations*, University of Nantes, France.

Morin E. 1986, La méthode t3. *La connaissance de la connaissance*, ed. Seuil, 1986, 244 p, ISBN 2-02-009257-3

Piaget J. 1967, *La psychologie de l'intelligence, 7ème ed.* Armand Colin, Paris, 1967, 192 p., ISBN2-200-32214-3

Piaget J., *La représentation du monde chez l'enfant*, éd de 1993, Paris, PUF

S. Falasconi, G. Lanzola, M. Stefanelli. 1997, An Ontology-Based Multi-Agent Architecture For Distributed Health-Care Information Systems, *Methods of Information in Medicine*, F.K. Schattauer Verlagsgesellschaft Mbh, 1997, pp. 20-29

Summons P.F. 2005 *A Methodology to Facilitate Knowledge Acquisition and Maintenance for Decision Support in the Medical Domain : Interpretation of Clinical Guidelines*, Thesis, University of Newcastle, Australia, 2005.

Turban, E., Aronson, J.E., Liang, 2005 T.P. *Decision Support Systems and Intelligent Systems (7th Ed)* Pearson Education, New Jersey.

Vermersch P. 1996*, L'entretien d'explicitation*, 1996, Paris, ESF 182 p

Zorman M., Lequette C., Pouget G. 2004, Dyslexies : intérêt d'un dépistage et d'une prise en charge précoce à l'école, évaluation du BSEDS 5-6, in *Développement cognitif et troubles des apprentissages : évaluer, comprendre, rééduquer et prendre en charge*, ed by Metz-LutzM-N, Demont E., Seegmuller C., de Agostini M., Bruneau N., 2004, Marseille, Solal p 245 à 270

Medical and pharmaceutical experiences: knowledge discovery with strategy by vote

Claude Dussart[1,3]
François-Xavier Magaud[1,2]
Claude Petit[1,2]

[1] European Center for Humanitary Health University Claude Bernard - Lyon 1 - France
[2] Laboratory LIRIS "modelling health decision" CNRS UMR 5205 University Claude Bernard - Lyon - France
[3] Military Hospital Desgenettes - Lyon - France
e-mail : claude.dussart@cesh.org

ABSTRACT
The meta-learning of medical and pharmaceutical experiences is a function of learning of meta-knowledge. The meta-learning operates on the results resulting from systems of learning based on distributed experiences. Three strategies aim at combining multiple predictions (resulting from separately learned classifier): strategy by vote, strategy by combination and strategy by arbitration. The strategy by vote was regarded as too simplistic.
Our work aims at rehabilitating this strategy which is rich of possibility. Beyond the vote majority and vote pondered, many voting procedures exist: plurality, procedure with fixed diary, procedure of KO with diary, procedures of Borda, Condorcet, Hare, Rubinstein, Shuhe Li, Greenberg, Diffo-Moulen. These procedures allow us to cast new interesting and innovative light on the meta-learning of medical and pharmaceutical experiences.

KEYWORDS
Medical, pharmaceutical, experiences; meta-learning, vote

1. INTRODUCTION

Every hospital is composed by a lot of medical and pharmaceutical experiences. These experiences are distributed in time or space, deal with a same subject and cast new light. For each experience, a prediction or a knowledge can be built according to various techniques of learning. The problem is to answer the following question: how automatically treat the results obtained by learning on distributed experiences?
The meta-learning answers this question. It is a question of building a meta-knowledge, metadata. It makes it possible to define a final prediction and to explain the variations observed on the predictions resulting from each learning.
Meta-learning is defined as learning of meta-knowledge. Meta-knowledge is knowledge on knowledge.

In our work, we concentrate on learning from the output of concept learning system in the field of the hospital distributed experiences. The experiences and the results obtained are carried out in a same site but at different times or on partitions from data bases obtained by chance, or on several sites. A distribution can be organized by a data base reduction.

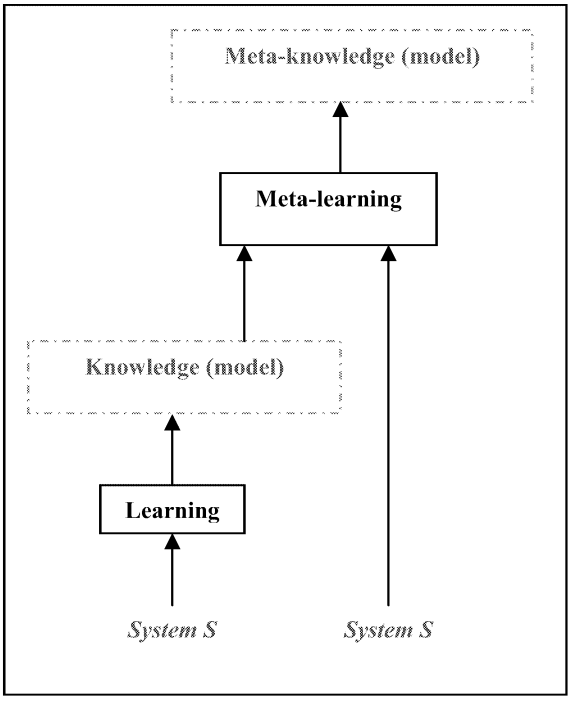


Figure 1. Meta-learning: concept

In our problem, meta-learning is learning from the predictions of classifiers on common training data. A classifier (or model) is the output of a concept learning system. A prediction (or classification) is the predicted class generated by the classifier when an instance is applied. In the phase of meta-learning, a learning algorithm is employed to learn how to integrate the results given by classifiers. Rather than having a predetermined and fixed integration rule, the integration rule is learned based on the behavior of the training classifiers.
Meta-learning requires several stages:
Step 1: base classifiers are trained from the initial training sets
Step 2: prediction are generated by the learned classifiers on the training sets
Step 3: a meta-level training set is composed from the predictions generated by the classifiers
Step 4: the final classifier (meta-classifier) is learned from the meta-level training set
Let us observe that this system is based on distributed and heterogeneous data bases. The learning algorithms can be different.

There are three distinct strategies for combining multiple predictions from separately learned classifiers.
The combiner strategy uses knowledge about how classifiers behave with respect to each other. Indeed if we observe that two classifiers envisaging a same class are always correct (relative to some test set), we can use this observation like a powerful predictive tool. We could then decide to ignore all other classifier when they predict a common outcome.
The aim of this strategy is thus to coalesce the predictions from the base classifier by learning the relationship between these predictions and the correct prediction.
The arbiter strategy requires the use of an objective judge whose own prediction is selected if the classifiers taking part in the meta-learning cannot make a consensual decision. The arbitror is itself a classifier. It can choose a final result based upon its own prediction but cognizant of the other classifiers'prediction. In this strategy, the meta-learners learn how to arbitrate among the predictions potentially different from the basic classifiers.
The strategy by vote : each classifier can exprim his preference. The classifier having a preferential vote are determined by observing the performance obtained by the classifiers with some statistical and mathematical tests. The goal of the vote then consists in choosing one of the predictions of one or more classifiers.
This paper proposes a state of sciences of voting methods that we can develop within the framework of the meta-learning of distributed experiences.

2. STRATEGY BY VOTE

2.1. The vote in the majority qualified

This is the simplest procedure. Let N be a set of n classifiers. Each classifier has a voice. Several candidates (indicated by A) are presented to the classifiers simultaneously. Each classifier votes for the candidate of his choice. The vote is anonymous and with secret bulletin. There is failure to gain absolute majority if several candidates obtain each one a number of votes equal or higher than q.
The choice of q (checking $n/2<q<n$) removes this type of failure. It's also possible that any candidate obtain a number of votes equal or higher than q. Nobody is then elected. Let us recall that for us a failure to gain absolute majority is not a case of indecidability, because a subsidiary rule makes it possible to decide. If n is even and $q=(n+1)/2$, there is the vote with absolute majority. In the general case, there is the vote with qualified majority.
In practice, it is often envisaged a subsidiary rule making it possible to decide between the candidates (relative majority for example).

2.2. Plurality

Plurality generalizes the case where there is more than two candidates. It consists in:
- each classifier has a voice and votes for the candidate of it choice
- the winner is the candidate who obtained the greatest number of voices. If several candidates are ex æquo with the greatest number of voices, we declare them elected or retained in the sense that we used this higher term.

It returns to same to define plurality in the following way:
- each classifier preferably classifies the candidates according to his command
- the winner is the candidate who is classified first the greatest number of times.

If N = { 1, 2, 3, 4 } and A = { a, b, c },
and if the classifiers vote according to the profile RN = (abc, abc, bca, bac),
each candidate a and b is classified first twice, the candidate c is classified first no time. Plurality (noted fp) declares a and b elected (or retained).
So, fp(RN) = { a, b }.

2.3. Borda's procedure

This procedure takes into account intensity of the preferences of the classifiers. Each classifier classifies the candidates according to a total ordered (translating their preference). One allots notes to the candidates: the last classified candidate receives note 0, the candidate before last receives 1., and the first receives the note p-1 (if p is the number of candidates). This operation is carried out for the classification of each classifier. We calculate then the total note of each candidate. The winner is the candidate who has the strongest total note.

2.4. Condorcet's procedure

By this procedure, the vote is organized in a succession of confrontations of the candidates two to two. The candidates are thus introduced to the classifiers per evens. All the evens are reviewed. The classifiers vote for the candidate of his choice. The winner of the first confrontation is the candidate who obtains the absolute majority. The final winner is the candidate who wins each confrontation. With each confrontation, the classifiers have to vote for one of the two involved candidates. They thus vote as many times as the number of confrontation. Each classifier can also exprim one vote (with a global order of preference). In this case, during confrontation between the candidates x and y, one considers that the classifiers (having classified x before y) vote for x against y.
The procedure of Condorcet does not manage to proclaim a winner when: a beats b; b beats c; and c beats a.

2.5. Hare's procedure

The procedure of Hare is based on successive elimination of the least desired candidates. Each classifier classifies the candidates according to a total order. If a candidate is classified first by at least half of the classifiers, it is declared elected and the procedure is stopped. The candidate classified first by the smallest number of classifiers is eliminated. This candidate (or these candidates) is eliminated from all the classifications establish by the classifiers.
With the lists thus modified, one looks at if a candidate is classified first in at least half of the lists. If so, this candidate wins. If not, candidate classified at the head in the smallest number of lists is eliminated etc… The procedure stops if a candidate is classified first by at least half of the classifiers. When the not eliminated candidates are the same number of times at the head of lists, they all are then declared elected.

2.6. Procedure of amendments with fixed diary

A diary is defined as any permutation of the ensemble candidates. It's a manner of drawing up the list of all the candidates in an order. This order isn't the expression of preference. It's the order in which the candidates will be examined. The concept of presentation of the candidates is clarified in what follows. Under this procedure, the vote is organized as a succession of confrontations of the candidates two to two, in the order determined by the fixed diary. The first confrontation opposes the first two candidates of the fixed diary in a vote in the majority absolute.

If the number of the classifiers is uneven, there is necessarily a single winner. If this number is even, there can be two elected officials. In this procedure, a tie braking rule laid down. So that for any confrontation of two candidates, there is a single winner. The loser is definitively eliminated from the competition. The survivor of this first confrontation is then opposed to the third candidate. The procedure continues with the last confrontation which opposes the last two candidates still not eliminated.

2.7. KO with fixed diary

Let us suppose that A = { a, b, c } and ABC is the diary. The procedure of the KO proceeds in the following way:

- candidate a is introduced to the classifiers who vote for or against his adoption. If a wins in the majority absolute this first vote, the procedure ends. If a loses this vote, a is definitively knocked out.
- if a loses the first vote, b is presented to the classifiers who vote for or against his adoption. If b wins this vote, the procedure ends and b is proclaimed elected. If b loses this vote, b is definitively except knocked out. c then is proclaimed elected and the procedure ends.

Thus, if a classifier chooses the order acb. In the vote concerning a, we will say « yes » for a. If a second vote is organized for b, it will vote against b. The classifiers thus vote as many times as necessary.

2.8. Dictatorial procedure

The dictatorial law designates a dictator among the classifiers. Each classifier preferably classifies the candidates according to an order. The candidate classified first by the dictator is declared elected. Instead of making all the classifiers vote, it's possible to directly require of the only dictator to vote, the other classifiers observing passively what occurs.

2.9. Vote by approbation

Each classifier registers on his bulletin one or more candidates of his choice. The order of the candidates is not taken into account. The winner is the candidate presenting the greatest number of inscription.

2.10. Classical procedure

This procedure is noted C_0 and is described in the following way:

1) there is a neutral promoter, which has no personal interest in the vote, which is not classifier, who takes care of the respect of the voting procedure and carries out a certain number of specified tasks in what follows.
2) a candidate, let us say y, is proposed for the vote. y can be the initial statu quo, or an unspecified candidate chosen by the promoter.
3) the coalitions having a veto counters are consulted there. Each one decides to apply yes or not the veto against y. If yes, the coalition has to propose a replacement.
4) two cases are possible:
   - no veto is proposed against y. then y is proclaimed there elected and the part is played
   - At least a veto is proposed against y. The promoter chooses then, among the vetoes suggested, the veto is applied. To carry out this choice, the promoter can possibly proceed by hard copy to the fate among the vetoes suggested.
5) if it's a veto of S against y which is applied, and if S proposed x to replace y then x is proclaimed elected and the part is played.

2.11. Rubinstein's procedure

This procedure is based on the concept of predominance. Predominance is an assumption of rational behaviour. It is described in the following way:

1. there is a neutral promoter, which has no personal interest in the vote, which is not classifier, who takes care of the respect of the voting procedure and carries out a certain number of specified tasks in what follows.
2. a candidate, let us say y, is proposed for the vote. y can be the initial statu quo, or an unspecified candidate chosen by the promoter.
3. the coalitions having a veto counters are consulted there. Each one decides to apply yes or not the veto against y. If yes, the coalition has to propose a replacement.
4. two cases are possible:
   - no veto is proposed against y. Then y is proclaimed there elected and the part is played
   - at least a veto is proposed against y. The promoter chooses then, among the vetoes suggested, that which is applied. To carry out this choice, the promoter can possibly proceed by hard copy to the fate among the vetoes suggested.
5. if it's the veto S counters y via x which is applied, then the promoter toss up for it:
   - if it's the reverse side, x is proclaimed elected and the part is played
   - if not, y is no more statu quo, but still remains a candidate in competition. x becomes the new the statu quo
6. the coalitions having a veto against x are consulted. All the proposals for a veto against x are put forward to the promoter. Two cases are possible:
   - No veto is proposed against x. Then, x is proclaimed elected and the part is played.
   - at least a veto is proposed against x. The promoter chooses then, among the vetoes suggested, the veto which is applied.
7. if it is the veto of S counters x via z which is applied, then z is proclaimed elected and the part is played.

2.12. Shuhe Li's procedure

This procedure is founded on a relation of predominance that we let us call Shuhe Li's predominance

1) There is a neutral promoter, which has no personal interest in the vote, which is not a classifier, who takes care of the respect of the voting procedure and carries out a certain number of specified tasks in what follows.
2) a candidate, let us say y, is proposed for the vote. y can be the initial statu quo, or an unspecified candidate chosen by the promoter.
3) the coalitions having a veto counters are consulted there. Each one decides to apply yes or not the veto against y. If yes, the coalition has to propose a replacement.
4) two cases are possible:
   - no veto is proposed against y. then y is proclaimed there elected and the part is played.
   - at least a veto is proposed against y. The promoter chooses then, among the vetoes suggested, that which is applied. To carry out this choice, the promoter can possibly proceed by hard copy to the fate among the vetoes suggested.
5) if it's the veto S counters y via x which is applied, then the promoter toss up for it:
   - if it's the reverse side, x is proclaimed elected and the part is played
   - if not, y is no more statu quo, but still remains a candidate in competition. x becomes the new statu quo
6) the coalitions having a veto against x are consulted. All the proposals for a veto against x are put forward to the promoter. Two cases are possible:
   - no veto is proposed against x. Then, x is proclaimed elected and the part is played.
   - at least a veto is proposed against x. The promoter chooses then, among the vetoes suggested, the veto which is applied.
7) if it's the veto of T against x via z which is applied, then the promoter toss up for it:
   - if it's the reverse side, z is proclaimed elected and the part is played
   - if not, x is no more statu quo, but still remains a candidate in competition. z becomes the new statu quo.
8) the coalitions having a veto against z are consulted. All the proposals for a veto against z are put forward to l.organisator. Two cases are possible:
   - no veto is proposed against z. Then z is proclaimed elected and the part is played.
   - at least a veto is proposed against z. The promoter chooses then, among the vetoes suggested, the veto which is applied.
9) if it is the veto of K counters z via u which is applied, then u is proclaimed elected and the part is played.

2.13. Suk-Young Chwe's procedure

1) there is a neutral promoter, which has no personal interest in the vote, which is not classifier, who takes care of the respect of the voting procedure and carries out a certain number of specified tasks in what follows.
2) a candidate, let us say y, is proposed for the vote. y can be the initial statu quo, or an unspecified candidate chosen by the promoter.
3) all the coalitions with or without veto counters are invited there to act in concert. This dialogue can lead to a unanimous decision. Among these coalitions, one at least must have a veto against y. Such a decision indicates that the coalitions concerned propose the application of a succession of vetoes against y. Such a succession of vetoes is described by the application of options a_0, a_1, …, a_i, …, a_m and of coalitions S_1, S_2, S_i, …, S_m, such as:

$$a_0 = y, \forall i = 1,\ldots,m, S_i \in \delta(a_i, a_{i-1})$$

All these decisions are given to the promoter.

4) two cases are possible:
   - no decision is proposed. Then y is proclaimed there elected and the part is played.
   - at least a decision is proposed. The promoter chooses then, among the decisions suggested, the decision which is applied.
5) if it's the decision a_0, a_1, …, a_i, …, a_m,S_1, S_2, S_i, …, S_m, such as:

$$a_0 = y, \forall i = 1,\ldots,m, S_i \in \delta(a_i, a_{i-1})$$

who is applied, then a_m is proclaimed elected and the part is played.

2.14. Diffo-Moulen's procedure

The last voting procedure that we present in this selection was described in 1996 like a natural manner to organize a vote.

1) there is a neutral promoter, which has no personal interest in the vote, which is not classifier, who takes care of the respect of the voting procedure and carries out a certain number of specified tasks in what follows.
2) a candidate, let us say y, is proposed for the vote. y can be the initial statu quo, or an unspecified candidate chosen by the promoter.
3) the coalitions having a veto counters are consulted there. Each one decides to apply yes or not the veto against y. If yes, the coalition has to propose a replacement.
4) two cases are possible:
   - no veto is proposed against y. then y is proclaimed there elected and the part is played.
   - at least a veto is proposed against y. The promoter chooses then, among the vetoes suggested, the veto which is applied. To carry out this choice, the promoter can possibly proceed by hard copy to the fate among the vetoes suggested.

5) if it's the veto of S counters y via x which is applied, y is eliminated there from the competition and x becomes the new statu quo.
6) the coalitions having a veto against x are consulted. They decide, with the unanimity of its members, if they propose or not a veto against x. If yes, they propose the replacement of x like statu quo. All the proposals for vetoes are put forward to the promoter.
7) two cases are possible:
   - no veto is proposed against x. Then x is proclaimed elected and the part is played.
   - at least a veto is proposed against x. The promoter chooses then, among the vetoes suggested, the veto which is applied.
8) if it's the veto of T against x via z which is applied, x is eliminated from the competition and z becomes the new statu quo.
9) the coalitions having a veto against z are consulted. Each one decides, with the unanimity of its members, if it proposes or not a veto against z. If yes, it has to propose the replacement of z like statu quo. All the proposals for vetoes are put forward to the promoter.
10) two cases are possible:
   - no veto is proposed against z. Then z is proclaimed elected and the part is played.
   - at least a veto is proposed against z. The promoter chooses then, among the vetoes suggested, the veto which is applied.
11) if it's the veto of S against z via u which is applied, z is eliminated from the competition and u becomes the new statu quo.
12) the vote continues as a same manner until a candidate is declared elected.

3. CONCLUSION

Currently, the literature proposes mainly two voting procedures in order to produce a meta-learning based on distributed experiences : the simple vote and the pondered vote. These procedures are exceeded by more sophisticated strategies: combiner strategy and arbiter strategy.

The voting procedures, we have just developed, allow to envisage then a new rise of the strategy by vote because only the simplest voting procedures were considered. The scientific rigour pushes us to explore this way completely, moreever these procedures are particulary adapted to the medical and pharmaceutical experiences. Actually, our work with a Borda's method gave good results without charging the processing times. We have indeed good reasons to think that an adapted and worked voting procedure will allow to extract robust hospital information (without bias).

REFERENCES

BORDA J.C. mémoire sur les élections au scrutin. Histoire de l'Académie Royale des sciences, Paris,1781.

CONDORCET. Essai sur l''application de l'analyse à la probabilité des décisions rendues à la pluralités des voix. Paris, 1785.

DIFFO LAMBO L. Comportements rationnels dans un jeu de vote. Thèse Doctorat d'Etat, Université de Yaoundé 1, 1999.

DUMMET M. Voting procedures. Clarendon Press, Oxford, New York, 1984.

DUSSART C, PETIT C, Méta-apprentissage d'expériences, ed. Vuibert, 124 p.

FARQUHARSON R. Theory of voting. Yale Univ. Press, New Haven, 1969.

GREENBERG J. The Theory of social situations. Cambridge U. Press, 1990.

LE BRETON M. On the core of voting games. Soc. choice and Welfare, 4, 295-305, 1987

MOULEN J., ANDJIGA N.G., BRISSAUD M. Préférences individuelles et choix collectifs. Hermès, Paris, 1999

RUBINSTEIN A. Stability of decision systems under majority rule. Journal Econ. Theory, 23, 150-159, 1980.

SHULE LI E.Stability of voting games. Soc. Choice and Welfare, 10, 51-56, 1992.

SUK-YOUNG CHWE M. Farsighted coalitional stability. Journal Econ. Theory, 63, 299-324, 1994.

TRUCHON M. Voting games and acyclic collective choice rules. Math. So. Sc. 29, 165-179, 1995.

BIOGRAPHY

Claude DUSSART holds a Ph. D. (2003) in medical decision support "meta-learning of distributed experiences" and a Pharm.D. (1998) by the pharmaceutical faculty of the Claude Bernard University (Lyon 1, France). He is a researcher in the European Center for Health Humanitary (CESH- University Claude Bernard Lyon 1 France), free teacher at the faculty of pharmacy (Lyon) and works at the military hospital Desgenettes.

François-Xavier MAGAUD is a PhD student at University Claude Bernard – Lyon 1 in France. His research topic is about meta-learning. He works on meta-learning strategies.

Claude PETIT: Habilitation for research director « approach for meta-model », Doctor of philosophy « Generativ modelisation », Doctor of philosophy « Global model for entreprise »
He is a research director in the European Center for Health Humanitary (CESH Lyon) and researcher in the laboratory LIRIS "modelling health decision" (University Claude Bernard – Lyon 1 - France)

EXPERIMENTAL STUDY WITH METRICS INDICATING THE INTERPRETATION OF CLINICAL GUIDELINES

Peter F. Summons
The University of Newcastle
University Drive, Callaghan, NSW 2308, Australia
Peter.Summons@newcastle.edu.au

Joël Colloc
University Lyon 2
CRIS-LEACM
ISH 14-16 Avenue Berthelot
69007 Lyon France
Joel.colloc@ish-lyon.cnrs.fr

KEYWORDS
Knowledge acquisition; Knowledge model; Clinical guideline; Ontology, Metrics

ABSTRACT

This paper outlines a research study investigating the creation of a knowledge model that is based on the clinical experts' interpreted knowledge of clinical guidelines. An expert's interpretation could be based on many aspects of the expert's experience and is often very hard for the expert to explain. This paper introduces some initial metrics that can be used to compare the expert's respective knowledge models, in terms of differences in knowledge between experts and the evolution of a single expert's knowledge model, and presents the results of those comparisons in terms of the present study.

1. INTRODUCTION

The application and acceptance of clinical guidelines remains problematic in many health and medical settings. Many reasons may be posited for this. One is that guidelines are not globally applicable - there is often a requirement for local adaptation of a guideline (SANPIS 2003). Other reasons are that guidelines tend to be static but medical and health knowledge evolves, and that guidelines have limitations in their ability to be maintained and updated in light of local best practice decisions and current clinical findings (Sanders, 1998). A further reason is that much of the expertise in medical and health areas is regarded as coming from experience – it is seen as a learned behaviour that is dependent on the local environment and that may become tacit over time.

This paper reports an experimental research study (Section 3) modelling the interpreted knowledge of a clinical guideline by four clinical domain experts. Section 2 describes the structured methodology used in the study. Section 4 introduces metrics that can be used to compare the expert's respective knowledge models and presents the results of such a comparison applied to the study. It shows that differences in the models might represent individual tacit experience which, if identified as such, may allow guidelines to be improved so they reflect and model the experience and expertise of the local expert clinical practitioners.

2. MDS METHODOLOGY

The knowledge acquisition bottleneck is still considered to be a significant problem in decision support and intelligent systems today. MDS is an Object-Oriented (OO) architecture that is suitable for modelling medical domain knowledge and that can support acquisition and change of knowledge (Summons, 2005). The architecture is based on objects that reflect a medical scenario (Summons and Villanueva, 1996). The objects have formalised behaviour and support formal definitions of temporal aspects and relationships that are normally required for modelling time in medical knowledge–based systems (Colloc and Summons, 2003) (Summons and Colloc, 2006).

In the context of the present study the MDS architecture can be broadly interpreted as having three major meta-classes:

- the Temporal Object class, which are described to clinicians as being Patient Properties consisting of physical things in the system that exist through time (such as Patient, Right Lung, Circulatory System);
- the Action Object class, which are described to clinicians as being Staff Interventions consisting of things or actions that staff would either perform, or have others (ward nurses etc) perform in relation to the system under consideration (such as tests, measurements, observations or signs);
- the Association Object class, which are described to clinicians as being Diagnoses consisting of labels given to the possibility of a Disease or a Pathological Process being present in an instantiation of a Temporal Object (such as Hypertension or Pain).

In the experimental study (Section 3) an additional category considered domain objects falling outside these three meta classes. Referred to as System Objects (the "Other" category), this class described the general environment and

captured clinician's references to people, places, instruments, and policy.

The MDS architecture supports a methodology to assist in knowledge acquisition, particularly tacit knowledge of an expert clinician (Summons, 2005). Tacit knowledge may be used in the interpretation of medical guidelines by an expert and not be explicit in the guidelines, such as when an expert uses a shortcut.

Typically in the development of a knowledge-based system (KBS) a knowledge engineer works with a domain expert to produce a Knowledge Model of the domain. This consists of a domain Ontology, describing the objects and structures of the domain, and also Rules for the domain indicating the system relationships and behaviours of domain objects.

Knowledge engineers may base their design on incorrect knowledge interpretations from the expert, or they may not know that they have missed knowledge from the expert. The interpretation differences of a system by stakeholders is referred to as *cognitive dissonance* (Cleal and Heaton, 1988). KBS developed for clinical applications also have an added problem in that the domain itself is very complex and specialised.

The MDS methodology addresses the interpretation problems associated with the elicitation of domain knowledge. The major stages of the methodology are:

- Knowledge acquisition is initially based on an interpretation of the expert's understanding of content in a clinical guideline;
- The domain expert and the knowledge engineer then refine the interpretation into a knowledge model of the content. The formulation of the ontology requires the identification of objects, which are then classified into categories prescribed by a supporting architecture. It is at this stage that object states are defined, causal relationships between objects are determined and temporal aspects of the domain are identified;
- System behaviour rules (existence rules, association rules and internal behaviour rules) are identified;
- The knowledge model, which is now fitted to the supporting architecture, is used as a prototype knowledge-based system with the expert;
- Iterations of the knowledge based model using best practice clinical test cases continue until expert validation or agreement is reached. The knowledge model is updated at each iteration. It is during this stage that the tacit expertise of the domain expert is elicited.

3. CLINICAL GUIDELINE STUDY

The clinical domain modelled is the management of complications for *patient controlled analgesia* (PCA). The study looked at the respective interpretations of the PCA clinical guideline by four expert clinical nurse consultants from the Hunter Area Pain Service who are consultants in both acute and chronic pain for the state hospital service in the Hunter Area of New South Wales, Australia. The experimental task was to acquire the interpreted knowledge of different experts about a common clinical guideline using the MDS methodology and to then produce knowledge models of the interpreted guideline knowledge in the MDS architecture. The participants identified domain objects as instances of MDS Classes, and their relationships in terms of production rules (IF...THEN... type rules) such as

- IF the Sedation Score > 2 THEN Administer 12 L/minute O_2 via mask
- IF the Sedation Score > 2 THEN Contact Anaesthetist or Medical Officer

The production rules may either be rules associated with internal class behaviour, with system event behaviour due to causal class interactions, or with system behaviour involving inter-class associations.

While the use of production rules in KBS is not new, the MDS architecture is based on specified meta-Classes and their instantiated objects, and the production rules model the behaviour of instantiated objects in response to an event (both internal to the object and external to it), and also provide a mechanism for an instantiated object to generate future events. The MDS architecture requires the antecedents and consequents of its production rules consist only of identified meta-class instantiations. A future event can be represented in a rule consequent by instantiating an Action Object that may schedule a test, or alert that a symptom has been recognised A negation of the existence of class instantiations is also allowed.

For the purposes of this experimental study the knowledge engineer played a passive role in the clinical experts' object categorisation, although in a normal clinical setting the knowledge engineer would validate classified objects more interactively.

The models were tested through iterations of case-based application data presented to each of the clinical experts participating in the experiment, and changes in their models produced from application of the MDS over the iterations were measured. The case based data was in the form of scenario data prepared by an expert clinician not involved in the MDS iteration process.

The suitability or completeness of the knowledge model can be expressed in metrics denoting the amount of change in the knowledge model that is required to accommodate new cases, where change will be measured in terms of new objects required, incorrect or redundant objects deleted or updated, new antecedents or pre-conditions added, incorrect or redundant antecedents or pre-conditions either removed or updated, new rule outcomes added and incorrect or redundant rule outcomes either deleted or updated.

While the experimental simulation involved identification of both objects and system behaviour (rules) the results reported in this paper to demonstrate the role of the developed metrics relate only to instantiations of identified meta-classes. The MDS architecture requires the antecedents and consequents of its production rules consist only of identified meta-class instantiations, so the results are related to possible changes in the system rules.

4. METRICS

This section overviews the notation and definitions of metrics that provide a course-grained comparison between knowledge models. Metrics can measure differences in the same clinical domain expert's interpreted knowledge model at two different times, reflecting either change in the expert's interpretation, or in the extent to which their interpretation has been iteratively captured by the MDS methodology. Differences between knowledge models at similar capture stages can indicate differences between domain experts' interpretations of a guideline, due to differences in the expertise and knowledge of the experts, or to ambiguity in the guideline.

Software was developed to implement some of the metrics by examining their underlying knowledge models where objects and attributes were represented in a relational database (Microsoft Access). Comparison between several experts' knowledge models was done by merging identified objects in their knowledge models into a single combined database, and then applying SQL queries to indicate commonalities and differences between individual models and the combined model. The objects in the combined data model formed a standard, or normalised, dataset of identified concepts for the clinical domain.

Counts of knowledge model artefacts represented in the MDS architecture provide simple metrics to identify and track differences between knowledge models. These artefacts include the classes and instantiated objects of the system. By implication they also reflect changes to the production rules representing the behaviour of these objects in the system, as the antecedents and consequents consist only of these instantiated objects.

4.1 Inter-Model Object Comparison

The *Same* class comparison metric: $Same\_KM_a\_P_b\_KM_c\_P_d\_N(X\_X)$ represents the number of concepts identified by two clinical experts, P_b and P_d, that are within the same meta-class X for P_b*'s* knowledge model at time t_a and also P_d*'s* knowledge model at time t_c. As mentioned in the last section, the time references may not have any significance apart from acting as indexes to indicate a specific knowledge model for a clinical expert. An example of this denotation is $Same\_KM_3\_D1\_KM_2\_D3\_N(Action\_Action)$ which indicates the number of common concepts of the Action Object meta-class that are in clinical expert D1's knowledge model KM_3 and also in clinical expert D3's knowledge model KM_2.

The Same class comparison metric represents a rough indicator of the similarity of the two knowledge models. Its complement is the *Different* Class Comparison metric $Different\_KM_a\_P_b\_KM_c\_P_d\_N(X\_X)$ which produces a count of the number of concepts identified and categorised into a class X in one model but not in the other.

Other metrics can be denoted analogously. For example, $Common\_KM_a\_P_b\_KM_c\_P_d\_N(X\_X)$ might denote a count of the number of identical concepts identified and categorised into a class X in two models produced by two experts.

These metrics can indicate the amount of cognitive dissonance that may be present between experts, and that may potentially be avoided, if the experts were consulting together in the development of a single consensual knowledge model.

Although the present study restricted categorisation to the meta-classes, further refinement involving sub-groupings of the above metrics are possible For example, a metric can be formed from the count of common concepts identified that are categorised in the same meta-class X, but in different subclasses (say, S and T) of X. This count is therefore of concepts which have not been interpreted in exactly the same way, but have had similar interpretations in the two models, and would be denoted $Similar\_KM_a\_P_b\_KM_c\_P_d\_N(X.S\_X.T)$.

4.2 Intra-Model Change Metrics

Longitudinal tracking metrics are formed by differencing a count metric applied to two knowledge models created by the same clinical expert at different times. The Δ symbol and the two knowledge models are adjoined to the metric name to denote differencing.

Thus
$\Delta Common\_KM_a\_KM_b\_P\_N(X)$ represents the number of *identical* concepts identified in meta-class X for clinical expert P*'s* knowledge model at time t_b that were also identified in P*'s* knowledge model at time t_a, where $t_a < t_b$.
Likewise for
$\Delta New\_KM_a\_KM_b\_P\_N(X)$ (the number of *new* concepts identified in meta-class X),
$\Delta Deleted\_KM_a\_KM_b\_P\_N(X)$ (the number of identified concepts in meta-class X for clinical expert P*'s* knowledge model at time t_b that have since been deleted from P's knowledge model at time t_a, where $t_a < t_b$; and
$\Delta Modified\_KM_a\_KM_b\_P\_N(X)$ (*modified* concepts, that is, concepts shared between the two knowledge models but the text label identifying them has been altered).

These metrics allow comparison of individual expert's development of knowledge models over time, as the metrics are directly influenced by the identification of the class and the choice of its granularity made by each individual.

5. DISCUSSION

A comparison between the categorised objects in each expert's final knowledge model was made. The metrics used were the counts of *Common* and *Different* objects for different participants. Pearson's correlation coefficient was used to investigate their commonalities, and the results are shown in Figure 1.

The process of identification of objects was well correlated between participants, consistent with the claim that each

participant is conducting a similar process of class identification. The MDS provides a consistent process for the formulation of a knowledge model (this does not imply that the MDS will produce exactly the same knowledge model for participants, but that the processes of the MDS are consistent between participants in producing a knowledge model and also that a knowledge model will be produced).

Correlations between counts of meta-class objects in the final knowledge models are reported in Figures 2a-c. This included concepts identified as *System Objects* by participants as this category contained potential Temporal Objects such as *People* and *Places*. Figure 2a shows the correlation for the total count of objects identified in these categories. The identified concepts were further broken down into categories dependent upon whether they represent a concept identified by more than one participant (*Same* category), which could be viewed as belonging to an ontology representative of shared knowledge between the clinical experts, or whether it was a concept only identified by one clinical expert (*Different* category), with a possibility of it being exclusive knowledge to that expert. The commonly identified concepts are reported in Figure 2b and the different concepts in Figure 2c. The total count of identified concepts (Figure 2a) for each participant is given by the number of *Same* (Figure 2b) + *Different* (Figure 2c) concepts in each MDS category.

Figure 2b shows a high degree of correlation between the *Same* concepts of participants. This indicates shared knowledge between two or more participants who described a commonly interpreted knowledge model of domain concepts, ie the guideline contains a commonly interpreted knowledge set.

The lower correlation values for the *Different* concepts in Figure 2c is explained by assuming there will be knowledge that is not shared by participants and that this will not be related between participants, ie there are different interpretations of the guideline by each domain expert.

The final knowledge model consisting of the commonly categorised objects by all participants could be used as a baseline model to judge degrees of change when new cases or other expert interpretations are introduced. The set of Different objects can be further investigated to see whether they represent differing expert experience or if they might identify tacit knowledge.

6. CONCLUSIONS

Metrics that provide an indication of the robustness of a system model to change and give a measure of its maintenance effort were proposed. An experimental study illustrating the use of these metrics was reported.

Although the sameness of objects identified by experts was qualitatively judged (after the MDS categorisation process) by the knowledge engineer in the experimental study, application of similarity (*Sim*) functions such as those proposed by Colloc (Colloc and Summons, 2006) would allow a more formal identification of similarity and differences in the categorisation of identified objects.

The application of metrics may allow clinical guidelines to be reviewed to identify the information they actually contain compared to the information they are intended to contain (and which domain novices might need to interpret). This would help in auditing the effectiveness and interpretation ability of specific guidelines for novices, where the MDS process is applied (using the novice in place of the expert) to gauge their interpretation of the guideline by building successive knowledge models. The metrics discussed would be allow comparison of the completeness of the novice's model compare to a previously completed expert's, and would also provide an indication of the evolution of the effectiveness of the novice's "learning".

The indication of the amount of commonality and difference in interpreted knowledge could also be used to audit a guideline that has been altered from a national guideline to reflect local demographics. This may allow guidelines to be improved so they reflect and model the experience and expertise of the local clinical practitioners. It may also help new practitioners in their familiarisation with local practices and expertise.

7. REFERENCES

Cleal D.M. and N.O. Heaton. 1988. *Knowledge-Based Systems: Implications for Human-Computer Interfaces* Ellis Horwood Ltd, Chitchester.

Colloc J. and P.F. Summons. 2003. "Modelling the Temporal Information in Medical Decision Support Systems" *Proceedings of the Computational Engineering in Systems Applications CESA'2003 IMACS Multiconference*, Lille, France. (Jun).

Colloc J. and P.F. Summons. 2006. "An Object Oriented Time Model of a Decision Support System for Intoxication Diagnosis" *Proceedings of the 4th International Industrial Simulation Conference ISC'2006*, Palermo, Italy. (Jun), 387-392.

Sanders G. 1998. "Automated Creation of Clinical Practice Guidelines from Decision Models: Gillian Sanders Doctoral Thesis", Stanford University, *CS-TR-98-1609 Technical Report*.

SANPIS. 2003. "South Australian Nurse Practitioner Information Sheet on Implementing Clinical Guidelines" http://www.dhs.sa.gov.au/nurselabour/documents/NuPrac/NursePractitionerKit/InfoKit_Sheet4_NPImplementingClinicalGuidelines.pdf electronic reference, 20th Aug, 2003.

Summons P.F. 2005. *A Methodology to Facilitate Knowledge Acquisition and Maintenance for Decision Support in the Medical Domain : Interpretation of Clinical Guidelines*, Thesis, University of Newcastle, Australia.

Summons, P.F. and J. Colloc. 2006. "Modelling Temporal Aspects of Clinical Objects" *Proceedings of the 4th International Industrial Simulation Conference ISC'2006*, Palermo, Italy. (Jun), 454-458.

Summons, P.F., and H. Villanueva. 1996. "Object Class Identification for a Fetal Health Decision Support System" *Proceedings of the Third World Congress on Expert Systems WCES'96,* Vol 1, Seoul, Korea. (Feb), 194-201.

BIOGRAPHY

Peter Summons was a Systems Engineer in industry. He holds degrees of Graduate Diploma in Education, Bachelor of Computer Engineering, Master of Computing and a PhD in Information Systems. He lectures at The University of Newcastle, where currently he is a member of the Faculty of Science and IT. His main research interests are health informatics, DSS and system design.

Joël Colloc is professor in computer sciences at the Lumière Lyon 2 University. His main research topics are DSS and knowledge bases in medicine, CBR and multi-agent systems, nervous system modelling using reactive MAS and genetic approaches. He holds a Ph.D. (1990) in computer sciences by the National Institute of Applied Sciences (INSA) of Lyon and a MD (1985) by the medical faculty of the Claude Bernard University (Lyon 1).

Correlation - Action Objects

	D1	*D2*	*D3*	*D4*
D1	1			
D2	0.95	1		
D3	0.94	1.00	1	
D4	0.86	0.91	0.94	1

Correlation - Association Objects

	D1	*D2*	*D3*	*D4*
D1	1			
D2	0.92	1		
D3	0.91	0.79	1	
D4	0.68	0.86	0.64	1

Correlation - Temporal Objects

	D1	*D2*	*D3*	*D4*
D1	1			
D2	0.42	1		
D3	0.93	0.51	1	
D4	0.86	0.59	0.97	1

Figure 1 Correlation of Identified Action, Association and Temporal Objects by All Participants

	D1 Total	D2 Total	D3 Total	D4 Total
Action	63	63	56	40
Association	16	20	18	12
Temporal	7	16	13	17
Other	32	43	52	54

Correlation	D1 Total	D2 Total	D3 Total	D4 Total
D1 Total	1			
D2 Total	0.98	1		
D3 Total	0.89	0.95	1	
D4 Total	0.64	0.76	0.91	1

Figure 2a Correlation - *Total* Concepts in Final Knowledge Models (All Participants)

	D1 Same	D2 Same	D3 Same	D4 Same
Action	51	51	50	35
Association	14	17	15	10
Temporal	7	8	10	11
Other	31	40	48	49

Correlation	D1 Same	D2 Same	D3 Same	D4 Same
D1 Same	1			
D2 Same	0.98	1		
D3 Same	0.92	0.98	1	
D4 Same	0.73	0.84	0.94	1

Figure 2b Correlation - *Common* Concepts in Final Knowledge Models (All Participants)

	D1 Different	D2 Different	D3 Different	D4 Different
Action	12	12	6	5
Association	2	3	3	2
Temporal	0	8	3	6
Other	1	3	4	5

Correlation	D1 Different	D2 Different	D3 Different	D4 Different
D1 Different	1			
D2 Different	0.76	1		
D3 Different	0.93	0.70	1	
D4 Different	0.05	0.53	0.27	1

Figure 2c Correlation - *Different* Concepts in Final Knowledge Models (All Participants)

Figure 2(a-c) Meta-Class Correlation in Each Expert's Final Knowledge Model

DISCRETE AND STOCHASTIC SIMULATION AND RESPONSE SURFACE METHODOLOGY: AN APPROACH TO A VARYING EXPERIMENTAL ERROR

Lucia CASSETTARI
Roberto MOSCA
Roberto REVETRIA
Flavio TONELLI
University of Genoa
DIPTEM Dipartimento di Ingegneria della Produzione,
Termoenergetica e Modelli Matematici
Via All'Opera Pia 15, 16145, Genoa, ITALY
E-mail: cassettari@diptem.unige.it, roberto@itim.unige.it

KEYWORDS
Discrete Stochastic Simulation, Response Surface Methodology, Experimental Error, Industrial Plants, Complex Systems

ABSTRACT

The aim of the Response Surface Methodology (RSM), usually applied to chemical, physical and biological systems, is to represent a reliable response surface with a strong adherence to the described reality. In order to achieve this target RSM researchers focused on experimental designs useful to calculate a regression meta-model featured by a good fit to the experimental data and a sort of stability of the prediction variance. Furthermore the design should allow to evaluate two experimental error components, namely pure error and lack of fit. Achieving this with the smallest number of experimental runs it is important to save experimentation time and costs. In classic RSM approach the Experimental Error (EE) is considered fixed and dependent only from the physical system under study. Several experimentations, though, have been performed on simulation models representing the physical system, as it isn't possible to experiment in a direct way. Complex industrial plants belong to this category because of the difficulty (i.e. time and investments required) to modify them in order to perform an experiment. In Discrete Event Stochastic Simulation (DEVS) the EE is simulation time or replications dependent. This requires a substantial change of approach for EE estimation. In this work Authors propose an improved approach and outline limitations of traditional ones.

1. INTRODUCTION

RSM researchers identify at least ten features to be considered when an Experimental Design (ED) has to be chosen particularly a reasonable distribution of data points throughout the region of interest, the possibility to provide an internal estimate of error and allow a precise estimates of the model coefficients and to ensure a good profile of the prediction variance throughout the experimental region (Box and Draper 1987; Myers and Montgomery 1995). The ED has to approximate, by using a regression meta-model, relationships between independent factors and one ore more dependent variables (namely target functions).

The EE aspect greatly influences the response surface quality. It should be distributed like a NID(0;σ^2) and its best estimator is the Mean Square Error (MS_E) (Montgomery 2005). The researcher only knows MS_E when the modelling phase has been completed through the squared sum of residuals, where residuals are the difference between the experimental value y and the estimated one $\hat{y}$ inside the k-dimensions domain:

$$e = y - \hat{y} \qquad SS_E = \sum e^2$$

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{\text{degrees of freedom}}$$

Theoretically, $\hat{\sigma}^2$ should be constant inside the experimental domain (Myers and Montgomery 1995). This value has great importance on the quality of the estimated response because of its influence on several fundamental elements of regression meta-models such as:

- Variance of individual regression coefficients $V(\beta_i) = \hat{\sigma}^2 (X'X)^{-1} \quad i = 0,1,...,p$
- Meta-model variance $V(\hat{y}(x)) = \hat{\sigma}^2 (\underline{x}_0'(X'X)^{-1}\underline{x}_0)$

These two elements are used to calculate the confidence intervals on the individual regression coefficients *(1a)*, the confidence interval on the mean response *(1b)* and the prediction interval for the future observation *(1c)* , like the following formula show:

$$b_j \pm t_{\alpha/2,N-p-1}\sqrt{\hat{\sigma}^2 Cjj} \quad (1a)$$

$$\underline{x}_0'\underline{\hat{\beta}} \pm t_{n-p-1,\alpha/2}\sqrt{\hat{\sigma}^2(\underline{x}_0'(X'X)^{-1}\underline{x}_0)} \quad (1b)$$

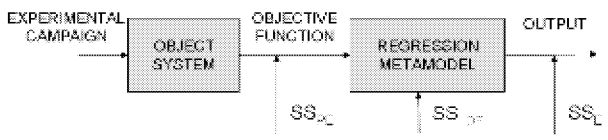
$$\underline{x}_0'\underline{\hat{\beta}} \pm t_{n-p-1,\alpha/2}\sqrt{\hat{\sigma}^2(1+\underline{x}_0'(X'X)^{-1}\underline{x}_0)} \quad (1c)$$

where C_{jj} is the jj elements of $(X'X)^{-1}$ matrix , p x p symmetric matrix, of which the multiplication for $\hat{\sigma}^2$ is the variance of $\hat{\beta}_j$.

In classic applications of Design Of Experiments (DOE) and RSM being $\hat{\sigma}^2$ specific for each analysed reality, and thus outside the control of researcher, and since the $V(\hat{y}(x))$ depends from the position inside the operability region (because function of $\underline{x_0}$), it is easy, according to these assumptions, to conclude there are not improvement margins of identified the response surface meta-model. Starting from these evident limitations several DOE and RSM researchers, since the '50, tried to improve the regression meta-model quality inside the experimental domain, by introducing concepts like optimal variance and orthogonality in order to obtain $V(\beta_i)$ smaller for first order models and rotatability for second order models. In particular by using rotatability, a reasonable stability of $V(\hat{y}(x))$ was searched for, in order to have an identical value of it for each concentric hyper-sphere inside the design domain, and avoiding, in this way, portions of the studied region in which the meta-model response would be degraded due to excessive variance values (Myers and Montgomery 1995). But since the EE is outside of control, as a matter of fact, it is not convenient trying to improve the response quality by acting only on optimal variance, orthogonality and rotatability, being present in them, as a multiplying term, the EE variance.
On the other side if the researcher has the possibility to improve the EE (and its variance), he/she should spend more time in running simulation model (and/or increase the number of replications), in order to reduce the MS_E value. In this work traditional approach will be reported and limitations outlined. A new approach will be discussed. Nowadays a complete experimental campaign of a simulated complex industrial plant is under study in order to show the validity of the proposed approach trying to overcome RSM limitations for complex simulated systems.

2. THE EXPERIMENTAL ERROR IN TRADITIONAL DOE APPLICATIONS

In traditional DOE and RSM applications the experimental campaign is consequent to the experimental design phase and it is performed on the real system. From the real system samples, corresponding to experimental conditions, are collected and target functions calculated. The Sum Squares Error (SS_E), in which all the errors in terms of experimental and modelling nature are included - respectively expressed like Sum Squares Pure Error (SS_{PE}) and Sum Squares of Lack of Fit (SS_{LOF}) – allows to put in evidence the share of the Sum Squares Total (SS_{yy}) (also called total variability) that is not understood from the experimenter (Montgomery 2005). In figure 1 has been reported the way in which the error is partitioned between experimental phase and meta-modelling phase.



$SS_E = SS_{PE} + SS_{LOF}$

Fig.1

The SS_{PE} is the component related to the pure experimental error, not controllable by the researcher, and can be expressed by:

$$SS_{PE} = \sum_{i=1}^{a}\sum_{u=1}^{n_i}(y_{iu} - \bar{y}_i)^2 \qquad MS_{PE} = \frac{SS_{PE}}{N-a}$$

where:

- y_{iu} are obtained experimental responses replicated for the same experimental level,
- $\bar{y}_i$ is the average of the observations done for the same i-level of independent variables,
- N is the total number of observations,
- a the number of levels of independent variables, and
- n_i the number of replications done for the i-level.

The SS_{LOF} is the component related to the lack of fit of the model :

$$SS_{LOF} = \sum_{i=1}^{a} n_i(\hat{y}_i - \bar{y}_i)^2 \rightarrow MS_{LOF} = SS_{LOF} / (a-p)$$

where p is the total number of regressors β appearing in the meta-model.

3. THE EXPERIMENTAL ERROR IN INDUSTRIAL DISCRETE EVENT SIMULATION APPLICATIONS

When the studied system is a complex industrial system, operating in a highly stochastic discrete event environment, a direct intervention on the system in order to experiment new configuration it is not possible due to costs and time required to implement it. In the last two decades, simulation models have been extensively used in order to perform not invasive experiments of industrial systems. Simulation adoption allows to test almost every type of process innovation and computers response time in running simulation model is so fast to test an high number of experimental configurations under many hypotheses.
If advantages of such approach are absolutely evident, the modelling phase produces, according to the previous paragraphs, a double additional error component about which the experimenter has to make some consideration. The first error component is directly connected to the translation of the reality into a simulation model while the second one is related to the transformation of the pure error (Figure 2). The pure error, indeed, from time invariant becomes variant on the simulated time (or replicated runs) according to the following formula:

$$SS_{PE}(t_i) = \sum_{j=1}^{n_o}\left[y_i(t_i) - \bar{y}(t_i)\right]^2$$

$$MS_{PE}(t_i) = \frac{SS_{PE}(t_i)}{n_o - 1}$$

where n_0 is the number of experimental replications or, in other words, the number of replicated runs of the simulation model (Mosca and Giribone 1982). According to the Central Composite Design theory, the pure experimental error can be calculated at the experimental design centre by replicating the so called central runs (Myers and Montgomery 1995) .

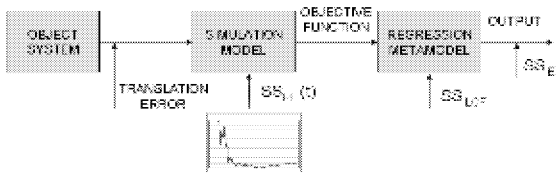


Fig.2

Under these conditions the MS_{PE} assumes, over the simulated time, a knee shape with asymptotic trend towards the noise error intrinsically related to the real system.
Starting from this assumption one can observe the first fundamental difference (from the experimental point of view), between the two before mentioned situations:

- for direct interventions on real system the pure error used by the RSM is an immutable value independently from its magnitude,
- for interventions mediated by a simulator the pure error is, within certain boundaries related to the intrinsic system stochasticity, controllable by the experimenter who can decide the amount acceptable for his/her experiment.

Concerning this last point, looking at the Fisher Test of Lack of Fit, featured by the following statistical summary:

$$F_0 = \frac{MS_{LOF}}{MS_{PE}}$$

It is possible to observe that whether MS_{PE} is high, so that the investigated reality is affected by high variability, the test is passed and thus H_0 hypothesis (a lack of fit is not existing) , is accepted even in presence of significant values of MS_{LOF} (corresponding to regression meta-models featured by low precision in terms of experimental data fit).
On the other hand as much MS_{PE} decreases as MS_{LOF} shall decrease in order to continue to accept the H_0 hypothesis because the F value, for the same confidence level α and degree of freedom, is equal.
As a matter of fact, simulated systems with low EE require, according to Fisher, meta-models able to express adequate fitting capabilities. So that a first fundamental point about RSM application on industrial plant simulators is to seek for an appropriate run length in order to achieve a low value of MS_{PE} but compatible with a reasonable fit of regression model to experimental data.

4. THE PROPOSED APPROACH

Starting from this last consideration it should be evident that the priority has to be assigned to a modelling and experimental phase in order to obtain MS_{PE} values, and thus MS_{LOF} and MS_E values lower than possible but compatible with fitting capabilities of regression meta-models. This approach allows a reduction of confidence and prediction intervals width over the entire domain improving the quality of the model. Only at this point experimenter can focus his/her attention on optimal variance, orthogonality and rotatability, features not always achievable in studying complex industrial systems and for which gain margins, in terms of response quality, are lower than those achievable with MS_E reduction. Figure 3 outlines that the choice of a well designed experimental phase, able to satisfy both Fisher's tests, brings to, if MS_E values are high, a low quality of response even if features like optimal variance, orthogonality and rotatability are achieved.

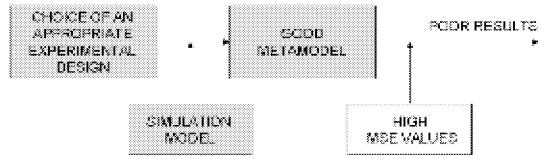


Fig.3

In order to have good results, in other terms confidence intervals width sufficiently close to average response, a small MS_E is required (Figure 4).

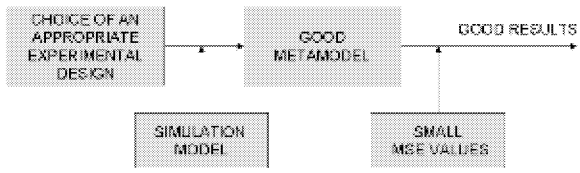


Fig.4

Figure 5 provides complete vision of the proposed approach.

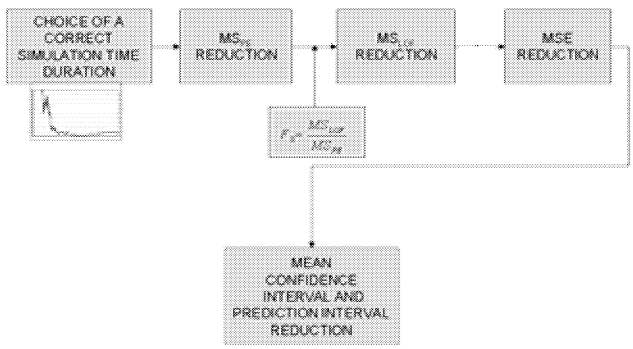


Fig.5

A remarkable point talking about RSM is the Scaled Prediction Variance V_x; it is interesting to outline that its application to industrial system simulation it looses part of significance at least with respect to the meaning given from experimental statisticians (Myers and Montgomery 1995). The Scaled Prediction Variance is defined as following:

$$V_x = \frac{NV(\hat{y}(x))}{\hat{\sigma}^2} \qquad (2)$$

An explanation to this is related to the need of design and management of the industrial plants to have high precision meta-models. As a matter of fact the quality performance

indicator of an experimental design calculated through the Scaled Variance is contrasting the target of this specific type of investigation because the N term (the number of runs) in (2) penalizes the most expensive and exhaustive experiments.
Thus, from the methodological point of view, three are the key steps that shall be considered to accurately plan a response surface design:

1. Build a simulation model in order to achieve the best adherence with respect to the studied reality. Errors done in this phase would bring, for sure, to incorrect or poor results even if the experimental campaign process would be perfect.
2. Chose a simulation time or a replication run number able to guarantee the smallest MS_{PE} but according to the real system intrinsic stochastic level, the available time to conduct the experiments, and the Lack of Fit (Frigato et al. 2001). This phase is determinant when experimenter is facing optimization problems and influences the design success because of its impact on both MS_{LOF} and MS_E. Moreover, through this phase, fundamental elements of regression meta-model like $V(\beta_i)$ and $V(\hat{y}(x))$ are estimated; from this estimation confidence intervals and prediction intervals are calculated. In an analogous way a benefit on rotatability is achieved with respect to its extensive meaning of $V(\hat{y}(x))$ stabilization feature of for second order models.
3. Since it is not possible to evaluate a priori the adopted experimental design quality, because before the beginning of experimental campaign it is impossible to define the order and the shape of regression meta-model, the experimenter should chose mainly canonical designs (factorial, central composite, etc…). They could be, if required, improved through the introduction of additional experiments in order to achieve the desired fit of appropriate regression meta-models. So the analysis on goodness of the adopted design in terms of variance distribution, rotatability, and orthogonality will be carried out after model validation.

5. RSM APPLICATION TO AN INDUSTRIAL PLANT SIMULATION MODEL

The case study used to test the before mentioned analysis refers to a production shop floor (iron components), in which the job routing is represented according to the below scheme:

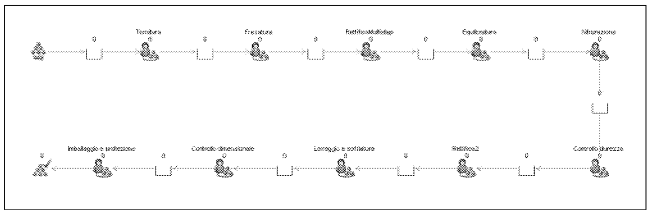


Fig.6

In the following table (Table 1) the number of machines for each job type has been reported; in particular statistical parameters such as frequency distribution, working times, setup times have been detailed. When setup time is not reported it has to be considered of hidden type (off-line set up).

Table 1

Machine Type	Machine number	Freq. Distrib. Job time	Statistical parameters	Set up Time
Lathe	8	Triangular	3-4-5	10' every 180'
Milling machine	8	Uniform	3-5	
Multistep grinding machine	3	Normal	μ=3,8; σ=1,5	
Balancing machine	6	Uniform	1-2	7' every 100 pz
Nitriding machine	14	Normal	μ= 4,5;σ=1	
Hardness control machine	5	Triangular	1-2-3	
Grinding machine	7	Normal	μ=2,4; σ=0,5	
Cleaner	5	Fixed	2,5	
Dimensional control device	6	Normal	μ=6,8;σ= 1,5	
Packaging	4	Uniform	1-2	

Items inter arrival statistical law follows a negative exponential distribution featured by an average value of 1 item per minute; the transfer process is realized by using a belt conveyor.
The line manager is going to study the influence of multi-step grinding machines and dimensional control devices on average daily production and considering a working shift of 8 hours.
According to the pre-set of machines and to the maintenance level it has been observed that the efficiency can vary continuously, with respect to the nominal value, over the following ranges:

- 2,90 ÷ 3,50 for multi-step grinding machines;
- 5,80 ÷ 6,20 for dimensional control devices.

By using Simul8, a stochastic discrete event simulation framework, a model of this production shop floor has been built. Hence Authors run three different experimental campaigns considering three simulation run times respectively 30 working days, 1000 working days, and 3650 working days, in order to reach a progressive decrease of pure experimental error as shown in the following graph (Figure 7):

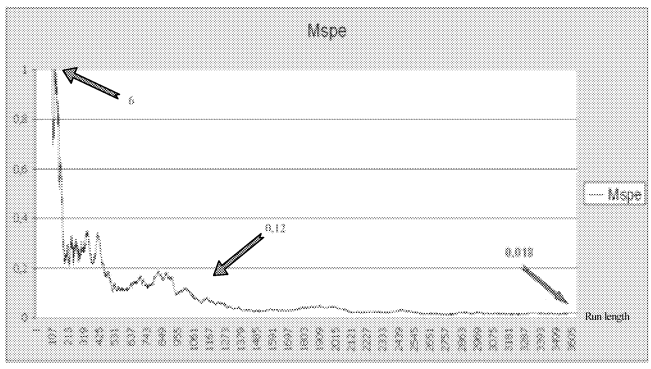
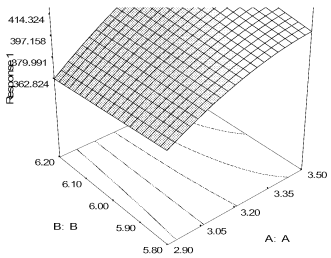


Fig.7

By observing the three different response surfaces, obtained with the Design-Expert statistic tool developed by Stat-Ease, Inc. the differences between them are not particularly evident (Figure 8).

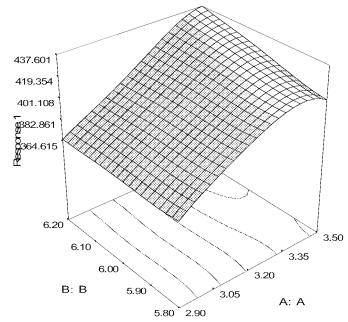
30 days quadratic model

$$\hat{y} = +401.21+27.13*A+3.48*B-7.37*A^2-0.28*B^2+7.21*AB$$



1000 days cubic model

$$\hat{y} = +403.99+36.11*A+0.52*B-9.07*A^2+7.07*AB-7.39*A^3+6.61*A^2B$$



3650 days quartic model

$$\hat{y} = +403.90+36.46*A+1.76*B-2.33*A^2-3.18*B^2+5.87*AB-7.72*A^3-1.75*B^3+7.04*A^2B-6.77*A^3+3.15*B^4+1.13*A^3B$$

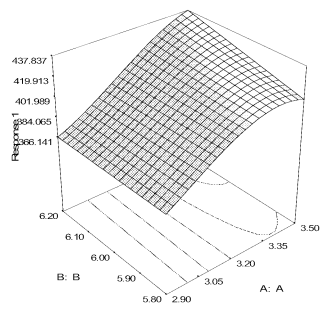


Fig.8

Considering the analysis of contour lines it is possible to observe that the trend with variable A varyng between 2.90 and 3.20 seems to be the same for the three models being lines parallel to B:B axis. Trend differences are evident inside 3.20 – 3.50 interval, in which it is possible to get evidence of a progressive curving phenomena of contour lines around the point P (3.50;6.20). Response differences, instead, depending on the different experimental error quantity (MS_E), featuring the three models, are significant. This particular behaviour can be quantitatively evaluated from confidence intervals on mean response as reported in figures 9, 10 e 11. Moreover, an other important aspect is that confidence intervals are not completely comprised each inside other, contrasting the common sense. For example in figure 10, representing the section where the variable A is fixed to 3.35, it is possible to see that the punctual responses are strongly different depending on the adopted model.

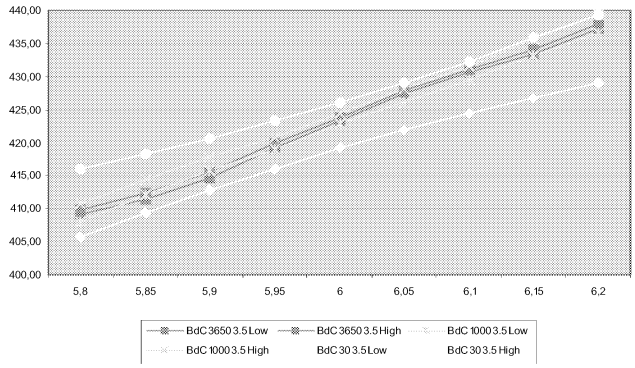


Fig. 9

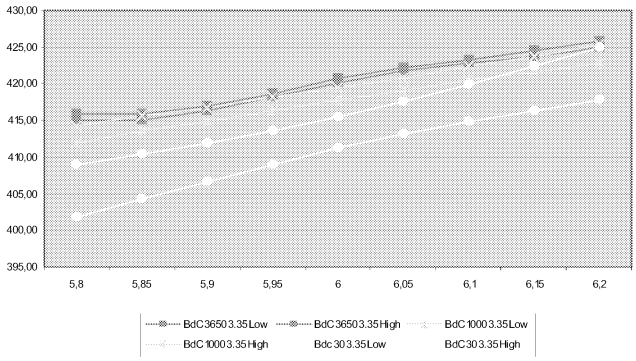


Fig. 10

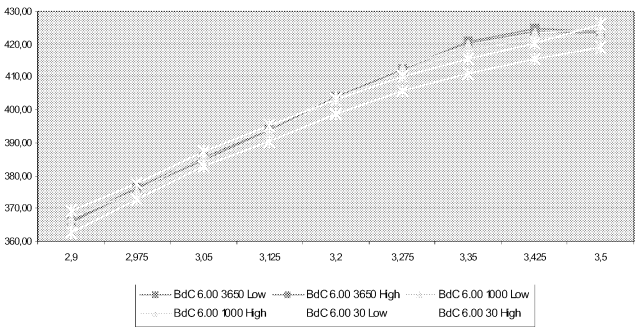


Fig. 11

From the comparative analysis of the three surfaces and related confidence intervals it is possible to conclude that the minimum of time varying error searching process allows, indeed, to identify behaviour about investigated reality with a further precision level, avoiding important estimation errors in determining the casual relationship between independent and dependent variables. Observing these results the main target of this study could be considered

achieved. In fact, in stochastic discrete event simulation context, the adoption of reduced MS_{PE} values and the sequential reduction of MS_{LOF} and thus of MS_E allows to achieve the relationship between the independent variables and the dependent one by using the peculiarity of a time varying experimental error. It is important to put in evidence that acting in this way the experimenter does not follow the typical fundamental assumptions featuring the canonical theory of RSM. RSM is constrained to the experimental error intrinsically related to every explored reality (the amount of error can only be taken into account), and focuses on searching a fair balance between the experimental points number and the regression meta-model quality (which usually has to be high). This can mainly be obtained through the achievement of maximum possible stability of model variance, for the observed domain, by seeking for specific properties like Optimal Variance, Orthogonality, Rotatability.

Authors don't consider the proposed view about simulation experiments like an heresy with respect to classic methodology; simply Authors would like to outline the need to investigate the behavior of high complexity systems, generally featured by high stochasticity levels, requiring model building process able to express a good reliability and precision since decision makers have to evaluate investments suitability. On the other side Authors would like to put in evidence that MS_E reduction, obtained through longer experimental campaigns with respect to those proposed by canonical RSM, doesn't influence the economic balance of experimental campaign because it can be translated in a longer calculation time.

5. CONCLUSIONS AND FURTHER RESEARCH

RSM provides over the same experimental domain, when simulated systems with time varying error are studied and in contrast with steady error related to physical systems, different opportunities to analyze the relationships between independent and dependent variables inside the same region of operability.

The experimenter should identify, according to the response accuracy requirements and starting from the MS_{PE} curve (consequently from a pre-set experimental level), a regression meta-model featured by an adequate fit level. The experimenter should also be aware about the fact that in order to validate Fisher's second test, smaller is MS_{PE} value, smaller has to be MS_{LOF}. This could lead to the need of high order meta-models, which are not suggested by canonical RSM philosophy and which denote a worrying lack of robustness in terms of sensibility. As a matter of fact over an experimental domain would happen to find the correct fit even with very different regression meta-models; this fact outlines the unreliability through which the target function could be estimated. Moreover, greater is the number of points available for the regression meta-model and greater is the difficulty of regression meta-model to behave according to its nature to find a trend of a cloud of points. For these reasons it seems to be correct to suggest experimenters to proceed carefully about results obtained from RSM, at least until a theoretical base for the identified problem will not be available. Nevertheless, before definitive conclusions are drawn, Authors will verify, through a deep analysis of other appropriate complex industrial plant simulation model, if the limitations described could be connectable to a specific case or if belong to RSM conceptual limitation.

REFERENCES

Box, G.E.P., Draper, N.R., 1987: Empirical Model-Building and Response Surfaces, John Wiley & Sons Inc.

Box, G.E.P., Hunter, W.G., Hunter, J.S., 1978: Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building, Wiley Inc.

Myers, R.H., Montgomery, D.C., 1995: Response Surface Methodology. Process and Product Optimization Using Designed Experiments, John Wiley & Sons Inc.

Frigato, A., Mosca, R., Tonelli, F., 2001: Designed Simulation (DOE) of Complex Systems: New frontiers in using neural networks for R.S. research part one: Theoretical Assumptions, Proc. of WMC'01, January, Phoenix, Arizona.

Frigato, A., Mosca, R., Giribone, P., Pozzi Cotto, S., 2001: Stocastic Optimal Design of Anaerobic Digester for Citrus Residuals, Proc. of ESS'01 – Simulation in Industry, October, Marseille, France.

Mosca, R., Giribone, P., 1982: Optimal Lenght In O.R. Simulation Experiments of Large Scale Production System , Proc. of IASTED International Symposium "Applied Modelling and Simulation", AMS'82, 29June - 2July, Paris, France, 78-82.

Montgomery D.C., 2005: Design and Analysis of Experiments, John Wiley & Sons Inc.

BIOGRAPHY

ROBERTO MOSCA is Full Professor of "Industrial Plants Management" and "Economy and Business Organization" at the DIPTEM (Department of Industrial Production, Thermoenergetics and Mathematical Modelling), University of Genoa. He has worked in the simulation sector since 1969 using discrete and stochastic industrial simulators for off-line and on-line applications. His research work focuses on original application of DOE and RSM for simulation experiment. He is author of about 200 scientific papers published for International Conferences and International Journals.

LUCIA CASSETTARI earned her degree in management engineering in 2004 at the University of Genoa. Currently she is a PhD student at DIPTEM at University of Genoa, in the field of simulator-based applications for industrial plants; particular attention is focused on the application of DOE and optimization techniques to industrial plant problems using simulation. She is also active in research projects involving BPR, Activity Based Management and Cost Control.

ROBERTO REVETRIA is an associate professor at University of Genoa. He earned his degree in mechanical engineering at the University of Genoa and completed his PhD in Mechanical Engineering in 2001. He is currently working on advanced modeling projects applied to ERP integration and maintenance planning applied to industrial case studies.

FLAVIO TONELLI is an industrial engineering researcher at University of Genoa where he teaches Operations Management, Business Management, Organization. He earned a PhD in “production systems and industrial plants management” from University of Parma, as well as degrees from Genoa University. Main research fields are manufacturing and supply chain management Other interests range on workflow and performance management as well as enterprise information systems (ERP, SCM, APS, MES). He is a member of SCS, POMS, EUROMA, SCC, ANIMP.

KRIGING METHODOLOGY

PERFORMANCE EVALUATION OF THE EVOLUTION CONTROL IN DESIGN OPTIMIZATION ASSISTED BY KRIGING SURROGATES

Gabriella Dellino
Università degli Studi di Bari, Dipartimento di Matematica,
Via E. Orabona 4, 70125 Bari, Italy
gdellino@dm.uniba.it

Paolo Lino, Carlo Meloni, Alessandro Rizzo
Politecnico di Bari, Dipartimento di Elettrotecnica ed Elettronica,
Via E. Orabona 4, 70125 Bari, Italy
{lino,meloni,rizzo}@deemail.poliba.it

KEYWORDS
Operations Research, Optimization, Approximation Techniques

ABSTRACT

This contribute deals with the use of Kriging surrogates in multi-objective engineering design optimization. The surrogate management issue to find a tradeoff between accuracy, efficiency and solutions quality is addressed. A Data Envelopment Analysis (DEA) approach to the performance assessment is proposed. The method is applied to carry out a comparative analysis of different strategies for a case study devoted to the design of a component of the injection system for compressed natural gas (CNG) engines.

A METAMODEL ASSISTED OPTIMIZATION SCHEME

In many different engineering areas, expensive computer simulation models are often employed in the design optimization process of complex systems. Aiming at reducing the high computational costs payed for running the simulation model, several approximation techniques (also called surrogates or metamodels) have been proposed in the literature. However, integrating metamodels in a computational optimization process based on classical, evolutionary or meta-heuristic optimizers is not straightforward. According to a recent research stream devoted to explore new perspectives and to advance the state-of-the-practice, this work addresses a pratical method to evaluate the performance of different model management schemes. At this aim, they are analyzed, in order to find a trade-off between metamodel accuracy, computational efficiency, and solutions quality; a comparative analysis is conducted — by following a Data Envelopment Analysis (DEA) approach — on a case study devoted to the design of a component of the injection system for Compressed Natural Gas (CNG) engines, enlightnening the usefulness of the proposed methods in the simulation-optimization based design practice. The optimization system we consider in this paper can be classified as a metamodel assisted multiobjective optimizer. The framework has been developed in the MATLAB environment, employing the Design and Analysis of Computer Experiments (DACE) Toolbox (Lophaven et al. 2002) for the construction of the metamodel. The system is composed of three main blocks, briefly described in what follows: the Design of Experiments (DOE) Tool allows studying the effects of multiple factors on design results; the Metamodel Constructor generates a mathematical model approximating costly computational functions, by using the Kriging method — which provides exact interpolation (i.e., the predicted results at input combinations already observed are equal to the simulated results values at those inputs); the Multiobjective Optimizer is mainly based on Deb's multiobjective algorithm NSGA-II (Non-Dominated Sorting Genetic Algorithm-II) (Deb et al. 2002) and it is adapted to solve constrained problems, by penalizing infeasible solutions. An external solutions archive is used to save the non-dominated solutions found during the optimization.
The proposed implementation supports the use of successive approximation models of a costly fitness function and enables different strategies — usually referred to as metamodel management — to integrate and manage the metamodel in the evolutionary optimization process. A starting set of data samples (i.e. input/output descriptions of computer simulation experiments) is obtained through the Latin Hypercube Sampling (LHS) DOE technique, chosen because of its flexibility and easy construction. These elements are given as inputs to computer simulation experiments; the computed outputs are used to construct a Kriging metamodel (van Beers and Kleijnen 2004, Santner et al. 2003). To integrate the approximated model into an evolutionary optimization process, many strategies can be adopted. The approach followed in this work is based on the evolution control (EC), according to which a fraction of individuals (the whole population, at worst) — in some generations —

is evaluated on the original fitness function. These evaluations are used to update the surrogate during the optimization process with the hope to obtain an accurate approximation in the search region explored by the algorithm. From the EC point of view, the model management in the proposed framework can be addressed by means of three main approaches: i) assuming the approximation model to be of high-fidelity and therefore, not using the original fitness function at all during the evolutionary computation process; ii) adopting the generation-based evolution control; and iii) using the individual-based evolution control. Generation-based evolution control requires to evaluate the entire population either on the surrogate model or on the original fitness function. In the individual-based control only a fraction of the population is evaluated using the original fitness function while the remaining part is evaluated on the surrogate model. In order to select the individuals to be controlled two main strategies are adopted: (a) a best selection strategy requires the best individuals (i.e. from the optimization point of view) in the current generation to be re-evaluated using the original fitness function; in this case, the fraction of individuals to be controlled can be set by the user before the algorithm's execution; (b) a worst selection strategy chooses the individuals to be controlled by the original function on the basis of some measure of the accuracy provided by the surrogate on the current generation; hence, the number of controlled individuals depends on the computed or estimated approximation error. Once the individuals have been selected and evaluated according to one of the previous strategies, the surrogate can be updated. The frequency of the update operation can be either a constant fixed by the user at the beginning of the optimization process, or it can be adaptive and alternatively based on the dynamic of the optimization process and on the fidelity of the surrogate model, estimated through some metrics properly chosen.
In this work, eleven different evolution control strategies are considered as candidates to be applied to an engineering design optimization problem. A set of computational experiments has been conducted to give a comparative analysis of the strategies behaviour in terms of both global and local accuracy ongoing the optimization process as well as in terms of computational efficiency, by keeping in mind that evolution control provides good results if it succeeds in finding a trade-off between reducing as much as possibile the computational cost of fitness functions evaluation and avoiding the risk of misleading the optimization using inaccurate approximation models.

Performance Evaluation with DEA

A performance evaluation of the proposed EC schemes requires to consider different aspects of the entire optimization process. In general, the performance assessment of a system is often a difficult task, because it is a multidimensional construct. One measure, for example a specific error measure, may not give a complete picture of the overall performance. Thus, it is commonly necessary to look at several experimental inputs simultaneously, along with the multiple outputs they produce. We address the study of the performances with a Data Envelopment Analysis (DEA) approach. DEA uses linear programming to search for the optimal combinations of inputs and outputs as revealed in the actual performance of systems, or decision making units (DMU) (Thanassoulis 2001). Hence, the comparative study of the performance of the different metamodel management strategies, considered as DMUs, is conducted under the DEA paradigm. We apply the method to the different EC schemes considering measures of inputs representing, in some sense, the costs associated to each strategy under study. On the other hand, measures of model accuracy and solution quality have been considered as outputs of the different candidate strategies.

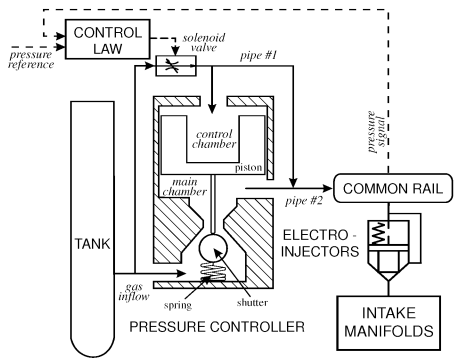


Figure 1: Block scheme of the CNG-Common Rail injection system.

THE CASE STUDY

The case study focuses on the design of an automotive compressed natural gas (CNG) injection system. The considered system is composed of the following elements (Fig. 1): a fuel tank, storing high pressure gas, a mechanical pressure controller, driven by a solenoid valve, and the fuel metering system, consisting of a common rail and four electro-injectors. The controller reduces the pressure of the fuel supplied by the tank, and sends it to the common rail, feeding the electronically controlled injectors. Then the injectors send the gas to the intake manifolds to obtain the proper air/fuel mixture. The injection flow depends only on the rail pressure, which is almost equal to the main chamber pressure, and injection timings.
The simulation-optimization scheme previously described has been employed in the design optimization of the pressure controller, trying to pursue both mechani-

cal and control objectives. From a mechanical point of view, it is necessary to reduce the friction force between the piston and the lateral surface of the pressure controller, in order to shorten the actuation delay, thus improving the efficiency of the system. At the same time, an accurate control of the injection pressure requires the minimization of the gas leaks between the control and the main chambers of the pressure controller, due to the pressure gradient across the piston surfaces. Some control specifications can be included in the optimization process, regarding system response and disturbance rejection. To this end, we have to analyze the main chamber, the control chamber and the common rail filling and emptying behaviours, and the piston dynamics: the former are approximatively described by a first order dynamics, while the latter can be represented by a second order model (Dellino et al. 2006, Lino et al. 2006). The optimization model has been developed on the basis of the targets exposed so far, thus stating a multi-objective constrained optimization problem. Another objective function has been added to those previously described, to take into account some dynamical aspects of the system that cannot be modelled otherwise. The evaluation of this function requires an accurate model and expensive computer simulations; for this reason, it has been approximated using a metamodel.

Performance Assessment

Aiming at evaluating different metamodel management strategies, an experimental campaign has been conducted. Our goals attain to compare the behaviour of the optimization process employing these strategies, and to individuate the configuration(s) able to solve the trade-off between model accuracy, computational efficiency, and solutions quality for the design problem under study. The model accuracy has been evaluated by computing the error in predictions related both to elements of the populations of the EA and to additional observations: both these aspects have to be taken into account in order to guarantee both a local accurate prediction — in the current search area — and a global fidelity of the metamodel. To evaluate these aspects of the metamodel accuracy, different metrics have been used: the maximum and mean Absolute Percentage Error have been computed, together with R^2, the Relative Average Absolute Error (RAAE), and the Relative Maximum Absolute Error (RMAE), thus providing a more complete evaluation of the global accuracy of the model. In order to evaluate the quality of the solutions obtained by the optimization process, several indicators can be adopted and, possibly, combined (Knowles et al. 2006). As for the metamodel management efficiency, the computational effort required for both constructing the surrogate and for predicting the response of the individuals in the entire optimization process has been considered. Moreover, for the different strategies, we are interested in evaluating the computational effort required by the metamodel management itself with respect to that strictly devoted to the optimization by the EA.

On the basis of the computational results on the case study, all tested metamodel management schemes based on Kriging predictions provide accurate approximations. However, schemes with both adaptive or passive metamodel update mechanisms are more suitable to be integrated in the considered optimization process. An adaptive update strategy based on the approximation error and adopting a best selection rule seems to be the scheme to prefer.

Further work directions may concern the use of different optimization algorithms, the metamodel management or the adopted Kriging technique. Concerning the metamodel management, other strategies can be experimented, e.g. based on estimated improvements in the approximation as well as in the optimization. Even if the Kriging technique has provided good metamodels in the case study, different related aspects deserve further analyses and research.

REFERENCES

Deb K.; Pratap A.; Agarwal S.; and Meyarivan T., 2002. *A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation*, 6, no. 2, 182–197.

Dellino G.; Lino P.; Meloni C.; and Rizzo A., 2006. *Multidisciplinary Design Optimization of a Pressure Controller for CNG Injection Systems.* In *Proceedings of the 2006 IEEE Conference on Computer-Aided Control Systems Design.* Munich, Germany, 2689–2694.

Knowles J.D.; Thiele L.; and Zitzler E., 2006. *A tutorial on the performance assessment of Stochastic Multi-objective optimizers.* TIK-report 214, 1–35.

Lino P.; Maione B.; Amorese C.; and Matthaeis S.D., 2006. *Modeling and Predictive Control of a New Injection System for Compressed Natural Gas Engines.* In *Proceedings of the 2006 IEEE Conference on Control Applications.* Munich, Germany, 1651–1656.

Lophaven S.N.; Nielsen H.B.; and Søndergaard J., 2002. *DACE – A Matlab Kriging Toolbox (version 2.0).* Technical Report IMM-REP-2002-12, 1–34.

Santner T.J.; Williams B.J.; and Notz W.I., 2003. *The design and analysis of computer experiments.* Springer-Verlag, New York.

Thanassoulis E., 2001. *Introduction to the theory and application of data envelopment analysis.* Kluwer Academic Publisher.

van Beers W.C.M. and Kleijnen J.P.C., 2004. *Kriging Interpolation in Simulation: a Survey.* In J.S. R.G. Ingalls M.D. Rossetti and B. Peters (Eds.), *Proceedings of the 2004 Winter Simulation Conference.* 113–121.

MODELING OF WATER PRESSURE DISTRIBUTION IN WATER NETS USING THE KRIGING ALGORITHMS[1]

Lucyna Bogdan, Jan Studzinski
Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
E-mail: studzins@ibspan.waw.pl

KEYWORDS

Mathematical modeling, kriging approximation, water net operation, engineering software.

ABSTRACT

While operating a municipal water network it is important for the network operator to know the current water net state, i.e. the water flows in the pipes and the water pressure values in the nodes of the net. With the help of a hydraulic model he can calculate these values but the huge number of them ordered mostly in the form of a table makes it quite impossible to recognize quickly in which part of the water net the pressure is too high or too low or the water flows are too slow. In the case of a badly working water net some improvement procedures are to be executed to avoid possible network failures or claims of the water consumers. In the paper a procedure to visualize quickly and qualitatively the distribution of water pressure in water nets using the kriging approximation method is presented.

INTRODUCTION

The crucial thing while operating and managing a water network is to know what is the state and what are the parameters of its work (Studzinski 2007). First of all it is necessary to have a monitoring system implemented on the network that transfers automatically the values of water pressure and flow from the measurements points located in the water net to the computer located in the room of the water net operator. Such the systems are more and more introduced into communal waterworks but their usefulness to get the whole knowledge about the water net is unfortunately limited because of a rather small number of the measurements points which are usually available. The cause for that is the high price of the measuring devices and of the whole infrastructure to be made to set up a measurement point and as a result only few nodes of a water net are commonly selected for getting the needed instrumentation while a communal water net consists usually of many thousands nodes.

Therefore a next approach for getting the current information about the network state is to develop and introduce in water net its hydraulic model that calculates the desirable pressure and flow values in all nodes and all pipes of the network. Such the models are under development in several Polish waterworks and there are commonly many problems with the exactness of the models and with their fitting to the real objects, i.e. with their calibration using the water net data. To realize the task of calibration the data coming from a monitoring system installed on the water net are very useful.

The good calibrated models are of a great importance for the water net operator and they make him possible to perform the simulation as well as optimization runs of the network under operation but the evaluation of the results received is quite difficult because of the huge number of the data. These data are ordered in a table with few columns containing usually the names of the water net objects, their space coordinates and the values of water flow in the case of network pipes and of water pressure in the case of network nodes, and with many rows whose number corresponds to the number of nodes and pipes of the water net. On the base of these data the operator can estimate very exactly the quality of the water net work but this estimation needs commonly a lot of time because of many data which are to analyse.

Therefore it is useful to give the operator the possibility of a quick estimation of the correctness of the water net work and only when he would state some incorrectness in the quality of this work then he could analyse more detailed the possible reasons of it on the base of the table with the results of hydraulic calculation. To create such the tool for fast recognition of the work quality of the water net we propose to approximate the results of the network

[1] The paper shows the results of the research project of Polish Ministry of Science and Higher Education No. 3T11A01026.

hydraulic calculation for getting a regular net of investigated values which are the water pressure and water flow and to draw then a contour map of them on which the water net parts with right or bad function features could be marked with different colours.

For getting the calculated pressure and flow values of a water net we have developed a program called OHIO and for making the 2D approximation of any parameter values we have developed a program called KRIPOS in which the kriging approximation method is applied. The calculation algorithms used in both programs and the program descriptions are showed in the following.

HYDRAULIC MODEL OF A WATER NETWORK

The hydraulic model of a water net is used to calculate the water flow and pressure values in pipes and nodes of the network. Its mathematical description consists of linear and nonlinear algebraic equations (Studzinski 2006).

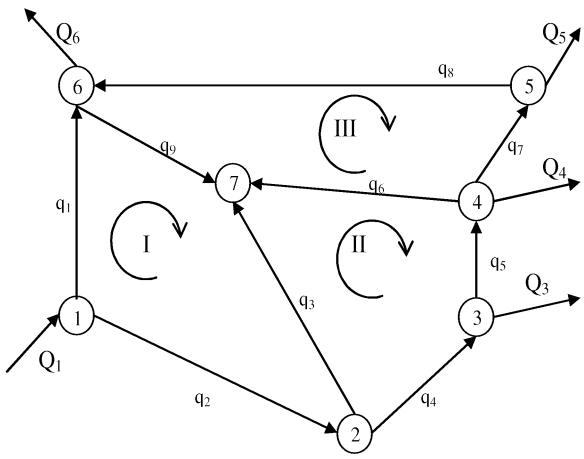


Figure 1. An exemplary water network.

For the network on Fig. 1 there are the following data or notations: notes number K = 7, pipes number R = 9, rings number S = 3, Q_i – water inflow or taking off, q_i – water flow. Our task is to calculate the water flows q_i in R pipes and the water pressures P_i in K nodes knowing the structure and Q_i-values of the water net. There are then (R + K) unknown variables to estimate where R + K = (K + S – 1) + K = (K – 1) + S + K. The model description is as follows:

Linear equations formulated on the base of I *Kirchhoff* law (K equations):

Node 1: $Q_1 - q_1 - q_2 = 0$

.....

Node 7: $q_3 + q_6 + q_9 = 0$

Nonlinear equations formulated on the base of II *Kirchhoff* law (S equations):

Ring 1: $h_{t1} + h_{t9} - h_{t3} - h_{t2} = 0$

..... (1)

Ring 3: $h_{t6} - h_{t9} - h_{t8} - h_{t7} = 0$

Linear equations formulated on the base of *Bernoullie* equation ((K – 1) equations):

Node i: $P_i = P_{i-1} - h_{ti}$

where:

$$h_t = \frac{\lambda l v^2}{2gD} \qquad v = \frac{4q}{\pi D^2}$$

and: h_{ti} – pressure decline in pipe *i*, λ – coefficient of water friction in a pipe; *l, D* – length and diameter of a pipe.

There are two main problems regarding the model solution and they are concerning:

- the choice of the most appropriate combination of the rings to formulate the nonlinear equations (1) (there are 28 rings combinations only for the simple network with 3 rings on Fig. 1),

- the choice of an appropriate numerical algorithm to solve the equations (1).

In our program OHIO the *Cross* algorithm has been used to get the solution of the nonlinear equations. The pressure decline h_t we can write down in a new form:

$$h_t = \frac{8\lambda l}{\pi^2 g D^5} q^2 = sq^2$$

If the equality conditions are not fulfilled in equations (1) then the variables h_{tij} are calculated with a flow correction Δq_i: $h_{tij} = s_{ij}(q_{ij}^2 + 2q_{ij}\Delta q_i + \Delta q_i^2)$ and then the equations (1) are getting the following new form:

$$\sum_{j=1}^{R_i} h_{tij} = \sum_{j=1}^{R_i} s_{ij}(q_{ij}^2 + 2q_{ij}\Delta q_i) = 0 \qquad (2)$$

for $\Delta q_i^2 = 0$. One can calculate now the flow correction from (2):

$$\Delta q_i = -\frac{\sum_j s_{ij} q_{ij}^2}{2\sum_j s_{ij} q_{ij}} = -\frac{\sum_j h_{tij}}{2\sum_j h_{tij} / q_{ij}}$$

The variable Δq_i is to modify until it will get the value under a given accuracy of calculation.

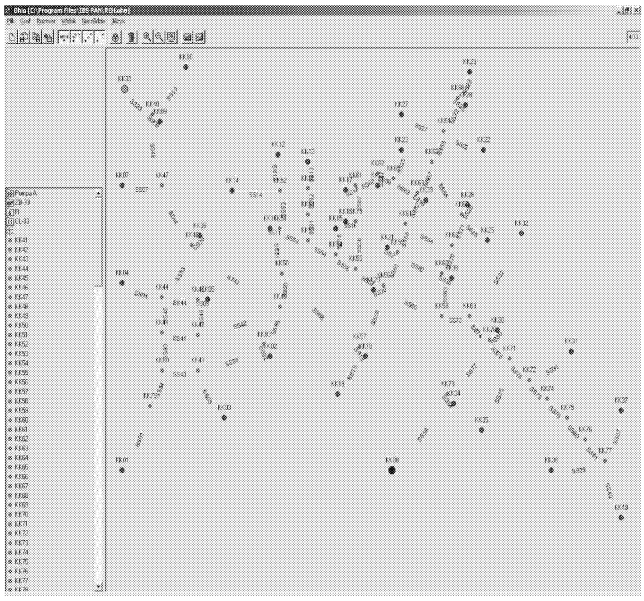


Figure 2. The main window of OHIO program with an exemplary water net graph.

There are many different programs to hydraulic modeling of water nets but our program OHIO written in DELPHI language has some good points which are missing in other programs and they are:

- an user-friendly graphical editor that enables to draw and change a water net graph with a mouse (see Fig. 2),

- the taking off values in the end user nodes can be given in the form of monthly or hourly distribution functions,

- an one-criterial optimization algorithm implemented in the program enables the improvement of the parameters of the modeled water net.

KRIGING APPROXIMATION

The kriging approximation consists of estimation of unknown values of a parameter at some determined points of an investigated area (so called basis points) with the help of the known values of this parameter collected from other points (so called measurement points) of this area.

For the parameter regarded as a random variable the following variogramm function being the main function of geostatistics is defined (Bogdan and Studzinski 2006 A):

$$D^2[Z(x+h) - Z(x)] = E[Z(x+h) - Z(x)]^2 = 2\,\gamma(h)$$

where: D, E – variance and the expected value, $2\gamma(h)$ – variogramm function, $\gamma(h)$ – semivariogramm function, $Z(x)$, $Z(x+h)$ – the random variable at the points x und $(x+h)$ for $h \in (0,+\infty)$.

The values of an empirical semivariogramm for a discrete net of points are to calculate from the formula:

$$\gamma(h) = \frac{1}{2n_h} \sum_{i=1}^{n_h} (z_{h+1} - z_i)^2$$

where: z_i, z_{i+h} – values of the random variable at the points removed by distance h from each other, n_h – number of the point pairs regarding the points removed by h. An exemplary semivariogramm is showed in Fig. 3.

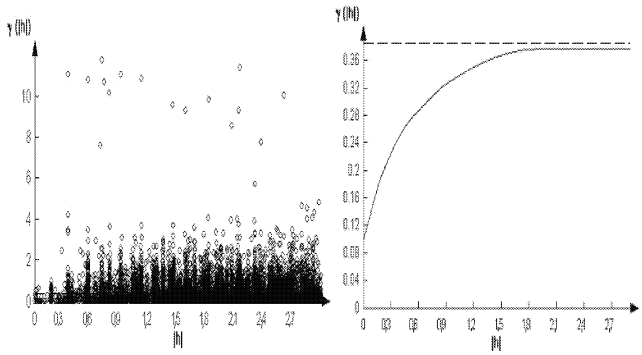


Figure 3. Examples of a semivariogramm and of its mathematical model.

The calculated points of the empirical semivariogramm are then to model with different analytical functions. For this modeling the methods of static optimization or mathematical programming can be used. The estimation of the mean value of the parameter investigated at a basis point occurs with the formula:

$$z_O = \sum_{i=1}^{N} \lambda_i z_i$$

where λ_i are weights coefficients and z_i are the parameter values at the known measurement points.

The main task of the kriging approximation is to calculate the values of the weights coefficients which fulfil the following conditions of the unbiasness and of maximal effectivity:

$$E(z_i - m) = 0 \qquad \text{and} \qquad \sigma^2 = E[(z_i - m)^2] = min$$

From these conditions the following formulas result:

$$\sum_{i=1}^{N} \lambda_i = 1 \qquad \text{and} \qquad \frac{\partial\left[\sigma^2 - 2v\sum_{i=1}^{N}\lambda_i\right]}{\partial z_i} = 0$$

where v is the Lagrange multiplier. From the last formula the system of equations results:

$$\gamma(x_j, x_0) = \sum_{i=1}^{N} \lambda_i \gamma(x_i, x_0) + v$$

for j=1,2,…,N, from which the demanded weights coefficients are to calculate.

PRESSURE DISTRIBUTION

The whole process leading to receiving a map with distributed pressure values in a water net is quite complex: firstly a hydraulic model of the water net must be developed, then the calibration procedure with the help of the real object data and especially these ones coming from a monitoring system implemented on the network must be executed, subsequently a simulation run of the water net to get a table with all its pressure and flow values and finally their approximation using the kriging method have to be made. Having a net with approximated pressure values we can draw a contour map on which the current work state of the operated water net is qualitatively illustrated.

Such the map we have developed for a communal water net shown in Fig. 2 (see Fig. 4). The hydraulic calculation of the network has been made with the self developed program OHIO while the kriging approximation and the contour map drawing were made with the self developed program KRIPOS (Bogdan and Studzinski 2006 B).

Looking on the map the water net operator can recognize very fast the work quality of the network. On our exemplary picture we can state that only a small part of the water net works satisfactory (this part is marked with the yellow, orange and pink colors) while on a big part of the net the water pressure is too high (what is marked with the red and violet colors) and on the end part of the network the pressure is to low (marked with blue and navy blue colors).

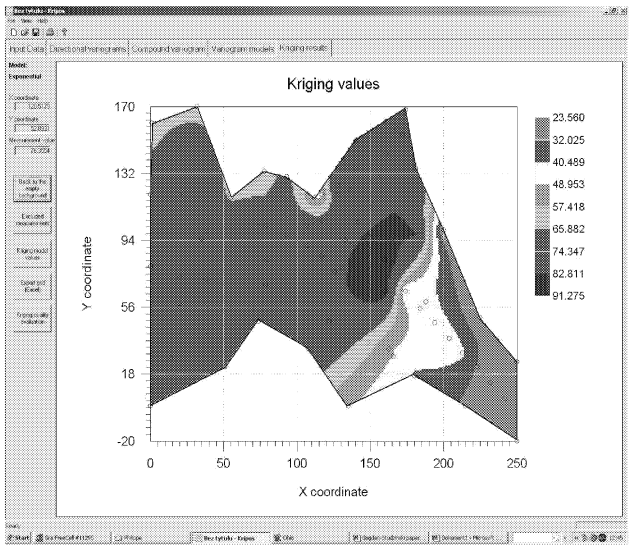


Figure 4. Pressure ditribution in the water net from Fig. 2.

In a real case of management of a water net the operator has to undertake in such the situation some improvement procedures for the low pressure in the water net means the lack of water by the end users and the high pressure can cause some damages to the network. While undertaking these improving procedures the operator shall analyse before the data tables received from the hydraulic model calculation, which will get him more detailed and quantitative information about the nodes with inappropriate pressure values.

CONCLUSIONS

In the paper a procedure to fast and qualitative estimation of the water net functioning has been presented and also the programs making possible the realization of this procedure have been described. In our example only the analyse of the correctness of the water pressure values is made while the work quality of a water net is determined also by the water flows which shall be in a strictly limited values range. When the water flow in a part of a water net is too small then some kinds of alga can arise there what will worsen the water quality.

Therefore it is also important for a water net operator to have a tool for drawing the contour maps of water flows and the work in this direction is currently made by the authors. A problem arising here is that by the pressure values they are given in defined space points what makes their direct approximation possible while by the flow values they are given for the pipe sections and this makes the direct approximation impossible. Because of that a transformation of the sectional flow values into punctual ones must be made before.

REFERENCES

Bogdan L., Studzinski J. (2006 A) *Anwendung von Kriging-Approximation zur Anfertigung der Regenfallkarten fuer Polen*. In: GI-Edition, Lectures Notes in Informatics (K.O. Wenkel, P. Wagner, M. Morgenstern, K. Luzi, P. Eisermann, Hrsg.) Koellen Druck Verlag Bonn, 33-36.

Bogdan L., Studzinski J. (2006 B) *The kriging approximation: algorithms, program and calculation results.* In: Eco-Info and Systems Research (J. Studzinski, O. Hryniewicz, eds.) PAN IBS Warsaw, 67-88.

Studzinski J. (2006) *Entwicklung von Modellen und Algorithmen zur Simulation und Optimierung von komplexen Wassernetzen*. In: Modellierung und Simulation von Oekosystemen. Workshop Koelpinsee 2004 (A. Gnauck, Hrsg.) Shaker Verlag Aachen, 1-12.

Studzinski J. (2007) *Computer aided management of waterworks.* In: Proceedings of QRM'2007, Oxford, 6th Intern. Conference on Quality, Reliability and Maintenance (R.A. Tomas, ed.) 254-258.

COMPLEX SYSTEMS SIMULATION

COMPLEX SYSTEMS SIMULATION METHODOLOGY

FINDNG RELATIONSHIPS THAT ARE FUNCTIONS

William Conley
Business Administration
University of Wisconsin at Green Bay
Green Bay, Wisconsin 54311, U.S.A.
E-mail: conleyw@uwgb.edu

KEYWORDS
CTSP statistic, multivariate correlation, functions.

ABSTRACT

Presented here is the CTSP multivariate correlation statistic applied to problems where the data are hypothesized to represent variables that form relationships that are functions. The examples will consist of a three variable problem (3 x 29 spreadsheet) where the multi stage Monte Carlo optimization (MSMCO) CTSP analysis reveals a nonlinear functional relationship between X_1, X_2 and Y. Then a second example involving six variables (6 x 59 spreadsheet) will be similarly analyzed with the MSMCO CTSP algorithm, showing another functional relationship between X_1, X_2, X_3, X_4, X_5 and $X_6 = Y$. A third example involving seven possibly correlated variables (a 7 x 55 spreadsheet of data) will find the hypothesis of the seven variables being correlated to be false. This is because the shortest route connecting the 55 seven dimensional real data points from the spreadsheet will have approximately the same length shortest route as the random 7 x 55 spreadsheets of data shortest routes (in the same ranges as the original data). The CTSP approach uses simulation to find the shortest route through the spreadsheet of data. If it is significantly shorter than the shortest route through similar size random data, then the variables are following a pattern and are therefore correlated.

INTRODUCTION

The new CTSP multivariate correlation coefficient was used in (Conley 2006) to discover relationships that are not functions. Presented here is an attempt to use the CTSP approach to discover relationships that are functions. The definition of a function that is being used here is that a unique Y value is produced for each input of the X values. An example would be X Y points taking the shape of a quartic equation and curving three times. However, X Y points tracing out a circle would be a perfectly good relationship, but not a function because the same X value could produce two Y values.

The CTSP statistic is an acronym for correlation using the traveling salesman problem of finding the shortest route to connect n points in a closed loop complete tour through each point. Therefore, the multi stage Monte Carlo optimization TSP algorithm (Conley 2006) is used and adjusted for k dimensions for k variables represented by k columns of data, to find a shortest route through k dimensional space. Then random data sets (of the same size) are created and similarly solved for shortest routes. Then if the real data has a significantly shorter shortest route than the random data, it means that the variables are correlated in some way. Note that we are using a more general definition of correlation here (detection of any pattern) than the traditional linear correlation discussed in (Anderson 2003) using r and big R.

A THREE VARIABLE FUNCTION

The 29 lines of three columns of data in Table 1 represent 29 observations of three variables that are thought to be correlated in some fashion.

Table 1: The Data

	X_1	X_2	X_3
1	3	13	17
2	69	74	97
3	68	95	95
4	40	100	92
5	20	41	71
6	10	11	43
7	81	48	91
8	53	5	98
9	22	24	75
10	85	56	86
11	90	61	78
12	32	100	85
13	53	84	91
14	4	37	22
15	54	0	98
16	100	95	52
17	26	46	82
18	58	50	100
19	21	98	69
20	39	1	94
21	100	66	55
22	34	49	92
23	26	31	82
24	34	46	92
25	21	100	68
26	49	31	99
27	68	92	95
28	85	77	85
29	17	81	62

Then the data is read into the multi stage Monte Carlo optimization (MSMCO) simulation traveling salesman

problem algorithm (Conley 2000) adjusted for three variables and 29 lines of data. Using the Pythagorean theorem distance measure (adjusted for three dimensions), the distances from each point to each other point are calculated and ranked in 29 columns from the closest point (at the top of the array) to furthest point away (at the bottom of each array) for each point. Then the MSMCO CTSP algorithm concentrates its simulation in pursuit of the shortest route at the top of this ranked 29 x 29 array. This makes for a faster, more efficient simulation to find a shortest route through the 29 points. About a one second run on a desktop PC produced a shortest route of A = 597.382 for the Table 1 data. Then four sets of 3 x 29 data (in similar ranges as the Table 1 data) were read into the MSMCO CTSP algorithm and also run. The four shortest routes produced for these random data sets were 821.176, 765.725, 837.246 and 841.489.

Now we test the null hypothesis of

H_0: no correlation between X_1, X_2 and X_3
versus
H_A: a correlation between X_1, X_2 and X_3

We calculate CTSP = A/B where B is the median of the four random data shortest routes. Therefore, CTSP = 597.382/((821.176 + 837.246)/2) = .72. Now taking the 12 A_i/B_i quotients with all combinations of the four random shortest routes, we see that the vast majority of the sampling distribution of CTSP is in the range of 765.725/841.489 = .91 to 841.489/765.725 = 1.10 if the null hypothesis of no correlation is correct. However, our calculated CTSP value is .72 using the real data for A. Therefore, the A = 597.382 distance is indicating that the real data in Table 1 is following a pattern. Therefore, the null hypothesis can be easily rejected because the variables are highly correlated. It turns out that Equation (1) fits the data pretty well.

$$X_3 = 100 - .00141[\left|(X_1 - 57)^{2.75}\right| + (X_2 - 41)^2] \qquad (1)$$

A SIX VARIABLE EXAMPLE

A spreadsheet of n = 59 lines of k = 6 columns of data is analyzed similarly to the previous three variable problem to see if the six variables (represented by the columns) might be correlated. The data is available upon request.

The null hypothesis is

H_0: no correlation between X_1, X_2, X_3, X_4, X_5 and X_6
H_A: a correlation exists between the variables

The MSMCO CTSP algorithm is adjusted for n = 59 lines of data and six variables. The simulation finds a shortest route on a complete closed loop tour through the six dimensional points of A = 2511.406. Then four similar random sets of 59 x 6 data are read into and run by the MSMCO CTSP algorithm finding shortest routes of 3073.851, 2871.029, 3076.121, and 3305.803. Again, CTSP = A/B = 2511.406/((3076.121 + 3073.851)/2) = .82. Now taking the 12 A_i/B_i quotients from the four random data spreadsheets shortest routes, it is seen that the vast majority of the sampling distributions of CTSP = A/B will be in the range 2871.029/3305.803 = .87 to 3305.803/2871.029 = 1.15 if the null hypothesis is true. Our CTSP = .82 is below this range so the null hypothesis of no correlation between the variables can be rejected. It turns out that Equation (2) fits the data pretty well.

$$X_6 = 1.83\exp(X_1 + X_2 + X_3 + X_4 + X_5)/125 \qquad (2)$$

A SEVEN VARIABLE EXAMPLE

A spreadsheet of n = 55 lines of k = 7 columns of data is analyzed similarly to the previous three and six variable problems to see if the seven variables (represented by the columns) might be correlated. The data is also available upon request.

The null hypothesis is

H_0: no correlation between X_1, X_2, X_3, X_4, X_5, X_6 and X_7
H_A: a correlation exists between the variables

The MSMCO CTSP algorithm is once again adjusted for n = 55 lines of data and seven variables. The simulation finds a shortest route through seven dimensional space of A = 3451.778. Then four similar random data sets of 55 x 7 data are read into and run by the MSMCO CTSP algorithm finding shortest routes of 3365.784, 3134.388, 3482.946 and 3489.064. Again, CTSP = A/B = 3451.778/((33645.784 + 3482.946)/2) = 1.01. Now taking the 12 A_i/B_i quotients from the four random data spreadsheet shortest routes, it is seen that the vast majority of the sampling distribution of CTSP = A/B will be in the range 3134.388/3489.064 = .898 to 3489.064/3134.388 = 1.11 if the null hypothesis is true. The CTSP value of 1.01 for the real data is right in the middle of the sampling distribution. Therefore, the null hypothesis of no correlation can not be rejected.

CORRELATION TODAY

The CTSP statistic approach allows us to create and use a more general definition of correlation than was possible in the era before the computer age began. Any pattern in k dimensional space will mean that certain areas of k dimensional space will not be very likely to have some of the n k dimensional points. This will mean that the n k dimensional points (the data in question) will be in a more compact region and hence the shortest route through them (in a complete closed loop tour) will be statistically significantly shorter than similar random data's shortest routes. This is easy to see in two and three dimensions. Examples would be a circle in two dimensional space. Researchers are fond of saying that X and Y are not correlated because the data is going around in circles (random). However, if the X Y data is on a circle, it is

highly correlated (a relation not a function) and its shortest route will be much shorter than random *X Y* points in the same general region as the circle.

Also, in three dimensions, think of *n X*, *Y*, *Z* points on the dome of the United States capitol building in Washington D.C. versus *n* random *X*, *Y*, *Z* points in the three dimensional region around and enclosed by the dome. Obviously, the shortest route through the points on the dome will be much shorter than the shortest route through the random points because the points on the dome are following a pattern and are hence correlated.

The amazing feature of the CTSP approach is that it also works very well in four, five, six and seven dimensions, etc. when the geometry is harder to picture. Seven, eight and nine dimensional problems are a little more difficult than two and three dimensional problems with CTSP, but they are not impossible.

FUTURE CTSP SIMULATION WORK

Work is now being done to create a software package for the CTSP multivariate statistic and a 300 page book to further explain and illustrate the multivariate analysis that is possible with the CTSP approach.

The traveling salesman computer programs (like MSMCO CTSP) can be used to do other things besides directing salespeople to their customers in an optimal fashion. They can direct advanced analysis of complex multivariate data to see if the variables (in a spreadsheet of data) represented by the columns are correlated in any way. Whether the correlation is a function like this current analysis or not a function (Conley 2006) or linear or nonlinear does not matter. It can analyze spreadsheets of multivariate data to see if the variables they represent are correlated in any way.

Twenty or thirty years ago this type of analysis would take 40 or 50 hours on a mainframe computer. Today, a desktop PC can handle the problems in a few seconds.

MULTI STAGE MONTE CARLO OPTIMIZATION

The multi stage simulation approach (MSMCO) is a general purpose multivariate solution technique to approximate the solution (or find it exactly) to any multivariate optimization problem. Repeated "random" searches are made in the sampling distribution of the optimization problem in question to locate and close in on the optimal solution region. It has applications in many areas of engineering science (Szarkowicz, 1995), (Wong, 1996) and business (Conley, 2000), (Conley 2003). The increase in speed and capacity and affordability of desktop PCs make its application easier and more practical in our 21st century.

REFERENCES

Anderson, T.W. 2003. *An Introduction to Multivariate Statistical Analysis*, 3rd edition, Wiley and Sons, New York.

Conley, W.C. 2000. "Statistical Simulation and a Six Dimensional Shortest Route Problem". In *Proceedings of the Japanese Society for Simulation Technology*, JSST 2000, Tokyo, 90-91.

Conley, W.C. 2003. "Linear and Nonlinear Programming in Manufacturing". In *Proceedings of the 8th International Conference on Computer Aided Optimum Design of Structure*, OPTI VIII, W.I.T. Press, Boston, 159-166.

Conley, W.C. 2006. "Discovering Relationships that are not Functions." In *Proceedings of 4th International Industrial Simulation Conference*, ISC 2006 (Palermo, Italy, June 5-7), EUROSIS-ETI, Ghent, Belgium, 163-167.

Szarkowicz, D.S. 1995. "Investigating the Brachistochrone with a Multi Stage Monte Carlo Method". *International Journal of Systems Science*, Vol. 26, No. 2, 223-243.

Wong, J. 1996. "A Note on Optimization in Integers". *International Journal of Mathematical Education in Science in Technology*, Vol. 27, No. 6, 865-874.

BIOGRAPHY

William C. Conley was born in Lansing, Michigan and went to Albion College where he studied mathematics and received a degree with honors in 1970. He then earned a masters degree in mathematics from Western Michigan University in 1971 and an M.Sc. and Phd. in mathematics and computer statistics from the University of Windsor in 1973 and 1976. He joined the faculty of the University of Wisconsin at Green Bay, U.S.A. in 1977 and is now a professor of business administration and statistics. The developer of multi stage Monte Carlo optimization and the TSP statistics, he has more than 190 publications worldwide. He is a senior member of the Society for Computer Simulation International and was named to *Who's Who in America* and the Albion College Athletic Hall of Fame in 1995 and 2005. He is a member of the American Chemical Society and a fellow in the Institute of Electronics and Telecommunication Engineers. He is a Michigan Scholar in College Teaching and was selected for the 2006 edition of Marquis *Who's Who in the World*. He teaches advanced business statistics to undergraduate students and does computer simulation research on multi stage Monte Carlo optimization (MSMCO) and the TSP class of statistics.

A DISCRETE MODEL FOR THE DYNAMICS OF SANDPILE SURFACES

Ignacio García-Fernández
Marta Pla-Castells
Rafael J. Martínez-Durá
Instituto de Robótica, Universitat de València
PO-Box 2085
46071 – Valencia, Spain
e-mail: ignacio.garcia@uv.es

KEYWORDS
Sandpile Surface Dynamics, Granular Systems

ABSTRACT

A new Cellular Automata model for the dynamics of sandpile surfaces is presented in this work. Following the BCRE model (Bouchaud et al. 1994) main ideas, we propose the use of two variables, instead of only the one that has been used in Cellular Automata models so far. The model predicts sandpile properties observed in other models, and in real systems, such as slope roughness decay or uphill propagation of avalanches. In addition, the introduction of the second variable allows the prediction of characteristics, such as the appearance of a second critical angle, that have been predicted by continuous models. The main advantage of the proposed model is that its formulation is simpler and easier to interpret than that of the continuous models. Also, the resulting numerical scheme is simpler and more efficient, allowing for a wider range of applications, including interactive simulation.

INTRODUCTION

Granular systems behaviour has been widely studied during the last decades, due to its many applications to industry. Many works have been devoted to study the dynamics of sandpile surfaces (Bouchaud et al. 1994; 1995, Aradian et al. 1999, Prigozhin and Zaltzman 2001; 2003), specially in the particular case when the mean slope of the surface is close to the so called *angle of repose* of the system.

In the last decade, the increase in computer capabilities has propitiated that sandpile models are used in more complex simulations, with applications to film industry or virtual reality simulation (Müller et al. 2003). However, classical models are computationally expensive, and provide an inadequate description of the system given that its graphical representation is necessary. Moreover, real-time or interactive simulation of such systems is far from being possible.

In this work we present a new Cellular Automata model that has been developed within a series of efforts aimed to obtain models that allow real-time, interactive, simulation of granular systems, for virtual reality applications.

In the next section the reader is briefly introduced in the field of the dynamics of sandpile surfaces. Following, the CA model is formally described and explained. Then the dynamic properties of the model are reviewed, and compared to those observed in empirical studies and in the continuous model, as studied by Bouchaud et. al. in (Bouchaud et al. 1994). Finally the results of this work are discussed.

DYNAMICS OF SANDPILE SURFACES

Schematically, sandpile surface dynamics can be described with very few ideas. Sand grains can be piled up until the surface of the heap is higher than the *angle of repose* of the system. When this happens, an avalanche occurs and some grains roll down the slope, until the slope of the surface is again reduced below the *angle of repose*.

This basic idea was used by Bak et. al. (Bak et al. 1988) to define the update rule of a Cellular Automata, in order to illustrate the dynamics of self-organised critical systems. Later on, other authors have proposed continuous models aimed to describe the dynamics of sandpile surfaces which reflect the same basic principles. Hwa and Kardar (Hwa and Kardar 1989; 1992) proposed a partial differential equation for the dynamics of the height of the surface.

Models based on this basic principle describe the system using a unique variable, the height of the system at each point, and reproduce the appearance of the critical angle and also predict a smoothing of the surface of the system, reaching an almost homogeneous slope along the system.

Later Bouchaud et. al. (Bouchaud et al. 1994; 1995) modified that model in order to introduce two variables, instead of just the height of the system. Their model considers two layers; a static layer, of resting material grains, and a rolling layer that contains the grains that are falling down the slope. They considered the inter-

change of matter among the two layers, and also added a convection term to the dynamics of the rolling layer. The main contribution of the work by Bouchaud et al. was the prediction of a second critical angle. If the mean slope is above this angle any perturbation leads to a catastrophic avalanche.

This work has become a referent in this field, and has motivated several other works, aimed both to the refinement of their model (Aradian et al. 1999, Hadeler and Kuttler 1999, Prigozhin and Zaltzman 2001; 2003) and to its application to industry problems (Hadeler and Kuttler 2001).

However, the use of continuous models (based on Partial Differential Equations (PDEs)) for the description of the dynamics of sandpile surfaces involves many technical considerations from the computational point of view.

The numerical schemes used in the solution of PDEs try to give a very precise solution to the equations. But, in contrast with the high precision that is sought with usual PDE methods, all the models mentioned above are phenomenological models; this means that the inner interaction between the particles of the system are nor considered individually.

In those works, the analysis of the results focus on qualitative aspects, and only medium to large scale behaviour of the system is considered. For this reason, the degree of precision that can be achieved with the solution of a set of PDEs is unnecessary. In other words, the effort to implement a PDEs method and its computational cost is somehow too hard for the analysis that will be done on the data.

The Cellular Automata approach proposed by Bak et. al. (Bak et al. 1988), which is computationally simpler and closer to the motivating ideas of the model, has not played a relevant role in the study of sandpile surface dynamics. Instead, this approach has been used in works that study statistical properties of granular systems (Prado and Olami 1992, Nerone and Gabbanelli 2001, Chen and Nijs 2002), and as a paradigmatic approach to critical systems modelling.

More recently, Pla-Castells et. al. (Pla-Castells 2003, Pla-Castells et al. 2004) have proposed a variation of Cellular Automata models in order to achieve a better description of the surface dynamics using Cellular Automata. The main motivation of these works was to obtain a computational model that can be used for real-time interactive simulation of granular systems, in applications such as driving simulators or virtual reality environments. With the requirement of real-time in mind, Cellular Automata models have proven a very reliable approach to the problem, that highly overcomes the performance of continuous models, and of Discrete Event Modelling (Müller et al. 2003).

In this work we go further and show that Cellular Automata models can provide the same predictions as continuous models do, by means of simpler and more efficient models. We analyse the qualitative behaviour of a two variable Cellular Automata model in order to show that its dynamics has the main properties predicted by the well known BCRE model. Specially, it is shown that the introduction of a second variable in the Cellular Automata model leads to the same behaviour as that observed in the continuous model.

PROPOSED MODEL

The model is described following the notation by Bouchaud et. al. (Bouchaud et al. 1994). For simplicity, it is presented for a one-dimensional Cellular Automata, for which two variables are considered at each cell; the height of the standing layer $h(i,t)$ above the centre of cell i, and the height of the rolling layer $R(i,t)$, at time t. Two parameters are considered for the dynamics of the rolling layer; the drift velocity of the rolling grains v, and a diffusion constant D. A parameter γ controls the exchange of matter among the two layers, together with the *angle of repose*, S_c.

Also, some additional parameters related to the Cellular Automata description of the system are considered. N is the number of cells of the system and d is a cell's length; the system will have a length of $N \times d$ length units. The neighbourhood of cell i (the set of cells that directly influence its evolution) is taken as the set of cells $\mathbf{V} = \{(i-1),(i+1)\}$.

Formal Cellular Automata Model

The Cellular Automata is described as

$$CA = \langle \mathbf{M}, \mathbf{V}, \mathbf{S}, \boldsymbol{\varphi} \rangle \tag{1}$$

where

- $\mathbf{M} = \{(i) \in \mathbb{N}, 1 \leq i \leq N\}$ is the set of cells that form the Cellular Automata.
- $\mathbf{V}$ is the neighbourhood of a cell, as described above.
- $\mathbf{S} = S_h \times S_R$ is the set of possible states for each cell:

  $S_h = \mathbb{R}$ is the height of the static layer;

  $S_R = \mathbb{R}$ is the height of the rolling layer.
- $\boldsymbol{\varphi} : \mathbf{S}^3 \rightarrow \mathbf{S}$ is the transition map of the Automata.

The Transition Map

Given a cell, the transition map, $\boldsymbol{\varphi} : \mathbf{S}^3 \rightarrow \mathbf{S}$, takes the value of its current state (the value of its variables) and that of its neighbours, and determines its new value after a time stepthe new state the state of a cell after a time step. The transition map is computed as follows.

For each cell $(i) \in \mathbf{M}$, let $h \in S_h$, the height of the static layer and $R \in S_R$ the height of the rolling layer above

the cell. The gradient of each layer is computed, using the difference of height with the neighbour cells:

$$dh(i) = \frac{h(i+1) - h(i)}{d} \quad (2)$$

$$dR(i) = \frac{R(i+1) - R(i)}{d} \quad (3)$$

Then, each cell is updated as follows

$$\begin{aligned} R(i) &\leftarrow R(i) + (-vR(i) + DdR(i)) \\ &\quad -R(i)\gamma(dh(i) - |S_c|) \quad (4) \\ h(i) &\leftarrow h(i) + R(i)\gamma(dh(i) - |S_c|) \quad (5) \\ R(i+1) &\leftarrow R(i+1) - (-vR(i) + DdR(i)) \quad (6) \end{aligned}$$

ANALYSIS OF THE MODEL

The introduction of a second variable in the model leads, in the continuous case, to the existence of a second critical angle S_d, for the static layer (Bouchaud et al. 1994). If the mean slope of the static phase S_0 is above this second critical angle, $S_0 < S_d$, any avalanche leads to a catastrophic phenomenon, in which the static layer looses all of its matter while the rolling layer grows along the whole slope.

This behaviour can also be observed in the proposed Cellular Automata model; with the introduction of a second variable in the model, a second critical value is observed in the slope of the system. The sandpile surface dynamics, in this case, is similar to that observed in the continuous model. Figure 1 shows the results of a simulation that illustrate this behaviour.

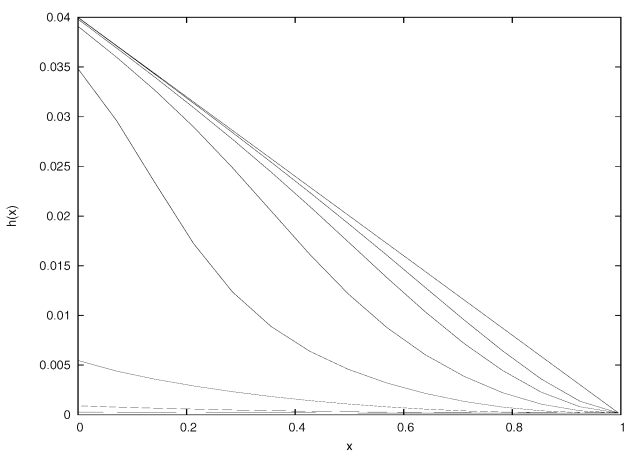


Figure 1: Evolution along time of a perturbation of the rolling layer at the bottom of the slope. The different lines indicate the profile of the static layer at successive instants of time. In this case, it can bee seen that the thickness of the static layer decreases until it is almost zero. The parameters of the model were $v = 0.1$, $D = 0.1$ and $\gamma = 0.01$

Also, the dynamics of the single-variable models can still be observed; the new model is an extension of previous ones. The numerical experiments performed so far reveal the following relevant properties.

1. If a system with slope close to the *angle of repose* is perturbed locally, then an avalanche occurs. The result is that the surface returns to an angle below the critical angle (see Figure 2).

2. Surface roughness tends to dissipate, due to the diffusion terms in the model. This makes that a perturbation spreads down the slope instead of just travelling across it.

3. Propagation of avalanches happens not only downwards, but also upwards, as predicted by continuous models and observed in real granular systems (Bouchaud and Cates 1998, Daerr and Douady 1999, Aranson and Tsimring 2001) (see Figure 3).

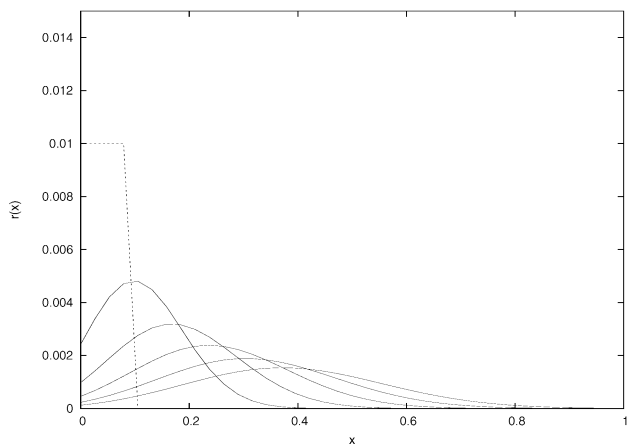


Figure 2: Evolution along time of a perturbation of the rolling layer on top of the slope. The dashed line indicates the initial profile of the rolling layer, while the different continuous lines indicate its successive distributions. The static layer has a slope to the right, slightly lower than the *repose angle*. The parameters of the model were $v = 2.5$, $D = 0.1$ and $\gamma = 0.01$.

Thus, the new model provides a proper description of the dynamics of sandpile surfaces, with the additional advantages observed in Cellular Automata models:

1. Cellular Automata models are discrete in space, thus the computation is not performed on a discretization of the data, but on the data itself.

2. Due to the statistical properties of the dynamics of granular systems, and more precisely, to the fact that they often reflect self-organised criticality, the implementation of two-dimensional systems allows an exhaustive optimisation of the model. This leads

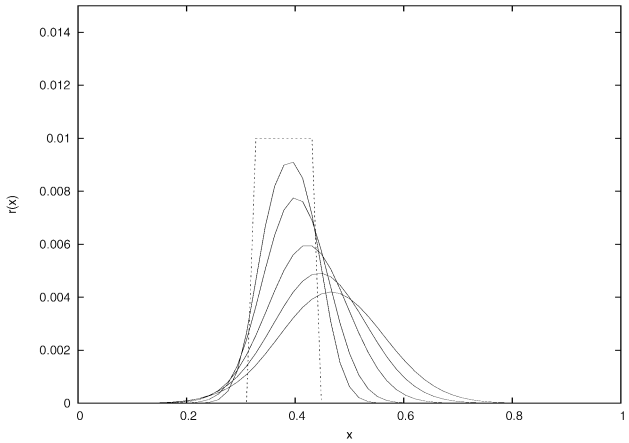


Figure 3: Evolution along time of a perturbation of the rolling layer in the middle of the slope. Again, the dashed line indicates the initial profile of the rolling layer, while the different continuous lines indicate its successive distributions. The region affected by the avalanche includes part of the cells above the initial perturbation. The parameters of the model were the same as in Figure 2.

to a computational cost that is linear respect to the side of the system, instead of quadratic.

3. The discrete description of the system is also very adequate for the graphical representation of a granular system. This makes Cellular Automata models more suitable for its use in computer graphics applications (such as virtual reality or simulation for training).

Part of the validation is still ongoing, and some characteristics observed in sandpile surfaces, such as the statistical properties of the model, have not been investigated in depth yet. However, preliminary simulations, together with previous results obtained with similar models (Pla-Castells 2003) give a good level of confidence that they will be met.

CONCLUSIONS

A Cellular Automata model for the dynamics of sandpile surfaces has been presented in this work. Previous Cellular Automata models already were able to reflect the main characteristics observed in the dynamics of sandpile surfaces, such as statistical properties of avalanches, slope roughness decay or uphill propagation of avalanches.
The new model extends such properties, by means of the introduction of a second variable; both the static and the rolling layer are considered, as in the well known BCRE model. The introduction of two variables, which is new to Cellular Automata description of sandpiles, allows to predict a second critical angle, as continuous models and experimental results did.
The success in the prediction of such property of the dynamics of sandpile surfaces, shows that proper models can be obtained without the need of PDEs, just using a simple model based upon a local phenomenological description of the dynamics. This description is simpler to understand and to implement than PDEs, and it avoids unnecessary technical considerations needed for the numerical solution of differential equations.
Although a lot of work is still to be done, the preliminary results of this work show that a reliable and efficient family of models can be available for the description of the dynamics of granular systems' surface. Together with other works, already presented by the authors in the field of interactive models of granular systems (Pla-Castells et al. 2006), the final results of this ongoing research open the possibility to develop interactive simulators involving granular systems in the short term.

This work has been partially supported by the University of Valencia, proj. UV-AE-20060238 and by Ministerio de Industria y Energia of Spain, proj. FIT-340000-2006-290.

REFERENCES

Aradian A.; Raphaël E.; and De Gennes P.G., 1999. *Thick Surface Flows of Granular Materials: Effect of the Velocity Profile on the Avalanche Amplitude. Phys Rev E*, 60, 2009–2019.

Aranson I.S. and Tsimring L.S., 2001. *Continuum Description of Avalanches in Granular Media. Phys Rev E*, 64, 020301.

Bak P.; Tang C.; and Wiesenfeld K., 1988. *Self-Organized Criticality. Phys Rev A*, 38, 364–374.

Bouchaud J.P. and Cates M.E., 1998. *Triangular and Uphill Avalanches of a Tilted Sandpile. Granular Matter*, 1, 101–103.

Bouchaud J.P.; Cates M.E.; Prakash J.R.; and Edwards S.F., 1994. *A Model for the Dynamics of Sandpile Surfaces. J Phys I France*, , no. 4, 1383–1410.

Bouchaud J.P.; Cates M.E.; Prakash J.R.; and Edwards S.F., 1995. *Hysteresis and Metastability in a Continuum Sandpile Model. Phys Rev Lett*, 74, 1982–1985.

Chen C.C. and Nijs M.D., 2002. *Directed Avalanche Processes with Underlaying Interface Dynamics. Physical Review E*, 66, 011306.

Daerr A. and Douady S., 1999. *Two Types of Avalanche Behaviour in Granular Media. Nature*, 399, 241–243.

Hadeler K.P. and Kuttler C., 1999. *Dynamical Models for Granular Matter. Granular Matter*, 2, 9–18.

Hadeler K.P. and Kuttler C., 2001. *Granular Matter in a Silo. Granular Matter*, 3, 193–197.

Hwa T. and Kardar M., 1989. *Dissipative Transport in Open Systems: An Investigation of Self-Organized Criticality. Phys Rev Lett*, 62, 1813–1816.

Hwa T. and Kardar M., 1992. *Avalanches, Hydrodynamics and Discharge Events in Models of Sandpiles. Physical Review A*, 45, 7002.

Müller M.; Charypar D.; and Gross M., 2003. *Procedural modeling and animation: Particle-based fluid simulation for interactive applications.* 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 154–159.

Nerone N.. and Gabbanelli S., 2001. *Surface Fluctuations and the Inertia Effect in Sandpiles. Granular Matter*, 3, 117–120.

Pla-Castells M., 2003. *Nuevos Modelos de Sistemas Granulares Basados en Autómatas Celulares para Simulación en Tiempo Real.* Master's thesis, Escuela Técnica Superior de Ingeniería, Universidad de Valencia. Spain.

Pla-Castells M.; Garcia I.; and Martinez R.J., 2004. *Granular System Models for Real Time Simulation.* Industrial Simulation Conference, Universidad de Málaga, Malaga, Spain, 88–93.

Pla-Castells M.; Garcia-Fernandez I.; and Martinez-Dura R.J., 2006. *Interactive Terrain Simulation and Force Distribution Models in Sand Piles. Lecture Notes on Computer Science*, 4137, 392–401.

Prado C. and Olami Z., 1992. *Inertia and Break of Self-Organized Criticality in Sandpile Cellular-Automata Models. Phys Rev A*, 45, 6665–6669.

Prigozhin L. and Zaltzman B., 2001. *Two Continuous Models for the Dynamics of Sandpile Surfaces. Phys Rev E*, 63, 041505.

Prigozhin L. and Zaltzman B., 2003. *On the Approximation of the Dynamics of Sandpile Surfaces. PORTUGALIAE MATHEMATICA*, 60, 117–137.

ALGORITHMS FOR RARE EVENT SIMULATION WITH MARKOV CHAINS

Dimitar Radev
University of Rousse
8 Studentska Street
7017 Rousse, Bulgaria
E-mail: dradev@abv.bg

Izabella Lokshina
SUNY Oneonta
324B Netzer Administration Bldg.
Oneonta, NY 13820, USA
E-mail: lokshiiv@oneonta.edu

KEYWORDS
Rare Event Simulation, Overflow Probability, Queuing Theory, Markov Chains.

ABSTRACT

An approach to overflow probability estimation in the queuing systems based on rare event simulation with Markov chains is proposed in this paper. Rare event probability that content of the second buffer exceeds some high level L before it becomes empty starting from a given state is considered. The approach is based on Markov additive representation of buffer processes, leading to exponential change of measure used in Importance Sampling (IS) procedure. The considered examples demonstrate that when the first buffer is finite, the relative error is bounded independent on some high level L. When the first buffer is infinite, a natural extension of exponential change of measure for finite buffer case is proposed. The relative error is shown to be bounded (independent on L) only when the second server is a bottleneck. However, when the first server is a bottleneck, experimental results confirm that the relative error is linearly bounded to level L. IS algorithms with cross-entropy are developed and used to speed up the overflow probability simulation. Numerical results are shown.

INTRODUCTION

Queuing networks are appropriate reference models being used in different methodologies and techniques to speed up rare event simulation in communication systems. Frequency of the rare events is very small, e.g. 10^{-6} or less. Estimation of the rare event probability using Monte Carlo simulation requires for a very long computer processing time, and cannot be implemented. Lately two basic methods of the rare event simulation were developed based on cross-entropy approach that can be applied to a wide range of optimization tasks: splitting of the sample path that for to reach definite intermediate level between the starting level and rare event; and Importance Sampling (IS) generation. The probability density function is used in the IS as the rare event evaluation measure, which can be compared and changed based on the likelihood ratio of the less rare event probability density function. One of the rare event simulation concerns is estimation of total network population overflow (Bolch et al. 1998). Exact large deviation analysis leading to asymptotically efficient change of measure is rather difficult. Instead, heuristic change of measure is proposed, which interchanges the arrival rate (to the first queue) and the slowest service rate (De Boer et al. 2002). Similar change of measure is suggested based on time reversal arguments. However, analysis shows that the IS estimator based on this change of measure is not necessarily asymptotically efficient. In fact, it has an infinite variance in some parameter regions (Radev et al. 2005). Another rare event simulation concern is estimation of buffer overflow that observed at individual network node. If the node of concern is a bottleneck relatively to all preceding nodes, then asymptotically efficient exponential change of measure can be obtained by interchanging the arrival rate and the service rate at this particular node; and the service rates at all other nodes are kept unchanged. However, this change of measure is not asymptotically efficient when overflow of buffer at the consequent node is considered. The effective bandwidth theory is used to derive heuristics for efficient feed-forward discrete-time queuing network simulation. This class of networks essentially resembles feed-forward fluid-flow networks. The analogous approach to continuous-time queuing networks has not yet been introduced even for Markov chains. Two-node queuing networks are considered in this paper; and the event of buffer overflow at the second node is studied. Discrete-time Markov chain (DTMC) with its regular structure is highly efficient model used for performance evaluation on the queuing system. On one hand, the states are easily arranged as a grid in DTMC (with as many dimensions as the number of queues). On the other hand, any transition in DTMC corresponds to a particular elementary event at one of the queues (arrival or service completion). These events are known as transition events, and they are defined separately from the states; i.e., there is only one transition event for a service completion at a given queue, and this particular transition event corresponds to a transition out of each state in the DTMC while this particular queue is non-empty. However, not all transition events are enabled in every state. For example, the service completion event of the particular queue is not possible in a state where the queue is empty.

MARKOV ADDITIVE CONTINUOUS TIME PROCESS REPRESENTATION

Markov additive continuous time process is a stochastic process (J_t, Z_t), where (J_t) is Markov chain with the denumerable state space, and (Z_t) has stationary and independent increments during the time intervals when (J_t)

is in any given state. That is, if given J_t has not changed in the interval (t_1, t_2), then for any $t_1<s_1<\ldots<s_n<t_2$, the increments $Z_{s_2}-Z_{s_1},\ldots,Z_{s_n}-Z_{s_{n-1}}$ are mutually independent, and the total increment during the interval [t_1, t_2] depends on t_1 and t_2 only through the difference t_2- t_1.

Moreover, the transition from state i to state j (J_t) has a certain probability (depending only on i and j) of triggering the transition of (Z_t) at the same time. The size of the transition in the process (Z_t) has fixed distribution, which depends only on i and j. Markov additive process (J_t, Z_t) is characterized by the family of the matrices ($M_t(s)$,$t\geq 0$), where (i,j)-th element of $M_t(s)$ is (1),

$$[M_t(s)]_{ij} = \mathrm{E}_i[e^{s(Z_t-Z_0)}I_{\{J_t=j\}}] \tag{1}$$

where E_i denote the expectation operator given to the initial MAP state $J_0=i$. Let's notice that $M_t(\cdot)$ is a generalization of the moment generating function for the ordinary random variables, as shown in (2).

$$\begin{aligned} &\mathrm{E}_i[e^{s(Z_{t+h}-Z_0)}I_{\{J_{t+h}=j\}}] = \\ &= \sum_k \mathrm{E}_i[e^{s(Z_{t+h}-Z_0)}I_{\{J_t=k\}}I_{\{J_{t+h}=j\}}] \\ &= \sum_k \mathrm{E}_i[e^{s(Z_t-Z_0)}I_{\{J_t=k\}}]\mathrm{E}_i[e^{s(Z_{t+h}-Z_t)}I_{\{J_{t+h}=j\}} \mid J_t=k] \\ &= \sum_k [M_t(\mathrm{s})]_{ik}\,\mathrm{E}_k[e^{s(Z_h-Z_0)}I_{\{J_h=j\}}] \end{aligned} \tag{2}$$

Consequently, if for all k and j (3) can be defined,

$$[A(s)]_{kj} := \lim_{h\downarrow 0}\frac{1}{h}\mathrm{E}_k[e^{s(Z_h-Z_0)}I_{\{J_h=j\}} - \delta_{kj}] \tag{3}$$

where δ is usual notation of Dirac, then (4) can be easily obtained,

$$\frac{d}{dt}M_t(s) = M_t(s)A(s), \qquad t\geq 0 \tag{4}$$

with $M_0(s)=I$ (the identity matrix). It is true as soon as (5) is true.

$$M_t(s) = e^{tA(s)}, \qquad t\geq 0 \tag{5}$$

The matrix $A(s)$ is known as the MAP (infinitesimal) generator. Let's consider a simple Markov chain that consists of two queues in tandem. The customers arrive to the first queue (buffer) according to Poisson process with the rateλ. The service time for the customer in the first queue is exponentially distributed with the rate μ_1. The customers that leave the first queue enter the second queue. The service time in the second queue has an exponential distribution with the rate μ_2. The queuing network stability is assumed, i.e. $\lambda<\min\{\mu_1, \mu_2\}$.

The size of the first buffer may be finite or infinite; in fact, let's consider both cases. Let X_t and Y_t denote the number of the customers in the first and the second queues at the time t, respectively. Let P_i denote the probability measure under which (X_t) starts from i at the time 0 (i.e., $X_0=i$, $i\geq 0$); and let E_i denote the corresponding expectation operator. Assuming that the second buffer is initially non-empty, say, $Y_0=1$, the probability that, starting from $(X_0,Y_0)=(i,1)$, content of the second buffer hits some high level $L\in\mathbf{N}$ before hitting 0, can be estimated. This probability is noted by γ_i and referred to it as the second buffer overflow probability, given that the initial number of customers in the first queue is i.

RARE EVENT SIMULATION

Let's consider the IS approach for the rare event simulation, where the probability density function is used as the measure of the rare event evaluation, which is compared and changed with the likelihood ratio of the probability density function of the less rare event. First, let's determine the rare event. Let $\mathbf{X} = (\mathbf{X}_1,\ldots, \mathbf{X}_N)$ be a random vector, which values belong to the certain state space χ. Let $\{f(\cdot, \mathbf{v})\}$ be the family of the probability density functions on χ, with respect to some base measure υ. Here $\mathbf{v}$ is a real-valued parameter (vector). Then, for any measurable function H is obtained as (6).

$$\mathrm{E}[H(\mathbf{X})\int H(\mathbf{x})f(\mathbf{x};\mathbf{v})\upsilon(d\mathbf{x})] \tag{6}$$

In many cases f is often called the Probability Mass Function (PMF), but in this paper the generic term density, or the Probability Density Function (PDF), is used. Let S be some real function onχ. The probability that $S(\mathbf{X})$ is greater than some real numberγ, under $f(\cdot;\mathbf{u})$ can be defined. Therefore, probability can be written as (7),

$$l = \mathrm{P}_{\mathbf{u}}(S(\mathbf{X})\geq\gamma) = \mathrm{E}_{\mathbf{u}}[I_{\{S(\mathbf{X})\geq\gamma\}}] \tag{7}$$

where $I_{\{S(\mathbf{X})\geq\gamma\}}$ is the indicator function. If this probability is very small, like 10^{-6} or less, then $\{S(\mathbf{X})\geq\gamma\}$ is a rare event. The simplest way to estimate l is to use the basic Monte Carlo simulation. Draw a random sample $\mathbf{X}_1,\ldots, \mathbf{X}_N$ from $f(\cdot;\boldsymbol{u})$, and use (8) as the unbiased estimator of l.

$$\frac{1}{N}\sum_{i=1}^{N} I_{\{S(\mathbf{X}_i)\geq\gamma\}} \tag{8}$$

However, it poses serious problems when $\{S(\mathbf{X})\geq\gamma\}$ is a rare event. In that case a large simulation effort is required in order to estimate l accurately. An alternative is based on the IS. Take a random sample $\mathbf{X}_1,\ldots, \mathbf{X}_N$ from the IS density g on χ, and evaluate l using the unbiased estimator, called the likelihood ratio estimator (9).

$$\hat{l} = \frac{1}{N}\sum_{i=1}^{N} I_{\{S(\mathbf{X}_i)\geq\gamma\}}\frac{f(\mathbf{X}_i;\mathbf{u})}{g(\mathbf{X}_i)} \tag{9}$$

It is well known that the optimal way to estimate l is to use the change of the measure with the density (10)

$$g^*(\mathbf{x}) = \frac{I_{\{S(\mathbf{x})\geq\gamma\}}f(\mathbf{x};\mathbf{u})}{l} \tag{10}$$

Specifically, applying this change of the measure to (9), (11) can be obtained for all i.

$$I_{\{S(\mathbf{X}_i)\geq\gamma\}}\frac{f(\mathbf{X}_i;\mathbf{u})}{g^*(\mathbf{X}_i)} = l \tag{11}$$

In other words, the estimator (9) has a zero variance, and only $N = 1$ sample need to be produced. The obvious difficulty is, of course, that the g^* depends on the unknown parameter l. Moreover, one often wishes to choose this g from the density family $\{f(\cdot,\mathbf{v})\}$. Now the plan is to choose the tilting parameter $\mathbf{v}$, such that the distance between the densities g^* and $f(\cdot,\mathbf{v})$ is minimal. A particular suitable measure of the distance between two densities g and f is the Kullback-Leibler distance, which is defined in (12).

$$D(g,f)=\mathrm{E}_g[\ln\frac{g(\mathbf{X})}{f(\mathbf{X})}]=\int g(\mathbf{x})\ln g(\mathbf{x})\upsilon(d\mathbf{x})-\int g(\mathbf{x})\ln f(\mathbf{x})\upsilon(d\mathbf{x}) \quad (12)$$

Therefore, minimizing the Kullback-Leibler distance between g in (11) and $f(\cdot,\mathbf{v})$ is the same as choosing $\boldsymbol{v}$, such that $-\int g(\mathbf{x})\ln f(\mathbf{x};\mathbf{v})\upsilon(d\mathbf{x})$ is minimized, or equally, such that $\int g(\mathbf{x})\ln f(\mathbf{x};\mathbf{v})\upsilon(d\mathbf{x})$ is maximized. Formally, it can be written as (13).

$$\max_{\mathbf{v}} D(v)=\max_{\mathbf{v}}\int g(\mathbf{x})\ln f(\mathbf{x};\mathbf{v})\upsilon(d\mathbf{x}) \quad (13)$$

Applying g from (10) to (13) as the substitution, the following optimization program can be obtained as (14).

$$\begin{aligned}\max_{\mathbf{v}} D(\mathbf{v})&=\max_{\mathbf{v}}\int\frac{I_{\{S(\mathbf{x})\geq\gamma\}}f(\mathbf{x};\mathbf{u})}{l}\ln f(\mathbf{x};\mathbf{v})\upsilon(d\mathbf{x})=\\&=\max_{\mathbf{v}}\mathrm{E}_{\mathbf{u}}[I_{\{S(\mathbf{x})\geq\gamma\}}\ln f(\mathbf{X};\mathbf{v})]\end{aligned} \quad (14)$$

Using the IS again, with the change of the measure $f(\cdot;\boldsymbol{w})$, (14) can be re-written into (15),

$$\max_{\mathbf{v}} D(\mathbf{v})=\max_{\mathbf{v}}\mathrm{E}_{\mathbf{w}}[I_{\{S(\mathbf{X})\geq\gamma\}}W(\mathbf{X};\mathbf{u},\mathbf{w})\ln f(\mathbf{X};\mathbf{v})] \quad (15)$$

for any tilting parameter $\mathbf{w}$, where the likelihood ratio at $\mathbf{x}$ between $f(\cdot\ ;\mathbf{u})$ and $f(\cdot;\mathbf{w})$ is W, and this can be presented according to (16).

$$W(\mathbf{x};\mathbf{u},\mathbf{w})=\frac{f(\mathbf{x};\mathbf{u})}{f(\mathbf{x};\mathbf{w})} \quad (16)$$

The basic idea of the IS method regarding the rare event is the acceleration of its frequency with iterative tilting of the unbiased estimation of the probability density function. The acceleration of the conditional probability of the rare event X_1 for the one-parameter function $f(x)$ is shown in Fig. 1.

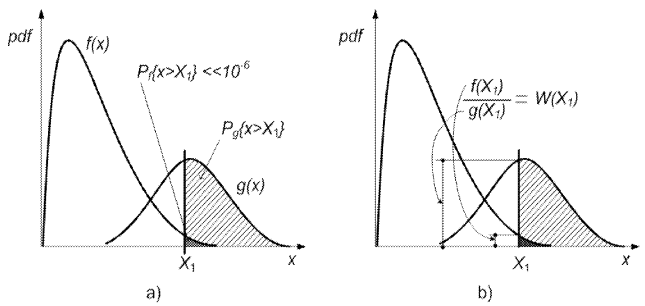


Figure 1: Conditional Probability of Rare Event $P_f\{x>X_1\}$ - (a) and Its Acceleration with Likelihood Ratio $W(X_1)$ - (b)

The conditional probability of the rare event $P_f\{x>X_1\}$ is changing with the conditional probability of $g(x)$ – $P_g\{x>X_1\}$. N independent samples, which distribution $g(x)$ can be evaluated with the likelihood ratio $W(X_1)$, are generated at each iteration of the IS simulation. The optimal solution of (15) can be written as (17).

$$\mathbf{v}^*=\arg\max_{\mathbf{v}}\mathrm{E}_{\mathbf{w}}[I_{\{S(\mathbf{X})\geq\gamma\}}W(\mathbf{X};\mathbf{u},\mathbf{w})\ln f(\mathbf{X};\mathbf{v})] \quad (17)$$

It can be obtained by solving the following stochastic program, which can be considered as a stochastic counterpart of (15) and written according to (18),

$$\max_{\mathbf{v}}\hat{D}(\mathbf{v})=\max_{\mathbf{v}}\frac{1}{N}\sum_{i=1}^{N}I_{\{S(\mathbf{X})\geq\gamma\}}W(\mathbf{X}_i;\mathbf{u},\mathbf{w})\ln f(\mathbf{X}_i;\mathbf{v}) \quad (18)$$

where $\mathbf{X}_1,\ldots,\mathbf{X}_N$ is a random sample from $f(\cdot\ ;\mathbf{w})$.

The solution of (18) can be obtained by solving the following system of equations with respect to $\mathbf{v}$ in (19),

$$\frac{1}{N}\sum_{i=1}^{N}I_{\{S(\mathbf{X}_i)\geq\gamma\}}W(\mathbf{X}_i;\mathbf{u},\mathbf{w})\nabla\ln f(\mathbf{X}_i;\mathbf{v})=\mathbf{0} \quad (19)$$

where the gradient is defined regarding $\mathbf{v}$. This confirms the expectation that the differentiation operators and the function $\hat{D}$ can be interchanged in (18) with respect to $\mathbf{v}$. The advantage of this approach is in the fact that often the solution of (19) can be calculated analytically. In particular, this happens when the random variable distributions belong to the natural exponential family. The cross-entropy program (19) is useful only when the probability of the target event $\{S(\mathbf{X})\geq\gamma\}$ is not too small, say $l\geq 10^{-5}$. Then, above program is useful in terms of potentially more accurate estimator determination. However, in the rare event context, (say, $l\geq 10^{-6}$), the program (19) is useless to the rarity of the events $\{S(\mathbf{X}_i)\geq\gamma\}$, $i=1,\ldots,N$ and the associated derivatives of $\hat{D}(\mathbf{v})$ vanish with high probability, as given in the right-hand side of (19), for reasonable sizes of N.

EXPONENTIAL CHANGE OF MEASURE

Let's first think that the first buffer has a finite capacity N. In this case the state space of the driving process (X_t) is finite in $\{0,\ldots,N\}$. Let's consider Markov additive process (X_t, Z_t). To construct the corresponding MAP generator (i.e., a matrix A(s) in (5)), the infinitesimal expectations $\mathrm{E}_i[e^{s(Z_h-Z_0)}I_{\{X_h=j\}}-\delta_{ij}]$ as $h\downarrow 0$, for all i, j in $\{0,\ldots,N\}$ have to be determined, where $Z_0=1$ and $\delta_{ij}=0$ for $j\neq i$. For instance, since the downward transition of (X_t) leads to the upward transition of (Z_t), (20) is used for $i=1,\ldots,N$, as $h\downarrow 0$.

$$\begin{aligned}&\mathrm{E}_i[e^{s(Z_h-Z_0)}I_{\{X_h=i-1\}}-\delta_{i,i-1}]=\\&=\mathrm{E}_i[e^{s(Z_h-Z_0)}\mid X_h=i-1]\mathrm{P}_i(X_h=i-1)=\\&=e^s(\mu_1 h+o(h))=\mu_1 he^s+o(h)\end{aligned} \quad (20)$$

Therefore, the $(i,i\text{-}1)$-th element of the matrix $A(s)$ exists and is equal to $\mu_1 e^s$. Other elements of the matrix $A(s)$ can be defined similarly. Consequently, (5) holds with $A(s)$ for Markov additive process (X_t, Z_t) with the given $(N+1,N+1)$-tri-diagonal matrix (21).

$$\mathrm{G}_N(s)=\begin{pmatrix}-\lambda-\mu_2+\mu_2e^{-s} & \lambda & & \\ \mu_1e^s & -\lambda-\mu_1-\mu_2+\mu_2e^{-s} & \lambda & \\ & \ddots & \ddots & \ddots \\ & & \mu_1e^s & -\mu_1-\mu_2+\mu_2e^{-s}\end{pmatrix} \quad (21)$$

Let's note that the MAP generator $G_N(s)$ is now equal to the matrix (22).

$$\hat{\mathrm{Q}}^{(n)}(u)=\begin{pmatrix}-\lambda-\mu_2+\mu_2u & \lambda & & \\ \mu_1/u & -\lambda-\mu_1-\mu_2+\mu_2u & \lambda & \\ & \ddots & \ddots & \ddots \\ & & \mu_1/u & -\mu_1-\mu_2+\mu_2e^{-s}\end{pmatrix} \quad (22)$$

Next, the change of the measure based on the family of the matrices $\mathbf{G}_N(s)$ is defined. For any $s\geq 0$,

$k_N(s) := \log(sp(M_t(s)))/t$ has to be defined, where $sp(M_t(s))$ denotes the spectral radius (or, the maximum eigenvalue) of $M_t(s)$. Using (5) $k_N(s)$ can be identified as the largest positive eigenvalue of $\mathbf{G}_N(s)$. Let $\mathbf{w}(s)=\{w_k(s), 0\le k\le N\}$ represent the correspondent right-eigenvector. When the first buffer has the infinite capacity, the process (X_t,Z_t) is still the Markov additive process, but the state space of the Markov process (X_t) is now infinite. Equation (5) is still used, but $A(s)$ is now given by the infinite-dimensional tri-diagonal matrix (23).

$$\mathbf{G}(s)=\begin{pmatrix} -\lambda-\mu_2+\mu_2 e^{-s} & \lambda & & \\ \mu_1 e^{s} & -\lambda-\mu_1-\mu_2+\mu_2 e^{-s} & \lambda & \\ & \ddots & \ddots & \ddots \end{pmatrix} \tag{23}$$

Let's note that the MAP generator $G(s)$ is now equal to the infinite tri-diagonal matrix (24).

$$\mathbf{Q}(u)=\begin{pmatrix} -\lambda-\mu_2+\mu_2 u & \lambda & & \\ \mu_1/u & -\lambda-\mu_1-\mu_2+\mu_2 u & \lambda & \\ & \ddots & \ddots & \ddots \end{pmatrix} \tag{24}$$

IMPORTANCE SAMPLING ALGORITHM

In this work the following IS algorithm is proposed:

1. Set t := 1 (initialization of iteration counter). Define the likelihood ratio v_1:=0 (in this case Monte Carlo simulation is appropriate).

2. Generate a sample $\mathbf{X}_1,\ldots,\mathbf{X}_k,\ldots,\mathbf{X}_N$, from the density $f(\mathbf{X}_k;\mathbf{v}_{t-1})$ in such way that for the ρ-th part of samples (ρ=0.01) the condition of the rare event $S(\mathbf{X}_k)>M$ is executed.

3. Determine the full paths and sort ascending in following way $S_{(1)}\le S_{(2)}\le\ldots\le S_{(N)}$.

4. Calculate the conditional probability $\gamma_t = S_{[(1-\rho)N]}$.

5. For each $S(\mathbf{X}_k)>\gamma$ define the rare event indicator $I_{\{S(\mathbf{X}_i)\ge M\}}=1$ and then determine the likelihood ratio for the next iteration $\mathbf{v}_{t+1}$ according to (15), (16) and (19).

6. If $\gamma_t<M$ then t:=t+1 and repeat steps 2, 3 and 4.

The Markovian network that consists of the tandem queues with feedback is shown in Fig.2 as a simulation example.

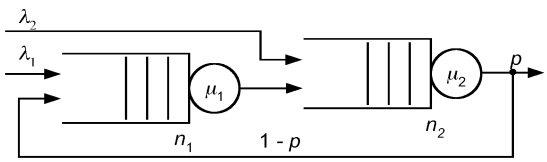


Figure 2: Two-Node Queuing Network with Feedback

The entry parameters follow: $\boldsymbol{\lambda_1}=\boldsymbol{\lambda_2}=1$, $\boldsymbol{\mu_1}=\boldsymbol{\mu_2}=6$, p=0.5. The simulation results for Poisson distribution and fixed customer numbers n_1 and n_2, for four different overflow cases are shown in Table 1. As could be seen, the overflow probability exponentially decreases with increase of the fixed number of customers in queues n_1 and n_2. The exponential behavior depends more on the number n_1.

Table 1: Simulation Results

Overflow level		λ_1	λ_2	μ_1	μ_2	p	Overflow probability
n_1	n_2						
25	25	0.13	0.11	0.29	0.18	0.42	2.22×10^{-8}
50	50	0.23	0.18	2.15	2.26	0.31	1.15×10^{-15}
60	40	1.18	1.42	4.91	3.58	0.44	6.29×10^{-25}
100	40	2.02	1.67	5.82	2.94	0.37	1.86×10^{-50}

CROSS-ENTROPY ALGORITHM

The multi-level cross-entropy algorithm is proposed in this work. The idea is to introduce a sequence of the reference parameters $\{\mathbf{v}_t, t\ge 0\}$ and a sequence of the levels $\{\gamma_t, t\ge 1\}$, and iterate in both γ_t and $\mathbf{v}_t$. The initialization is done with choosing a not very smallρ, say $\rho = 10^{-2}$, and defining $\mathbf{v}_0 = \mathbf{u}$. Next, we let γ_1 ($\gamma_1 < \gamma$) be such that under the original density $f(\mathbf{x}; \mathbf{u})$, the probability $l_1=\mathrm{E}_{\mathbf{u}}[I_{\{S(\mathbf{X})\ge\gamma_1\}}]$ is at least ρ. After, let $\mathbf{v}_1$ be the optimal cross-entropy reference parameter to estimate l_1, and repeat the last two steps iteratively with the purpose to estimate the pair $\{l, \mathbf{v}^*\}$. In other words, the iteration of the algorithm consists of two main phases. In the first phase γ_t is updated, in the second phase $\mathbf{v}_t$ is updated. Particularly, starting with $\mathbf{v}_0 = \mathbf{u}$, the subsequent γ_t and $\mathbf{v}_t$ are obtained as follows.

1. Adaptive Update of γ_t. For a fixed $\mathbf{v}_{t-1}$, let γ_t be a (1-ρ)-quintile of $S(\mathbf{X})$ under $\mathbf{v}_{t-1}$. That is, γ_t satisfies (25) and (26).

$$\mathrm{P}_{v_{t-1}}(S(\mathbf{X})\ge\gamma_t)=\mathrm{E}_{\mathbf{u}}[I_{\{S(\mathbf{X})\ge\gamma\}}] \tag{25}$$

$$\mathrm{P}_{v_{t-1}}(S(\mathbf{X})\le\gamma_t)\ge 1-\rho \tag{26}$$

where $X \sim f(\cdot\,;\mathbf{v}_{t-1})$. The simple estimator $\hat{\gamma}_t$ of γ_t can be obtained by drawing a random sample $\mathbf{X}_1,\ldots,\mathbf{X}_N$ from $f(\cdot\,;v_{t-1})$, calculating the performances $S(\mathbf{X}_i)$ for all i, ordering them from the smallest to the biggest: $S_{(1)}\le\ldots\le S_{(N)}$ and finally, evaluating the (1 - ρ) sample quintile as (27).

$$\hat{\gamma}_t = S_{[(1-\rho)N]} \tag{27}$$

Let's note that $S_{(j)}$ is called j^{th} order-statistic of the sequence $S(\mathbf{X}_1),\ldots,S(\mathbf{X}_N)$. Let's note also that $\hat{\gamma}_t$ is chosen such that the event $\{S(\mathbf{X})\ge\hat{\gamma}_t\}$ is not too rare (it has a probability of around ρ), and therefore, the reference parameter updated with a procedure such as (27) is not void of the meaning.

2. Adaptive Update of $\mathbf{v}_t$. For fixed γ_t and $\mathbf{v}_{t-1}$, let's derive $\mathbf{v}_t$ from the solution of the following cross-entropy program according to (28).

$$\max_{\mathbf{v}} D(\mathbf{v})=\max_{\mathbf{v}}\mathrm{E}_{\mathbf{v}_{t-1}}[I_{\{S(\mathbf{X})\ge\gamma\}}W(\mathbf{X};\mathbf{u},\mathbf{v}_{t-1})\ln f(\mathbf{X};\mathbf{v})] \tag{28}$$

The stochastic counterpart of (28) is shown as follows: for fixed $\hat{\gamma}_t$ and $\hat{\mathbf{v}}_{t-1}$, derive $\hat{\mathbf{v}}_t$ from the solution of the program according to (29).

$$\max_{\mathbf{v}} \hat{D}(\mathbf{v}) = \max_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^{N} I_{\{S(\mathbf{X}_i) \geq \hat{\gamma}_t\}} W(\mathbf{X}_i; \mathbf{u}, \hat{\mathbf{v}}_{t-1}) \ln f(\mathbf{X}_i; \mathbf{v}) \qquad (29)$$

Therefore, at the first iteration starting with $\hat{\mathbf{v}}_0 = \mathbf{u}$, the target event is artificially made less rare with temporarily use of the level $\hat{\gamma}_1$, which is chosen smaller than γ that for to get a good estimate for $\hat{\mathbf{v}}_1$. The value $\hat{\mathbf{v}}_1$ obtained in this way will make the event $\{S(\mathbf{X}) \geq \gamma\}$ less rare in the next iteration, so the value $\hat{\gamma}_2$ can be used in the next iteration, which is closer to γ itself. The algorithm terminates when the level is reached at some iteration t, which is at least γ and after the original value of γ can be used without getting too few samples. As mentioned before, the optimal solution of (28) and (29) can be often obtained analytically, in particular when $f(\mathbf{x}; \mathbf{v})$ belongs to the natural exponential family. The above rationale results are put in the following multi-level cross-entropy algorithm for the rare event simulation:

Algorithm:

1. Define $\hat{\mathbf{v}}_0 = \mathbf{u}$ *. Set t = 1. (Iteration = level counter).*

2. Generate a sample $X_1,...,X_N$ *with the density* $f(\cdot;v_{t-1})$ *and compute the sample (1 - ρ) - quintile* $\hat{\gamma}_t$ *performance according to (28) with* $\hat{\gamma}_t < \gamma$*. Otherwise, set* $\hat{\gamma}_t = \gamma$*.*

3. Use the same sample $X_1,...,X_N$ *to solve the stochastic program (29). Denote the solution by* $\hat{\mathbf{v}}_t$ *.*

4. If $\hat{\gamma}_t < \gamma$*, then set t = t + 1 and reiterate from Step 2. Else, proceed with step 5.*

5. Estimate the rare event probability l using (30),

$$\hat{l} = \frac{1}{N} \sum_{i=1}^{N_1} I_{\{S(\mathbf{X}_i) \geq \gamma\}} W(\mathbf{X}_i; \mathbf{u}, \hat{\mathbf{v}}_T) \qquad (30)$$

where T denotes the final number of iterations, or number of the levels used.

Table 2: Simulation Results

Iteration	Repetitive trails	λ	μ_1	μ_2	p	Estimation
1	10^4	0.2	0.8	0.2	0.5	-
2	10^4	0.216	0.643	0.258	0.364	$1.426e^{-15}$
3	10^4	0.198	0.621	0.287	0.322	$1.462e^{-15}$
4	10^4	0.196	0.614	0.279	0.318	$1.436e^{-15}$
5	10^4	0.195	0.614	0.282	0.320	$1.372e^{-15}$
6	10^4	0.196	0.612	0.284	0.322	$1.318e^{-15}$
7	10^6	0.196	0.612	0.284	0.322	$1.342e^{-15}$

As an example, let's apply this algorithm to the tandem queue with feedback shown in Fig.2, with the following entry parameters: λ=0.2; μ_1=0.8; μ_2=0.2; p=0.5. The simulation results for generating N=10,000 samples are shown in Table 2. As seen, overflow is obtained when there are j=6 iterations. The accuracy is increased up to N=1,000,000 in the seventh iteration, and the overflow probability is obtained equal to $\hat{\ell}_{IS}$ =1.342e^{-15}.

CONCLUSIONS

An effective approach to rare event simulation is demonstrated in this paper. Two simulation algorithms used to speed up the overflow probability simulation (that in queuing systems is a rare event) are developed; one with Importance Sampling procedure, and another with cross-entropy. Simulation results are shown for both algorithms. Recommended approach can be used to estimate quality of service in communication networks.

REFERENCES

Bolch, G., Greiner, S., Meer, H, Trivedi, K., 1998. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, New York, John Wiley& Sons, 726 pp.

Bueno, D. R., Srinivasan, R., Nicola, V., van Etten, W., Tattje, H., 2000. *Adaptive Importance Sampling for Performance Evaluation and Parameter Optimization of Communication Systems.* IEEE Transactions on Communications, 48(4), pp. 557-565.

Bucklew, J., 2004. *An Introduction to Rare Event Simulation.* Springer Series in Statistics, XI. Springer-Verlag. Berlin.

De Boer, P., Kroese, D., Rubinstein, R., 2002. "Estimating buffer overflows in three stages using cross-entropy." In *Proceedings of the 2002 Winter Simulation Conference*, San Diego, USA, pp. 301-309.

Heidelberger, P., 1995. *Fast Simulation for Rare Event in Queueing and Reliability Models.* ACM Transactions of Modeling and Computer Simulation, 5 (1), pp. 43-85.

Kalashnikov, V., 1997. *Geometric Sums: Bounds for Rare Events with Applications: Risk Analysis, Reliability, Queueing.* Kluwer Academic Publishers.

Keith, J., Kroese, D.P., 2002. "SABRES: Sequence Alignment by Rare Event Simulation." In *Proceedings of the 2002 Winter Simulation Conference,* San Diego,USA, pp. 320-327.

Radev, D., Lokshina, I., Radeva, S., 2005. "Clustering Procedure for Performance Analysis of Mobile Communication Networks." In *Proceedings of the EUROSIS-ETI 2005: 3rd Industrial Simulation Conference (ISC'05) at Fraunhofer IPK,* Berlin, Germany, pp. 83-87.

Radev, D., Rashkova, E., 2006. "Algorithms for traffic flows modelling with Monte Carlo simulation." In *Proceedings of TELEKOM'2006*, Varna.

Radev, D., 2003. "Splitting method in rare event simulation for high - speed communication networks." In *Proceedings of the International Conference Automatics and Informatics'2003*, Vol.1, Sofia, pp. 103-107.

BIOGRAPHY

DIMITAR RADEV, D.Sc. is Associate Professor at Department of Communication Technique and Technologies at University of Rousse, Bulgaria. His main research interests are Teletraffic Theory, Simulation and Modeling of Communication Networks, and Performance Analysis of Queuing Systems.

IZABELLA LOKSHINA, PhD is Associate Professor and Chair of Management, Marketing and Information Systems Department at SUNY Oneonta, USA. Her main research interests are Artificial Intelligence (Fuzzy Systems and Neural Networks) and System Modeling.

SOFTWARE MODEL DEVELOPMENT

A Component Framework for the Development of Complex Simulators

José A. Cortés, José A. Dianes, Manuel Díaz, Daniel Garrido and Enrique Soler
University of Málaga
Department of Languages and Computing Science
CP.29071 Málaga, Spain
E-mail: {jacortes, jdianes, mdr,dgarrido, esc}@lcc.uma.es

KEYWORDS
Component, Framework, Methodology, Code Generation, Simulator.

ABSTRACT

The development of complex simulators involves many hardware and software elements. The kernel of these simulators is composed of several hundreds of simulation models. The simulator kernel is also in charge of connecting with external applications providing simulation variables and responding to user actions. This paper presents a framework for simulators aimed at facilitating development of the simulator kernel and simulation models using software components. The use of the framework covers the whole life cycle of a simulation model: design, code generation and other aspects such as variables management and integration in the final simulation environment. In addition, the components are developed with Java*, a language subset of Java which allows platform independence to be obtained. In order to apply our proposal, we have developed a tool which allows these tasks to be carried out.

INTRODUCTION

Nowadays, simulation is used in very different situations such as aviation, the military, nuclear power plants (Díaz and Garrido, 2004), etc. Simulators allow the training of future operators to be carried out safely and cheaply. The development of these simulators is expensive and difficult and must take into account many different hardware and software elements as well as the coordination of development teams.

Often, the development of the simulators is carried out from scratch without applying modern methodologies provided by Software Engineering, which may be costly both in terms of time and money. Some new proposals can be found in (Dahmann and Fujimoto, 1997) (Object Management Group, 2002). More specifically, we think software components (Szyperski, 1999) are a useful technology for the development of simulators. Components have been used in many different software contexts and this allows reuse, adaptability and testing to be improved.

This paper presents a framework for the development of simulators through the use of software components. An important part of complex simulators are the simulation models. A complex simulator can have several hundreds of these models, responsible for the simulation of different parts of the modelled system. In our proposal, the simulation models are software components. In this way, the models can be easily integrated and used in different simulation environments. In addition, the simulator kernel is also a component and can be easily adapted to new simulation models that provide common interfaces.

The framework covers the whole life cycle of the simulation models. This way, the development of a model/component is carried out connecting primitive components developed with a neutral language, Java*. After this, automatic code generation is performed for the model/component allowing the integration of the model in a simulation environment, taking into account additional elements such as for example cycle, simulation variables, testing, etc. So, for example, we use introspection to test the model before its distribution.

In the following sections we present the operational environment of the simulators and the main elements of the presented framework. Next, we present the TEAM tool which allows the creation of components in the framework. In addition, we present two different situations where the framework has been used with the support of this tool. Finally, we present some conclusions.

OPERATIONAL SETTING

We distinguish between full scope simulators and interactive simulators. Full scope simulators are exact replics of the simulated system, where all the details are carefully reproduced, from physical objects to software,

simulating the applications running in the real system. Interactive simulators are simpler than full scope simulators. They are based on graphic applications which allow the same actions to be performed using graphical elements.

Both simulators use a large number of simulation models. These models are responsible for different parts of the system. Another important element in the simulators is the simulation variables. These elements represent physical parameters of the simulation such as for example, pressure, temperature, etc. Finally, the simulator kernel is responsible for the connection and interaction of these elements, providing an execution environment.

FRAMEWORK DESCRIPTION

The proposed framework is based on the intensive use of software components as a mechanism to improve productivity, reuse, etc. in both, simulator kernel and simulation models. External tools can communicate with the simulator kernel through the provided interfaces of these elements, allowing the querying of simulation variables and the performing of user actions.

Figure 1 shows a representation of the proposed architecture with the simulator kernel, simulation models and some external tools. Communication between these tools and the simulator can take place in many ways. In (Díaz and Garrido, 2004b) an approach based on the use of RT-CORBA (Schmidt and Kuhns 2000) is presented.

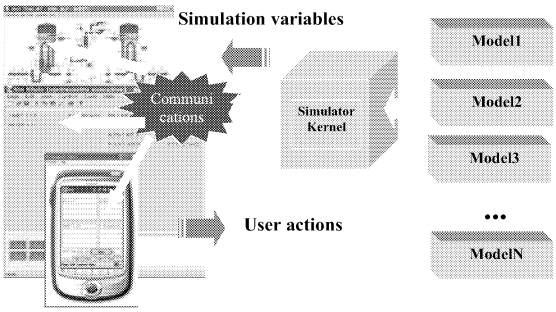


Figure 1. Framework architecture example

Simulation models

In the framework, every simulation model is a software component used by the simulator kernel. A simulation model is composed of input variables, behaviour and output variables. The developer of the simulation model component is responsible for creating the behaviour of the components. As regards simulation variables, the developer only has to indicate which variables the model needs and which variables the model updates.

The creation of new models is carried out through the connection of primitive components/models (described later) which use the values of the simulation variables, producing new values which are used by other primitive components. Finally, the output variables are obtained.

The framework also provides the possibility of creating model component containers. These containers can group several related component/models in such a way that externally they appear to be a single component.

The behaviour of the components is determined by the connection between the primitive components whose code is described in a new language, Java*. In addition, the primitive components can also be configured with the values of input variables, modifying their functionality. Finally, the model itself can be configured with several parameters such as for example the cycle.

Figure 2 shows a very simple example of a simulation model with three input variables. The model uses a lag function with the first variable. The other variables are added and finally the output variable is updated with the maximum of the lag function and the sum of the other two variables.

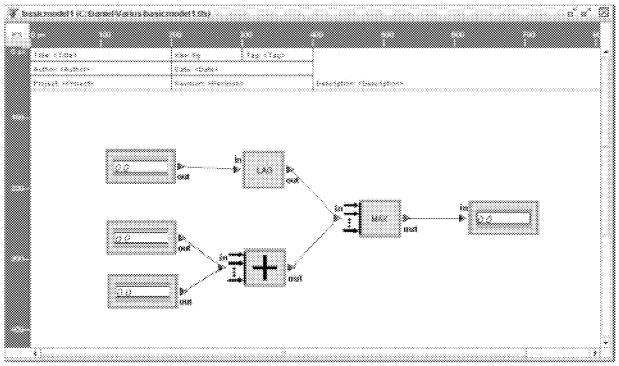


Figure 2. Composition example

Java* Language

Each simulator environment provides a set of primitive components developed in a neutral language, Java*, created with the purpose of achieving language independence.

Java* is a neutral subset of Java. The objetive of this neutrality is to obtain components which can be used in different simulation environments whose target language can be FORTRAN, Java, C, etc. Java* maintains most of the static constructions of Java and avoids some dynamic constructions which are not present in all the possible target languages considered. In addition, it adds modifications to some elements of Java such as for example array declarations (declaring the number of elements in a C way).

The following example shows the code of the LAG primitive component.

```
void LAG(double in, double out) {
  double output0[]={0.0};

  output0[0]=out;
  slag_1(in,output0,par1,deltat);
  out = output0[0];
}
/* Lag subroutine */
void slag_1(double input1,double output0[],
              double par1, double deltat) {
  double tf100;
  tf100 = deltat / par1;
  if (tf100 < 0.0)
     tf100 = 0.0;
     else if (tf100 > 1.0 )
     tf100 = 1.0;
   output0[0] = output0[0] + tf100 *
                    (input1 - output0[0]);
}
```

This way, a new component is developed combining primitive components and later, the component code is generated according to the desired target language such as for example FORTRAN or C#. The use of Java* provides the framework with great flexibility without loss of performance, because the component code generation is carried out at compilation time.

Simulation kernel

One important part of the framework is the existence of the simulator kernel which is the component responsible for the execution of the simulation models and their interaction with the other subsystems and tools. The simulator kernel provides an execution environment for simulation models. It is responsible for starting and stopping the simulation models, taking into account some of their properties such as the number of cycles per second. It also facilitates communication between different models in order to share simulation variables. The other responsibility of the simulator kernel is related to interaction with tools and subsystems. In this case, the simulator kernel must implement some communication interfaces which allow tool requests to be obtained and responses to be sent.

The following code shows part of the interfaces provided by a simulator kernel according to the framework:

```
interface ISimControl {
  // Start-stop methods
  void send_run();
  void send_freeze();
  ...
  void send_command(String command,
                    String params);
}
```

```
interface ISimVars {
  // Query of simulation variables
  void send_var_query(String varName,
                      double value);
  ...
  // Updating of simulation variables
   void send_var_update(String varName,
                        double value);
}
```

These interfaces provide methods for starting or stopping the simulation, send commands or querying/updating simulation variables.

Simulation variables

Simulation variables are an important issue in the development of the simulators. A simulation variable can represent physical properties of the modelled system or intermediate results of the models.

There are many different alternatives for the storing of simulation variables. In our solution, the components have interfaces indicating exported variables which can be queried by other components.

This way, during the initialization phase, component models ask for the component/model storing the desired variables, saving references for their later use.

The framework provides the possibility to use different simulation variable types such as binary, analog (single or double precision) or even defining variable arrays. In addition, simulation variables can form groups which allow users to better organize their management. So, for example, the variables of one subsystem can compose a simulation variables group.

Integration

One of the main features in a framework is the creation of "connection points" which allow the framework to be used by the developers. In this sense, the developed models / components provide interfaces which allow the integration and use of the components in different simulators. This approach allows a high degree of reuse for the models developed according to the framework.

The following code shows part of the public interfaces of the components/model.

```
interface IModel {
  // Initialization
  void init(String modelName,
            String configInfo);
  // Execution of the next frame
  void nextFrame();
  // Query of simulation variables
```

```
  void send_var_query(String varName,
                      double value,...);
  // Updating of simulation variables
  void send_var_update(String varName,
                       double value);
}
```

The developer does not need to know these interfaces, because the TEAM tool (described later) automatically generates and maintains these interfaces.

TEAM TOOL

The component generation process in our framework is clearly represented by the TEAM tool dynamics. This tool allows users to design, validate and integrate models with the simulation environment. Moreover, it offers mechanisms to create the basic composition units, the primitive models, such as introspective validation of composed components. The following are key elements in the TEAM tool: icons, libraries, sheets, autotest, verification and version control.

- *Icons* represent primitive components. They can include Java* code and a graphic interface, including interactive elements and data fields, for using with the component design tool. Data fields can represent simulation variables.

Icons have input and output ports that represent the interfaces with other components. An input port means that the component takes a parameter from another component output port. Moreover, icon designers can create interactive and dynamic elements, like buttons, bars or text fields, that the TEAM tool represents during execution of the component and autotest (described later). Figure 3 shows the ADD icon with input and output ports and associated code.

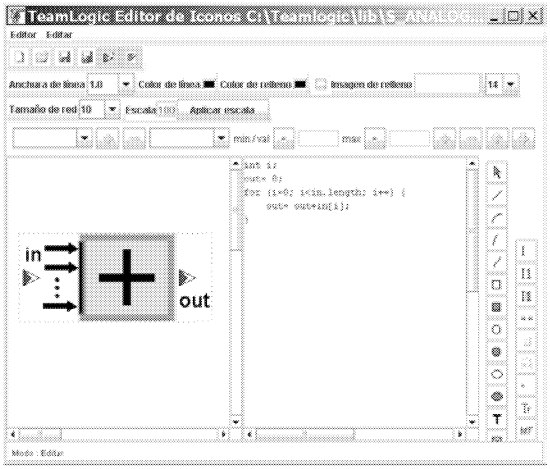


Figure 3. Icon Editor

- Icon *libraries* are repositories, normally grouped around some kind of criteria like functionality, such as logic gate families, instrumentation, etc.

- Icons are instantiated and interconnected in a *sheet*. These comprise the design view of a component-model.

- Component *autotesting* (Figure 4) is an introspection mechanism that allows component designers to simulate the component/model without connecting with the simulation environment. In this mode, users can interact with the graphical interface of a component as they do when integrated with the simulation environment.

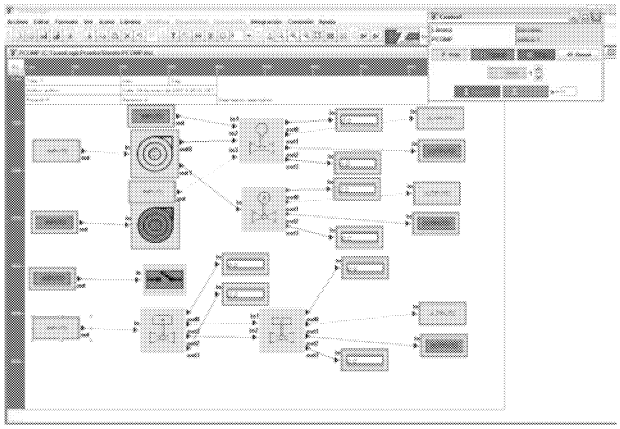


Figure 4. Autotest

- Component *verification* mechanism ensures that both component building tools, icon interconnection and variable instantiation, are used correctly. During connection verification, a test based on specific semantic rules is used to assure the correctness of data types and cardinality among others things. Icon instantiation via variable value assignment is also subjected to icon designer defined rules.

- Models also contain meta-information about the *version control* mechanism, keeping all the history of the component. The TEAM tool also offers *simulation environment connectivity*, making it easy for users to recover and modify simulation values, interacting directly or via graphical sheet representation.

APPLICATION CASES

The TEAM tool has been used satisfactorily to develop nuclear power plant simulator software, making it easy to adapt to different situations thanks to the many framework features. Here we are going to show two different TEAM application cases: TEAM-Logic (Figure 5) and TEAM-Trac (Figure 6).

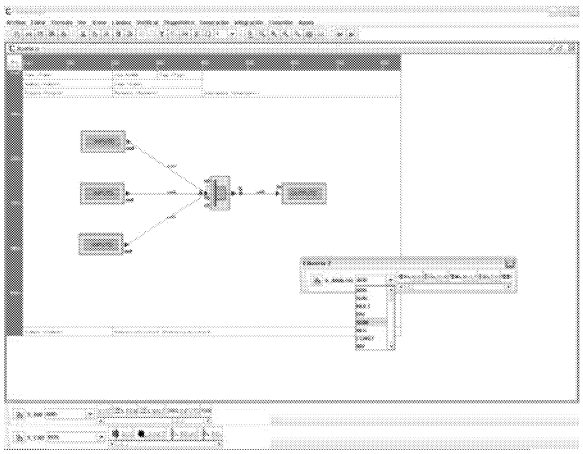


Figure 5. TEAM-Logic

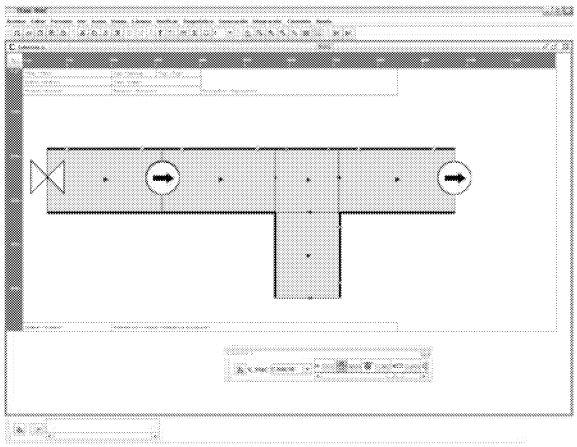


Figure 6. TEAM-Trac

TEAM-Logic allows simulator developers to design the different logic and wired models of the power plants. It offers up to six icon families including binary logic, analog, wiring, etc. The tool exploits framework features offering users an edition features, model verification and autotest as well as all the steps to generate and integrate the model-components with the simulation environment. During simulation or autotest, users can interact in both ways, graphically or getting and putting values directly into variables. This is a very powerful tool for models designers, that can test their models easily.

The TEAM-Trac component generation process produces thermo-hydraulic code that represents a configuration input file to the TRAC simulation model (Spore and Shumway, 1981). Icons are much more complex than those of TEAM-Logic, but are less numerous too. Verification rules are more restrictive, ensuring correctness of many of the compositions and constructions of thermo-hydraulic pipes during design time. This can be done in this way thanks to verification rules about connection and instantiation of different icons. In these constructions about 90% of the model information resides. TEAM-Trac icons has no Java* code, only many data fields. These data fields have specific instantiation dialogs and mechanisms to make it easier to fill values that are repeated, or even defined by default, in many situations.

CONCLUSIONS

This paper has presented a component-based framework for developing complex simulator systems. The framework allows simulator developers to design new models/components easily. The use of the framework covers the whole life cycle of a simulation model: design, code generation, integration. Components developed with the framework can also be integrated into different simulation environments. A new language, Java*, was developed to produce component code with platform independence. A tool, TEAM, for the development of the components has also been presented. The tool offers designers a graphical way to exploit all the features of the framework, including some extra functionalities such as introspection, validation mechanisms as well as a graphical interaction interface with simulations. Finally, two specific applications domains have been looked at.

REFERENCES

Díaz, M. and Garrido, D. “A Simulation Environment for Nuclear Power Plants”. 8th IEEE Intern. Symp. on Distributed Simulation and Real Time Applications, pp. 98-105, 2004.

Díaz, M. and Garrido, D. “An Experience Report: RT-CORBA in Nuclear Power Plant Simulators”. 2nd Industrial Simulation Conference, pp. 243-247, 2004.

Dahmann, J.S.; Fujimoto, R. and Weatherly, R.M.. “The Department of Defense High Level Architecture", Winter Simulation Conf., pp. 142-149, 1997.

Object Management Group, Distributed Simulation Systems Specification, version 2.0, formal/2002-11-11.

Schmidt D.C. and F. Kuhns. “An overview of the Real-time CORBA Specification”, IEEE Computer special issue on Object-Oriented Real-time Distributed Computing, June 2000.

Szyperski C. 1999. *Component Software: Beyond Object-Oriented Programming,* Addison-Wesley Longman.

Spore, J.W.; Shumway, R.W.; Giles, M.W. and Singer, G.L. TRAC-BD1: An Advanced Best Estimate Computer Program for Boiling Water Reactor Loss-of-Coolant Accident Analysis. NUREG/CR-2178, Volume 1-4. U.S. Nuclear Reg. Commission, 1981.

Modeling Complex Systems using the Web Ontology Language (OWL)

Tobias Lehmann and Andreas Karcher
Universität der Bundeswehr München
85540 Neubiberg, Germany
E-mail: {tobias.lehmann|andreas.karcher}@unibw.de

INTRODUCTION

Subject to research of this paper is how the W3C-Standard "Web Ontology Language" (OWL), which is means for formal and distributed knowledge representation, can natively provide help for system-modeling and system-analysis. Native in the context of this work means, that no class has to be defined by a user because it is already part of the language itself. For this, common characteristics of systems based on scientific literature are described. These characteristics are then mirrored on features the W3C-Standard OWL provides. As an important aspect of OWL, a variety of restrictions on classes and properties are integral part of the language. By using these restricting elements of OWL, important concepts for the modeling of systems are provided by default as shown in further parts of this work. After outlining OWL-modeling features selected analyzing-methods are presented which allow operationally-analysis and consistency-checking based on second order logic inferences. Concluding added-value and potentials as well as drawback of this approach are outlined.

MOTIVATION

Semantic Web Technologies are means for knowledge representation and reasoning in the field of heterogeneous and distributed environments such as the internet. They allow, amongst other features, not only modular and distributed modeling of subjects of interest (Niles and Pease 2001) but furthermore facilitate ease for the integration of similar models (Bouquet et al. 2004). For this, similarity of concepts within different models is inferred based on graph-theoretical algorithms (Diestel 2006), and used as a starting basis for mapping-operations. In addition to this, Ontolgies modeled in OWL have – compared to the Unified Modeling Language (UML) or most Data Bases – a flexible structure level (Bechhofer et al. 2004). That is classes and properties are defined and additionally enriched by descriptions (asserted structure) which are then resolved by Reasoners to new structures (inferred structure) (McGuinness and Harmelen 2004). Instances are mostly neglected here and primarily follow their classes.
For these beneficial reasons this paper asks the question, in which way the Standard-Language OWL is also feasible for modeling systems, which have particular requirements concerning model-elements. If OWL was feasible to model systems, then modelers would be able to kill two birds with one stone and be able to bring together two perspectives of modeling within one language. As far as the analysis methods are concerned, this paper focuses on the question, whether a given system is operational or not. What "operational" means in detail will be described later.

RELATED WORK: METAMODELS FOR SYSTEMS

This section first describes systems as they are seen in this paper based on common literature. Focus of this description is to find out, which meta-elements are to be considered system-specific – not depending on any technology or implementation. The meta-elements of the Web Ontology Language are outlined, based on its specification defined by the W3C (which is technology-independent as well). Then, the two sets of meta-elements are compared with each other, and – as the main conclusion of this paper – OWL's compliance to requirements of system modeling is estimated. Since native OWL does not cover all required elements, the last step is to complete all necessary elements determined by step one. An example for the application of systems-modeling in OWL is given over the different steps of the paper.
A system in the context of this work is (not exclusively) based on an (Luhmann and Baecker 2004), (Hill et al. 1994), (Foerster and Poerksen 2002) and (Foerster 2005) defined is by the following characteristics:

- It consists out of **System-Elements** and **Subsystems**, which **interact** with each other, or are connected in any way
- It has a **border**, which separates the inside of a system from its outside – the environment
- It is called **closed**, if there is no interaction of System-Elements with their environment, otherwise it is defined **open**
- A system serves particular **objective**(s), fulfills a special **function** or has a certain internal intent
- The **intent** of a system is of **generic** nature, it can change over time
- A system can, from different perspectives, follow **multiple objectives**, which do not necessarily have to be correlated.

In the following the characteristics defined above are mapped to elements necessary for a model of a system. The model presented is only one option and it does not claim to be exhaustive. As presented in Table 1, all keywords listed above are used as labels for criteria of systems and are connected with possible Meta-Elements. All numbers are used as anchors in describing text. Systems consist of System-Elements which can be (Sub-) Systems (1) again. Particular Systems can only contain certain Subsystems, which are of interest in the given model. This leads to "typed" systems, which are generally defined at structure level and created/modified and deleted at instance level.

The interaction of Systems (2) and their elements is modeled by relations between these. Relations within OWL are not exclusively allocated to structure or instances, but appear in various ways at both levels.
A system's border (3) is more complex and less directly definable. At least a model implicitly defines its border by having a limited set of elements. But separation of systems could also be achieved by systems' nature, which can be defined via inheritance. An infrastructural system is no weather system, although both might occur within one model. If constraints define two systems to be disjoint, then a border is created as well.

Table 1: Meta-Elements of abstract systems

#	Criteria	Meta-Element
1	System(element) / Subsystems	Class, Instance
2	Interaction	Relation
3	Border	Model, Constraint, Inheritance
4	Open-Closed	Inference on Relations
5	Objective, Function	Class-Relations
6	Generic intent	Instances, Inheritance
7	Multiple Views / Purposes	Multiple Relation Domains and Ranges, Multiple Inheritance

If a system is open or closed (4) can be derived from (3) and (2) depending on whether a relation both on structure or instance-level crosses the boarders defined. From the semantic point of view structure and instance level are an artificial construct, which – projected into real world – are often hardly distinguishable (Motik 2005).
In order to define a functionality or objectives for a system (5) different approaches can be taken into consideration. First, objectives can be defined and related to systems by means of relations (2). This lead to the direct inclusion of objective into system-models. Second, objectives can be kept out of the model and be implicitly connected to elements within the model by its user. Since objectives in general endeavors are usually interpreted strongly subjective and tend to be more vague than numerically definable (Doyle and Thomason 1999, Gregory 1988) a combination of both options with focus on the latter one seems most promising (Klein 1998).
Generic elements (6) within systems can be achieved by the usage of structure and instance-level. A particular class can be seen as placeholder for subclasses as well as for instances its own. So a restriction on relations at structure level lead to a restriction of selectable instances as well. This implies the necessity for restricted ranges and domains of relations.
Aspect #7, multiple view can be expressed by different models, which all contain an similar subject. Although different in properties and structure-relations (such as superclasses), one identical "thing" can occur in different shapes. A automobile can – from marketing standpoint – be something to count, without any interest in things such as internal state, technical data etc. in one model. It can be seen completely different during assembly by an engineer. But still in both models (of one or more enterprises) the car can be an identical thing.
The technologies examined here are limited to OWL. This is done for several reasons. First OWL is the latest Standard in W3C-Languages for Knowledge Representation. Second it is currently semantically the richest common language (Ding and Arroyo 2003, Gomez-Perez and Corcho 2002) and a direct successor to DAML + OIL (Mcguinness et al. 2002). This paper is meant to give an overview on how systems could be modeled, with one proof of concept, which is OWL here (and might be another language in different applications).

THE WEB ONTOLOGY LANGUAGE

Languages in the field of Semantic Web such as OWL are usually less connected to Knowledge Processing than to pure Knowledge Representation (Koide et al. 2005). OWL itself provides no executable methods, but requires external inference engines (Friedman-Hill 2003), rule systems (Horrocks and Patel-Schneider 2005) or wrappers to make structure and data accessible for applications (Knublauch et al. 2006). Still, encapsulating a knowledge structure in own language systems is advantageous, since it leads to more semantically powerful structures. Amongst others this is property-centric modeling (Bechhofer et al. 2004) or a dynamic and logic-based inference (Baader et al. 2004) of class structures introduce new features in Knowledge Representation. Besides this flexible mechanism to create structure by assertions and descriptions using Reasoners (Baader et al. 2003), OWL offers an fully URI-based referencing mechanism for elements on class level and instance level. Therefore an native implementation of distributed – but seemingly local – KBs (Patel-Schneider 2005) is possible. This capability for distributed modelling, creating and maintaining KBs can be combined (and boosted) with standardized so called upper-level and mid-level Ontologies. The latter ones work as reference for lower and more specific Domain-Ontologies. If done so, decentralized KBs can be achieved, where domain-experts maintain their special-field only. For usage all domain-Ontologies are then integrated into one single virtual Ontology. Sure, this sublime concept for integration will not work seamlessly, but compared to classical integration of different model (maybe in different languages also) and domains of application, this can lead to a dramatic reduction of transaction-costs (Froese 2004).

Matching Meta models

Due to the flexible structure of OWL, its Meta-model is compared to UML enriched by defined Restrictions, which can be resolved by Reasoners. Not all elements shall be presented here but instead referred to (Smith et al. 2004). Table 2 maps requirements derived from the description of systems to meta-elements of OWL. The identifier "From" hereby refers to the meta-elements of Table 1.

Table 2: Meta-elements of systems and of OWL

From	Meta-model Element	Meta- Element in OWL
1	Structure-Element	owl:Class
1	Instance	[className]
2	Relation at Structure-level	owl:ObjectProperty, owl:DatatypeProperty
2	Relation at Instance-	[PropertyName]

	level	
3	Constraints at Structure-level	owl:equivalentClass owl:disjointWith owl:allValuesFrom owl:someValuesFrom
3	Constraints at Instance-level	owl:allValuesFrom owl:someValuesFrom owl:hasValue
1	Relations on Relations at Structure-level	rdfs:subPropertyOf owl:equivalentProperty owl:inverseOf
1	Relations on Relations at Instance-level	owl:TransitiveProperty owl:SymmetricProperty
5	Objective, Functionality	OWL-Document
3	Constraints on Relations at Structure-level	owl:allValuesFrom owl:someValuesFrom
3	Constraints on Relations at Instance-level	-
6,7	Multiple Inheritance	rdfs:subClassOf
6,7	Multiple Relations at Instance-level	owl:FunctionalProperty owl:InverseFunctionalProperty

Restrictions required for modeling Systems as described above are so far almost all provided by OWL. But Reasoners that resolve these restrictions are not focused on Instances but on Classes only. In order to make use of this information at Instance-level it is necessary to connect OWL to different inference engines, which can deal with Instance-inferences using exactly the description from OWL.

SYSTEM-ANALYSIS

System-Analysis is a wide field, where no method is able to cover all imaginable requirements. Therefore many methods have developed over time. Subject to the System-Analysis of this paper is the question, if a system is operational or not. Operational in the context of this work is the meeting of constraints defined for systems, System-Elements and subsystems. As an example (depicted in

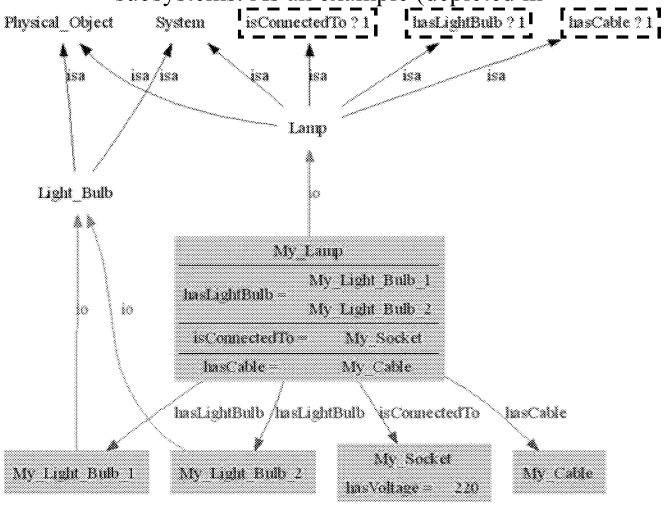


Figure 1) a lamp is defined operational, if "has" at least one operational light-bulb (sub-system), is connected to a socket which is operational and has no cable damage. All classes in

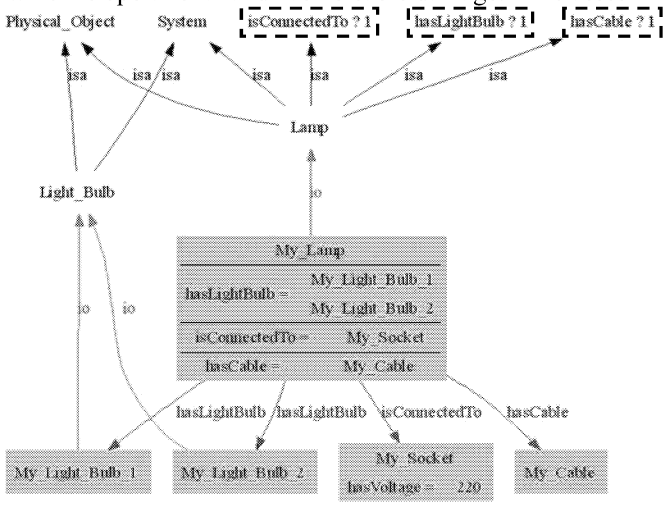


Figure 1 are rendered as black font, where relations are arrows carrying a brief description (*isa* = subclass of; *io* = instance of, or name of a property). Instances are visualized by grey rectangles. Minimum cardinality-restrictions are bordered with dotty lines. The idea behind this analysis based on OWL is, that each class can carry restriction for its properties (minimum cardinality *hasLightBulb* = 1). But instead of using these restrictions for Reasoning only they can also be taken into consideration, when dealing with instances.

The key point is now that, if our lamp has to be connected to a socket, it depends on the model, whether the socket is subject to further examination or not. If it is of interest to determine this, *Socket* can be defined as a *System* as well and is then also taken into account as a Subsystem. Or it might have simple constraints which are definable from the lamp's point of view such as a property *Voltage* with value *220V*.

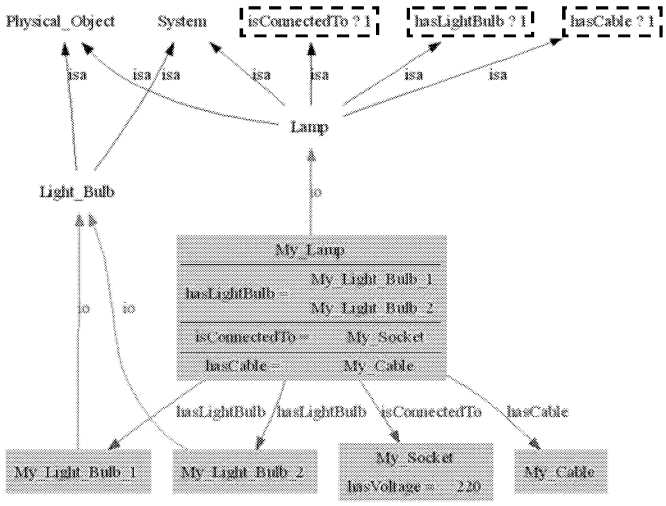


Figure 1: System example

This option for analysis is less determined, since it defines a *Socket* operational, if this restriction is met, no matter what a *Socket* consists of, is connected to, or any other relation.

Additionally for a better handling this implies a new classification which is Systems, System-Elements, System-Atoms and Subsystems.

Systems

A System is defined by inheriting from a class *System*. It may contain any property or have unrestricted number of subclasses. If systems carry restrictions defined in OWL such as in the *Lamp*-example, then these restrictions are evaluated, and the System-object is operational, if all of them are met.

System-Elements

Any object connected to a System via property is by definition a System-Element. System-Elements are operational by existence and require no further analysis. The object *MyCable* is a System-Element, since only its existence is necessary.

System-Atoms

If a System-Element is connected to an instance-related condition such as "value of a property A greater than X", then it is a System-Atom. System-Atoms have to meet their conditions to be counted operational.

Subsystem

Subsystems are systems which are connected to a system via property and inherit from the system class. They are operational, if an evaluation equal to a system-evaluation is successful. In the lamp-example, each *Light_Bulb* instance is a system and a subsystem of *MyLamp*. Restrictions to be fulfilled by a *Light_Bulb* could look like as depicted in Figure 2.

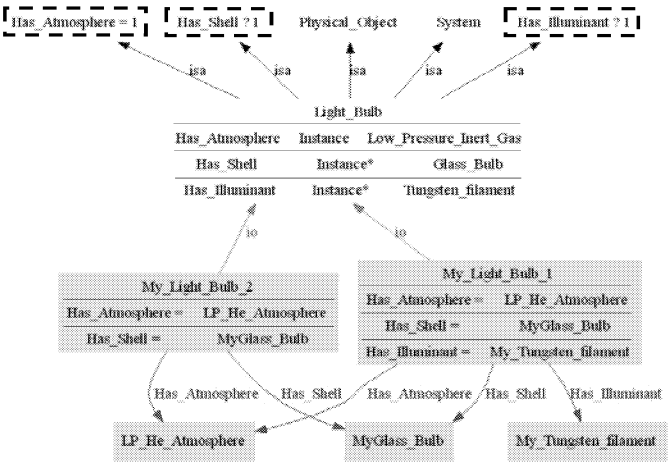


Figure 2 An example Light Bulb System

A Light Bulb requires at least one *Shell*, an *Illuminant*, and an *Atmosphere*, which are all System-Elements.

Conditionals

Conditions bridge Class-level and Instance-level restrictions. They associate System-Atoms with expressions that can be evaluated to Boolean values. Elements available are defined within a (mathematical) ontology, which is depicted in parts in Figure 3, and contain Classes for *Formulae* as well as *Numerical* and *Logical Comparators*.

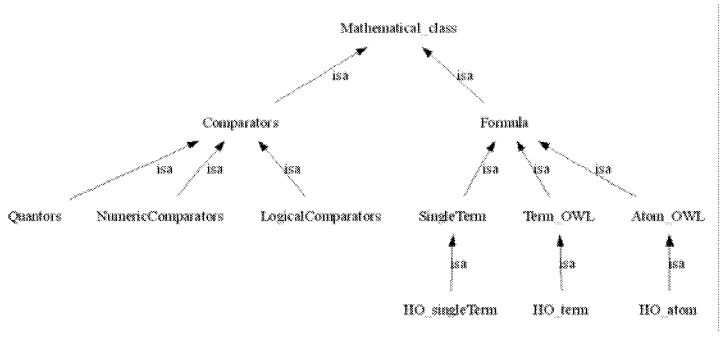


Figure 3 Mathematical Ontology – Excerpt

These statements can then be connected with classes of any ontology using a *ConditionLookUpTable*-Instance. Domain-Classes (Lamp) are connected with Range-Classes (Socket) and mean, that from the perspective of this relation System-Atoms need to fulfill the given Condition (see Figure 4).

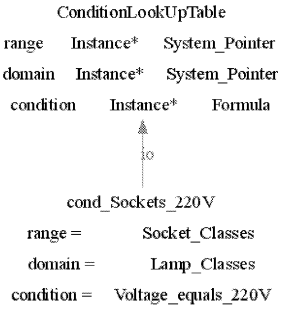


Figure 4 Conditional -Class and example instance

The construct described above is necessary in OWL, since it provides no possibility to define default values or default-instances for classes. Other languages for Knowledge-representation are more flexible in this aspect.

Evaluating DL-Statements at Instance-level

In order to evaluate the restrictions defined at class-level for the determination of operability of *System* instances, it is necessary to make use of external inference engines. It is technically of no importance which logical system is chosen, if Database, or logical languages such as Jess (Friedman-Hill 2003), Mercury (Somogyi et al. 1995) or Prolog (Wielemaker 2006). The implementation this paper refers to has chosen SWI-Prolog as technology (conversion and binding is described in (Lehmann and Karcher 2006), but all dependencies formulated are given as logical statements not depending on any language. Within the following expressions, a Knowledge Base is described by five Second-Order-Logic (Huth and Ryan 2004) structural predicates. These are:

- class(Cls) := A Class, that carries the name defined in Cls
- superClass(Sub, Sup) := The Class with name defined in *Sub* is Sub-Class of the Class carrying the name defined in *Sup*
- concept(Inst, Cls) := An Instance carrying the name defined in *Inst* is of Class carrying the name defined in *Cls*
- relation(Sub, Pre, Obj) := An arbitrary concept carrying the name defined in *Sub* is related to an arbitrary

concept carrying the name defined in *Obj* by a relation carrying the name defined in *Pre*

- restriction(RName, Cls, Pre, Val) := A Restriction carrying the name defined in *RName*, of value defined in *Val*, restraints a relation named by *Pre* for Class defined in *Cls*.

Logical dependencies

Based on the elements above the analysis-methods are introduces in two steps: first all functions are listed (see Table 3). Second by assembling elementary functions complex functions are generated which describe the system-analysis at a higher level of abstraction.

Table 3: Methods for System-Analysis

#	Method	Description
1	opSystem(Sy)	Is a system *Sy* operational or not.
2	isSystem(Inst)	Is an instance *Inst* of class System or not.
3	restOfASystem(Re, Sy)	Which restrictions *Re* does a system *Sy* have to meet?
4	subsystemOf(Sy, SubSy)	All subsystems *SubSy* for a System *Sy*
5	opSubSystem(Sy, SubSy)	Determines all operational subsystems *SubSy* of system *Sy*
6	restFulfilled(Re, Sy)	Determines all restrictions *Re*, that are met by system *Sy*
7	sysElements(Se, Sy)	All System-Elements *Se* of System *Sy*
8	condFulfilled(Se, Sy)	All System-Elements *Se* of System *Sy*, that fulfill their conditions
9	count(Set, NrElems)	Determines the number of elements *NrElems* of set *Set*
10	sysAtom(Sa, Sy)	All System-Atoms *Sa* of system *Sy*

The definition of an operating System is here:
If *Sy* is a System, then for all Restrictions *Re* of *Sy* are fulfilled.

opSystem(Sy) ← □ isSystem(Sy) ∩
□(restrictionsOfASystem(Re, Sy) ← restriction(Rn, Sy, Re, Rv) ∩ restrictionMet(Re, Sy))

Since Reasoners resolve Subclass-related Restrictions into new Class-Structures, only Restrictions aiming at Cardinalities of properties are considered here. A Restriction Res on Relation Re is counted as fulfilled for System Sy, if the sum of System-Elements, System-Atoms meeting their conditions and operational Subsystems is compliant to the restriction. This is here a distinction between minimum, maximum and exact cardinality.

restFulfilled(Res, Sy) ← relation(Sy, Re, Obj) ∩
restriction(Rn, Sy, Re, Rv) ∩
(
Count(sysElements(Obj), ElC) ∩
Count(sysAtoms(Obj, Sys), AtC)) ∩ condFulfilled(Obj, Sy)
∩
Count(opSubSystem(Sys, Obj), SC)
) ∩ (
(Rn = 'MaxCardinality' ∩ (ElC + AtC + SC) <= Rv) □
(Rn = 'MinCardinality' ∩ (ElC + AtC + SC) >= Rv) □
(Rn = 'Cardinality' ∩ (ElC + AtC + SC) == Rv) □
)

In addition to numerical restrictions OWL allows universal quantification and existential quantification which are definable by cardinality > 0 and cardinality = count(sysX) statements. Therewith all required elements for the resolution of OWL's Restriction at Instance-Level have been introduced. They are used for the inference of operational systems, which is only option for their usage besides classical DL-reasoning.

Benefiting from SW-Technologies

When using OWL as modeling-language, not only descriptions and restrictions are available which can be used for purposes such as mentioned in this paper, but within the "surroundings" of this language a set of beneficial technologies and methodologies is also available. This is fist of all to mention distributed modeling based on common high-level standards (Niles and Pease 2001). Domain-experts can create and maintain Knowledge Bases without being "distracted" by aspects of other domains – besides interfaces. Second, these interfaces can then be again integrated a semi-automated way by using alignment-tools as a support (Bruijn et al. 2004, Ehrig and Sure 2005). The third benefit is to be seen in Description Logics which not only help to build a flexible Class-structure but furthermore allow the querying for fitness of a concept within an existing structure. If a new concept should be added to an existing Ontology then descriptions will help to sort it in a better way (Bechhofer 2003).

CONCLUSION AND OUTLOOK

OWL has - in the context of this paper - proven to provide a meta-structure that contains many elements useful for modeling Systems. Besides relations that connect Classes and Instances in a default-value-way and mathematical expressions, every concept defined as prerequisite in this paper is available. Figure 5 depicts an overview locating introduced elements to OWL meta-elements (*owl:class* and *owl:restriction)*, to classes in an OWL model (*System*, *Conditional*) and to inferred concepts, that do not have to be defined explicitly (*System-Element*, *System-Atom* and *Sub-Systems*).

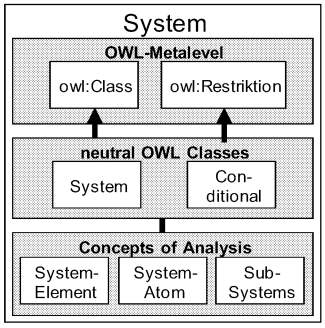


Figure 5 Dependencies between OWL and System-Analysis

In total, necessity for modification can be kept relative low, what leads to the conclusion, that OWL is – compared e.g. to

UML – a better means for modeling Systems. On the other hand, all analysis operations are to be carried out by third-party applications bringing additional costs for integration on the table. As future work, this integration should be made seamless not only for one inference system but for as many as possible by a generic integration architecture.

REFERENCES

Baader, F., Calvanese, D., L.McGuinness, D., Nardi, D. and Patel-Schneider, P.F. 2003. "The Description Logic Handbook". Vol. 1. Cambridge: Cambridge University Press.

Baader, F., Horrocks, I. and Sattler, U. 2004. "Description Logics", in *Handbook on Ontologies*, Stab, S. and Studer, R., Editors. Springer: Berlin Heidelberg New York

Bechhofer, S. 2003. "OWL Reasoning Examples". [cited 2006-08-15]; Available from: http://owl.man.ac.uk/2003/why/latest/

Bechhofer, S., Harmelen, F.v., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F. and Stein, L.A. 2004. "OWL Web Ontology Language Reference". [cited 2006-06-30]; Available from: http://www.w3.org/TR/owl-ref/

Bouquet, P., Ehrig, M., Euzenat, J., Franconi, E., Hitzler, P., Krotzsch, M., Serafini, L., Stamou, G., Sure, Y. and Tessaris, S. 2004. "Specification of a common framework for characterizing alignment, knowledgeweb deliverable d2.2.1v2." Universitat Karlsruhe: Karlsruhe.

Bruijn, J.d., Mart´ın-Recuerda, F., Manov, D. and Ehrig, M. 2004. "D4.2.1 State-of-the-art survey on Ontology Merging and Aligning V1". Digital Enterprise Research Institute, University of Innsbruck: Innsbruck. pp. 99.

Diestel, R. 2006. "Graphentheorie". Vol. 3. Berlin: Springer.

Ding, Y. and Arroyo, S. 2003. "State of the art on Semantic Web Languages". Institut for Informatics University of Innsbruck: Insbruck. pp. 49.

Doyle, J. and Thomason, R. 1999. "Background to qualitative decision theory". AI magazine. (**20**:2), pp. 55–68.

Ehrig, M. and Sure, Y. 2005. "FOAM - Framework for Ontology Alignment and Mapping; Results of the Ontology Alignment Initiative", in *Proceedings of the Workshop on Integrating Ontologies*, Ashpole, B., Ehrig, M., Euzenat, J. and Stuckenschmidt, H., Editors. CEUR-WS.org. p. 72-76

Foerster, H.v. 2005. "Einführung in den Konstruktivismus ". München: Piper Verlag GmbH.

Foerster, H.V. and Poerksen, B. 2002. "Understanding systems : conversations on epistemology and ethics". New York: Carl-Auer-Systeme.

Friedman-Hill, E. 2003. "Jess in action ". Greenwich, CT: Manning Publications Co.

Froese, A. 2004. "Ad-hoc Integration heterogener Dokumente". Berlin: Logos Verlag.

Gomez-Perez, A. and Corcho, O. 2002. "Ontology languages for the semantic web". IEEE Intelligent Systems. (**17**:1), pp. 54-60.

Gregory, G. 1988. "Decision Analysis". New York: Plenum Press.

Hill, W., Fehlbaum, R. and Ulrich, P. 1994. "Organisationslehre 1". Vol. 5. Stuttgart: UTB.

Horrocks, I. and Patel-Schneider, P.F. 2005. "SWRL 0.5 (aka Rules for OWL) Overview". Bell Labs.

Huth, M. and Ryan, M. 2004. "Logic in Computer Science". Cambridge: Cambridge University Press.

Klein, G. 1998. "Sources of Power - how people make decisions". Cambridge, MA: MIT Press.

Knublauch, H., Oberle, D., Tetlow, P. and Wallace, E. 2006. "A Semantic Web Primer for Object-Oriented Software Developers W3C Editor's Draft, $Date: 2006/03/09 14:58:26 $". [cited 2007-03-06]; Available from: http://www.w3.org/TR/sw-oosd-primer/

Koide, S., Aasman, J. and Haflich, S. 2005. "OWL vs. Object Orientated Programming". In *International Semantic Web Conference, Workshop1: Semantic Web Enabled Software Engineering*. Galway: Workshop paper

Lehmann, T. and Karcher, A. 2006. "Logic Based Agents and Ontologies Supporting Decision Making in Effects Based Operations". In *The 10th World Multi-Conference on Systemics, Cybernetics and Informatics*. Orlando, Florida

Luhmann, N. and Baecker, D. 2004. "Einführung in die Systemtheorie". Heidelberg: Carl-Auer-Systeme

Mcguinness, D.L., Fikes, R., Hendler, J. and Stein, L.A. 2002. "DAML+OIL: an ontology language for the Semantic Web". Intelligent Systems, IEEE. (**17**:5), pp. 72- 80.

McGuinness, D.L. and Harmelen, F.v. 2004. "OWL Web Ontology Language Overview". [cited 2005-11-24]; Available from: http://www.w3.org/TR/owl-features/

Motik, B. 2005. "On the Properties of Metamodeling in OWL". In *Proc. of the 4th Int. Semantic Web Conf. (ISWC 2005)*. Galway, Ireland: Springer

Niles, I. and Pease, A. 2001. "Towards a Standard Upper Ontology". In *2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*. Ogunquit, Maine

Patel-Schneider, P.F. 2005. "What is OWL (and why should I give a hoot)?" [cited 2007-05-08]; Available from: http://www.cs.bell-labs.com/cm/cs/who/pfps/talks/owl/index.html

Smith, M.K., Welty, C. and McGuinness, D.L. 2004. "OWL Web Ontology Language Guide W3C Recommendation". [cited 2005-09-25]; Available from: http://www.w3.org/TR/

Somogyi, Z., Henderson, F. and Conway, T. 1995. "Mercury: an efficient purely declarative logic programming language ". In *Australian Computer Science Conference*. Adelaide, South Australia: Glenelg

Wielemaker, J. 2006. "SWI-Prolog 5.6 Reference Manual - version 5.6.17". [cited 2006-08-04]; Available from: http://gollem.science.uva.nl/cgi-bin/nph-download/SWI-Prolog/refman/refman.pdf

HUMAN BEHAVIOURAL MODELLING

Multimodal Web based system for human emotion recognition

Dragos Datcu, L.J.M. Rothkrantz
Man-Machine Interaction Group
Delft University of Technology
2628 CD, Delft,
The Netherlands
E-mail: {D.Datcu ; L.J.M.Rothkrantz}@ewi.tudelft.nl

KEYWORDS
Automatic emotion recognition, speech analysis, face detection, facial feature extraction, facial characteristic point extraction, AAM, SVM.

ABSTRACT

The system being described in the paper presents a Web interface for a fully automatic audio-video human emotion recognition. The analysis is focused on the set of six basic emotions plus the neutral type. Different classifiers are involved in the process of face detection (AdaBoost), facial expression recognition (SVM and other models) and emotion recognition from speech (GentleBoost). The Active Appearance Model – AAM is used to get the information related to the shapes of the faces to be analyzed. The facial expression recognition is frame based and no temporal patterns of emotions are managed. The emotion recognition from movies is done separately on sound and video frames. The algorithm does not handle the dependencies between audio and video during the analysis. The methodologies for data processing are explained and specific performance measures for the emotion recognition are presented.

INTRODUCTION

Nonverbal communication plays an important role in everyday life of the people. As people have better understanding of the emotions related mechanisms and the correlation with human behaviour, it became even more necessary to design systems to automatically detect the human's emotional state. Nowadays it has been a challenging task to realize algorithms to identify meaningful clues and to learn machines to analyze human faces and utterances for extracting emotions.
In the current paper we present a Web based system that performs automatic recognition of emotions (Figure 1) from speech and video data. The users can upload audio and video files and can run full emotion analysis remotely.
According to Ekman et al. (Ekman and Friesen 1978) people are born with the ability to generate and interpret only six facial expressions: happiness, anger, disgust, fear, surprise and sadness. All other facial expressions have to be learned from the environment the person grows up.

Acknowledgments. The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

Our Web system provides the necessary algorithms for the analysis of the six basic emotions, given the face of the subject or the audio speech signal. For the case of facial expressions, the system is robust to the typical differences of the age, gender and race and culture. In the case of speech analysis, the emotions are restricted to the german language. This limitation was set by the availability of data sets used for training the emotion from speech classifier.

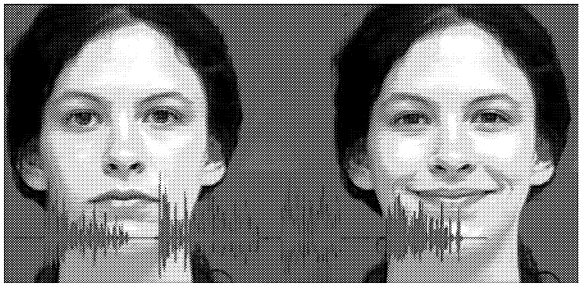


Figure 1. The Web based emotion recognition system performs analysis on both audio and video data.

RELATED WORK

The paper of (Saatci and Town, 2006) presents an approach to determine the gender and expression of faces by using Active Appearance Model. Four emotional states were employed for the analysis that was realized by using SVM classifier. The work shows an improvement of the recognition results by involving a first classification of the gender. The work of (Zhou et al., 2003) propose a Bayesian inference solution based on tangent shape approximation constructed in the form of Bayesian Tangent Shape Model. The work of (Lee and Elgammal, 2006) presents a novel nonlinear generative model using conceptual manifold embedding and empirical kernel maps for facial expressions. The algorithm deals with the complex nonlinear deformations of the shape and appearance in facial expressions and provides accurate emotional based synthesis.
The recognition of facial expressions was recently tackled in different ways involving the use of Viola&Jones features on input image data (Wong et al., 2006), static and temporal relations on facial characteristic points and by using different classifiers as BBN (Shan et al., 2006), (Datcu and Rothkrantz, 2004), HMMs (Aleksic and Katsaggelos, 2006), RVM and SVM (Datcu and Rothkrantz, 2005).
(Rothkrantz and Pantic, 2000) proposed a point-based face model composed of two 2D facial views, namely the frontal-

and the side view. Based on a point-based face model, expression-classification rules can be converted straight-forwardly into the rules of an automatic classifier.
The Facial Expression Dictionary (FED) (de Jongh, 2002) project aims at developing a non-verbal dictionary that contains information about non-verbal communication of people. The facial expression related analysis involves the user to select the region of the face and select the characteristic points of the face. The facial key points are predefined and the system helps to rapidly identify them on a new face sample. The face model used is that of Kobayashi and Hara (Kobayashi and Hara 1997). After the specification of all the FCPs, the system is able to run the emotion recognition process on the input face.
Face related analysis has been already incorporated in applications targetting e-learning (Loh et al., 2006), (Ben ammar and Neji, 2006) and smart meeting and conversation tracking systems (Zeng et al., 2006). The use of Active Appearance Model for extracting face shape information and the tracking of emotions in video sequences based on this shape data were recently researched in (Datcu and Rothkrantz, 2007).
For classifying human emotions in speech various attempts involved various algorithms. (Neiberg et al., 2006) have used a GMM to recognize emotions in spontaneous speech.
(Yu and al., 2004) applied a multilevel structure based on coupled hidden Markov models to estimate engagement levels in continuous natural speech. (Datcu and Rothkrantz, 2006) used the GentleBoost classifier to determine the optimal utterance segmentation for emotion recognition. The continuous speech signal is segmented into spoken utterances and the acoustic features are computed from each utterance portion. The extracted non-linguistic information is used for predicting the emotional states such as discrete emotion types or arousal/valence levels by employing SVM-based classifiers. The HMM uses the previous information to model the user's emotional state and engagement in conversation as a dynamic, continuous process. (Chateau et al., 2002) presents a study of the perception, the analysis and the modelling of styles or the 'emotional quality' of speech. The speech emotional quality is evaluated in terms of the emotional content that describes the listener's global impressions as elicited by their audition. criteria for evaluating the emotional quality are used to generate perceptive portraits of the speech. The evaluation is carried by using linear models to connect the perceptive portraits to physical data derived from signal analysis. Some work has also been focused on using additional information regarding speech.
The paper of (Lee and Narayanan, 2005) uses three sources of information - acoustic, lexical and discourse - for recognizing emotions. Linear discriminant and k-nearest neighbourhood classifiers are used to classify acoustical information to anger and frustration - as negative emotions and to neutral or positive emotions. The different features are extracted by using certain portions of the signal. A noticeable approach stands for multimodal analysis that aims at improving the recognition rates for the emotional state by fusing the results on separate modalities. The advantage of such methods relates to the overcoming the limited information that can be gathered from each single modality.
The work of (Busso et al, 2004) analysis the strengths and the limitations of systems based on the fusion of facial expression and acoustical information analysis at the decision level and in the case of feature level integration. (Kwon et al., 2003) provides a comparison on the emotion recognition performance of various classifiers. They obtained SVM and HMM based classifiers with significantly better results on SUSAS database from the previous approaches.
A recent research of (Rothkrantz et al., 2004) focuses on studying the effect of the workload on speech production by making use of a psychological experimental setup. A full analysis on each acoustic feature is conducted in order to create efficient models for stress detection.

MULTIMODAL APPROACH

The system provides the users with the capability of uploading audio and visual data for emotion analysis. For the audio data, the system accepts the processing of standard audio files. For the visual data there are two types of files that can be handled. The user can specify both static pictures and video sequences. In the case of processing a short video sequence, the user has a fixed limit for each video file to be processed.

Face detection

The detection of faces in both photos and video sequences is realized by using an implementation of an algorithm based on Viola&Jones features (Viola and Jones, 2001). These are visual features computed following simple addition and substraction operations on pixel intensities from rectangular areas (Figure 2). A strong classifier implementing Adaboost technique selects the most relevant Viola&Jones features that provide the best face detection rates. A combination of a fixed size sliding window and an multiresolution phyramid algorithm ensure the analysis of all possible areas for detecting faces in the frames. The input set of features for the classifier are selected from the fixed size of Viola&Jones features given the size of the sliding window.

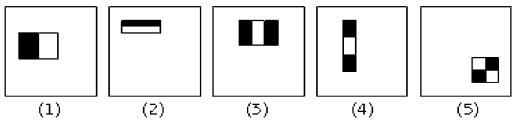


Figure 2. The five basic types of Haar-like features used for face detection. The summation operation is done on all pixel intensities from areas having the same color. The substraction is done between the two computed sums.

FCP model

The data set used for training the facial expression recognizer was Cohn-Kanade database (Kanade et al., 2000). The database contains a set of video sequences of recordings of several subjects acting on multiple scenarious. Each video sequence includes a subject showing a specific facial expression from the neutral state to the apex of the emotion. In the original database only the last frame of each sequence

is labelled using Auction Units AU codification. The process of creating the data set for training implied the selection of the last frame from each video sequence. The final set of selected samples has the structure as illustrated in
Table 1. The shape information extracted by the AAM from a face image is used to compute a set of suitable parameters that describe well the appearance of the facial features. The first step is the selection of the optimal key points on the face area from the shape data. The key points P_i are defined as Facial Characteristic Points (FCPs) and the FCP-set (Figure 3) is derived from Kobayashi & Hara model (Kobayashi and Hara, 1972).

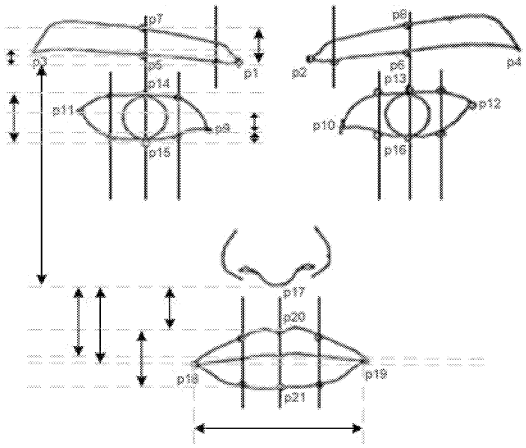


Figure 3:The Facial Characteristic Point FCP model.

In the second step a transform converts the FCP-set to some parameters v_i of an intermediate model. The parameterization has the advantage of providing the classifier with data that encode the most important aspects of the facial expressions.
Furthermore, it acts as a dimensionality reduction procedure since the dimension of the feature space is lower than the dimension of the image space. An advantage of the model is that it also can handle certain degree of asymmetry by using some parameters for both left and right sides of the face.

Table 1. The structure of the data set for facial expression recognition

Emotion	Fear	Surprise	Sadness	Anger	Disgust	Happy
#samples	84	105	92	30	56	107

The feature parameters are computed as the values of certain Euclidean distances between key points. The symmetry of the model is assumed to make the recognition process of facial expressions robust to occlusion or poor illumination i.e. if the left eye area is not directly visible do not use related information. The assumption is based on the supposition that the face detection procedure is also robust enough in such working conditions so as to be able to detect the face.
The parameters v_i model the variability of facial expressions in terms of distances among several pairs of FCPs. The complete list of such parameters is given in

Table 2.

Table 2: The set of visual feature parameters

		Visual feature			**Visual feature**			**Visual feature**
v_1	$(P_1, P_7)_y$	Left eyebrow	v_7	$(P_{14}, P_{15})_y$	Left eye	v_{13}	$(P_{17}, P_{20})_y$	Mouth
v_2	$(P_1, P_3)_y$	Left eyebrow	v_8	$(P_9, P_{11})_y$	Left eye	v_{14}	$(P_{20}, P_{21})_y$	Mouth
v_3	$(P_2, P_8)_y$	Right eyebrow	v_9	$(P_9, P_{15})_y$	Left eye	v_{15}	$(P_{18}, P_{19})_y$	Mouth
v_4	$(P_2, P_4)_y$	Right Eyebrow	v_{10}	$(P_{13}, P_{16})_y$	Right eye	v_{16}	$(P_{17}, P_{18})_y$	Mouth
v_5	$(P_1, P_{17})_y$	Left Eyebrow	v_{11}	$(P_{10}, P_{12})_y$	Right eye	v_{17}	$(P_{17}, P_{19})_x$	Mouth
v_6	$(P_2, P_{17})_y$	Right eyebrow	v_{12}	$(P_{10}, P_{16})_y$	Right eye			

CLASSIFICATION OF FACIAL EXPRESSIONS

The method used to encode the emotional patterns takes into account the subtle changes of the face shapes in different emotional postures. The vector $V = (v_1, v_2, ..., v_m)$ where m=17, encodes the set of parameters extracted from the Facial Characteristic Point - FCP model and accordingly has associated a certain emotional label.

For the classification of facial expressions, different classifiers have been taken into account. The method used for determining the results is 2-fold Cross Validation.
Table 4 illustrates the performance of a SVM classifier with polynomial kernel that works as a detector of Action Units AUs (Ekman and Friesen, 1978). The main algorithm for recognizing emotions is based on the classification done using the distance oriented model between FCPs. The second approach is the technique of detecting the AUs first

and then to recognize facial expressions from the AU sequences.

Table 3 presents the results in the case of the recognition of facial expressions by using Support Vector Machines – SVM as classifier.
The method used for determining the results is 2-fold Cross Validation.
Table 4 illustrates the performance of a SVM classifier with polynomial kernel that works as a detector of Action Units AUs (Ekman and Friesen, 1978). The main algorithm for recognizing emotions is based on the classification done using the distance oriented model between FCPs. The second approach is the technique of detecting the AUs first and then to recognize facial expressions from the AU sequences.

Table 3. The confusion matrix (%) for the facial expression recognition using SVM (polynomial kernel of degree 3)

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	**84.70**	3.52	3.52	4.70	1.17	2.35
Surprise	12.38	**83.80**	0.95	0	0	2.85
Sadness	6.45	3.22	**82.79**	1.07	3.22	3.22
Anger	3.44	6.89	6.89	**75.86**	6.89	0
Disgust	0	0	7.14	10.71	**80.35**	1.78
Happy	7.54	8.49	2.83	3.77	4.71	**72.64**

Table 4. The results of Action Unit detection using SVM (polynomial kernel of degree 4)

AU1	88.89%±2.08%	80.59%±0.00%
AU2	94.53%±0.25%	89.87%±2.39%
AU4	93.67%±0.47%	87.76%±1.19%
AU5	91.42%±1.13%	82.91%±2.09%
AU6	84.94%±5.93%	75.95%±3.58%
AU7	96.73%±1.92%	91.56%±1.79%
AU9	94.60%±1.88%	91.77%±2.09%
AU10	95.38%±1.82%	92.62%±2.69%
AU11	99.36%±0.30%	98.10%±0.30%
AU12	95.75%±0.41%	90.08%±2.09%
AU13	96.05%±2.44%	92.41%±1.79%
AU14	91.79%±1.51%	86.92%±5.37%
AU15	98.72%±1.20%	97.47%±0.60%
AU16	93.66%±0.69%	85.23%±1.19%
AU17	99.37%±0.90%	98.73%±0.00%
AU18	100.0%±0.00%	99.79%±0.30%
AU20	93.04%±2.60%	87.13%±2.09%
AU22	93.16%±3.18%	87.34%±2.98%
AU23	99.24%±0.38%	96.84%±1.49%
AU24	99.58%±0.60%	99.37%±0.30%
AU25	100.0%±0.00%	99.79%±0.30%
AU26	99.79%±0.30%	99.58%±0.00%
AU27	99.37%±0.90%	98.73%±0.00%
AU28	100.0%±0.00%	99.79%±0.30%

EMOTION ANALYSIS FROM SPEECH

The classifier chosen to model the emotion characteristics in speech is based on Gentle AdaBoost method for a maximum 200 training steps. The optimal classifiers are determined by employing ROC graphs to show the trade-off between the hit and the false positive rates. One important issue for the recognition of emotions in speech represents the segmentation of the speech signal. The way this process is done dramatically affects the subsequent results of the recognition of emotions.
One research question that rised was what is the optimal utterance segmentation method that gives the best results. Given the set of prosodic features, we determined the segmentation type and the utterance frame structure that leads to good recognition of emotions.
The data set used for emotion analysis from speech is Berlin (Burkhardt et al., 2005) – a database of German emotional speech. The database contains utterances of both male and female speakers, two sentences. The emotions were simulated by ten native German actors (five female and five male). The result consists of ten utterances (five short and five long sentences). The length of the utterance samples ranges from 1.2255 seconds to 8.9782 seconds. The recording frequency is 16kHz.
The final speech data set contains the utterances for which the associated emotional class was recognized by at least 80% of the listeners. Following a speech sample selection, an initial data set was generated comprising 456 samples and six basic emotions (*anger*: 127 samples, *boredom*: 81 samples, *disgust*: 46 samples, *anxiety/fear*: 69 samples, *happiness*: 71 samples and *sadness*: 62 samples).
In the case of emotion recognition from speech, the analysis is handled separately for different number of frames per utterance. In the current approach there are five types of splitting methods performed on initial data. Each type of splitting produces a number of data sets, according to all the frame combinations in one utterance.
The Praat (Boersma and Weenink, 2005) tool was used for extracting the features from each sample from all generated data sets. According to each data set frame configuration, the parameters **mean**, **standard deviation**, **minimum** and **maximum** of the following acoustic features were computed: Fundamental frequency (pitch), Intensity, F1, F2, F3, F4 and Bandwidth.
All these parameters form the input for separate GentleBoost classifiers according to data sets with distinct segmentation characteristics.

Results of emotion recognition from speech

The GentleBoost *committee* is trained for a maximum number of 200 stages. Separate data sets containing male, female and both male and female utterances are considered for training and testing the classifier models. The performance of each classifier is evaluated with the 5-fold cross validation. Depending on the number of sub-frames per speech frame, the different data sets are used to generate sets of classifiers. One curve on the graph stands for the set of representative GentleBoost strong classifiers generated by using the specific data set, associated with a certain split configuration. Each node on one curve relates to one

classifier in the set. The ROC graph in Figure 4 shows the tradeoff between the hit and the false-positive rates for all the GentleBoost classifiers generated from Berlin data set.
Each point on the figure stands for one GentleBoost classifier that is selected using the highest true-positive rate criterion. For each emotion class, a total number of 200 points is taken into account and only the ones with the highest scores are displayed on the same emotion curve. By analyzing each emotion curve separately, the final strong committee to be chosen is the one that is the closest to the north-west corner of the figure. In other words, the classifier in question is the one that has the highest true positive rate (*tpr*) while the false positive rate (*fpr*) is the lowest in the set of classifiers on the same curve.

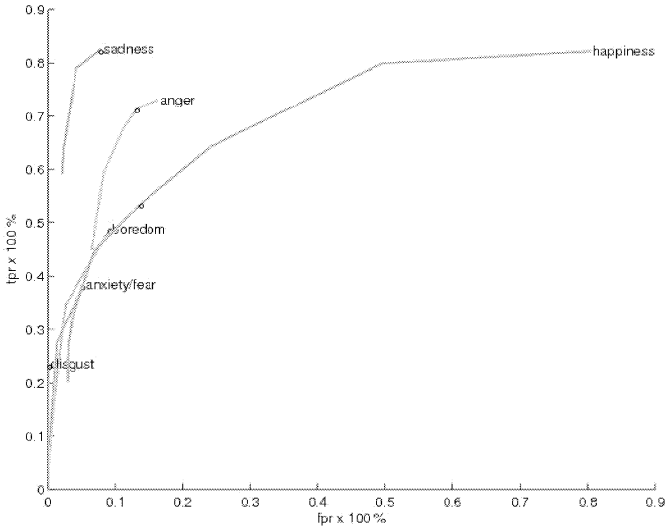


Figure 4. ROC graph that show the committees with the highest true positive rates for each emotion class.

Table 5 depicts the characteristics of each strong classifier that is selected for each emotion curve separately. The column *nr.stages* shows the number of stages required to train the associated strong committee.
An additional field (*ac*) in each table shows the accuracy rate achieved by the classifiers.
Each classifier is identified by the structure of the frames into the utterance sample (column *frames*). A digit from one binary sequence specifies that the correspondent frame contributes ('1') or not ('0') with features at the classification process.

Table 5: The results for the recognition of emotions using GentleBoost classifier.

emotion	nf	frames	nr. stages	ac (%)	tpr (%)	fpr (%)
anger	10	1101000001	5	0.83±0.03	0.72±0.16	0.13±0.06
boredom	2	10	58	0.84±0.07	0.49±0.18	0.09±0.09
disgust	10	0100001000	21	0.92±0.05	0.24±0.43	0.00±0.00
anxiety/fear	10	1110000011	86	0.87±0.03	0.38±0.15	0.05±0.04
happiness	10	1111010100	40	0.81±0.06	0.54±0.41	0.14±0.13
sadness	10	1011111101	13	0.91±0.05	0.83±0.06	0.08±0.06

An observation on the table proves that the majority of the strong classifiers lying on the emotion curves in the ROC graph clearly express the efficiency of using a ten frames per utterance configuration for the segmentation. The information presented in Table 6 is independent on the emotion class.

Table 6: The dependency of emotion recognition results on the number of frames per utterance for Berlin data set.

nf	ac (%)	tpr (%)	fpr (%)
1	0.85±0.11	0.36±0.63	0.07±0.17
2	0.83±0.31	0.44±0.67	0.10±0.41
3	0.84±0.17	0.46±0.62	0.09±0.23
5	0.84±0.13	0.50±0.63	0.10±0.22
10	0.77±0.33	0.58±0.64	0.20±0.45

One difference should be noted on the analysis methods used for choosing the best classifiers for Table 5 and Table 6. While for the first the criterion was to choose the classifiers with the best trade-off between hit rate and false positive rate, the last involved the choice for the classifiers with the highest true positive rate.

THE WEB BASED INTERFACE

The user can upload photos, audio and video files into the system. For each type of data files there is a size limit imposed for restricting the amount of processing necessary for running all the analysis. The numer of the users that can work remotely at the same time is also limited. From the functional point of view, the Web based emotion recognition system consists of a set of CGI applications that run behind the HTTP server. The Web interface is created using HTML and JAVA applets on the client side and MySQL database system, PHP scripting and CGI modules on the server side.
The files uploaded by the users are stored in a database on the server.
For facial expression recognition, the system first calls a CGI module that converts all photos to a common visual format. That makes sure that the other CGI components will be able to read the picture files without problems.
In case of video files, the video is decomposed so as to extract frame by frame and to store the collection of frames in a temporary location for further analysis. In case of videos, the frame sequence is scaled to a fixed rate of frames per second. In most of the cases this operation involves a downscaling to a lower rate so as to allow for a smaller amount of processing per whole sequence. The next step is the detection of faces into each frame. This is carried on by a face detection CGI application that is called from PHP environment. The results are passed back to the PHP script and written as parameters of the JAVA applet. In this way the user will be able to go through each frame, one for pictures, and to visualize the location of the detected faces in video or still picture. An additional step of the face detection program is the extraction of the areas of the faces and the storing of these as separate image files. The operation is useful for the next step that involves the running of the facial expression recognition CGI modules. Each face is analyzed several times by different CGI applications, one for each implementation of each facial expression classifier.
There are several classifiers implemented in this way and each of them uses the same image files containing the face to

be analyzed. The results are again, passed back to the script that formats the HTML response back to the client so as to include the information regarding the facial expression recognition data.

In the case of audio data, the uploaded files are analyzed first by a CGI script for extracting all the speech parameters such as pitch, frequencies, intensities and bandwidth. Another CGI application computes the certain parameters derived from those mentioned above, namely the mean, minimum, maximum and standard deviation. The last CGI application implementing the Gentle AdaBoost classifier, performs the recognition of emotions given the input data associated with the initial utterance file. The results are passed back in the same way as in the case of facial expression recognition.

The Web browser receives the PHP generated HTML response and the audio and video files back from the server. The user can visualize the initial audio-video data and the new information related to the recognition of emotions in speech and visual content. The mechanism implies that all the processing is done on the server and the client is no longer bothered with possible restrictions that may rise from the necessity of running data analysis on the local machine. Some experimental versions of the emotion recognition interface are illustrated in Figure 5 (for speech) and in Figure 6 (for video).

Some optimizations have been made on the server side in order to speed up the process of going though all the computations.

The implementation of the facial expression recognition and the emotion recognition from speech modules present a very high speed of frame and audio files processed per second. Practically, these can be integrated in a potential real time system.

Despite the fact that each CGI component – usually the ones implementing classification algorithms – needs very small time for processing the data, when a CGI is run, it requires a big amount of time for the initialization. Commonly, this comes from the need to load big data files of matrices required in the process of classification.

The improvement consists of an implementation of applications that are loaded in the memory and running continuously on the server or possible on other, more powerfull, computers, on the same network with that of the HTTP server. These kind of applications are idle during the time periods when the Web System is not accessed by users. When there are users attempting data analysis, the PHP script running on the HTTP server calls the CGI modules and takes back the provided results. Each CGI module keeps its own waiting list for handling the tasks according to the user requests.

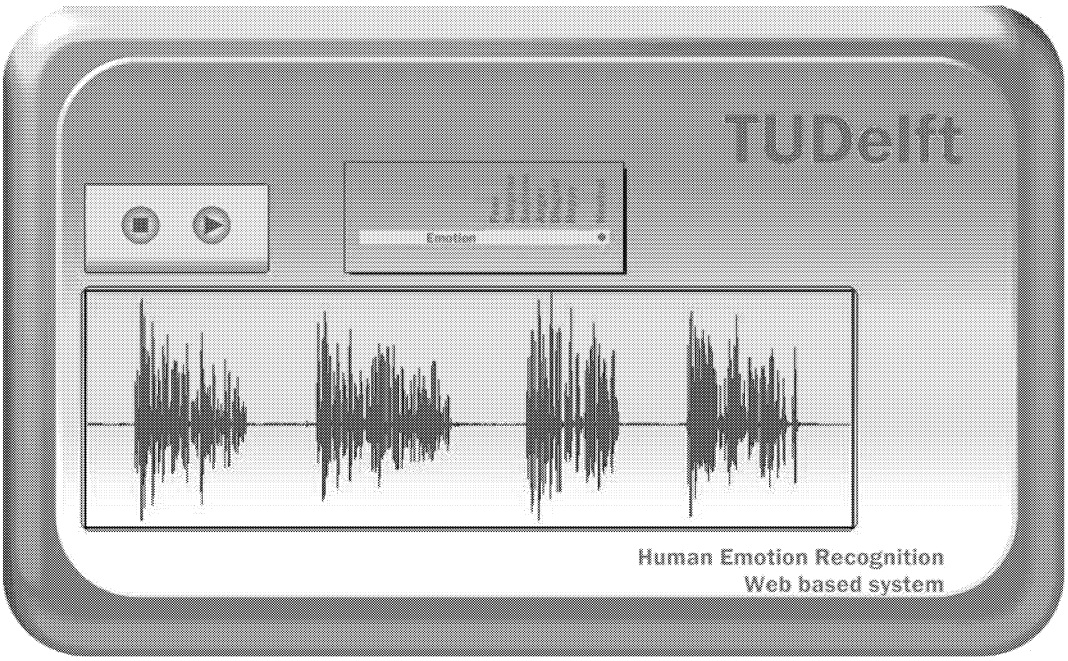


Figure 5. The Web interface for emotion recognition from speech

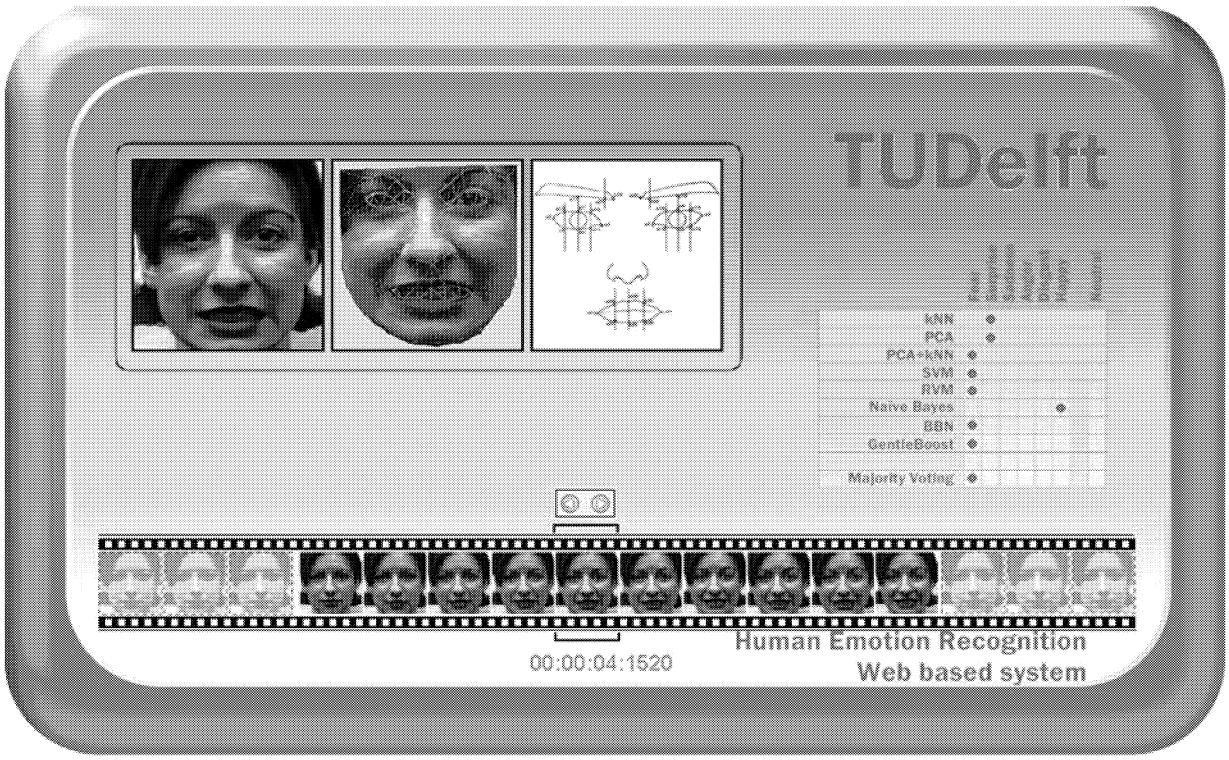


Figure 6. The Web interface for facial expression recognition

CONCLUSION

The system being described in the paper is able to perform automatic human emotion recognition. Beside the advantage of accessing a Web based interface for distance processing, the disadvantage of the system resides in the big amount of processing and time needed for running the complete human emotion analysis. In the case of multiple accesses the effects are considerable and it may lead to low performance. One more aspect is that the facial expression recognition is frame based and no temporal emotional patterns are used during the analysis of video sequences. Furthermore, a major drawback of the system is the lack of algorithms to handle the recognition of emotions in videos. If the user uploads a movie sequence, the system performs either audio or visual based emotion recognition. The emotion classification models do not take the combination of the two modalities into account at the same time In the case of emotion recognition from speech, the ability to run analysis only on german spoken utterances stands for a major restriction for the moment. In the research for the development of an algorithm for emotional speech analysis, we have conducted a set of analysis on different types of utterance segmentation. As a base technique, we used the GentleBoost classifier with a maximum of 200 training stages. The optimal strong classifier has been selected by making use of ROC graphs. Although the initial research included also separate analysis for male and female voices, the chosen solution is only on mixed male-female voices due to the limited amount of space.

The administration side of the website is enhanced with sections that provide statistics on the use of the system, to track the user processing tasks and to analyse the system performance. After a prior consent from the users, the multimodal database of the system is enriched with new audio and video samples as they are provided for analysis. In the current version, the emotion classification is restricted to the set of six basic emotions plus the neutral. Ongoing work is carried for extending the set of emotions that are to be recognized.

REFERENCES

Aleksic, P. S., A. K. Katsaggelos, “Automatic Facial Expression Recognition using Facial Animation Parameters and Multi-Stream HMMs”, ISSN: 1556-6013, in IEEE Transactions on Information Forensics and Security, 2006.

Ben ammar, M. and Neji, M. (2006), “Affective e-Learning Framework”, In T. Reeves & S. Yamashita (Eds.), Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2006 (pp. 1595-1602). Chesapeake, 2006.

Boersma, P., Weenink, D., “Praat: doing phonetics by computer (Version 4.3.14)” [Computer program]. 2005.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. Proceedings Interspeech, Lissabon, Portugal 2005.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S., “Analysis of emotion recognition using facial expressions, speech and multimodal information”, *ICMI*, State College, Pennsylvania, 2004.

Chateau, N., Maffiolo, V., Ehrette, T., s’Alessandro, C, “Modelling the emotional quality of speech in a telecommunication context”, *Proceedings of the International Conference on Auditory Display*, Kyoto, Japan 2002.

Datcu, D., Rothkrantz L.J.M., "Automatic recognition of facial expressions using Bayesian Belief Networks", *Proceedings of*

IEEE SMC 2004, ISBN 0-7803-8567-5, pp. 2209-2214, October 2004.

Datcu, D., L.J.M. Rothkrantz, "Machine learning techniques for face analysis", *Euromedia 2005*, ISBN 90-77381-17-1, pp. 105-109, April 2005.

Datcu, D., L.J.M. Rothkrantz, "The recognition of emotions from speech using GentleBoost Classifier", CompSysTech'06, June 2006.

Datcu, D., L.J.M. Rothkrantz, "Facial expression recognition using Active Appearance Model in crisis environments", ISCRAM'07, Delft, The Netherlands, May 2007.

Ekman, P. and W. Friesen. 1978. "Facial Action Coding System." Consulting Psychologists Press, Inc., Palo Alto California, USA.

Freund, Y. and R.E. Schapire. 1995. "A decision-theoretic generalization of on-line learning and an application to boosting." In *2nd European Conference on Computational Learning Theory*.

Harris, C.G. and M. Stephens. 1988. "A combined corner and edge detector". Proceedings 4th Alvey Vision Conference, Manchester, 189-192.

Jongh de, E.J. 2002. "FED: An online facial expression dictionary as a first step in the creation of a complete nonverbal dictionary." TU Delft.

Kanade, T., J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis,". *Proc. of the 4th IEEE Int. Con. on Automatic Face and Gestures Reco.*, France, 2000.

Kearney, G.D. and S. McKenzie. 1993. "Machine interpretation of emotion: design of a memory-based expert system for interpreting facial expressions in terms of signaled emotions." (JANUS). Cognitive Science 17, Vol. 4, 589–622.

Kobayashi, H. and F. Hara. 1997. "Facial Interaction Between Animated 3D Face Robot and Human Beings". IEEE Computer Society Press, 3732-3737.

Kwon, Oh-Wook, Chan, K., Hao, J., Lee, Te-Won, "Emotion recognition by speech signals", *EUROSPEECH'03,* Geneva, pp. 125–128, 2003.

Lee, C., M., Narayanan, S., S., "Toward detecting emotions in spoken dialogs", IEEE Transactions on speech and audion processing 13(2), pp. 293–303, 2005.

Lee, C.S., A, Elgammal, "Nonlinear Shape and Appearance Models for Facial Expression Analysis and Synthesis", *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, volume 01, pp. 497 – 502, 2006.

Loh, May-Ping, Y.P. Wong, C.O. Wong, "Facial Expression Recognition for E-learning Systems using Gabor Wavelet & Neural Network", *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06*), pp. 523-525, 2006.

Morishima, S. and H. Harashima. 1993. "Emotion Space for Analysis and Synthesis of Facial Expression". IEEE International Workshop on Robot and Human Communication, 674-680.

Neiberg, D., K. Elenius, K. Laskowski, "Emotion Recognition in Spontaneous Speech Using GMMs", INTERSPEECH 2006 – ICSLP.

Rothkrantz, L.J.M. and M. Pantic. 2000. "Expert Systems for Automatic Analysis of Facial Expressions". Elsevier, Image and Computing, Vol. 18, 881-905.

Rothkrantz, L., J., M., Wiggers, P., van Wees, J., W., A., van Vark, R., J., "Voice stress analysis", *Proceedings of Text, Speech and Dialogues*, 2004.

Saatci, Y., Town, C. 2006, "Cascaded Classification of Gender and Facial Expression using Active Appearance Models", *The 7th Conference on Automatic Face and Gesture Recognition*, FGR'06.

Shan, C., S. Gong, P. W. McOwan, "Dynamic Facial Expression Recognition Using A Bayesian Temporal Manifold Model", BMVC06, 2006.

Tipping, M.E. 2000. "The Relevance Vector Machine. Advances in Neural Information Processing Systems". Vol. 12, 652-658.

Treptow, A. and A. Zell. 2003. "Combining Adaboost Learning and Evolutionary Search to Select Features for Real-time Object Detection". University of Tuebingen, Department of Computer Science, Germany.

Viola, P. and M. Jones. 2001. "Robust Real-time Object Detection." Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling.

Zhao J. and G. Kearney. 1996. "Classifying facial emotions by back-propagation neural networks with fuzzy inputs". International Conference on Neural Information Processing, Vol. 1, 454–457.

Zhou, Z.-H. and X. Geng. 2002. "Projection Functions for Eye Detection". State Key Labaratory for Novel Software Technology, NU, China.

Yu, C., Aoki, P. M., Woodruff, A., "Detecting user engagement in everyday conversations". *8th International Conference on Spoken Language Processing (ICSLP'04)*. 2004.

Zeng, Z., Y. Hu, Y. Fu, T. S. Huang, G. I. Roisman, Z. Wen, „Audio-visual emotion recognition in adult attachment interview", in Proceedings of the 8th international conference on Multimodal interfaces, ISBN:1-59593-541-X, pp. 139-145, 2006.

Zhou, Y., Gu, L., Zhang, H-J. 2003, "Bayesian tangent shape model: estimating shape and pose parameters via Bayesian inference", ISBN: 0-7695-1900-8, *Computer Vision and Pattern Recognition, Proceedings,* 2003.

Wong, W.S., W.Chan, D.Datcu, L.J.M. Rothkrantz, "Using a sparse learning relevance vector machine in facial expression recognition", *Euromedia'2006*, ISBN 90-77381-25-2, pp. 33-37, University of Ghent, April 2006.

Christiansen Grammar Evolution for the Modelling of Psychological Processes

Emilio del Rosal, Alfonso Ortega, Manuel Alfonseca and Marina de la Cruz
Departamento de Ingeniería Informática
Universidad Autónoma de Madrid
C/Tomás y Valiente, 11 - Madrid 28049-Spain
e-mail: emilio.delrosal@uam.es

KEYWORDS

Evolutionary computing, Associative learning, Modelling, Model design.

Abstract

Psychologists have developed models of associative learning for more than 30 years. Despite the strong efforts made, they still suffer many shortcomings. We have tried to build an integral model of habituation, the simplest type of learning within the area of associative learning and the basic support for other types. To overcome the deficiencies of traditional models, we have made used of Christiansen Grammar Evolution. This evolutionary technique is capable of automatically search for a target expression (the model) in a given formal language (the formalism of the model). Under this perspective, that we call *Automatic Modelling*, we have found models of habituation with interesting characteristics.

1 INTRODUCTION

The present research has a multi-disciplinary character. We introduce the psychological area of associative learning (AL in the remainder of the paper), and the phenomenon of habituation, as well as the computer science procedure called grammatical evolution (GE from now on).

1.1 Models of associative learning

Since the pioneer work of Ivan Pavlov (Pavlov, 1927), animal associative learning has been widely studied by psychologists, mainly under the methodological perspective of behaviourism (O'Donohue and Kitchener, 1999). From this approach, the main object of the researcher is to find the functional relation between the stimuli of the environment and the responses of the animal.

AL recognizes four groups of phenomena. The first two are sensitization and habituation. Although they are actually non-associative processes, their role is essential for the functioning of AL. The other two, classical conditioning and operant conditioning, conform the core of associative learning (see (Mazur, 2002) for a more detailed explanation).

Current models of AL suffer from two main limitations. Firstly, most of them take into account only one of the four types presented above, isolating it from all the others. It has been previously stated (Alonso et al., 2005; del Rosal et al., 2006) that this lack of integration is a big weakness in the current state of the art, specially if we consider that one kind of learning can affect the functionality of the others. Secondly, even after taking into account only one of the types, they cannot reproduce all its main characteristics.

Thus, our perspective to create an integral model of associative learning is the following: we understand that habituation, given its filtering role (see below) of stimuli and its simplicity, is the most appropriate type of learning to begin with. Once a satisfactory model of habituation has been built, the more complex types of learning could be modelled, integrating the previously modelled functionality of habituation. Finally, in our opinion, this challenge can be better confronted with support from the latest developments in computer science for the automatic resolution of problems (i.e. GE and its extensions).

Habituation acts by decreasing the innate response of an organism to a given stimulus. This decrease is a consequence of the repeated presentation of the stimulus. For example, the response of many organisms after hearing a novel sound decreases when the sound is presented many times. Habituation has an important function in the organism's learning, its role consisting of filtering irrelevant stimuli which should not be processed. See (Hall, 1991) for a fuller account. As an example, figure 1 shows the typical habituation curve with an inter-stimulus interval (ISI) of thirty seconds and the following *spontaneous recovery* after a period with no stimulus presentation.

To assess our modelling work we identified the defining characteristics of habituation. Each of our models was evaluated in terms of its ability to reproduce them. The list was built following (Thompson and Spencer, 1966) for the first 6 characteristics (the rows in table 1). The seventh row was supported by more modern empirical evidence (Byrne, 1982; Rankin and Broster, 1992). The main previously existing models (columns) were compared according to those characteristics, to make us able to discuss the value of our models within the current state of the art.

1.2 Evolutionary automatic programming and GE

Evolutionary automatic programming (EAP) follows the same principles as any other evolutionary computing tech-

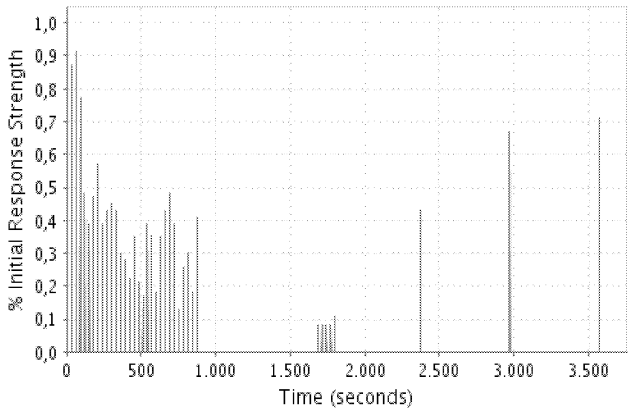


Figure 1: A stimulus is presented every 30 seconds (between seconds 0 and 1800 approximately). During this period the response of the animal to the stimulus decreases progressively. After some time with no exposure to the stimulus (seconds 1800-2400 approx.) three new stimulus presentations show a clear recovery of the animal's response. Data taken from (Rankin and Broster, 1992).

Property/Model	(Stanley, 1976)	(Wang, 1994)	(Staddon and Higa, 1996)	(del Rosal et al., 2006)
Response's exponential decrease	yes	yes	yes	yes
Spontaneous recovery	yes	yes	yes	yes
Lower intensity, faster habituation	yes	X	X	yes
Shorter ISI, faster habituation	X	yes	yes	yes
Dishabituation	X	X	X	yes
Long-term and short-term habituation	X	yes	X	X
Shorter ISI, faster recovery	X	X	yes	X

Table 1: *Comparison of habituation models.*

nique with the aim of creating programs automatically from a high-level description. From the beginning, genetic programming (Koza, 1992) (the main and first branch of EAP) has faced two main weaknesses. The first is getting over the closure problem, i. e., the generation of syntactically valid expressions (or programs) and the preservation of their validity under the action of the evolutionary operators (mutation, crossover...) The second is building a system independent of the different languages used.

In the last fifteen years, many advances in genetic programming have been related to the effort to overcome the closure problem (O'Neill and Ryan, 2003). In this context, Grammar-guided genetic programming (GGGP) has been used successfully (Rodrigues and Pozo, 2002). GE also plays an important role in this line of advances (O'Neill and Ryan, 2003). Like GGGP, it has the ability to produce only valid expressions in a general and parsimonious way, and manages arbitrary languages.

The GE defining features are (for a detailed explanation refer to (O'Neill and Ryan, 2003)):

- Individuals have two explicitly separate representations: genotype and phenotype. The former is a string of integers subject to the action of the evolutionary operators, the latter is written in the target programming language and is confronted to the fitness function.
- A predefined free-context grammar specifies the target language.
- A genotype-phenotype mapping transform the integer string in a ***valid*** expression of the programming language.

The main limitation, for our purposes, of the original GE system (as well as GGGP) is the use of context-free grammars. As we will see below, these algorithms based on context-free grammars have not enough expressive power to be used as an *automatic modelling tool*. However, recent extensions to GE make use of more powerful grammars (Ortega et al., 2007; Echeandia et al., 2005). We have employed Christiansen grammars (CG) (Ortega et al., 2007) an adaptable extension to attribute grammars (Knuth, 1968) whose first attribute is the CG that contains the rules applicable to each non- terminal in every derivation step. CG are adaptable because this attribute can change as words are generated. Like attribute grammars, CG can fully describe the solutions of any computable task, but their adaptability allows more comfortable approaches to some problems.

2 MODELLING METHOD

It is easy to show that there is no reason to restrict the set of languages to be used in CGE to the subset of *programming* languages. In fact, an evolutionary search can be done in any formal language which can be defined as a Christiansen grammar. Therefore, if we define the modelling task as finding an expression, written in a given language, which describes the behaviour of a real system, CGE can be understood as an *automatic modelling tool* whose grammar is the formalism in which the model is expressed and the fitness function determines the features of the target model. The original system GE has been used in previous works with the same general perspective: (O'Neill et al., 2002), economics and finance, or (Moore and Hahn, 2004), hierarchical Petri net modelling of complex genetic systems. As far as we know, this is the first time that CGE has been applied to a real problem.

With this in mind, our *modelling method* goes through the following steps:

1 **Grammar**: the grammar introduced in the CGE system determines the expressions landscape to be searched by the evolutionary algorithm. It implements the following aspects of the model.

   (a) **General architecture of the model**: the formalism used to express the model.

   (b) **Researcher's assumptions and bias**: due to theoretical reasons or model assumptions, some restrictions to the grammar may be introduced.

   (c) **Other practical or technical constraints**.

2 **Fitness function**: used to assess the quality of the expressions generated by the algorithm. This function has to sort models according to their quality, to make comparisons possible.

3 **Parameter tuning and search**

4 **Interpretation and study of the model**: in some contexts, the fitness function may not provide enough information to conclude that the model found is appropriate. In other cases, the model found may fit the conditions imposed, beside bringing some other knowledge in terms of the way it performs the task, thus encouraging theoretical discussion.

We consider that this scheme can be applied to any modelling context. In our case, it has been used to model habituation.

2.1 Prior assumptions and constraints

To represent our models of habituation we have used the formalism and conditions of one of the latest habituation models (del Rosal et al., 2006), together with other practical considerations.

1. Since habituation features happen in time, the variables of the model must be functions of time. We have used the iterated functions formalism, where time is the independent variable. Iterated functions have the form f(t+1) = g(f(t)), with initial condition f(0).

2. As this model is to be integrated with a more general model of associative learning, its architecture should have the ability to manage more than one stimulus and their interactions.

3. Each stimulus is represented by three variables:

   - $S_i(t)$ represents the intensity of stimulus i. Its value is a number in the $[0, 1]$ interval. Any value greater than zero indicates that the stimulus is present.
   - $A_i(t)$ represents the strength of the response of the organism to stimulus i.
   - $D_i(t)$ is an auxiliary variable.

4. For each pair of stimuli (i, j), $T_{ij}(t)$ represents their interaction.

5. **Initial conditions**: we decided to maintain the same initial conditions as in (del Rosal et al., 2006). $\forall i$, $A_i(0) = T_{ij}(0) = 0$ and $D_i(0) = 1$.

6. **Complexity**. Since the model is to be used by the researcher as a representation of the mechanism that produces habituation, it must not be too complex, so that the researcher can manipulate it as a prediction tool and for theoretical discussions.

2.2 The Christiansen Grammar

Before presenting the grammar, we will explain its main features.

All the iterated functions are defined by means of clauses such as $A_i(t) =< arithmetic - expression >$. For clarity, the time step specification remains implicit. The arithmetic expressions are written in prefix notation. The following operators can be combined in the expressions:

- Binary operators: **+**, **-**, *****, **/**, **max**, **min** and **pow** (with the usual meaning).
- Unary operators: **log**, **abs**, **int**, **sqrt** and **exp** (with the usual meaning).
- **sum<n>**: a summation operator with some constraints. It adds all the values in an array. For instance, the expression $sum_k, D_i, sum_l, A_k, A_l$, is equivalent in classic notation to $\sum_{k=0}^{n} D_i \sum_{l=0}^{n} A[k]A[l]$ (where n is the number of stimuli).

Variables may be subscripted by referring to a single stimulus, as in i or j, or to all of them, as in $< n >$. The following are legal subscripted variables: A_i, D_i, S_i, $A_{<n>}$, $D_{<n>}$, $S_{<n>}$, $T_{i<n>}$, T_{ij} and $T_{<n>j}$.

It can easily be seen below that the grammar we have used is a context-free grammar, except for two elements, which require the use of the CG's capabilities:

- **Summation indices**: a variable with subscript index of type $< n >$ does not make sense outside the scope of a summation operator. Our CG has rules to generate this kind of variables only when the non-terminal is inside the scope of a summation operator. This is a context-dependent property.
- **Model complexity**: as mentioned before, the complexity of the model should not be too large. The Christiansen grammar includes an attribute (↑c) to store the number of operators in each expression. If this number gets larger than a given threshold, the expression is dismissed (function *tooComplex*(c) in the grammar).

We present the grammar below. Our notation follows (Watt and Madsen, 1977), where inherited attributes have down-arrows symbols (↓) and synthesised attributes have up-arrows (↑). (↓g) stands for the Christiansen Grammar attribute. More details in (del Rosal, 2006).

- <Model>(g) ::= A_i(t) = <ExpTypeAD>$_A$(↓g) : D_i(t) = <ExpTypeAD>$_B$(↓g) : T_{ij}(t) = <ExpTypeT>(↓g)
  {
  <ExpTypeAD>$_A$.↓g = <Model>.g
  <ExpTypeAD>$_B$.↓g = <Model>.g
  <ExpTypeT> . ↓g = <Model>.g
  }

Below, rules for the non-terminal <ExpTypeAD>:

- <ExpTypeAD>(↓g, ↑c) ::= <BinaryOp>(↓g), <ExpTypeAD>$_A$(↓g, ↑c), <ExpTypeAD>$_B$(↓g, ↑c)
  {
  <BinaryOp>.↓g = <ExpTypeAD>.↓g

<ExpTypeAD>$_A$.↓g = <ExpTypeAD>.↓g

<ExpTypeAD>$_B$.↓g = <ExpTypeAD>.↓g

<ExpTypeAD>.↑c = <ExpTypeAD>$_A$.↑c + <ExpTypeAD>$_B$.↑c + 1

tooComplex(<ExpTypeAD>.↑c)

}

- <ExpTypeAD>(↓g, ↑c) ::= <UnaryOp>(↓g), <ExpTypeAD>$_A$(↓g, ↑c)
  {

  <UnaryOp>.↓g = <ExpTypeAD>.↓g

  <ExpTypeAD>$_A$.↓g = <ExpTypeAD>.↓g

  <ExpTypeAD>.↑c = <ExpTypeAD>$_A$.↑c + 1

  tooComplex(<ExpTypeAD>.↑c)

  }
- <ExpTypeAD>(↓g, ↑c) ::= sum<indexAD>(↓g, ↑new_g), <ExpTypeAD>$_A$(↓g, ↑c)
  {

  <indexAD>.↓g = <ExpTypeAD>.↓g

  <ExpTypeAD>$_A$.↓g = <indexAD>.↑new_g

  <ExpTypeAD>.↑c = <ExpTypeAD>$_A$.↑c + 1

  tooComplex(<ExpTypeAD>.↑c)

  }

As we are inside the scope of a summation operator, new rules are include to allow indices of type <n>.

- <indexAD>(↓g, ↑new_g) ::= a | b | c | d | e | f | g | h | k | m | o | p | q | r | s | t | u | w | x | y | z
  {

  rule1 = <ExpTypeAD>(↓g) ::= A$_{<n>(\downarrow g)}$ | D$_{<n>(\downarrow g)}$ | S$_{<n>(\downarrow g)}$ | T$_{i<n>(\downarrow g)}$
  {

  <n>.↓g = <ExpTypeAD>.↓g

  }

  rule2 = <n> ::= *"THE ACTUAL INDEX"*

  <indexAD>.↑new_g = <indexAD>.↓g + rule1 + rule2

  }
- <ExpTypeAD>(↓g, ↑c) ::= <RealNumber>(↓g)
  {

  <RealNumber>.↓g = <ExpTypeAD>.↓g

  <ExpTypeAD>.↑c = 0

  }
- <ExpTypeAD>(↓g, ↑c) ::= $S_i|A_i|D_i$
  {

  <ExpTypeAD>.↑c = 0

  }

Rules for <ExpTypeT> are equivalent to those for <ExpTypeD> except for the variables allowed. We omit equivalent rules:

- <indexT>(↓g, ↑new_g) ::= a | b | c | d | e | f | g | h | k | m | o | p | q | r | s | t | u | w | x | y | z
  {

  rule1 = <ExpTypeT>(↓g) ::= A$_{<n>(\downarrow g)}$ | D$_{<n>(\downarrow g)}$ | S$_{<n>(\downarrow g)}$ | T$_{i<n>(\downarrow g)}$ | T$_{<n>j(\downarrow g)}$
  {

  <n>.↓g = <ExpTypeT>.↓g

  }

  rule2 = <n> ::= *"THE ACTUAL INDEX"*

  <indexT>.↑new_g = <indexT>.↓g + rule1 + rule2

  }
- <ExpTypeT>(↓g, ↑c) ::= $S_i|A_i|D_i|S_j|A_j|D_j|T_{ij}$
  {

  <ExpTypeT>.↑c = 0

  }
- <BinaryOp> := + | - | * | / | max | min | pow
- <UnaryOp> ::= 1 | abs | int | sqrt | exp
- <RealNumber>(↓g) ::= <IntPart>(↓g).<DecPart>(↓g)
  {

  <IntPart>.↓g = <RealNumber>.↓g

  <DecPart>.↓g = <RealNumber>.↓g

  }
- <IntPart> ::= 0 | 1 | 2 | ..... | 8 | 9 | 1<IntPart>$_A$ | 2<IntPart>$_A$ | .... | 9<IntPart>$_A$
  {

  <IntPart>$_A$.↓g = <IntPart>.↓g

  }
- <DecPart> ::= 0 | 1 | 2 | ..... | 8 | 9 | <DecPart>$_A$1 | <DecPart>$_A$2 | .... | <DecPart>$_A$9
  {

  <DecPart>$_A$.↓g = <DecPart>.↓g

  }

2.3 The fitness function

The fitness function compares our models to empiric data with respect to the characteristics mentioned in table 1. (Rankin and Broster, 1992), experiment 1, is used for properties 1, 2, 4 and 7. (Rankin et al., 1990) is used for property 5. The comparison is done in terms of the absolute differences between the points of the simulated and empiric curves. Therefore, the values go from zero to ∞. The values are in terms of the percentage of the initial response to the stimulus, as usually in the psychological literature. For those models with negative values, their curves are scrolled up to avoid them, since negative values has no meaning as an activation value to a stimulus. The mathematical expression is:

$$Fitness = \sum_{i=0}^{i=n} \sum_{k=1}^{k=m} \left| \frac{P_i[k]}{P_i[0]} - \frac{E_i[k]}{E_i[0]} \right| * weight(i,k)$$

Leaving apart the *weight* function, E stands for the empirical points' vector. P symbolize the simulated points' vector. The subscript index "i" represents each experiment in the data set and "k" represents each point. Finally, "n" is the number of experiments, while "m" is the number of points of the given experiment.

The *weight* function increases the influence of points that are specially important. Every point has a *weight* of 1 by

default, except for a few points that strongly characterize habituation. These points are those that represent recovery in (Rankin and Broster, 1992), experiment 1, (weight = 3) and the one that reflects dishabituation in the experiment taken from (Rankin et al., 1990), (weight = 6). These values are somehow arbitrary and could be subject to further tuning.

3 EXPERIMENTS AND RESULTS

During our experiments we tried to find the best combination of evolutionary parameter values, in terms of their influence to reach better fitness values. Only a subset of them varied: mutation, cross-over and the generational gap (in a steady-state population model). The rest remain constant with the following values: *population* = 10.000, fixed *number of codons* = 1.000, *wrapping* = 3, *codon range* = [0,256], *parent selection* = "fitness proportional", and *survivor selection* = "replace worst scheme". The termination condition was the existence of a fitness below 1 (a very small number that reflects an almost perfect matching between the empiric and simulated data) or reaching the maximum number of generations, which is determined by the maximum number of fitness evaluations, an amount between 30000 and 120000 depending on the CPU performance. With this we tried to avoid very long experiments (a normal experiment lasted 4-6 days).

We tried to find the best parameter combination considering the following values: mutation rates of 10%, 50% or 100%, generational gaps of 1% and 50% and cross-over rates of 10%, 50% or 100%. The best combination of values was mutation=50%, cross-over=10% and GG=1%. We ran at least four experiments for each combination. It is worthy to note at this point that, in our program, mutation rate is not defined in the traditional way. Rather than applying mutation to every bit with a given probability, our 50% rate gives the probability that a single codon in the genotype will be mutated. This corresponds to a much smaller rate under the traditional interpretation.

This program of experiments gave two main results: a) most experiments ended with a big convergence in the population, b) the models with better fitness needed to be studied in detail, as they showed some of the features listed in table 1. They thus represent interesting novel models of habituation. We will look in depth at two of them. We have translated the models to a more usual notation. We have also omitted the $T_{ij}(t)$ function since is irrelevant in both models. The first is represented by the expression $A_i(t) = D_i : D_i(t) = S_i - (D_i * ln2.67)$. Figure 2 shows its behaviour during the habituation process (with the same conditions as in (Rankin and Broster, 1992), experiment 1, ISI 30). We can see how it reproduces the two main features of habituation: exponential response decrement and spontaneous recovery.

The second model $A_i(t) = D_i : D_i(t) = (\sum_k S_k * 4.7) - max(D_i, sqrt(abs(D_i/sqrt(343.8))))$ was confronted to a habituation process followed by a tentative of dishabituation (same conditions as in (Rankin et al., 1990)), corresponding to features 1 and 5 in table 1. Dishabituation is a sudden recovery of the response after the presentation of a novel stimulus (second 397 in our case). Figure 3 shows how this model exhibits a basic exponential response decrement and dishabituation. This last feature is specially important, since it has been neglected by all the existing models in the literature, except for (del Rosal et al., 2006).

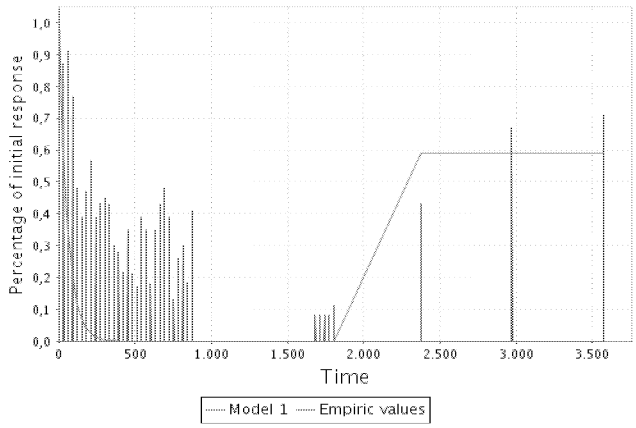


Figure 2: Empiric data Vs Model 1. Habituation and spontaneous recovery

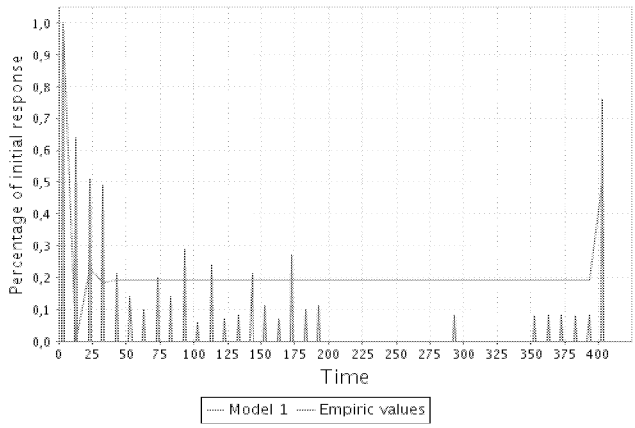


Figure 3: Empiric data Vs Model 2. Habituation and dishabituation

4 DISCUSSION AND FURTHER WORK

The curves presented in this paper show that our models have, only partially, some of the characteristics that these models have to exhibit. This circumstance becomes evident when confronting to the fitness function our model and some of the literature: one of the last models of habituation Alonso et al. (2005) gets a value of 22.11, on the other hand, the best models we have found never gets a fitness value below 24. We have, therefore, further work to do in terms of parameter tuning and other elements that could improve the evolutionary automatic technique we have developed.

Some of the future improvements concern the maintenance of diversity throughout the experiment, which can be done by modifying the genetic algorithm running in the background of the CGE system with mechanisms to preserve diversity. Other improvements will affect the grammar: adding new semantic constraints, in the form of new variables and their features, might make it possible to find better models than those

in the literature, specially if these new conditions are based on knowledge from experts in the field of habituation. In fact, our grammar includes the assumptions and constraints of the latest models in the literature (del Rosal et al., 2006). A more flexible grammar (for instance, with no limitation on the number of variables) could also open new possibilities for novel models.

Nevertheless, the results prove that the methodology we have followed is capable of automatically building models of habituation with interesting features. This encourages us to continue refining the algorithm, and supports the idea that CGE can be used as an *evolutionary automatic modelling* tool, since it has been able to tackle, with a significant degree of success, a modelling problem as difficult as associative learning.

Acknowledgments

This work has been partially sponsored by the Spanish Ministry of Education and Science (MEC), project number TSI2005-08225-C07-06.

.

References

Alonso, L., Moreno, R., Vázquez, M., del Rosal, E., and Santacreu, J. (2005). Spontaneous recovery of conditioned response in an autonomous agent. *Estudios de Psicología*, 26(3):365–376.

Byrne, J. H. (1982). Analysis of synaptic depression contributing to habituation of gill-wthdrawal reflex in aplysia californica. *Journal of Neurophysiology*, 48:431–438.

del Rosal, E. (2006). Automatic modelling; grammatical evolution applied to the problerm of modelling habituation in psychology. Master's thesis, Escuela Politecnica Superior. Universidad Autonoma de Madrid.

del Rosal, E., Alonso, L., Moreno, R., Vázquez, M., and Santacreu, J. (2006). Simulation of habituation to simple and multiple stimuli. *Behavioural Processes*, 73:272–277.

Echeandia, M. D., de la Puente, A. O., and Alfonseca, M. (2005). Attribute grammar evolution. *Artificial Intelligence and Knowledge Engineering Applications: a Bioinspired Approach, PT 2, Proceedings*, 3562:182–191.

Hall, G. (1991). *Perceptual and associative learning*. Oxford: Clarendon.

Knuth, D. E. (1968). *Mathematical Systems Theory*, volume 2, chapter Semantics of Context-Free Languages, pages 127–145.

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA.

Mazur, J. E. (2002). *Learning and behavior*. Upper Saddle River(New Jersey). Prentice-Hall.

Moore, J. H. and Hahn, L. W. (2004). An improved grammatical evolution strategy for hierarchical petri net modeling of complex genetic systems. *Applications of Evolutionary Computing*, 3005:63–72.

O'Donohue, W. T. and Kitchener, R. (1999). *Handbook of behaviorism*. San Diego: Academic Press.

O'Neill, M., Brabazon, A., and Ryan, C. (2002). *Genetic Algorithms and Genetic Programming in Economics and Finance*, chapter Forecasting market indices using evolutionary automatic programming. A case study. Kluwer Academic Publishers.

O'Neill, M. and Ryan, C. (2003). *Grammatical Evolution. Evolutionary Automatic Programming in an Arbitrary Language*. Kluwer Academic Publishers.

Ortega, A., de la Cruz, M., and Alfonseca, M. (2007). Christiansen grammar evolution: grammatical evolution with semantics. *IEEE Transactions on Evolutionary Computation*, 11-1:77–90.

Pavlov, I. P. (1927). *Conditioned reflexes*. London: Oxford University Press.

Rankin, C. H., Beck, C. D. O., and Chiba, C. M. (1990). Caenorhabditis-elegansa new model system for the study of learning and memory. *Behavioral Brain Research*, 37(1):89–92.

Rankin, C. H. and Broster, B. S. (1992). Factors affecting habituation and recovery from habituation in the nematode caenorhabditis-elegans. *Behavioral Neuroscience*, 106(2):239–249.

Rodrigues, E. and Pozo, A. (2002). Grammar-guided genetic programming and automatically defined functions. *Advances in Artificial Intelligence, Proceedings*, 2507:324–333.

Staddon, J. E. R. and Higa, J. J. (1996). Multiple time scales in simple habituation. *Psychological Review*, 103(4):720–733.

Stanley, J. C. (1976). Computer-simulation of a model of habituation. *Nature*, 261(5556):146–148.

Thompson, R. F. and Spencer, W. A. (1966). Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychological Review*, 73:16–43.

Wang, D. L. (1994). A neural model of synaptic plasticity underlying short-term and long-term habituation. *Adaptive Behavior*, 2:111–129.

Watt, D. A. and Madsen, O. L. (1977). Extended attribute grammars. Technical Report 10, University of Glasgow.

MANUFACTURING SIMULATION

STEEL MANUFACTURING SIMULATION

PROCESS RULES FOR SHEET METAL HYDROFORMING

Antonio Del Prete
Teresa Primo
Gabriele Papadia
Barbara Manisi

Department of Engineering Innovation, University of Salento
Via per Arnesano, 73100 – Lecce - Italy
e-mail: antonio.delprete@unile.it, teresa.primo@unile.it, gabriele.papadia@unile.it, barbara.manisi@unile.it

KEYWORD
Computer Aided Engineering (CAE), model analysis.

ABSTRACT

It is said many times that sheet hydroforming [1] is considered a good opportunity for industrial applications related to niche and medium-low volumes productions. Nowadays this technology has not found a specific application contest like it is for tubes hydroforming. In this last case, its large application to industrial cases has allowed to define through appropriate experimental validations "best practice" rules for the process design [2] and for its tryout [3][4][5].
In this specific case, a "shape factors" set has been defined with the proper goal to use it in order to classify metal components production through the application of sheet hydroforming. Finite Element Analysis (FEA) has been extensively used in order to investigate and define each shape factor with a proper comparison to the macro feasibility of the chosen component geometry. These shape factors have been used to track the process performances through their variation thanks to the usage of the numerical simulation which will be later validated with an appropriate experimental campaign which will be executed thanks to the usage of a specific equipment properly designed.

INTRODUCTION

Among all the possible components classes "producibili/ottenibili" thanks to sheet hydroforming specific applications related to the automotive industrial segment have been chosen. Typical examples are represented by external panels (Figure 1) and fuel metal tanks with complex geometries.

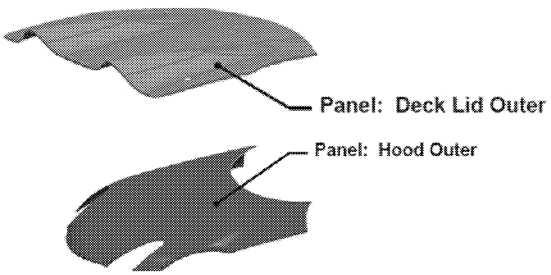


Figure 1: some industrial examples of components obtained with sheet hydroforming.

The feasibility configuration definition for a given shape can be done through an appropriate study of the influence of each process variable for sheet hydroforming on the process performances. The n-process variables define a process feasibility space made by n dimensions. With a proper evaluation of the influence of each process parameter on the process performances through the extensive usage of Finite Element Analysis a feasibility window can be defined. A subdomain of the feasibility window is the Process Window which takes into account all the possible constraints due to the available equipment characteristics like: maximum fluid chamber pressure, maximum blankholder force and its distribution, maximum pre-forming height, etc. Even if it is not possible to represent the n-dimensions space representative of the process feasibility a schematic representation can be adopted in order to understand which the process sensitivity to the variation of its variables is (Figure 2).

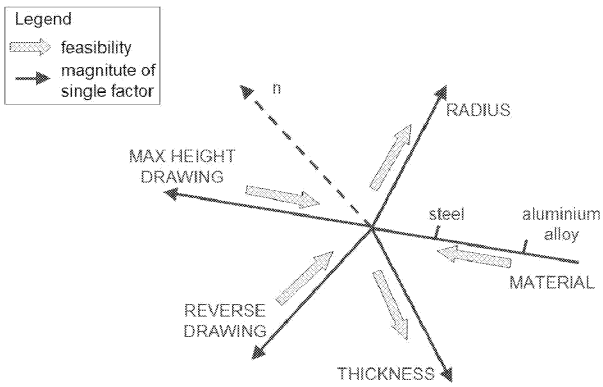


Figure 2: schematic feasibility space representation.

In this first schematic representation of each process variable a reference axis has been defined in order to understand the influence of its variation on the process feasibility which, in this case, is independent from the geometry profile. In order to be completely representative of the approached physical problem, this schematic representation has to be related to the specific geometry "producible by hydroforming". Any geometry can be characterized by a certain number "m" of adimensional shape factors. Then the feasibility space will change for the original n dimensions to the n+m dimensions due to the fact that in this case it is taken into account the formulation of the formed shape (Figure 3). In this case, the process feasibility is related to the specific shape factors for the given geometry. Authors have investigated which the

possible adimensional shape factors needed to characterize the external panels components class are.

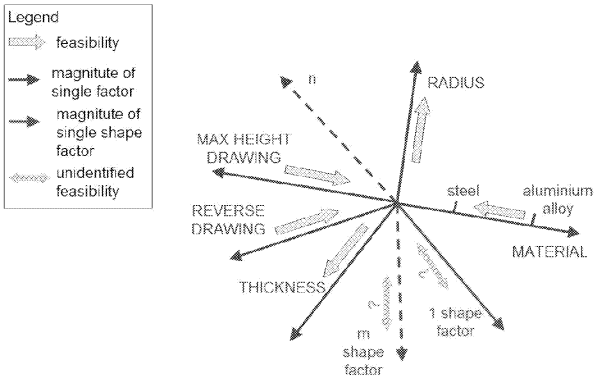


Figure 3: feasibility space for a specific component taking into account the proper shape factors.

The application of sheet hydroforming for external panels productions has already few remarkable industrial applications while on the contrary in the case of components for automotive fuel tanks this application looks like a very promising one because the most important automotive OEM (**O**riginal **E**quipment **M**anufacturing) is involved in specific research projects having as objective to promote the design and production of fuel tanks components through the plastic deformation of sheet metal panels (**S**trategic **A**lliance for **S**teel **F**uel **T**anks SASFT) [6]. Hydroforming could be a possible promising option, for this reason authors have considered also this class components in order to define appropriate shape factors.

SHAPE FACTORS DEFINITION

Having as reference the chosen industrial cases authors have proceeded with the shape factors definition in order to have adimensional coefficients representative for the given geometries.
Representative parts of the external panels "class" for automotive applications have been chosen, in particular two typical geometries were analyzed: a hood (Figure 4 and 5) and a fender components (Figure 6 and 7).

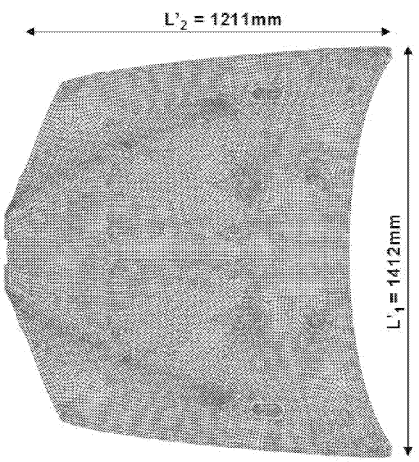


Figure 4: example of panel component: hood (top view).

For each one of them appropriate adimensional parameters can be defined trying to correlate the maximum hydroforming depth with the most critical dimension of the original blank in top view.

Figure 5: example of panel component: hood (side section view).

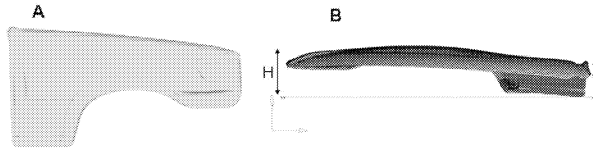


Figure 6: example of panel component, fender A: cad model (top view); B: FE model (side view).

Considering L_1' and L_2' (
Figure 7) and H (Figure 6), you can define the following shape factor given from the ratio:

- $R_2 = L'_2 / H$ (where L'_2 is the smallest blank dimension in top view and H is maximum drawing depth).

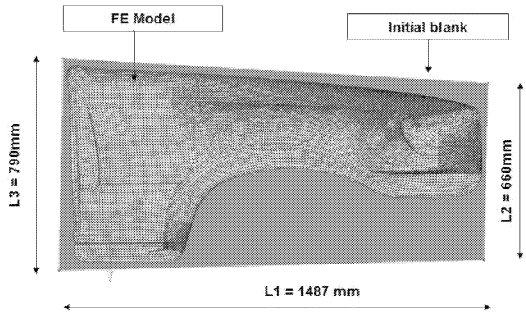


Figure 7: example of panel component: fender, top view of FE model and its initial blank.

The defined adimensional characteristic is only one of the possible ratios that can be useful to characterize a certain class of components.
In the case of components used for the fuel tanks construction, characteristic geometric elements have been also considered in order to define the relative shape factors. Elements such as redrawing height and top view dimensions were considered as significative as well as the presence of more than one drawing depth in the same shape.
Thanks to these elements new shape factors have been defined:

- $R_h = L_2 / H_2$ (where L_2 is the minimum dimension of the blank in the top view related to the difference between two adjacent drawing heights, Figure 11 (b)).
- $R_c = L_c / H_c$ (where L_c is the maximum dimension in top view of a redrawing area and H_c is the maximum depth of this area).

Figure 8 illustrates some of the aspects which make the difference between a plastic fuel tank and a metal fuel tank. In the present production, upper and lower metal shells are realized separately after the forming operations the two parts are welded together. This process may lead to a high percentage of defected parts due to the bad quality of the welding process. This type of component can be produced thanks to sheet hydroforming avoiding the welding phase with some design reviews in comparison to the original geometry used for the plastic one [7].

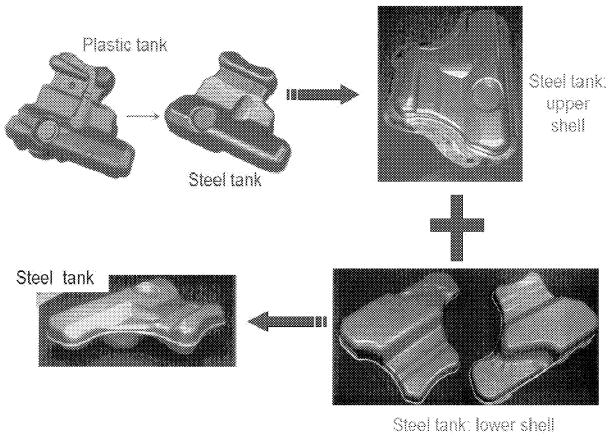


Figure 8: from the plastic tank to the parted steel tank.

For a general component, obtained by hydroforming, it is possible to assume that its general sections could appear like the one reported in
Figure **9**. For this profile it's possible to apply the shape factors already defined:

- $R_2 = L_2/H = L_2/1$
- $R_h = L_2/H_2 = L_2/6$ or $L_2/2$, or $L_2/7$
- $R_c = L_c/H_c = 4/2$.

If it's possible to define more than one shape factor the most critical one in terms of its influence on the process feasibility will be considered.

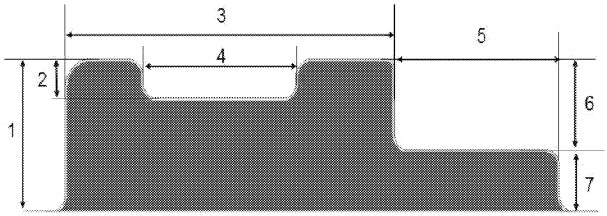


Figure 9: generic section shape for a sheet hydroformed component.

NUMERIC ADOPTED PROCEDURE

Equivalent geometric shape

The qualitative and quantitative analysis of the identified factors for the components classes taken into account will require a numerical-experimental activity which needs a relevant amount of resources if real parts geometries have to be investigated. For this reason similar but simpler components were chosen with the precise goal to reproduce the behavior of the defined real components, in order to use them for the numerical and experimental campaigns related to the "Process Performances" investigation. Three reference models have been defined and named as MOD1, MOD2 and MOD5 as it is reported in Figure 10, 11 and 12, respectively.
For each one of them it can be possible to identify the generic shape factors defined for the profile reported in Figure 9.

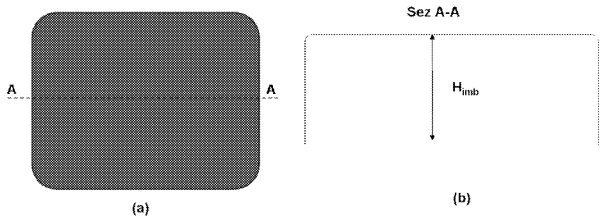


Figure 10: MOD1, shape factor definition. (a): top view, (b): section view.

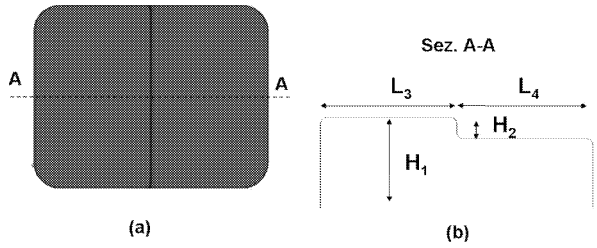


Figure 11: MOD2, shape factor definition. (a): top view, (b): section view.

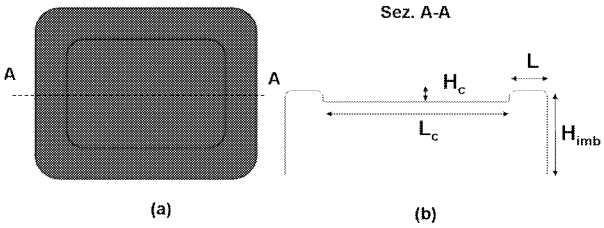


Figure 12: MOD5, shape factor definition. (a): top view, (b): section view.

Analysis of the equivalent shapes

For the given three models a numerical investigation has been developed in order to evaluate the influence of some geometric and process parameters on the process performance. Here the analyzed factors of each model are described:

TAB 1: MOD1's factors			
		LEVELS	
N° FACTOR	Name Factor	Lower Level (LL)	Upper Level (UL)
1	$H_{preforming}$ [mm]	15	45
2	Thickness [mm]	0.7	1
3	A1 [ton]	10	22.5
4	A2 [ton]	10	40
5	A3 [ton]	10	22.5
6	H_{imb} [mm]	100	150
7	R_p [mm]	10	50
8	R_m [mm]	10	20

where the Ai values are the blankholder forces applied by each actuator. In fact, in the developed models it has been considered a total number of twelve independent actuators which can be considered not all independent for the chosen shapes. For MOD1 and MOD5 due to the double symmetry of the models the independent actuators are just three while they become six for MOD2 where the geometry has just one symmetry plane. $H_{preforming}$ is the value of the preforming height, H_{imb} is the maximum drawing depth, R_p and R_m are punch and die radius, respectively.

TAB 2: MOD2's factors.			
N° FACTOR	Name Factor	LEVELS	
		LL	UL
1	$H_{preforming}$ [mm]	15	45
2	Thickness [mm]	0.7	1
3	A1 [ton]	12	25
4	A2 [ton]	10	30
5	A3 [ton]	10	20
6	A4 [ton]	10	25
7	A5 [ton]	15	35
8	A6 [ton]	10	25
9	H_{imb} [mm]	100	150
10	H_2 [mm]	20	35
11	R_m [mm]	10	20
12	R_1 [mm]	10	50

For MOD2 and for MOD5, H_2 (Figure 11), R_1, and L (Figure 12) are added geometric parameter in order to fully define the geometry profile.

TAB 3: MOD5's factors.			
N° FACTOR	Name Factor	LEVELS	
		LL	UL
1	$H_{preforming}$ [mm]	15	45
2	Thickness [mm]	0.7	1
3	A1 [ton]	10	18
4	A2 [ton]	8	20
5	A3 [ton]	12	18
6	H_{imb} [mm]	100	150
7	H_2 ($H_{cavità}$) [mm]	20	30
8	L [mm]	65	130
9	R_p [mm]	10	25
10	R_m [mm]	10	20

For each given shape the appropriate metal forming set up has been developed for the different process conditions taken into account. Different process responses have been evaluated for each model. As example it is reported (Figures 13, 14) the FLD contour map and plot obtained for one of the considered process conditions related to MOD5.

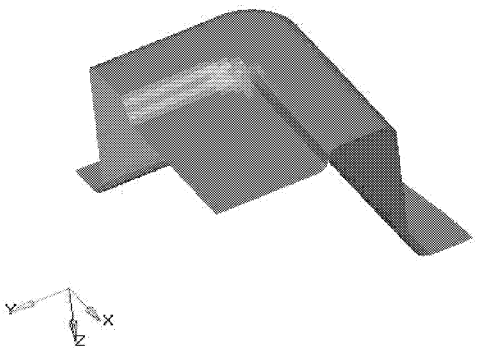


Figure 13: MOD5 hydroformed part, iso view.

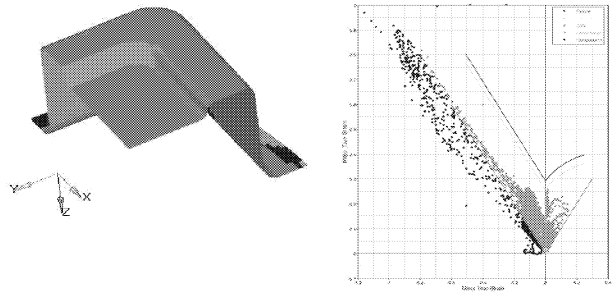


Figure 14: MOD5 hydroformed part, FLD contour map and plot.

In order to have a graphical immediate description of the interactions among the main factors of the sheet metal hydroforming, specific plots have been created in order to formalize the relationships between some of the process responses and some of the shape factors is reported in the case MOD1 (Figure 15) where it is reported the thickness reduction (%) trend having as independent variable the shape factor R_2.

For two different combinations of the thickness of the initial blank (0,7 mm and 1 mm) and constitutive material (low carbon steel FeP04 and aluminium alloy Al6061). The dimensions of the blank are always the same, the different depth drawing determines different R_2 values.

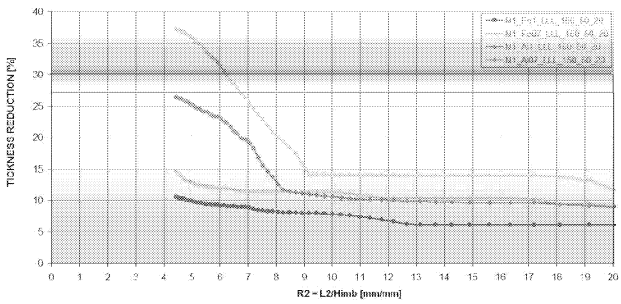


Figure 15: MOD1 first defined plot for process feasibility evaluation

Following the same philosophy in Figure 16 is reported the distance of the areas with maximum major strain in FLD (**F**orming **L**imit **D**iagram) from the FLC (**F**orming **L**imit **C**urve) versus R_2 parameter.

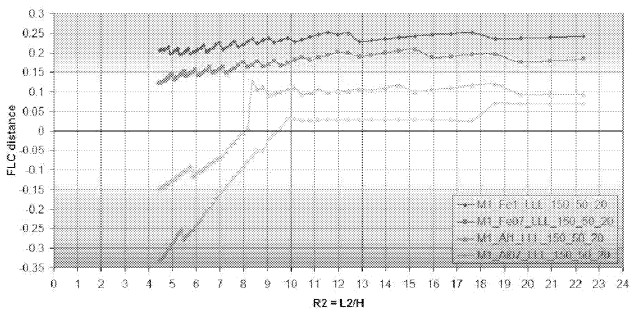


Figure 16: second graphic topology, for MOD1.

A coloured representation has been used in both cases in order to immediately identify the feasibility areas (the green ones).

CONCLUSIONS AND FURTHER DEVELOPMENTS

The defined plots reported in Figure 15 and 16 can be considered a significant aid for process feasibility definition for sheet metal forming. In fact, through their usage it is possible to find out the critical process parameters combination and to evaluate the safe or non safe margins of the obtained results in comparison to the solution/s suggested like the feasible one/s, according to the defined quality standards parameters. The defined diagrams are a first attempt to support process designers with effective solutions in order to understand the process performances for each designed configuration. At the same time, these diagrams can help to understand the robustness level of the implemented solutions.
The implemented shape factors give the chance to analyze, in a very early stage of the process development, the macro feasibility of a particular shape manufactured by a sheet metal hydroforming process.
In the next future, a specific experimental activity will be developed to validate the implemented procedure, taking advantage of the fact that a dedicated equipment has been assembled for this scope in accordance with the simulation set up developed for the different analyzed models (Figure 17).

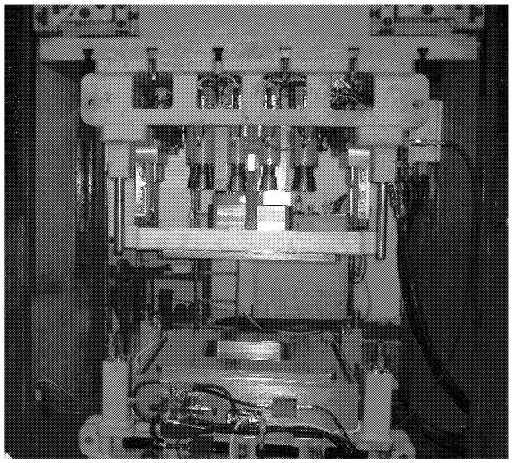


Figure 17: hydroforming cell.

ACKNOWLEDGEMENTS

Authors are very grateful to "MUR: Ministero dell'Università e della Ricerca" for funding this work recognized as I.T.Idro: innovative solutions for sheet hydroforming.
Special thanks goes to Stamec srl which has the role of industrial partner of the project.

BIBLIOGRAPHY

[1] *K. Siegert* - Hydroforming of Tubes, Extrusions and Sheet Metal vol. 2 – Institute for Metal Forming Technology of the University of Stuttgard.

[2] *M. Vermeulen, K. Siegert, A. Schwager* - Sheet Metal Formability Test for Hydroforming - Papers of the International Conference on Hydroforming, Fellback (Germany), 6/7 November 2001.

[3] *T. Sokolowski, K. Gerke, M. Ahmetoglu, T. Altan* - Evaluation of tube formability and material characteristics: hydraulic bulge testing of tubes - J. Mater. Process. Technol. 98 (2000) 34–40

[4] *W. Rimkus, H. Bauer, M.J.A. Mihsein* - Design of load-curves for hydroforming applications - J. Mater. Process. Technol. 108 (2000) 97–105

[5] *B. Carleer, G. van der Kevie, L. de Winter, B. van Veldhuizen* - Analysis of the effect of material properties on the hydroforming process of tubes - J. Mater. Process. Technol. 104 (2000) 158–166

[6] Evaluation of the Corrosion Durability of Steel Systems for Automobile Fuel Tanks – SASFT.

[7] *Ekkehard Friebe, Torsten Lanchammer, Arndt Birkert, Jorg Neubert, Matthias Kleiner, Axel Wellendorf* - Hydromechanical Deep Drawing of Passenger Car Fuel Tanks – International Conference on Hydroforming, Germany 6th-7th November 2001.

[8] www.sasft.com.

PROCESS AUTOMATION TOOLS DEVELOPMENT FOR SHEET METAL HYDROFORMING SIMULATION

Antonio Del Prete
Alfredo Elia
Teresa Primo
Barbara Manisi
Department of Engineering Innovation,
University of Salento – Lecce – Italy
e-mail: antonio.delprete@unile.it, teresa.primo@unile.it, alfredo.elia@unile.it, barbara.manisi@unile.it

KEYWORDS

CAE, Process Oriented Models, Software Engineering.

ABSTRACT

For a no conventional forming process metal forming simulation has been extensively used in order: to design an appropriate experimental tooling and, at the same time, to understand the influence of each parameter on the process performances. Large scale computation corresponding to a very high number of simulated models has led authors to take consideration the option to develop a customized software application. A general purpose explicit-implicit solver and a commercial software suite have been joined to form the proper base for the customization process. The developed application introduces interesting advantages not only in terms of time saving but also in terms of improvement of the process development procedure giving the chance to manage properly engineering process data.

INTRODUCTION

Nowadays hydroforming techniques have been largely accepted by industries for the production of components characterized by good surface quality, high-dimensional accuracy together with high drawing ratio and complex shapes [1][2]. The main distinction among the large variety of hydroforming techniques is usually based on the raw material adopted: tubular or sheet material. Today tubular material is predominantly adopted for the mass production; actually sheet hydroforming is used for small batch production because both its cycle time and clamping forces (expansive press) are higher than tube hydroforming ones [3]. However, sheet hydroforming process offers the possibility of no restrictions in the final shape of the part (on the contrary tube hydroforming is simply used for the production of space frame structures or longitudinally extended parts) together with a large number of variants [4].
In order to achieve the increasing technical demands of automotive components, complex industrial design leads to difficult forming processes. In certain cases, conventional sheet metal forming techniques have reached their limits. Optimized application of dies and presses for sheet hydroforming leads to a reduction of product specific costs and thus to lower costs per part, which is extremely important for low volume productions. As generally known, there are great differences between conventional deep drawing and sheet hydroforming, resulting from liquid forming medium. This leads to frictionless force transmission to the sheet and to a high contact pressure between sheet and die. This results in nearly no friction between sheet metal and fluid and high friction forces between rigid punch and sheet when deep drawing is realized with counter pressure. Due to that, higher forming forces can be transferred to the forming zone, which results in greater draw ratios (parts with a higher drawing height). Parameters such as lubricant, surface treatment of the die and coating on the sheet as well as the forming behaviour of the sheet metal play an important role in the hydroforming process [5].
The present work highlights some of the obtained results of a research program whose object is to understand the influence and the management of the process variables in sheet metal hydroforming.
This project has as one of its focus a specific experimental phase, which is necessary to understand the reliability of the numerical defined models and their output data. For these reasons, it has been necessary to develop an experimental toolkit able to accomplish the research program requirements in order to test the process procedures on different shape components in the shortest time and in the cheapest way. The developed experimental kit has been designed to test different shape components using the major part of the equipment and changing just few elements in order to adapt the tooling to each specific shape. The used methodology can be easily extended to industrial applications allowing possible end users to face with the hydroforming technology without requiring large investments.
The project philosophy is to investigate the relationships between material and geometrical characteristics of the product and process parameters. This investigation can be developed only throughout the usage of numerical and experimental campaigns, to define quality functions and to analyze new strategies for sheet hydroforming process tryout.
In the same Program, it is also foreseen the use of the developed knowledge in a specific numerical simulation environment in order to give, even to no-expert users, a useful and simple way to design and to manage the hydroforming process.
Before the software architecture definition, the proper software environment has been individualized to verify the

effective possibility to develop the customized applications. Authors have chosen a multi-purpose suite named Altair HyperWorks due to the fact that this environment is particularly oriented to customization thanks to the fact that it has specific internal tools for process automation development mainly based on Tcl/Tk language.
The specific developed application has been defined as HydroForming Manager (HFM). This tool is an environment in which the user can set up models in order to perform hydroforming process simulation, as well as manage results by using automated reporting options. In order to address properly the system architecture particular attention has been paid to some aspects like:

- intercommunication
- platform independency
- user friendly
- modular structure.

Having these precise requirements as mandatory, the developed application let users analyze and design sheet processes hydroforming, in the shortest time, reaching the product/process feasibility with a high level of accuracy.
It is noteworthy to mention that in order to improve the knowledge in sheet hydroforming, four different geometrical shapes have been investigated (Figure 1).

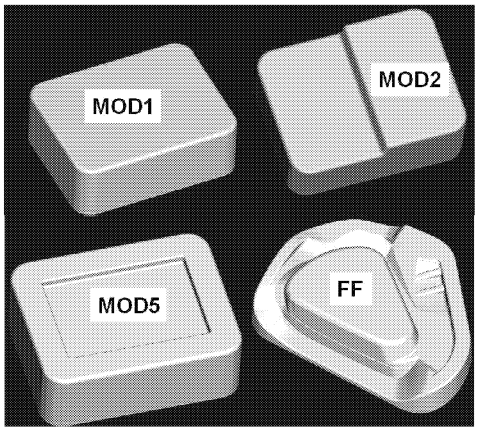


Figure 1: four different study cases: MOD1, MOD2, MOD5 and industrial case "Fondello Fanale" (FF)

SHEET HYDROFORMING ENVIRONMENT

Actually, different commercial software, having as capability the hydroforming process simulation (both tubes and sheets), are available such as Autoform® or PamStamp®, but these tools do not offer the chance to be customized following the user requirements. On the contrary, HyperWorks®, among its strength points, has the peculiarity to be customizable in all its applications (pre, solver, post, reporting interfaces) and it has the advantage to be fully integrated with Ls-Dyna®, one of the most used explicit-implicit solvers for metal forming simulation. The numerical definition for FEA input file is the first difficulty normally encountered for hydroforming process analysis. In fact, due to the fact that Ls-Dyna is an explicit-implicit general purpose code for a wide range of applications, it does not offer a specific user interface for hydroforming simulation [5]. In the input file for FEA it is necessary to define some control parameters related to the numerical solution, such as: the mathematical formulation of the used elements, number of integration points for each element, the material law definitions, contact formulations, etc.
Each of the set-up operations is carried out manually for every input file (Figure 2). The numerical simulation requires the application of practical procedures which are not strictly related with the forming process thus many times users face with problems not related to the manufacturing aspect but to the numerical one.

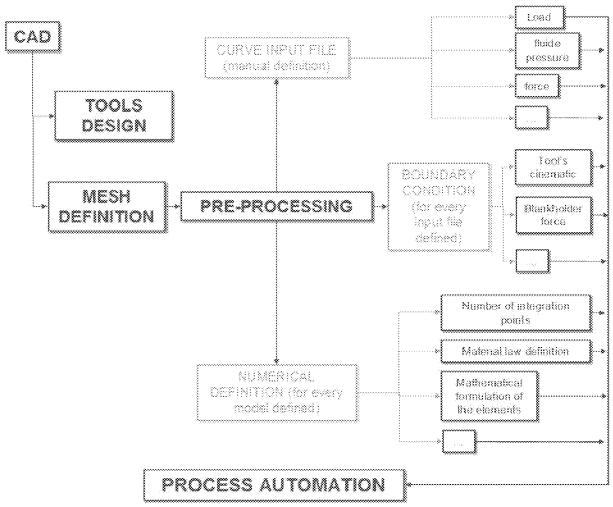


Figure 2: process outline

As example it is noteworthy to describe one of the common set up problems such as the fluid pressure law definition used for the hydroforming process simulation. In this case it is necessary to refer to the *AIRBAG card with the *linear fluid* option. The card is defined by:

- load curve identification (LCID) for fluid pressure time variable;
- fluid density.

The fluid pressure operates in a volume defined as *control volume*. The FE model's elements, pertaining to blank and fluid chamber/die, define the control volume. The selection of these elements is very complex due to the mesh aspect (Figure 3).

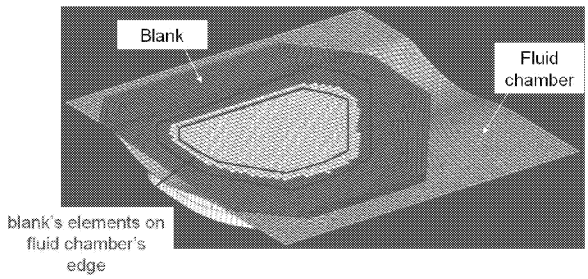


Figure 3: elements of the control volume for FF

The geometry complexity increases the difficulty for the definition of the control volume.
The aspect related to the fluid pressure law definition has been solved thanks to the process automation implementation like for the following other aspects always related to the sheet hydroforming simulations set up: A) In order to optimize the layout mesh for control volume definition (Figure 4) a special automatic partition has been implemented for the surface of the blank.

As example, in MODi, the different patches for blank partition are shown in Figure 5 (a). Every patch has a different elements average dimension, smaller in areas which will be interested by radius forming action and larger in the other parts of the model (Figure 5 (b)). The same approach has been applied to all the considered models.

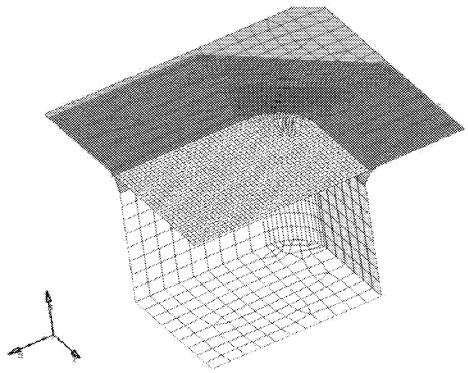


Figure 4: elements of the control volume for MOD1

The automatic creation of the fluid chamber is another software tool specifically implemented. In order to optimize the contour shape, the fluid chamber is created starting from punch shape, considering its external perimeter.

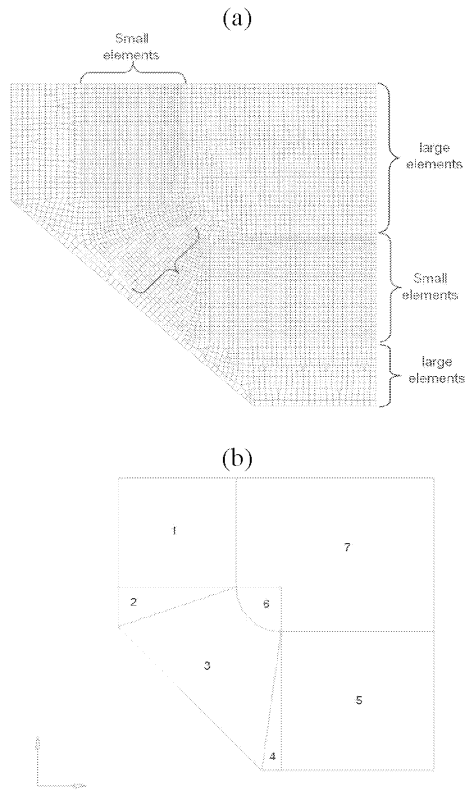


Figure 5: (a): mesh aspect; (b) patch surface

B) Like it has been done for the above mentioned cases, also for the simulation parameters set up, such as the elements mathematic formulation, number of integration points and hourglass energy control an automation strategy has been defined in order to obtain the proper customization.
C) The material description is another step in the input file definition. In order to define the material characteristics all its principal mechanical properties are necessary: Young's Module, Poisson's Module, yield stress, density, anisotropic hardening parameter and true strain versus true stress curve. All these parameters need to be set for any calculation; a library with principal material used for deep drawing has been implemented in order to minimize the set up necessary time for the material definition.

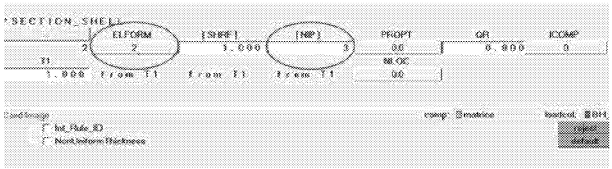


Figure 6: blank section definition.

D) Another time consuming aspect is due to: contacts definition, materials and section assignment to the components, process curves definition, blankholder forces and pressure load curves assignment, punch motion definition. In particular, all the assignment operations foresee a sequence where all the parameters are set into correspondent cards. For example for process curves definition it is necessary to characterize: a card image, a load curve, the component identification and load's degree of freedom.

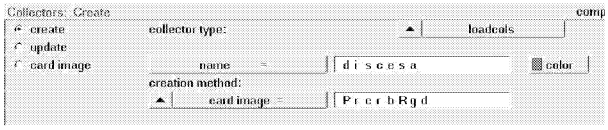


Figure 7: first step for cinematic punch definition

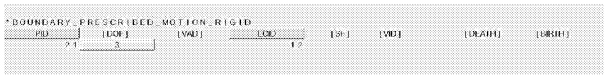


Figure 8: second step for cinematic punch definition

PROCESS AUTOMATION TOOLS

The developed environment (HFM) can be referred as a framework strictly connected to the sheet metal hydroforming process application. Thus, as it has been already said, a benchmark activity has been conducted, in order to choose proper tools to integrate in HFM, which has been focused on two main requirements: customization and intercommunication. As result HyperWorks® (HW) environment and LS-Dyna solver have been resulted as the optimal solution.
HW open architecture and its extended interfaces to popular analysis software provide a solution-centric, customization-friendly, solver-neutral environment to be used for design process automation and manufacturing process simulation.
Anchored by the HyperWorks product line, these software tools enable designers and engineers to model complex structures and assemblies, study their physical behaviour and improve and optimize design performance.
The powerful of this package resides in the technology it uses, based on platform independent programming languages as Java, Tcl/Tk (Tool Command Language/Tool Kit) and C [7][8][9]. Each product included in HyperWorks package has

its own libraries of Tcl commands and an integrated Tcl console. Thus, customized commands have been developed by implementing Tcl scripts that collect Hyperworks specific commands as well as Tcl specific commands, and by launching them from inside HyperWorks Tcl console. Moreover, whenever limitations of HW APIs (Application Programming Interfaces) have been found in communication with other parties applications, the ability of Tcl has been used to directly launch Operating System commands.
In order to develop HFM GUI (Graphical User Interface) and its connection with solver and data server, an IDE (Integrated Developing Environment) has been used, referred as Process Studio ®, available inside HW package. This product enables rapid development and deployment of process automation applications in the CAE domain, by offering a customizable Java class named Process Manager, a programmable personal workflow manager that guides users through standard work processes. This class implements a GUI made by two main components (Figure 9):

- process automation tree;
- panel area.

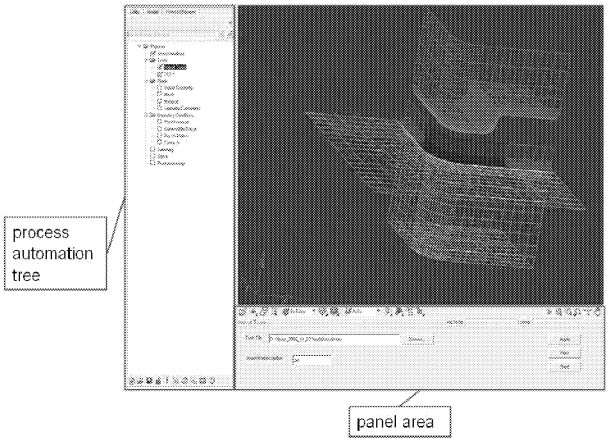


Figure 9: HFM GUI

Each node of a process tree represents an atom action (i.e. import a file, launch solver, make surface meshing, etc.) referred as a task. Each task has its own panel that hosts task specific GUI controls. Moving through process tree nodes will cause panel area to switch to task specific panel. Following this strategy a user friendly GUI has been obtained, by asking user only for those data referred to a specific context.
By using Process Studio, Hydroforming Manager Process Tree Class has been developed. This Java class should be instanced by user to access to Hydroforming Manager Tree. Figure 9 shows an example of a process tree opened in an Hyperworks pre-processor instance.

HYDROFORMING MANAGER ARCHITECTURE

In order to be integrated in Hyperworks framework and to communicate with all its tools, a specific customization of the Process Manager class has been made Figure 10 shows the interfaces dependences tree.
Since the IHWMControl interface is implemented, an instance of the IHWMFrameWork interface is passed on to the Java bean. This is the point where beans gain access to various classes (as interfaces) in the Process Manager architecture, including:

- DataModel
- Workflow
- Communication Manager
- Tool Bar Customizer

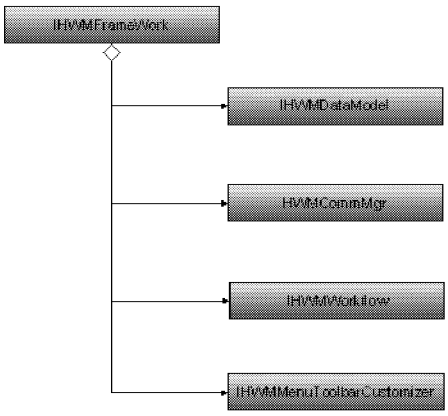


Figure 10: interface architecture.

The DataModel is a placeholder for the beans to store data, and to share data across beans. Any bean can get the DataModel from the IHWMFrameWork handle. The DataModel persists any information that is placed in it (unless it is specified as transient).
The Workflow Manager serves as a handle to the workflow. With the handle to the Workflow, the beans can switch to the next, to the previous, or to any specific task.
The Communication Manager is the gateway to spawning third party applications and establishing a socket connection with them, or connecting to and communicating with servers (ex. Database servers).
This interface can:

- Spawn another application and create a socket connection (for communication), and send/receive strings.
- Connect to another application that has created a server socket and communicate with that application using strings.
- Connect and send SQL queries to a relational database server using JDBC drivers.
- Connect to servers that use SOAP protocols to communicate (ex. eCompute server).

Any bean can make calls into the Communication Manager and spawn and/or connect to an external application.
The Tool Bar Customizer interface allows to customize HW application tool bars and menus, by adding, deleting or modifying default controls.
In conclusion, what has been implemented in HFM is the right answer to the formalized needs which can be summarized as follows:

- build a unique integrated system to be used throughout all sheet hydroforming project activities;
- launch simulations locally as well as remotely, on a server;
- interface with many solvers to perform implicit and explicit calculations;

- be friendly for users with a technological background, even if they don't have FE modelling experience;
- be ready to connect to an EDM (Engineering Data Management), in order to keep trace of historical data, allowing the growth of corporate knowhow.

PERFORMANCE EVALUATION OF THE IMPLEMENTED CAE ENVIRONMENT

A specific process tree has been defined in the HFM GUI interface and it has been used by the user not only as step by step procedure but also as a reference during the process simulation.
Figure 11 shows the panel for the blank partition generation. Traditionally, this operation can be manually performed through a series of geometric operations. Thanks to HFM, once imported the FE tools models, it is only necessary: to import the blank cad model, to select the reference (the die) and to define the thickness. Applying this procedure, a new blank cad is generated; it is characterized by the appropriate card and the partition, with remarkable time saving.

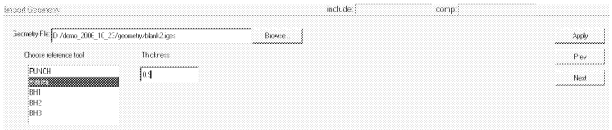


Figure 11: HFM panel for blank partition generation

For blank materials a database containing all necessary information for material definition has been implemented. In HFM, it's possible to filter the material by topology: steel, High Straight steel and aluminium alloy. In this way it's possible to update parameters like density, Young's Module, etc. by the material topology choice in a single operation.
In the same step, the HFM allows to assign material, section properties and numerical definition (hourglass, element formulation, integration point number) to each component of the FE model (Figure 12).

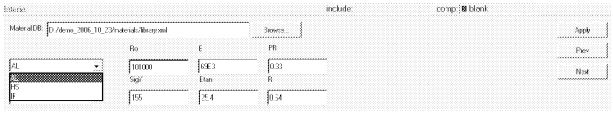


Figure 12: material definition

Having as reference what has been said for the *AIRBAG card definition, it can be said that in HFM the control volume elements are automatically selected. Only the fluid pressure in the fluid chamber is editable (Figure 13).

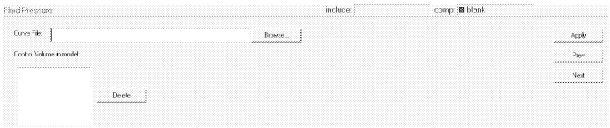


Figure 13: *AIRBAG automatic definition

The developed application allows to manage FE models with partitioned blankholder like in the case of the one developed for the experimental campaign which has twelve different actuators able to differentiate the force during the process time and along the blankholder profile. This option is well modelled in HFM and it is easy to update thanks to the automatic option which enables to assign the right force profile to each blankholder segment. For the specific equipment it has been implemented a space variable blankholder force.

CONCLUSIONS AND FURTHER DEVELOPMENTS

Process Automation application has demonstrated its benefits when applied in the development of a simulation environment for sheet metal hydroforming process calculation. Thanks to the developed tool, users can concentrate their attention to the technological aspects of the problem they are facing with. Authors believe that another main advantage could be obtained if the actual developed tool it will be implemented giving the chance to connect it with proper optimization tools in order to let users to optimize process parameters for a given solution. It is also noteworthy the possible advantage coming from a structured data management which could be helpful in order to evaluate which is the corporate know how for each specific application.

ACKNOWLEDGEMENTS

Authors are very grateful to "MUR: Ministero dell'Università e della Ricerca" for funding this work recognized as I.T.Idro: innovative solutions for sheet hydroforming. Special thanks are addressed to Stamec srl which has the role of industrial partner of the project.

BIBLIOGRAPHY

[1] K. Siegert, M. Haussermann, B. Loesch, R. Rieger, "Recent developments in hydroforming technology", *Journal of Materials Processing Technology* 98 (2000) 251-258.
[2] S. H. Zhang, Z. R. Wang, Y. Xua, Z. T. Wang, L. X. Zhou, "Recent developments in sheet hydroforming technology", *Journal of Materials Processing Technology* 151, (2004) 237-241.
[3] Ch. Hartl, Research and advances in fundamentals and industrial applications of hydroforming, *Journal of materials Processing Technology* 167 (2005) 383-392.
[4] L. Lang, J. Danckert, K. B. Nielsen, X. Zhou, "Investigations into the forming of a complex cup locally constrained by a round die based on an innovative hydro mechanical deep drawing method", *Journal of Materials Processing Technology* 167 (2005) 191-200.
[5] M. Vermeulen, K. Hertschap, K. Siegert, A. Gehle, A. Schwager, "Sheet Metal Formability Tests for Hydroforming", Hydroforming of Tubes, *Extrusions and Sheet Metals*, volume 2, Institute for Metal Forming Technology of the University of Stuttgart.
[6] LS-DYNA Keyword User's Manual – version 970 LSTC.
[7] M. Harrison, M. McLennan. "Effective Tcl/Tk Programming: Writing Better Programs with Tcl and Tk", Addison-Wesley Professional (1997).
[8] S. Horstmann, G. Cornell "Core Java 2" , volume 1, Prentice Hall (2001).
[9] R. Jaeschke, H. Schildt, "C: The Complete Reference (Osborne Complete Reference)", McGraw-Hill Education (2000).

DETAILED MODEL OF SPOT-WELDED JOINTS TO SIMULATE THE DYNAMICAL FAILURE OF CAR ASSEMBLIES

Emmanuel H.J Lamouroux
DaimlerChrysler AG
Group Research & Technology
Benz-Straße
D-71063 Sindelfingen, Germany
E-mail:Emmanuel_Lamouroux@yahoo.fr

Daniel Coutellier
Valenciennes University
Laboratory L.A.M.I.H UMR CNRS 8530
F-59313 Valenciennes Cedex 9, France

Falko Seeger
DaimlerChrysler AG
Group Research & Technology
Benz-Straße
D-71063 Sindelfingen, Germany

KEYWORDS
Finite element method, spot-weld, material properties' identification, Gurson material model, dynamical loading.

ABSTRACT

This paper presents the construction and the validation of a spot-weld's detailed model using the finite element method. The goal is to provide, using the detailed model, input data for the validation of a simplified model, which will be later on implemented in the complete car model.
First, by studying a spot-weld's cross-section, three different sub-areas (spot-weld, heat affected zone and basis material) have been identified and then characterized using a Vickers hardness test. Moreover, the geometry of the weld has been carefully analyzed in order to build an accurate and very fine numerical model.
To reflect the mechanical properties of each zone, a conversion method based on the extrapolation of the tensile strength from the Vickers hardness value has been used coupled to a scaling method providing the strain at rupture using the "newly" extrapolated tensile strength value. Finally, the strain rate effect has been taken into account by considering the stress/strain curves of the different sub-zones for three strain rates. To compute the damage's evolution and the specimen's rupture, the Gurson material model implemented in the explicit time integration method "LS-DYNA3D" has been used. The global accuracy of the detailed model is estimated by comparing not only the simulated and experimental force/displacement curves but also the weld's rupture modus under dynamical load cases.
This study shows that the detailed models of the "Tensile/Shear" and "Coach-Peel" specimens are validated for the steel H320LA considering a symmetrical material thickness of 1.50 mm.

INTRODUCTION

For a few years, the offer on the automotive market has grown and diversified considerably due to the emergence of constructors coming from "new countries" such as China and India for example. To withstand to this new concurrence and consequently preserve their market share, the "established" constructors should be extremely competitive and must develop products for the different continents, which leads obviously to a diversification of their products. Parallel to that, the constructors must also nowadays renew their models more frequently in order to integrate permanently the newest technological innovations. As a matter of fact, the dynamism and the reactivity of the automotive market do not make it possible anymore to spread out the conception of a new vehicle over several years. Thus, during the last years, several constructors reconsidered their conception's processes to eliminate the redundant steps and all superfluities without sacrificing the quality of the final product. The remaining conception's key-steps have been often parallelised to reach an important profitability in terms of time and finance.

For example, it is of high importance to finalise rapidly and precisely the mechanical characteristics of a piece without building it physically. The use of numerical tools offers the possibility to carry out easily parametric studies which are very useful in the conception phase. Due to the very important progress of informatics in terms of rapidity, reliability and flexibility, numerical methods are widely used in the industry. Thus it is possible nowadays to simulate using finite element method the crash of a whole car with dummies and airbags against a rigid barrier. However, to evaluate the quality of a numerical model, it is necessary to compare the numerical results with the experimental one and to observe the same trends.
In terms of numerical simulation of crash behaviour, the most recent comparisons between numerical/experimental

testiness have shown that the main source of results´ dispersion is due to a bad prediction of the mechanical behaviour and fracture of spot-welded structures or other joining techniques such as adhesive bonding, clamping or laser welding.

As a matter of fact, in the complete car model, it seems that the whole frame and all the different joining techniques have to be modelled in order to represent as best as possible the reality. For crash simulation, it is very important to know how many joints will fail and to be able to predict not only their location but also the force they can withstand. In fact, during the crash the spot weld may fracture, which induces that the complete structure will have different deformations and intrinsic mechanical properties than before the impact. Since, a car's frame contains several thousand spot-welds, it is impossible to have a detailed representation of each of them according to the huge calculation time they require. One possibility for a reliable and practicable description of spot-welds in the crash simulation is to developed replacement elements based for e.g. on connectors or hexa-elements. To define the mechanical properties of this simplified element, the deformation and failure behaviour of the spot-weld has to be precisely characterised.

The classical way is to carry out numerous experimental campaigns for different materials, thicknesses, load angles and load speeds to develop a general failure criterion leading to the creation of a simplified element. Thus, the immense number of experiments needed to validate the simplified model takes a lot of time and money. As a matter of fact, the goal of this paper is to develop a method for the determination of the load-carrying capacity of the spot-weld (for different material combinations and for different load conditions) using few standard experiments and detailed simulations. Some preliminary investigations using this method based on "detailed-simplified" models are available in the literature. For example, in 2001, using the explicit FE-code PAM-SOLID™, Markiewicz et al. (Markiewicz et al. 2001) have identified a material behaviour law for the various zones of a spot-weld under quasi-static loadings. They analysed and studied, for a metal sheet assembly structure of 0.7 mm/0.7 mm, the rupture of spot-welded tensile-shear, cross-tensile and peeling specimen using the Gurson material model. Currently, different activities on the topic using also the Gurson damage model and the FE-code ABAQUS/STANDARD are done at the Fraunhofer Institute from Freiburg in Germany (Sommer and Sun 2003).

To have access to the local mechanical properties of the spot-welded structures, a method based on simple tests such as micro-hardness measurements and tensile tests will be proposed. With the previous mechanical data and with the pre-processor MEDINA a very fine numerical model composed of eight nodes volume elements with a length of around 0.2 mm will be constructed. Using LS-DYNA3D as processor and post-processor, the method's validation will be done for the "Tensile/Shear" and the "Coach-Peel" specimens under dynamical loadings.

CHARACTERISATION OF THE GEOMETRICAL AND MECHANICAL PROPERTIES OF THE SPOT-WELD

Basically, welding is a process used to permanently join two, usually metallic, components by the localized coalescence that occurs under certain combinations of temperature, pressure and metallurgy (Panontin et al). The components are welded together as a result of the heat created by electrical resistance. This is provided by the work pieces as they are held together under pressure between two electrodes. Spot welding may be performed manually or robotically and may take between 0.75 and 1 second per weld.

The range of pressure and temperature used is quiet broad, although heating and cooling are integral parts of most welding processes. The particular combination of these variables results in a unique joint in terms of material variations, potential flaws and residual stresses. Some material variations occur across a weld joint because each position in the weld is subjected to a different thermal history, with some temperatures rising above, those required for phase transformations and grain growth (Panontin et al.). Residual stresses are created during welding by the solidification, phase transformation and thermal shrinkage strains associated with molten weld metal as it cools. Existing without external loading, residual stresses are self equilibrating, and often reach magnitudes of yield level. They can cause cracking and distortion in weld joints and even premature failure of structures under certain conditions (ASM 1983).

Consequently, it is of high importance to know in detail the geometrical and the mechanical properties of the weld. Thus, the observation on the microscope of a joint's cross-section will be used to have a precise representation of the local specimen's geometry and by the way building an accurate finite element model. To show the presence of different sub-zones along the joint, a Vickers hardness test will be done along the previous cross-section.

Local geometry of the spot-welded specimen

A cross section in the middle of two spot-welded metal sheets has been done as presented in the figure 1. The figure shows the presence of a curvature on the upper and lower surface of the spot-weld, which results from the pressure applied by the electrodes to the two material sheets to hold them together before starting the welding process. The lateral notches are very important from a mechanical point of view since they will create stress concentrations when submitted to a load.

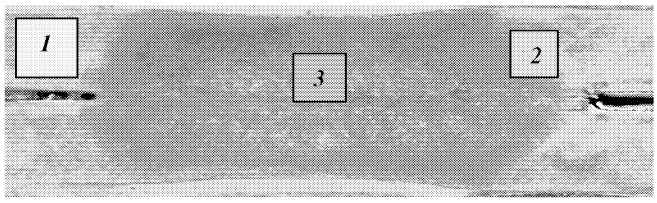


Figure 1: Spot-weld's cross-section – H320LA 1.50 mm / H320LA 1.50 mm

From an optical point of view, it is possible to recognize three phases along the specimen: the spot-weld in the middle (area 3), the basis material on the edge (area 1) and in between the heat affected zone (area 2). Since this seems to be different sub-zones, they should have different mechanical properties. Consequently, along this cross section a Vickers hardness test will be done in order to show physically the existence of these zones.

Vickers hardness measurement

The Vickers hardness test method consists of intending the tested sample with a diamond intender, in form of right pyramid with a square base and an angle of 136 degrees between opposite faces subjected to a given load for a predefined duration. The two diagonals of the indentation left in the surface of the material after removal of the load are measured using a microscope and their average calculated. The area of the sloping surface of the indentation is calculated. The Vickers hardness is the quotient obtained by dividing the kgf load by the square area of indentation. The distance between two prints or the one between the print and the extremity of the metal sheet should be at least equal to 2.5 times the print's diagonal in order to eliminate any boundary effect (Barralis and Maeder 1997).

In order to build in the next step a fine and accurate finite element model of the different specimens, a very detailed hardness measurement will be done to capture the change in hardness along the joint. In fact, the distance between two prints is equal to 0.06 mm.

In the figure 2, the Vickers hardness is plotted for 41 measurement's points distributed along the joint; it shows clearly the existence of three different sub-zones: spot-weld, heat affected zone (HAZ) and basis material as supposed previously from the cross-section. The basis material and the spot weld can be characterized as homogenous area since the hardness seems to be almost constant along the area. At the opposite, the heat affect zone can be characterized as highly heterogeneous since the Vickers hardness changes drastically in a short distance. Despite as first approximation, for the finite-element model, a constant hardness value will be considered in each sub-area, which means that the heat affected zone will be considered as homogenous.

Consequently, using the hardness measurement, it is possible to come with the following mean hardness' values for each sub-zone considering an assembly H320LA/H320LA with a symmetrical sheet thickness of 1.50 mm.

- Mean hardness value for the spot-weld: 386 HV
- Mean hardness value for the HAZ: 269 HV
- Mean hardness value for the basis material: 195 HV

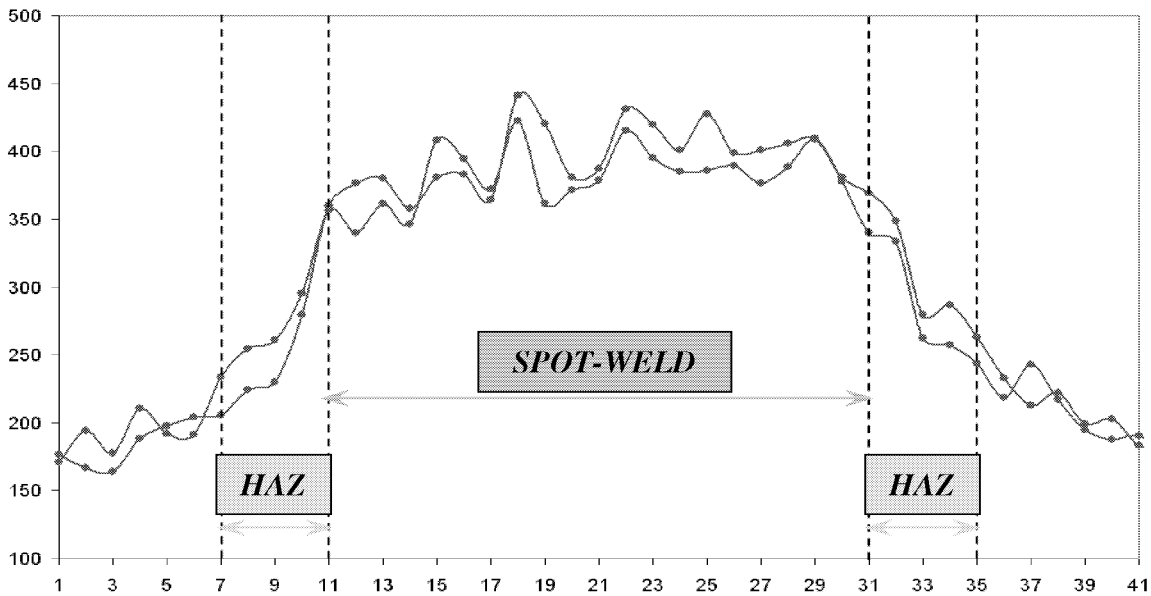


Figure 2: Hardness diagram along the spot-weld - H320LA 1.50 mm / H320LA 1.50 mm

By correlating the results from the cross section's observation under microscope and the results from the hardness measurement, it is possible to determine accurately the width of the different sub-zone as basis for the finite element model's construction. For the studied steel and the given thickness combination, it leads to:

- Diameter of the spot-weld: 5.60 mm
- Width of the heat-affected zone: 1.00 mm
- Gap between the upper and the lower sheet: 0.16 mm

Mechanical characterization of the different weld's sub-zones

As mechanical input for the numerical simulation, the strain/stress curves should be defined for each sub-zone of the specimen. The basis material's curve can be obtained simply from a tensile test. At the opposite, any direct experimental method is available so far to determine simply the stress/strain curve of the heat affected zone and of the spot-weld. Consequently, a scaling method has been used to determine the characteristics of these two zones.
In fact, according to the norm DIN 50150 (Deutsches Institut für Normen 2000), it is possible, for low alloyed steels, to convert the value of the Vickers hardness into an "equivalent" tensile strength value. Then, by dividing the "equivalent" tensile strength of the heat affected zone (respectively of the spot-weld) by the one of the basis material, a scaling factor is calculated. From a theoretical point of view, the heat brought during the welding process induced not only an increase of the tensile strength for the heat affected zone and the basis material but also a decrease of the strain at rupture for the same zones. To evaluate the strain at rupture of the heat affected zone and of the spot-weld, a graphical method based on a curve from the company Thyssen will be used. According to the figure 3, for a given tensile strength's value, it is possible to determine the corresponding strain at rupture.

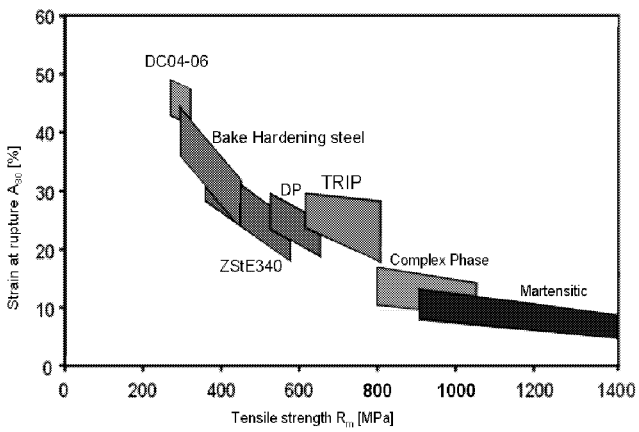


Figure 3: Tensile strength versus strain at rupture for different steels' families

Consequently, using the described method and the mean hardness value of the different sub-zones it leads to:

- Tensile strength value for the spot-weld: 1241 MPa
- Tensile strength value for the HAZ: 862 MPa
- Tensile strength value for the basis material: 625 MPa
- Strain at rupture for the spot-weld: 10,5 %
- Strain at rupture for the HAZ: 15,3 %
- Strain at rupture for the basis material: 24,6 %

In case of dynamical loading, it is important to consider the strain rate effect which describes the evolution of the mechanical properties with the increase of the loading velocity.

In this study, three strain rates will be considered ($4*10^{-3}$ 1/s, 1 1/s and 250 1/s) which leads to the implementation of nine strain/stress curves to describe entirely the problematic. For the basis material, the influence of the strain rate effect will be investigated by doing several tensile tests for the given loading velocities. These experiments show that the strain rate effect is similar to a shift of the overall stress level, which means that it can be approximated applying a given scaling factor on each stress value of the quasi-static curve.
Consequently, to take into account the strain rate effect in the study, for each sub-zone, the stress' values of the quasi-static stress/strain curve (strain rate: $4*10^{-3}$ 1/s) will be multiplied by 1.169 (respectively by 1.375) to obtain the curve equivalent to the strain rate 1 1/s (respectively to the strain rate 250 1/s).

Construction of the different numerical models

The capability of the proposed method to represent accurately the reality will be tested on the spot-welded "Tensile/Shear" and "Coach-Peel" specimens (Figure 4).

As represented on the figure 5, the "Tensile/Shear" specimen allows testing the spot-weld according to four different load angles: 0° for a shearing solicitation, 90° for a tensile solicitation and 30° or 60° for combined load cases.

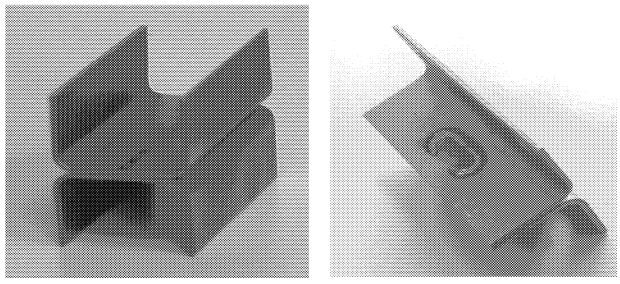


Figure 4: "Tensile/Shear" and "Coach-Peel" specimens

The "Coach-Peel" specimen is used to characterize the bending behaviour of the spot-weld. As a matter of fact, using these two specimens, it is possible to describe in a complete manner the weld behaviour for a given assembly.

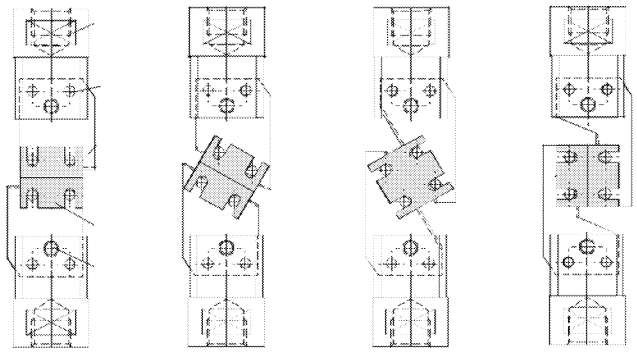


Figure 5: *Experimental load angles for the "Tensile/Shear" specimen (90°, 60°, 30° and 0°)*

The numerical models of the different specimens (figure 6) have been constructed and meshed with the CAD software Medina using only volume elements. In order to avoid numerical instabilities and important modelling influences on the calculus, each sub-area of the specimen (basis material, heat affected zone and spot-weld) has been modelled with elements having approximately a constant characteristic size. Thus, the smallest elements have a length of 0.20 mm while the largest ones have a length of 1 mm. Of course, the coarser mesh is located at the border of the specimen, far away from the surrounding of the heat affected area, where the fracture will probably occurs. As a matter of fact due to this high discretization, the numerical model of the "Tensile/Shear" specimen for the material gage 1.50mm/1.50mm contains approximately 320 000 volume elements.

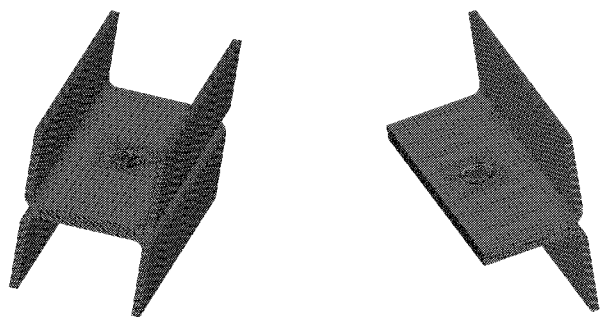


Figure 6: Finite element models of the "Tensile/Shear" and "Coach-Peel" specimens

The numerical simulation will be performed using the Gurson Tvergaard-Needleman material model, classified as *MAT_120 in LS-DYNA3D [4]. As theoretical basis, the phenomenon leading to the ductile fracture as well as the theoretical model from Gurson Tvergaard-Needleman will be presented in the following paragraph.

MICRO-MECHANISM OF DUCTILE FRACTURE INITIATION AND MODELS OF DAMAGE

Basically, the ductile fracture's phenomenon is initiated by void formation around non-metallic inclusions and second-phase particles in metal matrix that is subjected to plastic strain under influence of external loading. The size of the second phase particles and non-metallic inclusions in engineering alloys may range from approx. 0.01 µm to values that by far exceed 1 µm (Thomason 1990). Their shape varies from spherical to lamellar or even irregular form. Depending on the size, shape and quantity of these particles as possible spots for initiation of ductile fracture, numerous models have been developed in an effort to describe these complex micro-mechanisms. Some models evaluate a critical stress, others use a critical strain. Both type of criteria are based on the fact that a critical stress (respectively strain) at the interface of an inclusion or in the centre of an inclusion must be exceeded to cause debonding or cracking of the particle (Achenbach et al. 1998).

In 1977, Gurson (Gurson 1977) proposed for porous materials a theoretical model in which the plasticity criteria and the plastic potential depends on the macroscopic stress and on the void volume fraction of the material's cavities (i.e. the porosity). After that, numerous studies have shown that the original Gurson's material model was able to describe properly the initial stage of the damage but it does not permit to represent accurately the material's behaviour during the coalescence and the rupture phases. In fact, the initial Gurson's formulation does not take into account the interaction's effects between the micro-cavities because the theory is based on the behaviour of one representative volume element (RVE) containing only one cavity loaded under hydrostatic stresses. Thus, using these experimental results, Tvergaard and Needleman (Needleman and Tvergaard 1984) modified the original Gurson's plastic flow to obtain a better fit between the experimental and the simulation results.

According to this modified Gurson-Tvergaard-Needleman (GTN) model, the plastic potential is given by:

$$\phi(\sigma_{kk},\sigma_M,f)=\left(\frac{\sigma_{Eq}}{\sigma_M}\right)^2+2.q_1.f^*.Cosh\left(\frac{3.q_2.\sigma_H}{2.\sigma_M}\right)-1-(q_1.f^*)^2=0 \quad (1)$$

Where σ_M denotes the actual flow stress of the material's matrix, σ_{Eq} is the von Mises equivalent macroscopic stress calculated from deviatoric part of the Cauchy's stress tensor, σ_H is the hydrostatic part of the macroscopic stress tensor and q_1, q_2 are correlating parameters introduced by Tvergaard and Needleman to improve the ductile fracture prediction's quality of the initial Gurson's model. It is of

interest to note that in case of the absence of porosity ($f^*=0$), the GTN plastic potential is reduced to the Von Mises' yield surface.

The function f^* was added by Tvergaard and Needleman to account for the accelerating damage rate associated with void coalescence, such that:

$$f^*(f, f_c, f_F, q_1) = \begin{cases} f & \text{if } f \le f_c \\ f_c + \dfrac{\dfrac{1}{q_1} - f_c}{f_F - f_c} \cdot (f - f_c) & \text{if } f > f_c \end{cases} \qquad (2)$$

Where f is the "current" void volume fraction, f_c is the critical value at which void coalescence starts and f_F is the void volume fraction at which there is a complete loss of stress carrying capacity.

The void volume fraction growth rate can be obtained by adding the effects of the nucleation of new voids as a function of effective plastic strain with the growth of existing voids according to:

$$\dot{f} = \dot{f}_{nucleation} + \dot{f}_{growth} \qquad (3)$$

With:

$$\dot{f}_{nucleation} = D.\dot{\varepsilon}_{eq}^{PL} \quad - \quad \dot{f}_{growth} = (1-f).tr\left(\dot{\varepsilon}^{PL}\right)$$

And $$D = \frac{f_N}{S_N.\sqrt{2.\pi}}.\exp\left[-\frac{1}{2}.\left(\frac{\varepsilon_{eff}^{PL} - \varepsilon_N}{S_N}\right)^2\right]$$

Where f_N is the volume fraction of void-nucleating particles, $\dot{\varepsilon}_{eq}^{PL}$ represents the equivalent plastic strain rate, ε_{eff}^{PL} represents the effective plastic strain, S_N represents the standard deviation of nucleating strains assuming a normal distribution, ε_N represents the mean strain value at which 50% of the second phase particles ruptured.

DETERMINATION OF SPECIFIC MATERIAL MODEL'S PARAMETERS

In the GTN material model, the parameters ε_N and S_N are more or less constant for a given material's family. For steels, in the literature the standard values for ε_N and S_N are respectively 0.3 and 0.1. At the opposite the following parameters: f_0, f_N, f_c, f_f, q_1 and q_2 are material's specific. As a matter of fact, these damage parameters should be identified and optimised for the studied steel. Their identifications will be done manually based on a numerical inverse method. The goal of the procedure is to fit, using the finite element simulation, the global shape (global stiffness, maximal stress level), the accumulation of damage (loss of stress carrying capacity) and the strain at rupture of both experimental and numerical curves.

For the version of LS-DYNA3D used in this study, it is of importance to note that the parameter f_f varies with the element's size of the finite element mesh (Feucht and Fassnacht 1999). As a matter of fact, four tensile specimen's numerical models have been meshed using volume elements respectively with a characteristic length of 0.25 mm, 0.33 mm, 0.50 mm and 1.00 mm. The graphical result of the parameter's optimization (and respectively the associated numerical value of the parameters) is presented for the studied material combination (H320LA 1.50 mm / H320LA 1.50 mm) on the figure 7 (respectively. in the table 1) where the black curve is the experimental one and the colored curves are the simulated one for the four different element sizes.

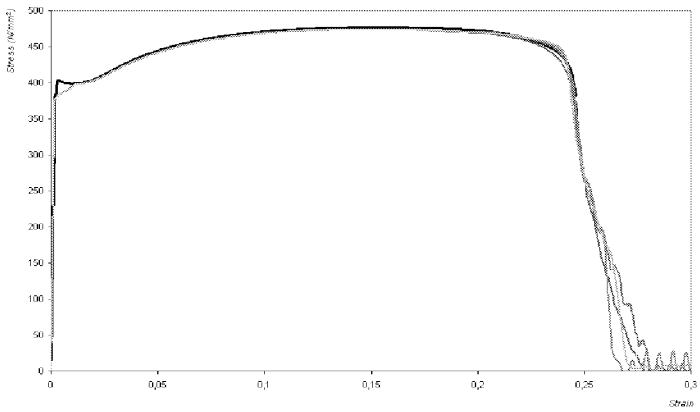


Figure 7: Optimized Gurson's material parameters based on an inverse numerical method using tensile specimens – Basis material H320LA

Table 1: Value of the optimized Gurson's parameters – Basis material H320LA

f_0	f_c	F_f *(0.25mm)* *(0.33mm)* *(0.50mm)* *(1.00mm)*	f_N	q_1	q_2
0.0030	0.0042	0.277 0.273 0.265 0.250	0.0075	1.5000	1.0000

The previous curves are in perfect accordance not only for the damage evolution but also for the strain at rupture of the tensile specimen between the simulated and the experimental curves. Consequently for the studied material, the damage parameters are considered as optimized for the basis material. Of course, the same procedure has been followed to calibrate the Gurson's parameters for the heat affected zone and for the spot-weld itself.

DESCRIPTION OF THE EXPERIMENTAL SET-UP FOR DYNAMICAL LOAD CASES

The experimental set-up used for the quasi-static loading has been reinforced in order to sustain to velocities between 2 m/s and 5 m/s. Moreover, the most of the clearance between the different sub-parts of the fixations has been eliminated to reduce the propagation of oscillatory phenomenon leading to unstable experimental results.

Due to the large number of sub-systems between the specimen and the tensile machine's cross beam, it would have been very approximate for the results' accuracy to measure the specimen's displacement in a global way (i.e. directly on the machine's cross beam). In fact, it seems necessary to measure the specimen's displacement, in a local way, as close as possible from the weld itself. Consequently, it has been decided to use the GOM optical measurement protocol. To use it, the tested specimen must be first covered with a black/white pattern. It consists of applying a white paint's coat called "priming coat" on which a black paint's coat, named "contrast coat", has to be superimposed as shown on the figure 8.

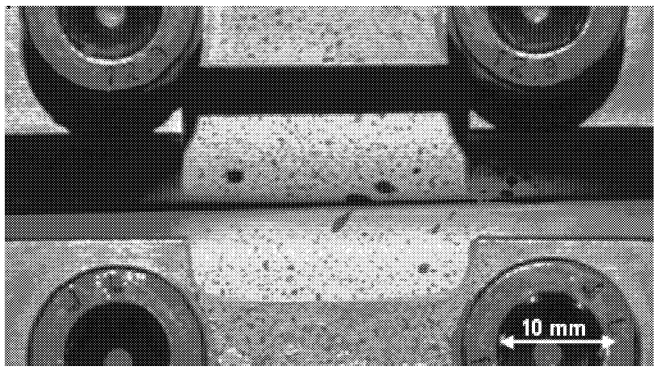


Figure 8: Side view of a "Tensile/Shear" specimen with its associated pattern required for GOM measurements

Then, on the side of the "coated" specimen, two measurement's points, one on the upper part and one on the lower part of the specimen, are marked. Finally, the tensile machine with its specimen is surrounded by a high speed camera and two very high intensity's lamps needed for the correct operation of the camera. Concerning the data's acquisition, a trigger ensures the coupling between the machine, which delivers the force's evolution versus the time and the camera which records optically the path of the two measurement's points previously marked on the specimen's side. This recording is then exploited using the software ARAMIS in order to evaluate the relative displacement between the two points versus the time. After synchronizing temporally the two previous curves, the final force versus displacement curve measured locally is obtained. In this study, all experimental curves presented in the following paragraph (black curves) have been obtained using this GOM measurement protocol. In order to secure the experimental results, for each load case, a first test will be done to calibrate properly the different equipments and then as "real" experimental basis for the detailed model's validation five specimens will be crashed.

COMPARISON BETWEEN EXPERIMENTAL AND SIMULATED FORCE/DISPLACEMENT CURVES

As mentioned previously, in order to validate the ability of the detailed model to represent with accuracy the reality, two different specimens ("Tensile/Shear" and "Coach-Peel") with their associated load cases have been numerically simulated for the material combination H320LA 1.50mm / H320LA 1.50mm . As validation basis, the locally measured experimental force/displacement curve will be compared with the simulated one. For each load case, a first simulation with an elastic/plastic material model (*MAT_24) will be carried out. Then to compute the accumulation of damage and the specimen's rupture, the Gurson-Tvergaard-Needleman material model with the previous optimized damage parameters and with consideration of the strain rate effect will be used. The numerical simulations have been performed using the software LS-DYNA3D on a cluster which allows us modelling the whole specimen without any symmetry.

The comparison between the simulation and the experimental curves are plotted for the "Tensile/Shear" specimen 90° on figure 9, for the 60° on figure 10, for the 0° on figure 11 and for the "Coach-Peel" specimen on figure 12. In these figures, the black curves represent the experimental joint's characteristic, the blue curve represents the simulation's result delivered by the elastic-plastic material model and the green curve represent the simulation's results using the GTN material model (i.e. with damage and failure's computation).

Firstly, it is of interest to notice that the "Tensile/Shear" specimen loaded under 0° withstand to a higher force than the same specimen loaded respectively under 60° or 90°. This means that it is more difficult and it requires more energy to break a spot-weld by pure shearing impacts than by pure tensile impacts.

As illustrated through the figures 9 to 12, despite the local measurement method of the specimen's force/displacement characteristics, numerous oscillations can be observed in

the experiments especially for 60° and 0°. Numerical investigations carried out in parallel have shown that the top and the bottom of the peaks are not relevant as comparison basis for the simulation. In fact, only the mean value between the top and the bottom of the peaks should be considered. From a macroscopic point of view, the simulations and the experiments are in good correlation for the studied load cases and material combination. In fact, the simulated global stiffness is in accordance with the measured one. Moreover, the simulated maximum force level that the spot-weld can withstand and its associated displacement are in good accordance with the experiments.

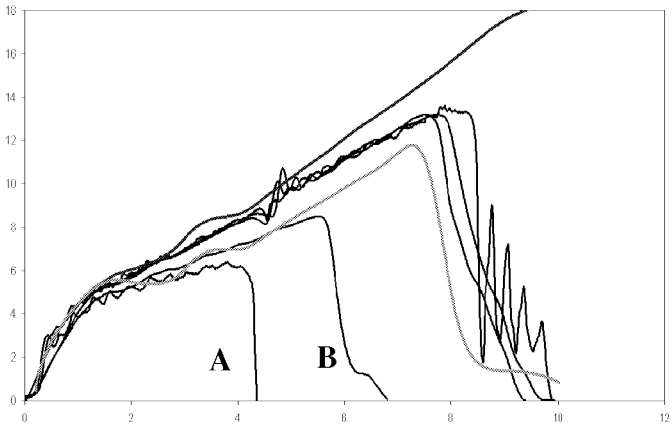


Figure 9: Comparison of the force versus displacement curve between experiments (black) and simulations (blue and green) – H320LA 1.50 mm / H320LA 1.50 mm – "Tensile/Shear" 90°

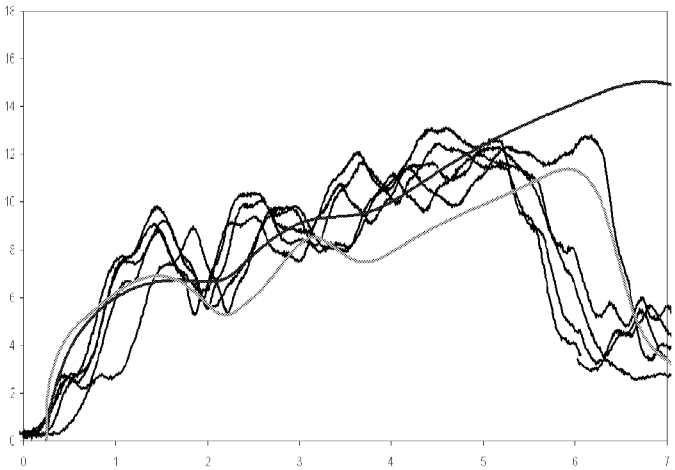


Figure 10: Comparison of the force versus displacement curve between experiments (black) and simulations (blue and green) - H320LA 1.50 mm / H320LA 1.50 mm – "Tensile/Shear" 60°

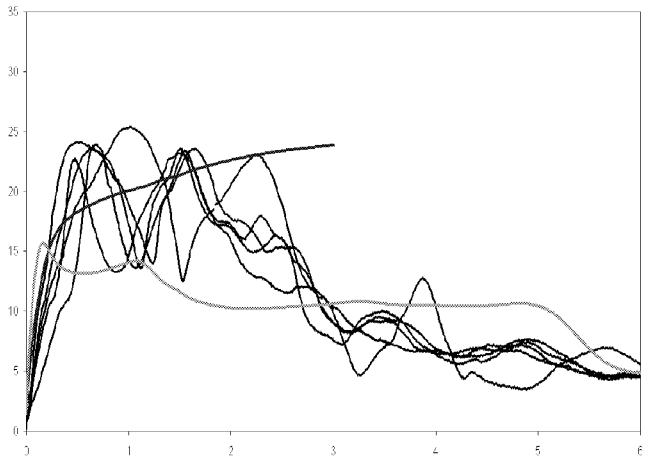


Figure 11: Comparison of the force versus displacement curve between experiments (black) and simulations (blue and green) - H320LA 1.50 mm / H320LA 1.50 mm – "Tensile/Shear" 0°

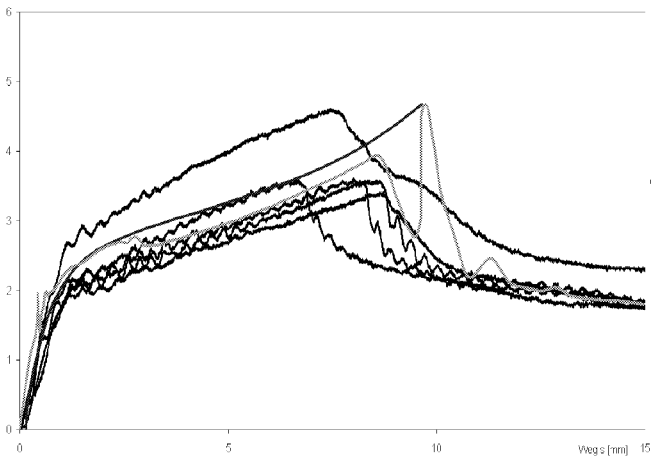


Figure 12: Comparison of the force versus displacement curve between experiments (black) and simulations (blue and green) - H320LA 1.50 mm / H320LA 1.50 mm – "Coach-Peel" 90°

For all studied load cases, the failure occurs via a spot-weld's pull-out in both experiments and simulations as shown in the figure 13. On this figure, it can be seen that the numerical detailed simulation represents properly the local rupture's modus of the specimen.

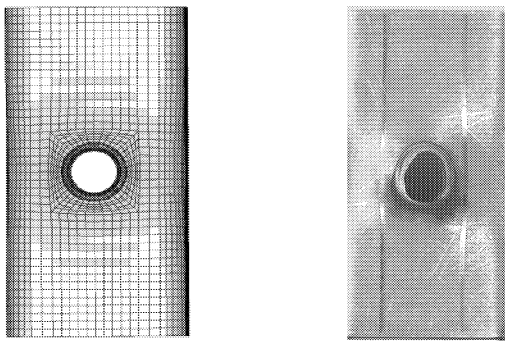


Figure 13: Comparison between the simulated (links) and the experimental (right) rupture modus of the specimen

For the "Tensile/Shear" specimen loaded under 90°, an earlier rupture may be observed for two specimens (figure 9 - curves A and B) in the experiments. By studying one of the crashed specimens, it can be seen that only a part of the spot-weld rupture (figure 14) while the surrounding of the spot-weld especially at the interface between the heat affect area and the weld itself remains intact. Such rupture modus is due to a poor spot-weld's quality which is often due to the use of non-optimized welding parameters for the welded material (respectively material's thickness)

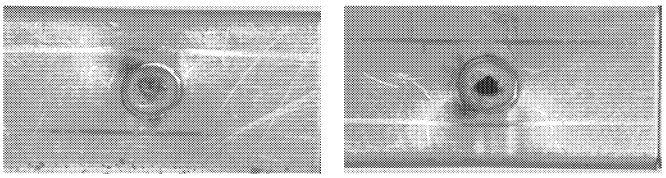


Figure 14: Tested "Tensile/Shear -90°" specimen presenting non-optimized welding parameters- Specimen A

CONCLUSIONS

Using the Vickers hardness measurement and a specimen's cross section, the mechanical properties of the three different sub-zones existing along the spot-welded specimen (basis material, heat affected zone and spot-weld) have been identified and characterized. Using a scaling method based on the estimation of the local tensile strength from the Vickers hardness associated to a graphical determination of the strain at rupture's values, the material properties have been obtained for each sub-zone of the spot-weld. The local geometry of the spot-weld itself (curvature, notches, radius ...) has been obtained using the previous cross-section. With these data different finite element models have been built using eight nodes volumic elements and a very fine mesh.

To analyze the spot weld's rupture the Gurson material model has been used. In order to compute properly the damage's accumulation and the failure, the Gurson's parameters have been first optimized using an inverse method based on a force/displacement curve fitting from a tensile specimen. With the optimized parameters, more complex specimens ("Tensile/Shear" and "Coach-Peel") have been simulated using LS-DYNA3D.

The simulated force vs. displacement curves are in good correlation with the experimental ones in term of global stiffness, maximal stress carrying capacity and rupture description. As a matter of fact, the detailed model of spot-welded specimen is validated for the studied steel (H320LA) for the symmetrical combination 1.50 mm – 1.50 mm.

Finally, in combination with small experimental campaigns the validated spot-weld detailed model may be used as input data for developing a simplified spot-weld's model. Such a model, suitable to describe the spot-weld's failure in car's body in white, was developed in 2005 by F. Seeger et al. The methodology requires the associated force vs. displacement curve of the "Tensile/Shear" and "Coach-Peel" specimens for each material or gage combination. With the presented detailed model the number of experiments and consequently the costs can be considerable reduced.

REFERENCES

Achenbach, U.; S. Klingbeil; U. Prahl; W. Dahl; and W. Bleck. 1998. "Fracture from defects". In Proceedings of ECF 12, Editors E.R. De la Rios, K.J. Miller, M. W. Brown, EMAS Publishing, Sheffield, Volume 2, pp. 877-881.

ASM Handbook. 1983. "Welding, Brazing and Soldering". Volume 6.

Barralis, J. and G. Maeder. 1997. "Précis de métallurgie, élaboration, structure, propriétés et normalisation", AFNOR Nathan.

Deutsches Institut für Normen e.V. 2000. "Prüfung metallischer Werkstoffe Umwertung von Härtewerten – DIN 50150".

Feucht, M. and W. Fassnacht. 1999. "Simulation der Rissbildung in Aluminiumgussbauteilen". DaimlerChrysler AG. Internal Report.

Gurson, A.L. 1977. "Continuum Theory of Ductile Rupture by Void Nucleation and Growth: Part 1-Yield Criteria and Flow Rules for Porous Media". Journal of Engineering Materials and Technology, N°99, pp. 2-15.

Markiewicz, E.; T. Fourmentraux; P. Ducrocq; P. Drazetic; G. Haugou; and J. Y. Berard. 2001. "Material behavior law identification for the various zones of the spot-weld under quasi-static loadings". International Journal of Materials and Product Technology. Volume 16, N° 6-7, pp. 484-509.

Needleman, A. and V. Tvergaard. 1984. "An analysis of ductile rupture in notched bars". Journal of Mechanics and Physics of solids, N°32, pp. 461-490.

Panontin, T.L.; O. Nishioka; and M. R. Hill. "Fracture Assessments of Welded Structures", Internet Publication.

Seeger, F.; M. Feucht; T. Frank; B. Keding; and A. Haufe. 2005. "An investigation on spot-weld modelling for crash simulation with LS-DYNA", LS-DYNA User's Forum 2005, Bamberg, Germany.

Sommer, S. and D. Z. Sun. 2003. "Untersuchung des Einsatzes von Schädigungsmodellen bei Punktschweißverbindungen unter Crashbelastung", Fraunhofer IWM.

Thomason, P.F. 1990. "Ductile fracture of metals", Pergamon Press, Oxford.

AUTHOR BIOGRAPHY

EMMANUEL H.J. LAMOUROUX was born in Vichy (France) in 1981. In 2001, he went to the engineering school of Valenciennes (France) named ENSIAME (École Nationale Supérieure d'Ingénieurs en Informatique, Automatique, Mécanique et Énergétique), where he studied mechanical and structural engineering. He studied also one semester at the Chalmers University of Gothenburg (Sweden) attending the lectures of the Master of Automotive Engineering. Since 2004, he works as PhD student in the research department of DaimlerChrysler AG in Stuttgart (Germany). His task focuses on the detailed numerical modeling of weld's damage and failure under dynamical load cases. Parallel to that, he works as independent expert for the SINTEF (Norwegian Scientific Foundation) in Trondheim, where he takes part of the development of a new technology of rapid prototyping and tooling called "Metal Printing Process".

GRINDING AND MILLING SIMULATION

A Study on the Development of High Precision Cam Profile CNC Grinding Machine with Maximum Grinding Speed to obtain Constant Surface Roughness

Younghwa Roh
Sangheon Lim
Sunghwan Ahn
Choonman Lee
Jongyun Jung
Mechanical Design and Manufacturing Department
Changwon National University
53203, #9 Sarim-dong, Changwon, Kyeongnam
South Korea
E-mail : cmlee@sarim.changwon.ac.kr

Sangdae Yoon
Fine Ace Technology Co. Ltd
#853-10 Woe-dong, Changwon, Kyeongnam
South Korea

KEYWORDS
Cam, Constant surface roughness, Maximum grinding speed, CNC grinding machine.

ABSTRACT

Cam mechanisms are one of the most popular devices for generating irregular motion and are widely used in many automatic equipments, such as textile machinery, internal combustion engines and other automatic devices. In order to obtain the positive motion of follower by rotating cam, its shape should be correctly designed and manufactured. The manufacture of high precision cam is necessary to develop a cam profile CNC grinding machine with maximum grinding speed to obtain constant surface roughness.

The purpose of this study is manufacturing of a cam with constant surface roughness established from general grinding theory and calculates the maximum grinding speed by the algorithm which it derived. And three types of disk cam are successfully manufactured by using developed machine with proposed algorithm.

INTRODUCTION

A machine is consist of various elements of machine. A cam as a mechanical component of a machine is in contact with follower when the cam is rotating and the follower carry out various motion. So, cams are widely used in various automation machinery. Especially, a disk-typed cam produce irregular positive motion with component which have simple mechanism.

Recently, demands on precision manufacturing of machine elements are increased by the development of high-technology products and an advancement of industry. According to the trend of high precision manufacturing for machine elements like this, a grinding process has been developed with fast speed. A grinding process is substituted by a machining of materials with high speed and high precision. However, the cam is required by a high precision product, the surface roughness of sub-micrometer unit(10-7m), and high productivity(H.S. Kim et al. 1995).

In order to obtain smooth motion of a follower by the cams rotation, cam profile should be designed accurately and process of precise machining is essential to obtain a high precision cam. Although the cam is perfectly designed, the process of machining is not performed in the machining step, the desired cam will be not obtained. Gao(Y. Gao and J.A Webster. 1991) proposed a prediction model of cam shape with considering deformation and vibration of the workpiece. A CAD/CAM system is developed by the coauthors of this paper for shape design of disk cams using relative velocity method and NC code generate apply the biarc curve interpolation(S.H. Lim et al. 2006).

Even though many studies have been carried out to suggest various techniques, cam grinding most relies on experience. As a shape of cam profile is completely not a circle, the cutting force of a workpiece is continuously changed by grinding wheel.

In order to obtain maximum grinding speed with constant surface roughness, an algorithm of calculation of the maximum grinding speed is proposed from grinding theory. And three types of disk cam are successfully manufactured by using developed machine with proposed algorithm.

MAXIMUM GRINDING SPEED CALCULATION ALGORITHM

Cutting height of cutting edge

As a result of studies on surface roughness through a grinding theory, it has been thought that the process to calculate a grinding volume should be considered above all by using the sectional area of grinding chips. The volume could be derived from the common equation to have a sectional area of grinding chips(D.W. Kim. 1991), but the equation is come from a hypothesis the cutting edge of a grinding wheel does not have shedding and regenerating, so this study carried out a modeling for the cutting edge operated in actual grinding by the kinematics of machines in general grinding theories.

In addition, equations on surface roughness were formulated through relations with the subsequent grain and the equation for the maximum grinding speed could be achieved for having a constant surface roughness from the correlation between grinding volume and surface roughness equations.

First, through the kinematics in general grinding theories, it can be considered that grinding is done not only by the

cutting edges on the side of the circumference at the most external angle of a grinding wheel but also by some cutting edges inside the wheel. In such a situation, the center for the rotation of the grinding wheel is O and the location of a random cutting edge is (ρ,θ) represented as polar coordinates, and then the cutting height of that cutting edge can be calculated as the following equation.

In this study, modeling about grains for actual grinding is carried out by using grinding theory, and a equation of surface roughness is defined from references related to a subsequent grain(C.W. Park. 2002 and D.W. Kim. 1991).

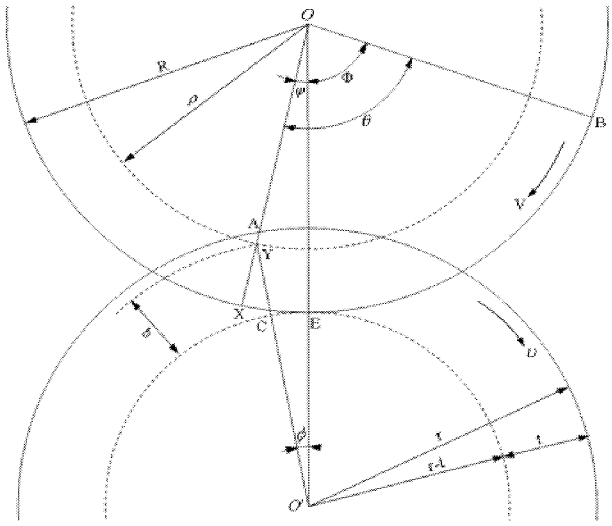


Figure 1 Schematic diagram of a flange grinding mechanism

Fig. 1 shows a condition of the moment a random cutting edge $Y(\rho,\theta)$ passes through the part of AC. That is, it assumes that in short time, the point C on a workpiece coincides with the cutting edge B on the outermost surface of a grinding wheel at the point E. Fig. 1 can be formulated as follows;

$$\angle YO'O = \phi \qquad \angle YOB = \varphi + \Phi = \theta$$
$$\overline{OB} = R$$
$$\angle YOO' = \varphi \qquad \overline{OY} = \overline{OX} - \overline{XY} = \rho$$
$$O'A \approx r$$
$$\angle O'OB = \Phi \qquad \overline{YC} = h$$

From above, the cutting height h is calculated as follows;

$$h=(R+r-t)\cos\phi-(r-t)-\sqrt{\rho^2+(R+r-t)^2}\sin\phi \quad (1)$$

It can be deduced from above equations there is an geometric relation between ϕ and θ as follows;

$$\theta = \frac{\Omega}{\omega}\phi - \phi + \sin^{-1}\left\{\frac{(R+r-t)}{\rho}\sin\phi\right\} \quad (2)$$

Where Ω is the angle velocity of the grinding wheel and ω is that of the workpiece. From equations of (1) and (2) above, the cutting height h of the cutting edge is formulated as a function of the coordinates (ρ,θ) of the cutting edge. But, the depth of notch t is significantly small comparing with R, r, ρ, so ϕ^2 can be neglected comparing to 1. Under such a condition, the equations of (1) and (2) are re-formulated as the following equation (3) and (4).

$$h = R - \rho + \frac{1}{2}\cdot\frac{r}{R}(R+r)\phi^2 \quad (3)$$

$$\phi = \frac{\theta}{\left(\frac{\Omega}{\omega}+\frac{r}{R}\right)} = \frac{v}{V}\frac{R}{r}\left(1+\frac{v}{V}\right)^{-1}\theta \quad (4)$$

The equation (5) can be drawn from the equations of (3) and (4) as follows;

$$h = R - \rho + \frac{1}{2}\left(\frac{\upsilon}{V}\right)^2\left(\frac{R+r}{Rr}\right)\left(1+\frac{\upsilon}{V}\right)^{-2}\times R^2\theta^2 \quad (5)$$

In the equation (5) above, the coefficient of $R^2\theta^2$ in the 3rd term

$$\frac{1}{2}\left(\frac{\upsilon}{V}\right)^2\left(\frac{R+r}{Rr}\right)\left(1+\frac{\upsilon}{V}\right)^{-2}$$

is a value to be determined in case radiuses and velocities of the wheel and workpiece are settled, so it can be formulated for G^2 as follows;

$$G = \sqrt{\frac{1}{2}}\left(\frac{\upsilon}{V}\right)^2\sqrt{\frac{R+r}{Rr}}\left(1+\frac{\upsilon}{V}\right)^{-1}$$
$$= \left(1+\frac{\upsilon}{V}\right)^{-1}\frac{\upsilon}{V}\sqrt{\frac{1}{D}+\frac{1}{d}} \approx \frac{\upsilon}{V}\sqrt{\frac{1}{D}+\frac{1}{d}} \quad (6)$$

where G is a function to be determined by operating conditions, and for the surface grinding, $d \to \infty$ is applied and for the internal grinding, $d \to -d$ and $\left(1+\frac{v}{V}\right) \to \left(1-\frac{v}{V}\right)$ (D.W. Kim. 1991).

The equation (5) is re-defined through the equation (6) as follows;

$$h = R - \rho + (GR)^2\theta^2 \quad (7)$$

The equation (7) above shows an equation to present the cutting height of a random cutting edge $Y(\rho,\theta)$.

Subsequent grain

As a grinding surface is formed by the tool mentioned in the preceding paragraph, Fig. 2 illustrates the finishing process of a finishing surface. The hatching part, as shown in Fig. 2, is the profile of the finishing surface and other cutting lines not related with the hatching can be thought that the cutting of each location proceeds prior to the final grinding, not being related to the final grinding.

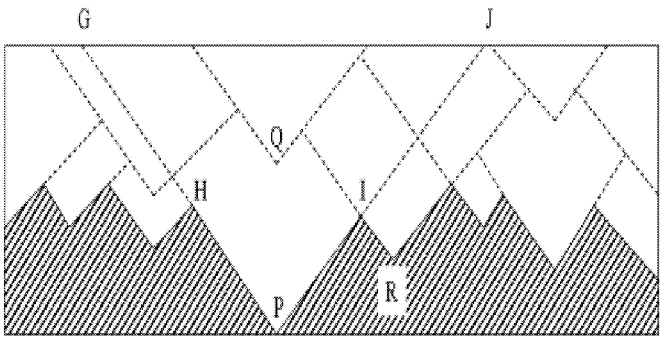


Figure 2 A forming process of a grinding surface

For 1 cutting groove, $GHPIJ$ as an example shown in the Fig. 2 (But, the lines GH and IJ are virtual cutting lines.), the summit P of the groove becomes the name of $GHPIJ$, the external line a cutting edge passes through. In case considering two cutting grooves Q and R, Q is included in P, yet R is not. The cutting edge P generating such a cutting groove P is defined as a subsequent grain for the cutting edge Q.

Those cutting edge P and Q can have coordinates of (ρ,θ) and (ρ_0,θ_0), respectively. And as shown in Fig. 3, the cutting heights of two grooves formed by two cutting edges are presented as (ρ,θ) and (ρ_0,θ_0), respectively, as their coordinates in the grinding wheel. In case h and h_0 are indicated as the cutting heights of two grooves produced by two cutting edges, the following equation (8) can be drawn from the equation (7).

$$\begin{aligned} h_0 &= R-\rho_0+G^2R^2\theta_0^2 \\ h &= R-\rho+G^2R^2\theta^2 \end{aligned} \tag{8}$$

As shown in Fig. 3, the cutting lines of Q, QM' and QN' are extended and the intersection points with the line of the cutting height h are M and N, and then ΔQMN has no cutting groove other than P.

If it is assumed that there is a cutting groove like R shown in Fig. 3, the groove R includes the groove Q and the gap of the cutting heights of R and Q is less than that of Q and P, so Q comes to be the subsequent grain for P and Q and it is against the definition of a subsequent grain mentioned previously.

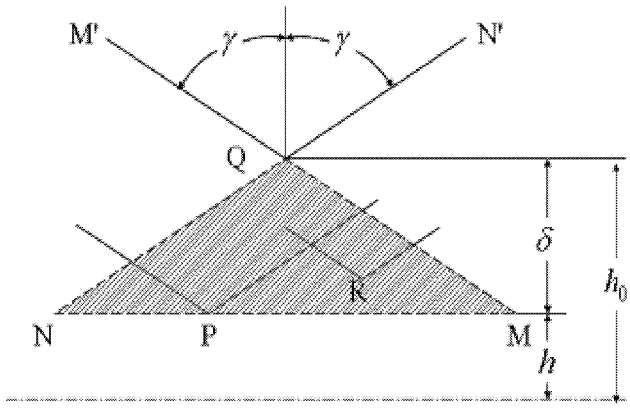


Figure 3 Schematic diagram for the relation of subsequent grain

Thus, it is examined that only one cutting edge is in the volume of the grinding wheel the area ΔQMN among sections of the workpiece.

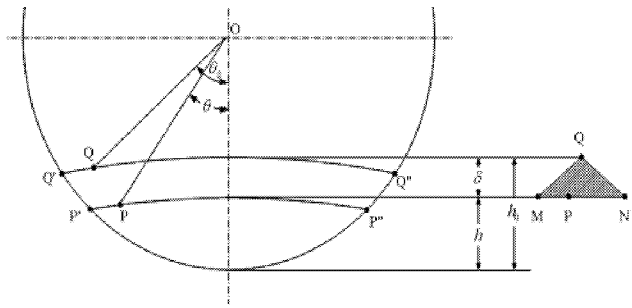


Figure 4 Schematic diagram of a grinding volume by the subsequent grain

Fig. 4 shows a grinding volume by the cutting edge Q and the subsequent grain P, and the lines of $Q'Q''$ and $P'P''$ are the cutting heights of h_0 and h, respectively. Besides, the volume u in the grinding wheel which corresponds to ΔQMN of the workpiece has a band with the sides of $Q'Q''$ and $P'P''$, and it can be considered to be the volume of a trigonal prism. Such u is formulated as follows;

$$\begin{aligned} u &= \int_h^{h_0} 4\frac{\sqrt{h}}{G}(h_0-h)\tan\gamma\, dh \\ &= \frac{8\tan\gamma}{15G}\left[2h_0^{\frac{5}{2}}-5h_0h^{\frac{3}{2}}+3h^{\frac{5}{2}}\right] \end{aligned} \tag{9}$$

The following equation (10) can be defined because the number of cutting edges in the unit volume of the wheel is $1/v^3$.

$$\frac{8\tan\gamma}{15Gv^3}\left[2h_0^{\frac{5}{2}}-5h_0h^{\frac{3}{2}}+3h^{\frac{5}{2}}\right]=1 \tag{10}$$

The grinding surface is finished by several cutting edges having a cutting height close to 0. There remain only cutting grooves by cutting edges having no subsequent grains in grinding surfaces based on the fact that factors with subsequent grains disappears completely as stated above.

Thus, in case the heights of two grooves existing in a grinding surface are h_{v_1} and h_{v_2}, respectively, the left side of the equation (10) can be set to be less than 1.

$$\frac{8\tan\gamma}{15Gv^3}\left[2h_{v_1}^{\frac{5}{2}} - 5h_{v_1}h_{v_2}^{\frac{3}{2}} + 3h_{v_2}^{\frac{5}{2}}\right] < 1 \qquad (11)$$

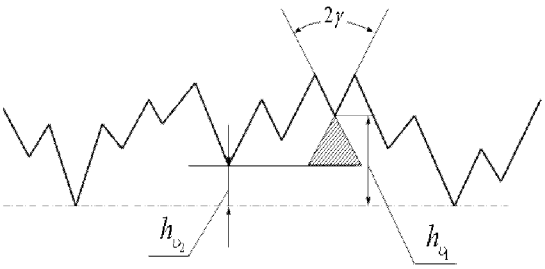


Figure 5 An interaction of random groove between h_{v_1} and h_{v_2} on the shape of the grinding surface

When having h_{v_1}, h_{v_2} $(h_{v_1} \geq h_{v_2})$ as the heights of two grooves, the triangle with a contact point of h_{v_1}, a height of $(h_{v_1} - h_{v_2})$ and a settled angle of 2γ has no cutting grooves as presented in Fig. 5, and it is two grooves for which h_{v_2} and h_{v_1} are carved and the following equation (10) can be formulated through determining the highest point $(h_{v_1} = H_v)$ and the lowest point $(h_{v_2} = 0)$.

$$\frac{16\tan\gamma}{15Gv^3}H_{\upsilon}^{\frac{5}{2}} = 1 \qquad (12)$$

$$H_{\upsilon} = \left(\frac{15}{16}\right)^{\frac{2}{5}} v^{\frac{6}{5}}(\cot\gamma)^{\frac{2}{5}}G^{\frac{2}{5}} \qquad (13)$$

where G can be defined as $G = \frac{v}{V}\sqrt{\frac{1}{D} \pm \frac{1}{d}}$, drawn from the equation (6) and the equation above is re-defined as follows;

$$H_{\upsilon} = 0.97v^{\frac{6}{5}}(\cot\gamma)^{\frac{2}{5}}\left\{\frac{\upsilon}{V}\sqrt{\frac{1}{D} \pm \frac{1}{d}}\right\}^{\frac{2}{5}} \qquad (14)$$

Where V, velocity of grinding wheel(m/s); D, diameter of grinding wheel(mm); γ, end cutting edge angle of grain(°); v, number of grains; d, diameter of workpiece(mm).

Cutting Height of grain(H_v) is changed into surface roughness and center line-average-height is used in this paper. And H_{cp} is 0.256 H_v (C.W. Park. 2002) in accordance with the correlation between H_{cp} and H_v. The equation (14) can be rewritten as;

$$\frac{h_{cp}}{0.256} = 0.97v^{\frac{6}{5}}(\cot\gamma)^{\frac{2}{5}}\left\{\frac{v}{V}\sqrt{\frac{1}{D} \pm \frac{1}{d}}\right\}^{\frac{2}{5}} \qquad (15)$$

From equation (15), the maximum grinding speed to obtain constant surface roughness is proposed as ;

$$v^{\frac{2}{5}} = \frac{h_{cp}}{0.24832v^{\frac{5}{6}}(\cot\gamma)^{\frac{2}{5}}} \times \frac{1}{\left(\sqrt{\frac{1}{D} \pm \frac{1}{d}}\right)^{\frac{2}{5}}} \times V^{\frac{2}{5}} \qquad (16)$$

EXPERIMENTS AND RESULT

A NC-code Generated by the developed CAM system. A created dialog box for the cam machining is shown as Fig. 6. There is the source selection radio button on the top of dialog box. And It can be selected 「From Design Data」 as shown in Fig. 6 for NC code generation of Palanquin cam button.

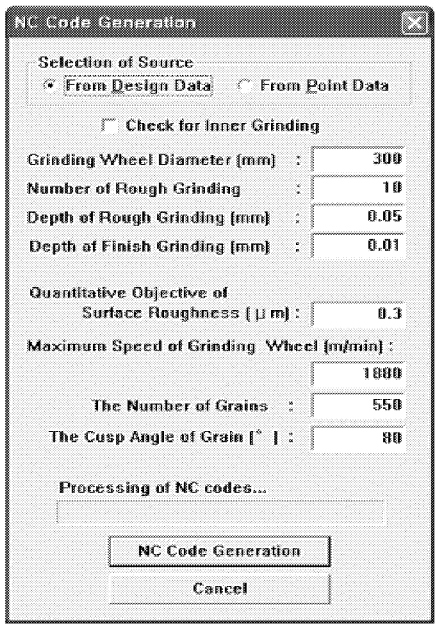


Figure 6 NC code Generation

The check box under the radio button is unchecked, because the experiment is limited to an external grinding. Each conditions and parameters that the griding wheel speed, the rough grinding number, the rough ginding depth, the finish grinding depth, the quantitive objective of surface roughness, the maximum speed of grinding wheel, the grain number and the grain cusp angle are input in below the check box. Than, maximum grinding speed algorithm with constant roughness applied NC-code can be obtained through the CAM system.

Expriments are carried out by the developed high precision cam profile CNC grinding machine for three types of cam as shown in Fig. 7.

As cam is made of steel (SS41), CBN grinding wheel is used in the experiments. The number of grains in the equation (16) is 550 and the end cutting edge angle of grain is 80°. As result, the maximum grinding speed can be obtained to maintain constant surface roughness. The NC-code is obtained by the developed CAD/CAM system(S.H. Lim et al. 2006).

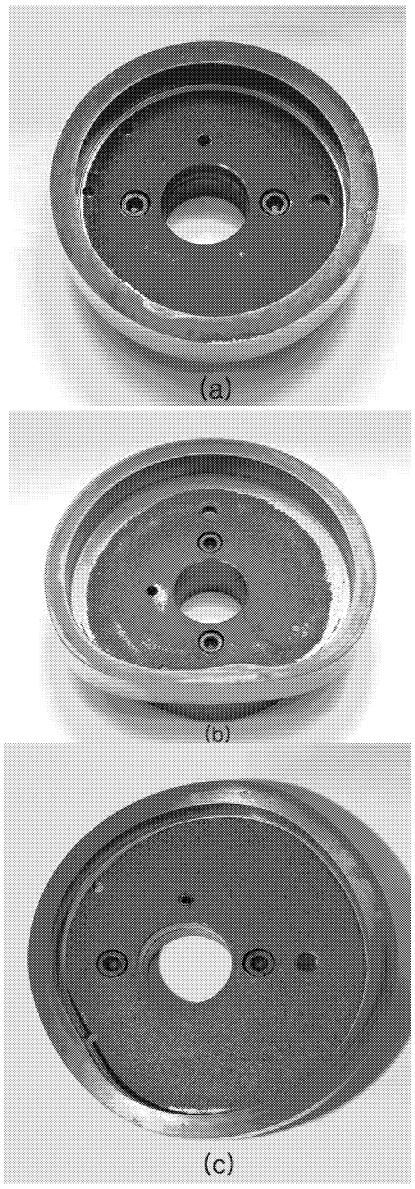


Figure 7 Result of grinding machining for the Palanquin cam(a), Blanket cam(b) and Nassel needle cam(c)

Table 1 Results of the manufacturing time using the only biarc curve fitting versus the biarc curve fitting with developed algorithm

Cam type	Palanquin cam	Blanket cam	Nassel needle cam
Objective of Surface Roughness[㎛]	0.3	0.2	0.2
Only biarc curve fitting	56' 03"	59' 12"	59' 38"
Biarc curve fitting with developed algorithm	40' 52"	49' 48"	50' 10"
Variation of manufacturing time	27.1% reduction	15.9% reduction	15.9% reduction

As result of experiments, the reduction ratio of machining time is 27.1% and leads higher productivity. In order to confirm the objective of surface roughness is achieved and the surface roughness(R_a) of the cams is measured by a instrument of Mitutoyo's Surftest SV-624. The measuring points is selected 4 times by 90 degrees rotation. Table 2 shows the results of surface roughness.

Table 2 Result of the surface roughness (R_a) evaluation [Unit : ㎛]

Cam type	Palanquin cam	Blanket cam	Nassel needle cam
Objective of Surface Roughness	0.3	0.2	0.2
Point 1 (90°)	0.29	0.19	0.20
Point 2 (180°)	0.25	0.19	0.16
Point 3 (270°)	0.22	0.20	0.21
Point 4 (360°)	0.28	0.20	0.19

As shown in table 2, the objective of constant surface roughness is achieved by the proposed algorithm.

CONCLUSIONS

In this study, an algorithm of maximum grinding speed to obtain constant surface roughness is proposed. Three types of the cam(Palanquin cam, Blanket cam, Nassel needle cam) in the spinning machines are designed and manufactured by using the developed CAM system.

As result, the machining time is decreased to 27.1% by the proposed algorithm which estimates the maximum grinding speed for the constant surface roughness, and the objective of constant surface roughness is achieved.

REFERENCES

Kim. H.S.; K.S. Jeong.; D.G. Lee. 1995. "Structural Design and Evaluation of the Three Axis Ultra-Precision CNC

Grinding Machine". *Transaction of KSME*, vol.19, No.12, 3392-3402.

Gao. Y. and J.A. Webster. 1991. "Computer Simulation of the Deformation of Slender, Multi-diameter Rollers During Grinding". *International Journal of Machine Tools and Manufacturing*, vol.31, 83-93.

Lim. S.H.; S.H. Shin.; C.M. Lee. 2006. "A study on Structural Design and Evaluation of the High Precision Cam Profile CNC Grinding Machine". *Journal of the KSPE*, vol.23, No.10, 113-120, ISSN 1225-9071

Park. C.W. 2002. "Theory of Grinding Machining". *Science Technology*, Korea, 11-23, 27-47, 77-130

Kim. D.W. 1991. "Method of Workshop Practice". 516-521

Ohmori. H. and T. Nakagawa. 1988. "Grinding of Silicon Using Cast Iron Fiber Bonded Diamond Wheel(4th Report ; Mirror Surface Utilizing Super- Abrasive Wheel)". *Proc. of the Spring Conf. of JSPE*, 521-552

Gal. T.Z.; M. Shpitalni.; S. Malkin. 1989. "Design and Manufacturing Analyses for Integrated CAD/CAM of Cams". *ASME Journal of Engineering for Industry*, vol.111, 307-314

Tsay. D.M. and H.M. Wei. 1996. "A General Approach to the Determination of Planar and Spatial Cam Profiles". *ASME Journal of Mechanical Design*, vol.118, 259-265

Shin. J.H.; C.M. Lee.; J.S. Kim. 1996. "Shape Design of Disk Cam Mechanisms Using Instant Velocity Center". *Proceedings of 6th International symposium on Transport Phenomena and Dynamics of Rotating Machinery*, vol.1, 178-186

Weck. M. 1984. "Machine Tools". *John Wiley & Sons,* vol.2, 129-157, 167-173, 203-205

Shigley. J.E. and J.J. Uicker. 1980. "Theory of Machines and Mechanisms", 226-241

Hahn. R.S. 1954. "On the Theory of Regenerative Chatter in Precision Grinding Operations". *Trans. ASME*, vol.76, No.1, 593-597

ADVANCED KINEMATIC-GEOMETRICAL MODEL OF GRINDING PROCESSES

Błażej Balasz
Tomasz Królikowski
Koszalin University of Technology
Department of Fine Mechanics
75-620 ul. Raclawicka 15-17 Koszalin, Poland
e-mail: blazej.balasz@tu.koszalin.pl

ABSTRACT

Grinding is a very complex process depending on large number of correlated factors. In precise grinding it is very important to select optimal conditions and to preserve stable conditions during the process. The model of grinding process comprise usually a few elementary models: model of a grain, model of a grinding wheel topography, model of surface roughness, model of the process kinematics, model of a chip formation, forces and energy, thermal and vibration. The author of this paper undertook a study on developing algorithms and programs for complex simulation of grinding process. This paper presents assumptions, schemes, examples of models, and results of the advanced kinematic-geometrical model of grinding processes.

KEYWORDS: grinding, surface, modeling, simulation

INTRODUCTION

Grinding is a very complex process depending on large number of correlated factors. In precise grinding it is very important to select optimal conditions and to preserve stable conditions during the process. All physical and conventional parameters characterizing grinding process could be divided up among two groups: one comprising those which are experimentally determined, and second which determination during the experimental process is impossible or too difficult to achieve and results of such measurements are unreliable. As a solution to the problem of second group of parameters a modeling and computer simulation of grinding process is one of the possible answer.

In the last 15 years the importance of modeling and simulation of grinding processes has significantly risen. Due to increase of processor power the new category of models based on physical process description has arisen. Each of that categories resolves different aspects of the process. Fine Element Analysis (FEA) which is based on description of physical process by differential equations is targeted in macroscopic scale to calculate the heat transfer, forces and energy during the process and they influence on grinded surface pressure. The grinding wheel is modeled as heat source and surface pressure moved over workpiece, but the plastic material behavior and the chip formation are not considered (Brinksmeier et al. 2006). The microscopic FEA models most accurately calculate the elasto-plastic material deformation but are limited to the interaction of a single grains and due to still not sufficient computer power it is not possible to simulate the whole grinding wheel interaction with material in microscopic scale. The most advanced modeling and simulation methods to understand the microscopic material behavior during grinding are Molecular Dynamic Models (MDM), based on atomistic models of materials, but due to complexity of the models simulation are often limited to the initial contact and chip formation which covers only a few nanometers of contact length. It limits an application of FEA and the MDM in the area of a new grinding tool designing and grinding process optimization.

The authors of this paper undertook a study on developing kinematic-geometrical models of the grinding tools and grinding processes which enable to set assumptions for the creation of new abrasive tools with parameters to facilitate obtaining the expected results of machining, and increase of the productivity of the process and a much better use of the machining potential of grinding wheels.

MODELING

Fundamentals of the models

Models of the process geometry where developed based on the experimental results of microcutting process carried out with a single grain. The aim of the experiments was to obtain the conditions for chip formation according to grain shape, depth of cut and value of cutting speed for different type of the grains and machined material. To achieve that objectives the experimental stand has been developed, based on plane grinding machine equipped with grinding tool with a grip for mounting a single grain (cp. figure 1a), material sample mounted on dynamometer (cp. figure 1b), what enables the measurement of grinding forces during cutting. The results of each experiment where gathered in databases for father analysis.

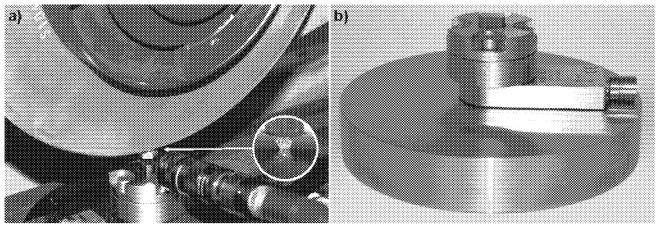


Figure 1: Experimental stand for microcutting process: a) grinding tool with single grain, b) material specimen

Each material sample after cutting experiment was measured with profilometer in order to obtain the data on depth of cut, size of pile-ups and deviation of that on the length of the cut. A microscopic pictures of a grains and the scratches made by a cutting edges identified on the grains were also obtained with the use of scanning electron microscope (cp. figure 2).

The model of grinding process comprise usually a few elementary models: model of a grain, model of a grinding wheel topography, model of surface roughness, model of the process kinematics, model of a chip formation, forces and energy.

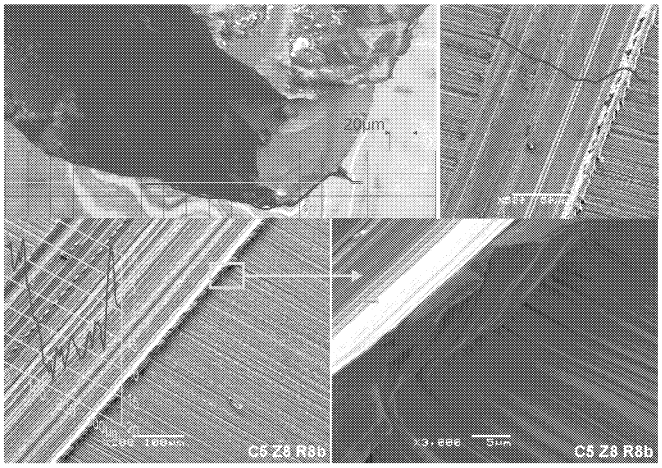


Figure 2: SEM-pictures of the grain and scratches made by single grain during microcutting

Model of grains

Analyses of the stereometry of real grains on the basis of research formed the basis for the development of models of abrasive grains. In the simulation method developed it was assumed that what is important for the machining process is the grains' contours protruding over the surface of the grinding wheel as well as their shape and size above the level of the binding material, as it is only those fragments of the grain that have an influence on the grain's contact with the material and its wear. For this reason, the models develop describe only the stereometry of the part of the grain located above the geometrical surface of the grinding wheel. It was assumed in the modeling of the grains that the shape of the grain is described on a convex solid, with the local concavities of the grain's surface being taken into account and modeled in the form of micro-roughness on the surface.
A numerical notation of the shape of the grain obtained is done with the use of a matrix of real numbers Z_k (1), whose size [m, n] is determined on the basis of assumptions concerning the size of the grain modeled. The size of the matrix increases together with the growth of the grain's sizes.

$$Z_k(x,y) = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ z_{m1} & z_{m2} & \cdots & z_{mn} \end{bmatrix} \qquad (1)$$

A numerical notation of the shape of the grain's topography facilitates a modification of its shape during the simulation process caused by the grain's contacts with the material machined, and also as a result of the dressing process of the grinding wheel.
Modeling of the surfaces of abrasive grains is provided with application of an elastic neural network. In the neuron model developed, the output parameters are the number of the grain vertices, the apex angle and the vertex radius. As a result of the work of the system, a random model of a grain with set parameters is obtained. In the network developed, the weights of individual neurons represent the coordinates of points on the surface of the grain generated. The work of this neuron network consists in the change of the values of neuron weights, as a result of which the coordinates of points describing the surface of the modeled grain are obtained. The elastic neuron network consists of N neurons, where each one of them has a vector of weights w_n to determine its location in space R^N. Between the neurons in the network, there exists a system of elastic connections. For each neuron n, a set of neurons is assigned with which it is directly connected. Each connection is assigned function $f_e(d)$, called the function of elasticity. This function depends on the distances of weight vectors of connected neurons. The system of M nodes constituting characteristic points on the grain's surface constitutes the output data for the network. These nodes, in the case in question, constitute a system of characteristic points of the surface of the simulated abrasive grain. During the network's adaptation process, the weight vectors of individual nodes affect simultaneously all the neurons located in the neighborhood determined by a certain radius. With the progress of the network's adaptation both the neighborhood radius and the impact factor are subject to a reduction to lead to the network's stabilization. The purpose of the network's adaptation in the case in question is to obtain such a final form of the network, i.e. such vectors of the weights of neurons w_n and such vectors of connections N_N that it should map the abrasive grain's surface (cp. figure 3).

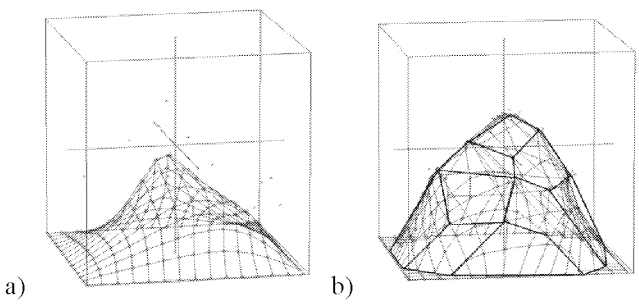


Figure 3: a) Adaptation process of an elastic neuron network. b) Network's final form depicting the surface of the simulated abrasive grain

Each modeled grain is saved in the grains database of simulation system in the matrix form. During a grinding wheel modeling the grains are randomly selected from database and are located on the surface model.

Model of grinding tool

Structure of grinding tool is composed of grains located randomly on its surface. Both a grain size and its locations have a great influence on quality of machined surface. In the developed model of grinding tool surface, one of the most significant factor of optimization of grinding process is achieved with optimal location of grinding grains on the surface. In the process of modeling a grinding tool surface, every single grain is randomly located on the surface with specified grain concentration C_{gr} (cp. figure 4a).

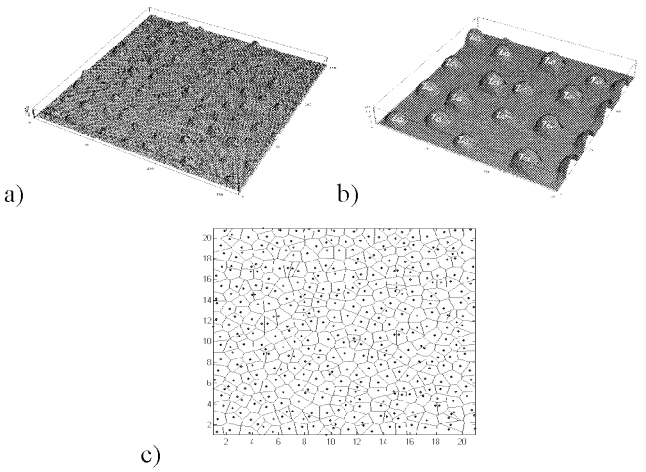


Figure 4: Grinding wheel topography a) model; b) indexes of grains; c) assessment of grain location with Voronoi diagrams

With every generated grain there is associated vector of grain parameters, describing temporal states of the grain during the whole process (e.g. number of contacts with workpiece material, volume of removed material, normal and tangential forces etc.). After grain generation, the working surface of the grinding wheel is generated by the aggregation of single grains into one surface, where each grain has a unique index (cp. figure 4b). Thanks to that, the characteristic of behavior of contact during the process could be thoroughly discovered. On generated surface the model of the bond is placed on. As a completion to this task, models of grain displacement and removal and the dressing process are also elaborated.

Model of grinding kinematics

Developed model describes the relative displacement of the single grains in relation to material displacement in Cartesians co-ordinations in surface grinding. Fundamentals of grinding process model description are mathematical equations describing single grain trajectory displacement on the path of cutting in the grinding zone. Taking in to account superposition of grinding wheel rotational movement and reciprocating movement of workpiece relative movement a single grain in grinding zone is described by trochoidal path (cp figure 5). Equations for the cutting path in the functions of grinding wheel angular displacement relative to x-y coordination system with its origin fixed to the workpiece are defined for horizontal (2) and vertical (3) displacement of cutting point.

$$y = \frac{D_s}{2}\sin\psi \pm \frac{D_s}{2}\frac{v_{ft}}{v_s}\psi \qquad (2)$$

$$z = \frac{D_s}{2}\left(1-\cos\psi\right) \qquad (3)$$

Developed model of grinding process takes into consideration reciprocal interpenetration of material and grinding wheel profiles in the grinding zone. Base on this kinematic-geometrical relation during the simulation process, a successive cross-sections, grinding forces and energy for a single grain can be calculated.

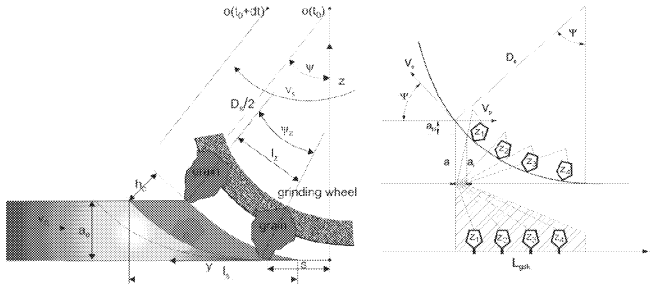


Figure 5: a) model of the grain displacement in the grinding zone; b) schema of discretisation

Simulation calculations are performed in a discrete intervals of the time Δt with constant increment. The value of Δt_s is based on kinamatic-geometrical parameters of the process. Simulation calculations are carried out with the discretisation of the grinding zone and grinding wheel with a precision of 1 micrometer. In every step of calculations the position of every grain in grinding zone is settled and calculations of contacts of all grains in zone is performed and results are saved in the database. With the correlation of grinding wheel rotation motion, the longitudinal move of workpiece profile are performed. The interdependence of both motions are calculated with the relation of grinding wheel velocity v_s and workpiece velocity v_{ft}.

Model of cutting and pile-up formation

Forming of the workpiece surface during the simulation process is based on material removal and pile-up formation of single grins on the path of cutting calculated with step of 1 micrometer. For an every contact of grain with material profile, the contact parameters are calculated and determinations of type of contact is specified. There are three possible contact types: cutting, ploughing and rubbing. It was modeled in case of cutting, the calculated cross-section of material is partially remove in the form of chip and the other part is forming the pile-ups on the both sides of the grains depending on the material properties, grain types and conditions of the grinding process. In case of ploughing all material is moved sidewise of the grain and forms a pile-ups, when the rubbing take place no material removal occurs.

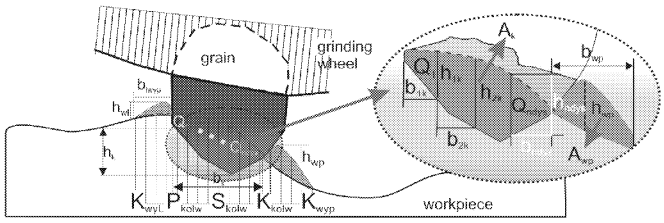


Figure 6: Model of the grain contact, forces calculation and pile-up creation

Determination of types of contact is performed for an every single grain in every step of simulation based on its calculated depth h_k, width b_k , and attack angle α_k (cp figure 6). There was proposed a parameter of *grain cut ability k* (4) specifying the influence of grain contact on type of material removal.

$$k = \alpha_k \sqrt{h_k \cdot b_k} \tag{4}$$

For a determination of contact type, the boundaries values of parameter k has been established base on experimental researches. Two constant value W_s and W_p, which determines the transition from rubbing to ploughing and cutting have been determined for different types of material. Based on its value during simulations the type of contact is selected with the relation (5):

$$type\ of\ contact(k) = \begin{cases} rubbing, & k \le W_p \\ ploughing, & W_p < k \le W_c \\ cutting, & k > W_c \end{cases} \tag{5}$$

Values of constants W_p, W_c depends on workpiece material type, grinding wheel type and process parameters and are established as input parameters of simulation. During the simulation, when condition of grinding wheel deteriorates vales of constants are also changed to reflects different condition of cutting (e.g. taking into account a blunting a grains or increased temperature in grinding zone, causing increased contribution of plastic deformation of workpiece).

As results of grain interaction with workpiece only a part of material is being removed in the form of chip, the rest part is plastically deformed on both sides of the groove and forms a pile-ups. Model of pile-up formation assumes that cross-section area A_p of pile-ups is calculated as the part of the groove cross-section area of contact with consideration of randomness of grinding process. For every grain model created on grinding wheel during simulation a random coefficient of cutting efficiency c_e is assign which values are determined for different materials type, using the uniform distribution for intervals:

$$c_e = \begin{cases} [0, 0.5), & for\ ploughing \\ [0.5, 1), & for\ cutting \end{cases} \tag{6}$$

The total pile-up cross-section area can be then express as:

$$A_p = (1 - c_e) A_k \tag{7}$$

The model of cut allows to create pile-ups which are not symmetric on both side of groove, and changes along the track of cut. The shape of pile-ups and its partition on groove sides depends on location of lowest point of grain determined in the cross-section of cut (cp. figure 6). For each grains contact, the outer p_{kolw}, k_{kolw} and lowest s_{kolw} points position are calculated. The ration of left and right pile-up area depends on distance $dist(p_{kolw}, s_{kolw})$ and $dist(p_{kolw}, s_{kolw})$. For each of pile-up its height h_w and width b_w is calculated and the shape of pile-up can be establish with possible description as triangular, parabolic, or compound functions (cp. figure 6).

Model of forces and energy

During the grinding process only a small percent of the grain located on active grinding wheel surface and having contact with workpiece participate in the work of material separation in the form of chip. The other part of active grains performs a work of plastic deformation and friction. Total energy of grinding depends on elastic and plastic deformation of material, internal friction of chip formation, friction on the contact surface of active grains and workpiece. Depending on workpiece and grinding wheel topography, vibrations during grinding and stiffness of machine tool and grinding wheel, the work of cutting is variable on the length of cut made by single grain. For more precise estimation of grinding forces and energy the proposed model allows the calculation of normal and tangential forces and specific energy in relation to single grains as well as total forces and energy during the whole process of simulation.

In the modeling of dependences of forces on parameters and conditions of grinding, it is assumed that the normal F_{nk} and tangential F_{tk} component of force affected a single grain depends on elementary cutting resistance k_l and cross section area of cut A_k (cp. figure 7).

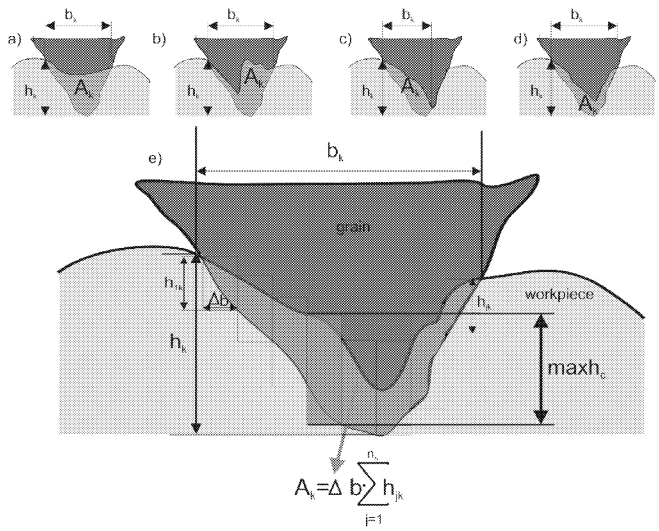


Figure 7: Cross-section area of the cut

The elementary cutting resistance depends on cross section area of cut, radius of grain tip ρ, angle of grain tip 2ε, and mechanical properties of material expressed by constant C_p. Equations for calculating the grinding forces can be expressed by the formulas:

$$F_{nk} = k_l(A_k, \rho, \varepsilon, C_p) \cdot A_k \tag{8}$$

$$F_{tk} = k_\mu(v_s, \rho, \varepsilon) \cdot F_{nk} \tag{9}$$

where: k_μ proportionality coefficient of grinding forces.

The cross-section area (cp. figure 7) is defined as:

$$A_k = \Delta b \cdot \sum_{j=1}^{n_h} h_{jk} \tag{10}$$

The grinding energy related to single active grain is a function of tangential component of grinding forces acting on the discreet length of grain interaction with workpiece in the grinding zone:

$$E_k = F_{tk} \cdot \Delta l_k \tag{11}$$

Energy variability during the grinding process is conditioned mainly by cross section area shifting, caused by different depth of cut and randomness of grain shape.

MODELS IMPLEMENTATION

Developed models enable precise analysis of phenomena taking place in grinding zone during grinding. It is possible due to information registration of individual grain contacts with its time and location in the grinding zone. In order to present a potential of grinding models the results of a few results from simulation experiments are presented.
During a simulation, intermediates values of material profile in the grinding zone are registered. Thanks to that the character of topography profiles lengthwise a grinding zone is examined (cp. figure 8).

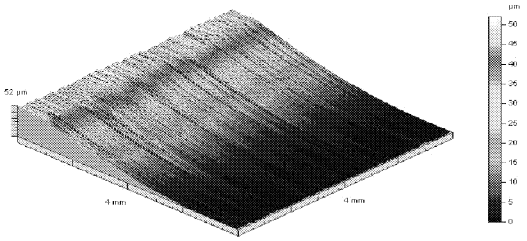


Figure 8: Grinding zone obtained by simulation

Important knowledge for designing a new optimal grinding tools is determination of grains load and its activity. It the presented models it is achieved due to ability of modeling a grinding tools with known locations of grains and then with registration its load during simulation. Figure 9 presents an average grain load obtained for simulation of plain grinding within the range of v_s and v_{ft} variation.

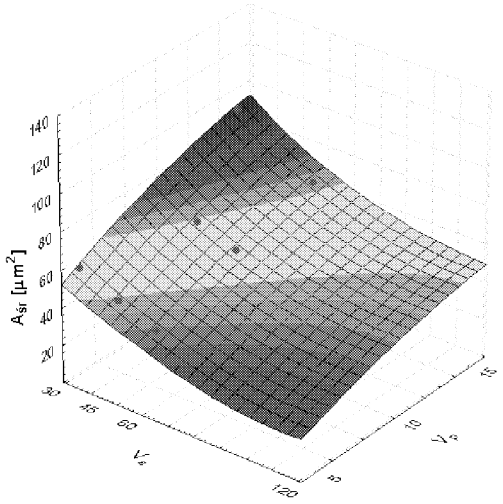


Figure 9: The average grain load

Figure 9 presents the influence of the grinding speed value V_s and longitudinal feed of the object V_{ft} upon the variations of grinding work E_s. The mean values of grinding work rise alongside with increasing speed V_{ft}, The load of grains in the grinding zone rises, as well, and grinding is characterized by high levels of energy consumption.

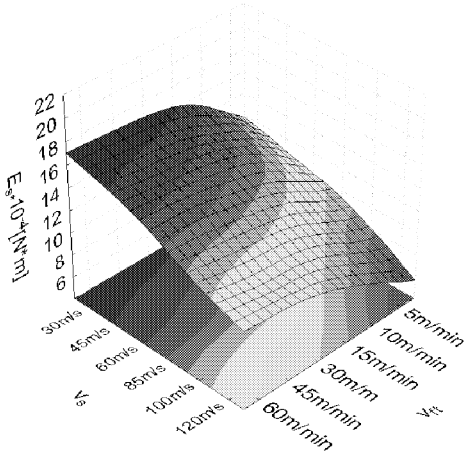


Figure 10: The influence of grinding process parameters on grinding energy

CONCLUSION

The developed models of grinding processes reveals features which enables designing a new models of grinding tools with optimal grains shape and size, and its orientation on the grinding tool surface. The optimization process is feasible due to possibility of the models to carry out the simulation within a vast range of process parameters variability and exact gathering data concerning individual contact of grains. The innovation solution of presented models depends on isolation of individual grains during the simulation process and analyzing the phenomena in the grinding zone in relation to single grain. The most significant analyzes concern: the grain activity and its load, the average cut layers, flotation of single grain depth of cut along the grinding zone, and the influence of grains shape, size and arrangement on afore mentioned phenomena.

ACKNOWLEDGEMENTS

This work was supported by grant: KBN Nr 4 T07D 033 28 form Polish Ministry of Science and Higher Education

REFERENCES

Balasz B., Królikowski T., Kacalak W. 2001. "Method of Complex Simulation of Grinding Process". THIRD INTERNATIONAL CONFERENCE ON METAL CUTTING AND HIGH SPEED MACHINING Metz, France, pp. 169-172

Balasz B. Królikowski T.. 2002. "Utility of New Complex Grinding Process Modeling Method." PAN Koszalin, pp. 93-109

Brinksmeier E. , et al. 2006. "Advances in Modeling and Simulation of Grinding Processes. Annals of the CIRP, vol. 55/2/2006, 667-696.

Królikowski T., Balasz B., Kacalak W. 2001. "The Influence of Micro- And Macrotopography of the Active Grinding Surface on the Energy Consumption in the Grinding Process". 15TH EUROPEAN SIMULATION MULTICONFERENCE, Prague, Czech Republic, pp. 339-341

Stępień P., Balasz B. 2006. "Simulation of the Formation Process of Regular Grooves on Surface Ground", Industrial Simulation Conference, Palermo, Italy s. 269 – 276.

Szatkiewicz T., Balasz B., Królikowski T. 2005. "Application of an elastic neural network for the modeling of the surfaces of abrasive grains. Artificial Neural Networks in Engineering ANNIE 2005, ASME Press, New York, pp. 793-800.

A STUDY ON THE EFFECT OF SURFACE ROUGHNESS AND CUTTING FORCE IN MQL END-MILLING

Su Hyun Park
Jong Yun Jung
Choon Man Lee
Young Kug Hwang

Mechanical Design and Manufacturing Department
Changwon National University
#9 Sarim-Dong Changwon, Gyeongnam, 641-773, South Korea
E-mail : cmlee@sarim.changwon.ac.kr

KEYWORDS
MQL(Minimum Quantity Lubrication), Cutting force, Surface roughness, ANOVA

ABSTRACT

At present, industry and researchers are looking for ways to reduce the use of lubricants because of ecological and economical reasons. Therefore, metal cutting is to move toward dry cutting or semi-dry cutting. This paper presents an investigation into MQL(Minimum Quantity Lubrication) machining for the end-milling process of aluminum. To reach these goals, finish milling experiments were carried out with MQL. The surface roughness and cutting force results of tests were measured and the effects of cutting conditions were analyzed by the method of Analysis of Variance(ANOVA). From the experimental results, spindle rotation, feed per tooth and the interaction between RPM of spindle and feed per tooth and between axial depth of cut and feed per tooth are significant factors on the surface roughness.

INTRODUTION

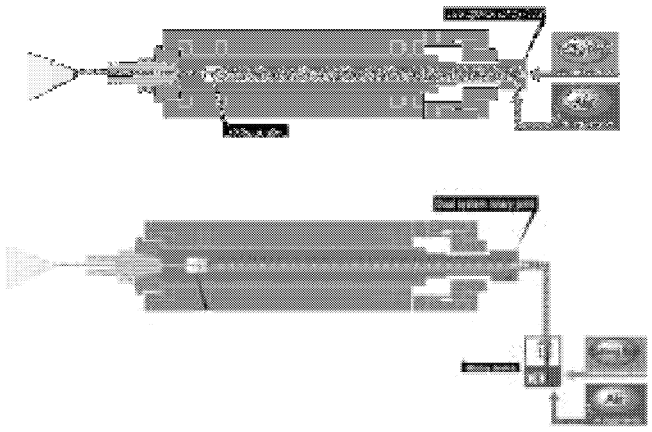
Lubricants in industry are furified by amount of expense as untreated waste of lubricants affects human being and ecosystem. Therefore, researchers are looking for ways to reduce the use of lubricants because of ecological and economical reasons. At present, metal cutting is moving toward dry cutting or semi-dry cutting.

This paper presents an investigation into MQL(Minimum Quantity Lubrication) machining with the objective of presenting data of groundwork by analyzing cutting force and surface roughness to identify the optimum cutting condition for the end-milling process of Al6061 with the MQL machining technology. Simulations of end-milling in MQL are carried out by using AdvantEdge 4.72 to compare experiments and the surface roughness results and cutting force of tests are measured. The effects of cutting conditions are analyzed by the method of Analysis of Variance (ANOVA). Better surface roughness is obtained with decreasing oil quantity and cutting force is similar to each other.

MQL TECHNOLOGY

MQL machining technology was first invented in 1994 in Germany and is also called by various names such as semi-dry machining, near-dry machining.

The MQL machining technology can be classified into external and internal supply methods as shown in Figures 1 according to the place where the lubricant is mixed with air (C. M. Lee and Y. K. Hwang. 2005). By using mixing inside spindle method, the lubricant and the carrier gas are separately supplied inside spindle, and MQL is created near the cutting tool. In this method, all supplied lubricant can be discharged. On the other hand, in mixing outside spindle, MQL is created at the upper stream from the spindle of machine tool and supplied to tool tip via piping and spindle. In this case, the only deliverable smoke type lubricant can be supplied.



Figures 1 Internal and external MQL supply methods

EXPERIMENTAL PROCEDURES

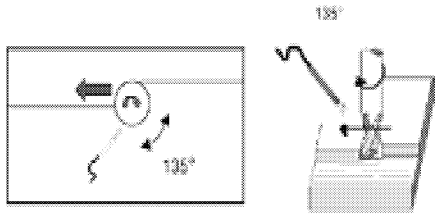
Machining System

The machining center used in the experiment is the Sirius-2 manufactured by Hwacheon. Co. Ltd. The MQL supplier and mist collector are attached to it. The MQL supplier is Vario UFV 1—001 produced by VOGEL, the German manufacturer and is applied both to the external and internal supply methods. Specific features of supplier are shown in table 1.

Table 1 Specification of MQL supplier

Oil droplet size [•]	Aerosol quantity [• /h]	Air consumption [Nl/min]
0.5	5~150	140~300

Because machinability varies according to the direction of the MQL supply and diameter and shape of the nozzle, in this experiment the distance between nozzle-tips is fixed at 20mm and the MQL is supplied at 45 degree angle as shown in Figures 2. The radial depth of cut is fixed with 2mm. The air pressure supplied to the MQL supplier is fixed at 6 bar, and LubriFluid F100, its specifications are shown in Table 2 is used as cutting fluid.

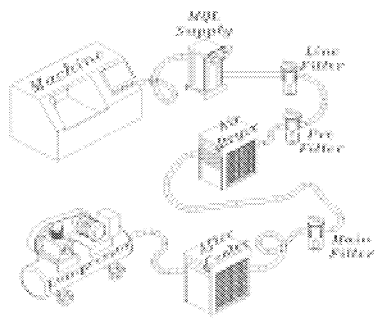


Figures 2 Spray angle

Table 2 Specifications of MQL lubricant

Type	Base	Density at +20˚C [g/• 3]	Viscosity at +40˚C [• /s]	Flash point [˚C]
LubriFluid F100	higher alcohol	0.84	25	184

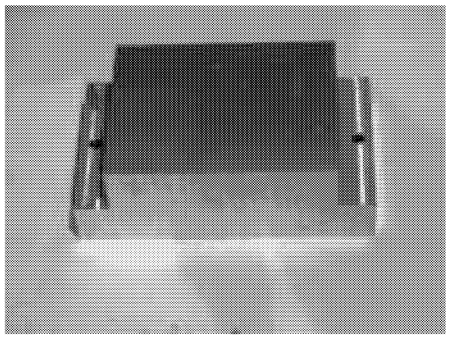
In order to assure stable air pressure and satisfy the pressure (6bar) recommended by the MQL supply company, a compressor is set-up. And in order to exclude malfunction of the MQL supplier by moisture generated at the compressor, an air cleaning system is additionally set to remove it. Total experimental systems are shown in Figures 3.



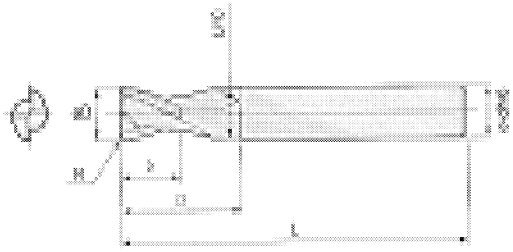
Figures 3 Experimental set up

Specimen and Cutting tool

In the experiment, as shown in Figures 4, Al 6061 is workpiece material. The size of the specimen is 80• •×35• •



Figures 4 Shape of test specimen



Figures 5 Shape of two flutes end-mill

The shape of cutting tool (TaeguTec manufactured) is shown in Figures 5. The cutting tool (8mm diameter, 2 flutes, helix angle= 30••, nose radius= 0.6•) is used in this study.

Machining Tests and Measurements

All the machining trials were carried out on sirius-2 three-axis machining center with factors. The experiments were randomly arranged by full factorial design. The factors and levels are shown in table 3. The workpiece is clamped onto KISTLER 9256A2 dynamometer that in turn is fixed on the machine table of sirius-2 machining center. In order to see the surface roughness in center-line-average-height(Ra), the specimens are measured using Surftest SV-624 manufactured by Mitutoyo. In order to ascertain the effects of individual factors on the surface roughness and the cutting force, the data obtained from the experiments are statistically analyzed by MINITAB software.

Table 3 Factors and values

Factors	levels	values
Nozzle diameter	2	Ø 2, Ø 6
RPM	2	1000, 3000
Feed per tooth(mm/z)	2	0.04, 0.05
Axial depth of cut (mm)	2	0.8, 1.4
MQL line length (m)	2	4.20, 6.20

REESULTS AND DISCUSSIONS

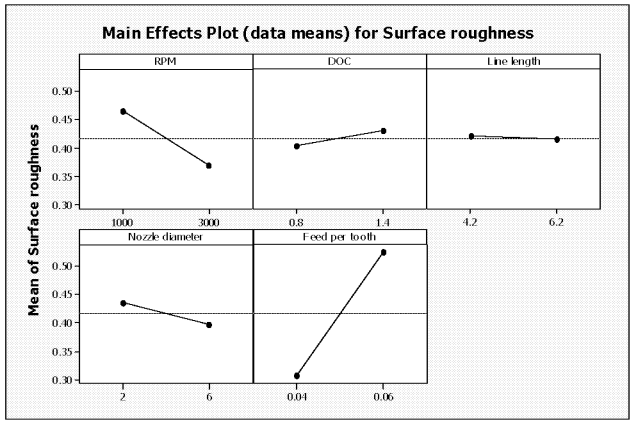
In this study, ANOVA (Analysis of Variance) is carried out to analyze effects on characteristic. The level of significance is 95% and value P is less than 0.05, it will be estimated that the factor for value P is significant on characteristic. The experiment design and the results are shown in table 4.

Table 4 Result of experiments

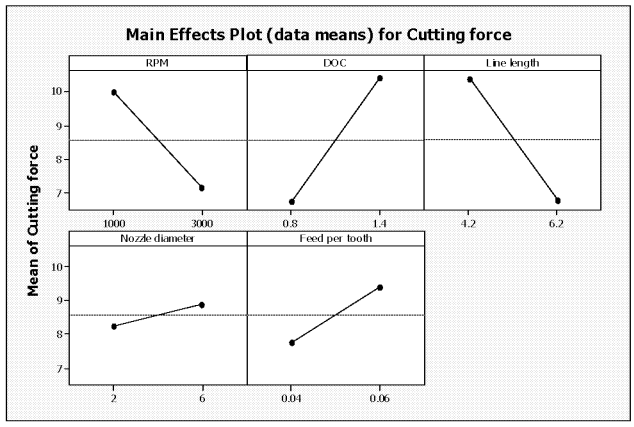
Run Order	RPM	DOC (mm)	Line length (m)	Nozzle diameter (••)	Feed per tooth (mm/z)	MQL		Flood coolant	
						Roughness (Ra)	Cutting Force (N)	Roughness (Ra)	Cutting Force (N)
1	1000	0.8	6.2	2	0.04	0.163	8.19	0.325	3.66
2	1000	0.8	6.2	6	0.06	0.263	3.01	0.553	3.09
3	1000	1.4	4.2	6	0.04	0.175	27.49	0.302	14.41
4	1000	1.4	6.2	2	0.06	0.351	9.46	0.605	9.53
5	3000	0.8	4.2	6	0.04	0.129	2.74	0.229	2.8
6	1000	0.8	6.2	2	0.06	0.331	3.16	0.622	3.18
7	3000	1.4	4.2	6	0.04	0.146	3.12	0.247	3.16
8	1000	0.8	4.2	2	0.04	0.149	8.79	0.455	8.9
9	1000	1.4	4.2	2	0.06	0.364	15.17	0.561	15.35
10	3000	0.8	6.2	2	0.06	0.258	2.03	0.398	2.08
11	3000	0.8	4.2	2	0.06	0.286	8.79	0.478	9.06
12	3000	1.4	4.2	2	0.06	0.32	2.55	0.504	2.52
13	3000	1.4	6.2	2	0.04	0.114	8.02	0.258	7.99
14	3000	0.8	6.2	2	0.04	0.124	9.61	0.28	3.54
15	3000	1.4	6.2	2	0.06	0.287	10.41	0.556	10.4
16	1000	1.4	4.2	6	0.06	0.423	31.33	0.643	25.73
17	1000	1.4	6.2	2	0.04	0.114	11.27	0.266	13.09
18	1000	1.4	6.2	6	0.04	0.182	9.55	0.329	9.61
19	3000	0.8	4.2	2	0.04	0.133	8.78	0.43	8.93
20	3000	1.4	4.2	2	0.04	0.164	12.37	0.33	12.71
21	3000	1.4	4.2	6	0.06	0.301	3.71	0.431	3.74
22	3000	0.8	6.2	6	0.06	0.229	4.29	0.468	4.23
23	1000	1.4	4.2	2	0.04	0.17	15.73	0.407	5.61
24	1000	0.8	6.2	6	0.04	0.129	2.35	0.287	2.57
25	1000	0.8	4.2	6	0.06	0.3	10.33	0.437	10.57
26	3000	0.8	4.2	6	0.04	0.131	8.39	0.23	8.48
27	3000	1.4	6.2	6	0.06	0.206	13.34	0.419	13.26
28	1000	1.4	6.2	6	0.06	0.379	10.55	0.815	10.06
29	1000	0.8	4.2	6	0.04	0.147	9.19	0.333	9.10
30	1000	0.8	4.2	2	0.06	0.31	15.24	0.484	15.25
31	3000	0.8	4.2	6	0.06	0.235	11.83	0.43	12.01
32	3000	1.4	6.2	6	0.04	0.216	9.08	0.197	9.20

Flood Coolant

The main effects on surface roughness and cutting force are shown in Figures 6 and 7 to compare the magnitude of each main effect. As is seen from the results of the ANOVA, surface roughness is mainly influenced by feed per tooth and spindle revolution in turn affects it. However, surface roughness is vastly not affected by axial depth of cut and nozzle diameter. Moreover, cutting force is mainly influenced by axial depth of cut and MQL line length and is not affected by feed per tooth and nozzle diameter. As surface roughness and cutting force are different according to nozzle diameter from the results, cutting condition should be selected with proper standards of machinability.



Figures 6 Main effects plot for surface roughness



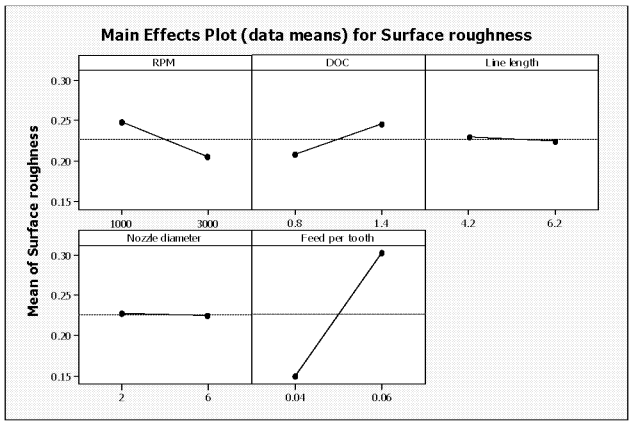
Figures 7 Main effects plot for cutting force

Minimum Quantity Lubrication

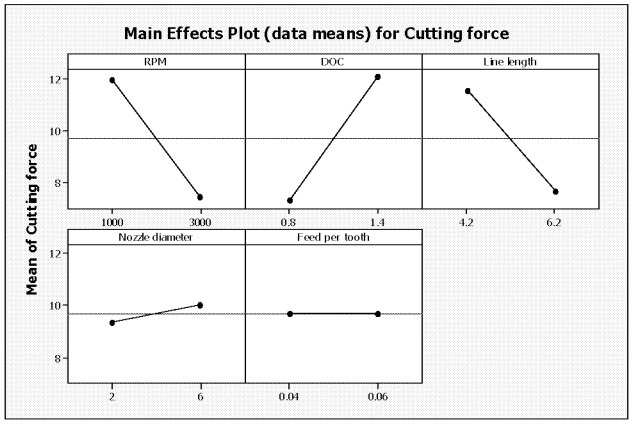
Surface roughness and cutting force in MQL end-milling are measured by the same method in flood coolant. The analysis results indicate that feed per tooth, spindle revolution are significant on surface roughness. Also, it shows that interaction between RPM of spindle and feed per tooth and between axial depth of cut and feed per tooth are significant. As can be seen from the results, cutting force is mainly influenced by axial depth of cut and MQL line length. Feed per tooth in surface roughness is significant but is not significant in cutting force.

FEM ANALYSIS

FEM analysis is carried out to identify the cutting force in MQL machining by using AdvantEdge 4.72. Factors able to control are selected. The Factors and levels are shown in table 5and it shows simulation conditions. A friction coefficient in MQL machining is fixed at 0.2.(I. O. Hwang 2005). Tool geometry is the same.



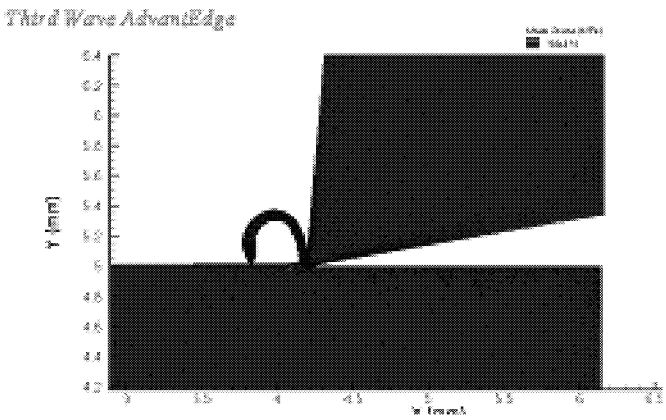
Figures 8 Main effects plot for surface roughness



Figures 9 Main effects plot for cutting force

Table 5 Simulation conditions

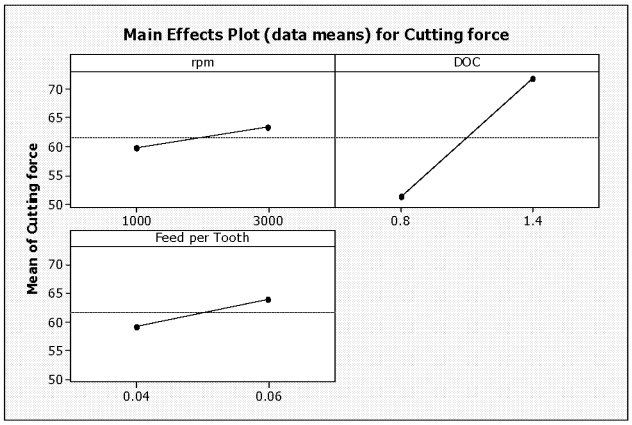
Length of cut	8.3776 mm
Max. number of nodes	12,000
Max Element Size	0.1 mm
Min Element Size	0.02 mm



Figures 10 Simulation

Table 6 Result of FEM analysis

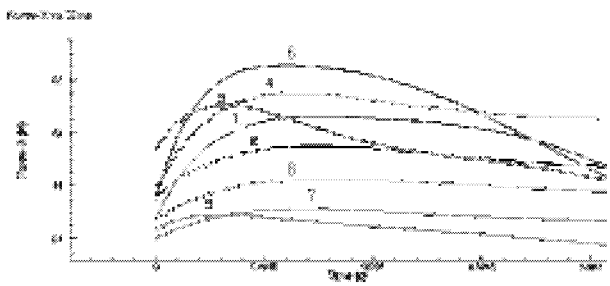
Run Order	RPM	Axial DOC (mm)	Feed per Tooth (mm/z)	Cutting Force (N)
1	3000	1.4	0.06	72.29
2	1000	1.4	0.04	53.74
3	3000	0.8	0.06	66.01
4	1000	0.8	0.06	42.12
5	1000	0.8	0.04	68.13
6	3000	1.4	0.04	85.87
7	3000	0.8	0.04	29.04
8	1000	1.4	0.06	42.68



Figures 11 Main effects plot for cutting force

Figures 10 and 11 show the results in cutting force according to various factors. As can be seen from the results of the ANOVA, cutting force is decreased in RPM 3000, axial depth of cut 0.8mm, feed per tooth 0.4mm/z.

In experiments, cutting force is decreased according to increase of spindle rotation and spindle rotation is important. In FEM analysis, spindle rotation is not significant and cutting force is increased according to increase of spindle rotation. The cutting force according to run orders is shown in Figures 12



Figures 12 Main effects plot for cutting force

CONCLUSION

In order to investigate the effects of cutting conditions on the surface roughness in MQL end-milling, the experiments of

Al 6061 end-milling in MQL and the FEM analysis while varying the cutting conditions is carried out. And the conclusions are summarized as follows.

1. In MQL end-milling operations, spindle rotation, feed per tooth, the interaction between RPM of spindle and feed per tooth and between axial depth of cut and feed per tooth are significant factors on the surface roughness.

2. In FEM analysis in MQL condition, spindle rotation is not significant and cutting force is increased according to increase of spindle rotation but spindle rotation is important in experiment.

In this research, experiments are carried out at 2 levels according to factors. However, it needs that additional research on the effects of MQL quantity on more specified levels should be conducted in the future. Also, the nozzle shape and the supplied air pressure should be considered in the future.

ACKNOWLEDGEMENTS

This work was supported by grant No. RT104-01-03 form Resional Technology Innovation Program of the Ministry of Commerce, Industry and Energy(MOCIE).

REFERENCES

M. Rahman; A. Senthil Kumar; M.U. Salam. 2002. "Experimental evaluation on the effect of minimal quantities of lubricant in milling". *International Journal of Machine Tools & Manufacture,* 112-116.

S. W. Shin; C. M. Lee; J. Y. Jung; Y. K. Hwang. 2006. "A study on improvement of the surface in MQL turning". *IMECS international conference.* 683-687

Y. Beauchamp; M. Tomas; J. Masounave; A. Youssef. 1996. "Investigation of cutting parameter effects on surface roughness in lathe boring operation by use of a full factorial design". *Compute Ind. Eng.* 645-651.

P. S. Sreejith; B. K. A. Ngoi. 2002. "Dry machining: Machining of the future". *Journal of Materials Processing Technology.* 287-291

Park T. J.; Yang S. H.; Lee Y. M. 2001. "Temperature Analysis of Each Coolant Level by a Multiple Comparison in Turning Process". *Journal of KSPE.* Vol. 18, No. 12. 110-115.

Ronan Autret; Steven Y. Liang; Ceorge W. *Minimum Quantity Lubrication in Finish Hard Turning.* HNICEN.

C. M. Lee; Y. K. Hwang. 2005. "A Study on the Effect of Cutting Conditions on the Surface Roughness in MQL Turning". *Proceedings of the 14th IASTED International Conference.* pp. 449-454.

I. O. Hwang; 2005. "Microscopic precision evaluation of machined surface according to the variation of cooling and lubrication method". *Pusan national university,*

ROBOTICS SIMULATION

SERVO-ENGINE SIMULATION

INFORMATICS PLATFORM FOR DETERMINING THE DYNAMIC PERFORMANCES OF A TWO STAGES SERVO VALVE

Alexandru Marin
University POLITEHNICA from Bucharest
313 SPLAIUL INDEPENDENTEI, 060042, Bucharest,
Romania
E-mail: marin@hydrop.pub.ro

Petrin Drumea, Marian Blejan and Ioana Ilie
Hydraulics and Pneumatics Research Institute from Bucharest
14, Cutitul de Argint, Bucharest,
Romania
E-mail: ihp@fluidas.ro

KEYWORDS
Test equipment, Dynamic modeling, Filtering, System identification, Linear model

ABSTRACT

Servo valves offer a proportional law between the command current and the controlled flow, at constant pressure loads. The present paper is devoted to the presentation of an informatics platform, installed on an experimental set-up, destined to make the identification of an equivalent transfer function of a two stage servo valve. The obtained data and the determined specific dynamic parameters were validated by comparing them with to the data mentioned in the manufacturer catalogue. As a conclusion, the developed experimental procedure is considered valid and able to be integrated in a more complex technology for identification the steady state and dynamic performances of modern servo valves.

INTRODUCTION

Electro-hydraulic servomechanisms with servo valves have good static and dynamic performances, used in applications such as machine-tools, molding machines, robots, aerospace industries and the list can easily be continued. The dynamic performances are characterized by high pass values higher than 20 Hz and the power controlled larger than 15 kW (Jelali 2003 & MOOG).
Servo valves offer a proportional law between the command current and the controlled flow, at constant pressure loads. Their dynamic behavior depends on the supply pressure, input current amplitude, fluid temperature etc. Variable loads don't influence the stability, as it happens in the case of proportional valves. A major problem is the fluid contamination, imposing special necessities for filtration. For normal operation installing and exploitation, with relative high costs, the useful life of servo valves supplied for most applications should exceed 1000 hours.
The paper is devoted to the presentation of an informatics platform, installed on an experimental set-up, destined to make the identification of an equivalent transfer function of a two stage servo valve. The obtained data and the determined specific dynamic parameters were compared with the specifications, mentioned by the manufacturer in its technical bulletins (MOOG Technical Bulletins).

EXPERIMENTAL SET-UP

The experiments were made in the laboratories of the Hydraulics and Pneumatics Institute from Bucharest, the test equipment and facilities following the standards imposed by the manufacturer (MOOG Technical Bulletins, Jones 1997, D'Amore 2001, DeRose 2003). In Figure 1 is presented the experimental set-up, which contains the servo valve SV, the supply pressure being ensured with a pump P and a pressure valve S; load on the output stage of the servo valve is simulated with a variable hydraulic resistance R. The command signal is done by a function generator GF and the measurement equipment contains four pressure transducers TP (one max. 400 bar supply pressure P, 2 max. 160 bar on A and B ports and one max. 25 bar on tank port T), connected to all ports of the servo valve and a flow transducer Q on the input supply chain (with max. 67 l/min). All electrical signals are gathered with a TEST POINT acquisition device SA, having USB interface, 16 analogical inputs (sample rate 100 ks/s), 2 analogical outputs (analogical resolution 12 bytes), 8 digital input/output ports, 2 counters on 24 bytes. For dynamic measurements of the servo valve exit flow we used the diaphragm method, respectively an equation containing the pressure drop on the variable hydraulic resistance R:

$$Q = k * sign\,(P_A - P_B) * \sqrt{|P_A - P_B|} \qquad (1)$$

where: Q is the exit flow of the servo valve; k – calibration constant determined in steady state conditions with the flow transducer Q; PA and PB are the pressure values in the ports A and B of the servo valve.

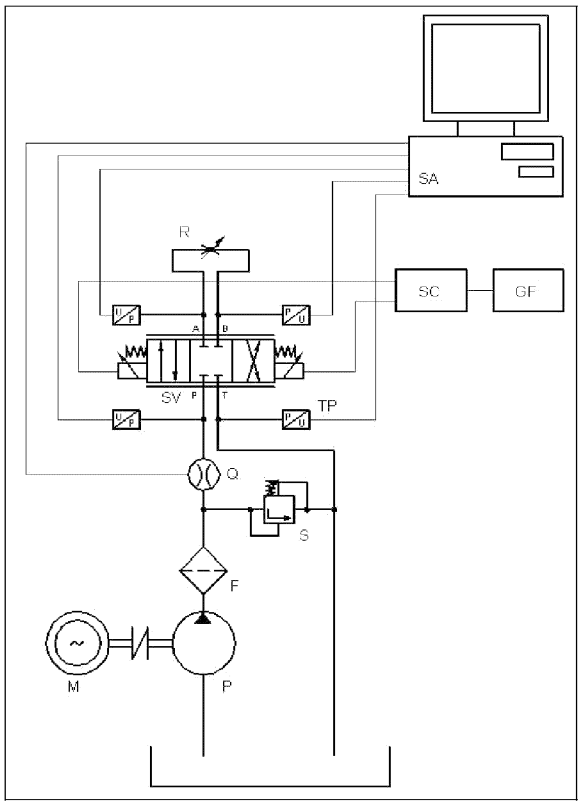


Figure 1: The experimental set-up

The informatics platform of the experimental set-up is presented in Figure 2.

Figure 2: Informatics platform of the experimental set-up

DYNAMICAL TESTS

For the step response, the time duration of the transient regime was of 5 seconds, the sample time being 0.1 ms, respectively 50000 acquisition points for each monitored parameter (input current and flow). The data were stored in EXCEL files and imported to MATLAB – SIMULINK, this software program package having all necessary facilities for analyzing, interpretation and graphical representations of the obtained results.

Some representative step responses of servo valve are presented in Figures 3÷8.

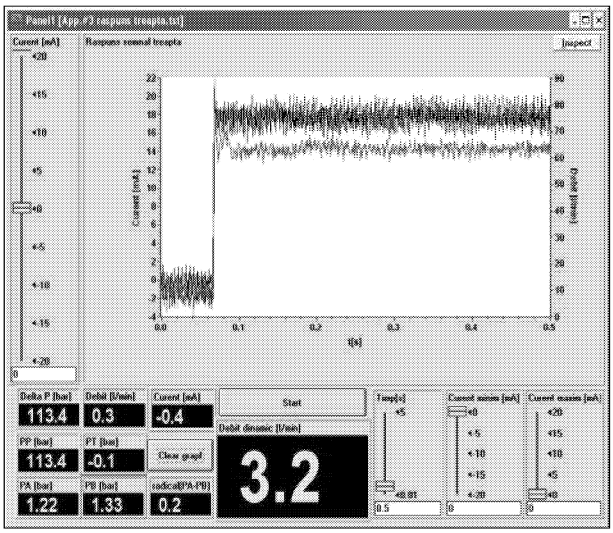


Figure 3: Step current 0...I_n

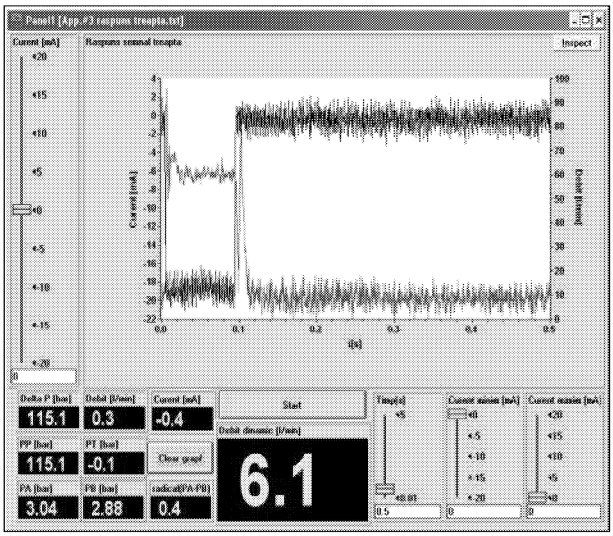


Figure 4: Step current 0... -I_n

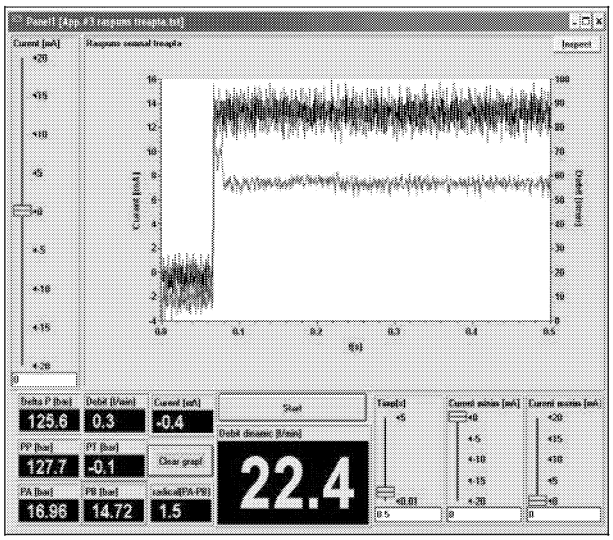


Figure 5: Step current 0... 0,75 I_n

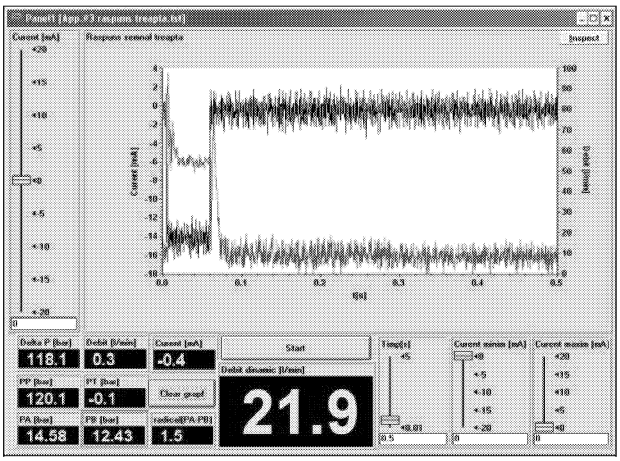


Figure 6: Step current 0...- 0,75 I_n

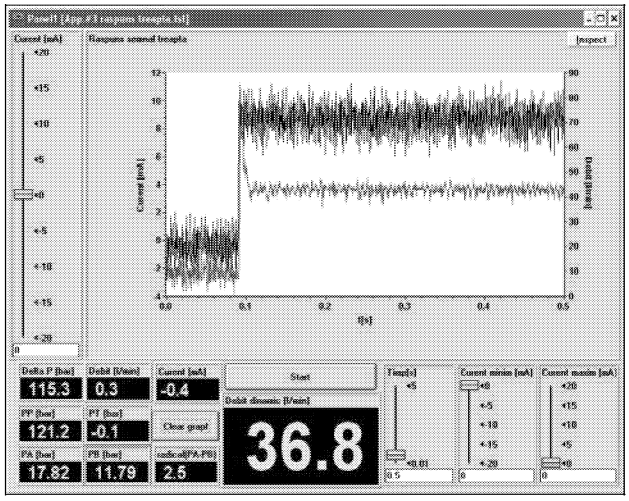


Figure 7: Step current 0... 0,5I_n

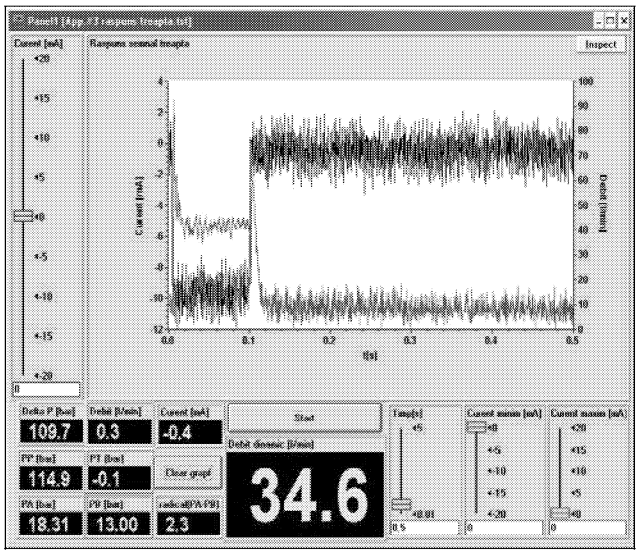


Figure 8: Step current 0...- 0,5 I_n

For the frequency response, we applied at the input of the servo valve a constant amplitude signal, with linear continuous increasing frequency (0,2 ÷ 200 Hz), correspondingly measuring the flow through the servo valve. The time duration of the experiment was of 10 seconds, the sample time being 0.1 ms, respectively 100000 acquisition points for each monitored parameter (input current and flow). The data were stored in EXCEL files and imported to MATLAB – SIMULINK, this software program package having all necessary facilities for analyzing, interpretation and graphical representations of the obtained results.

The frequency responses of the servo valve are presented in Figures 9 & 10.

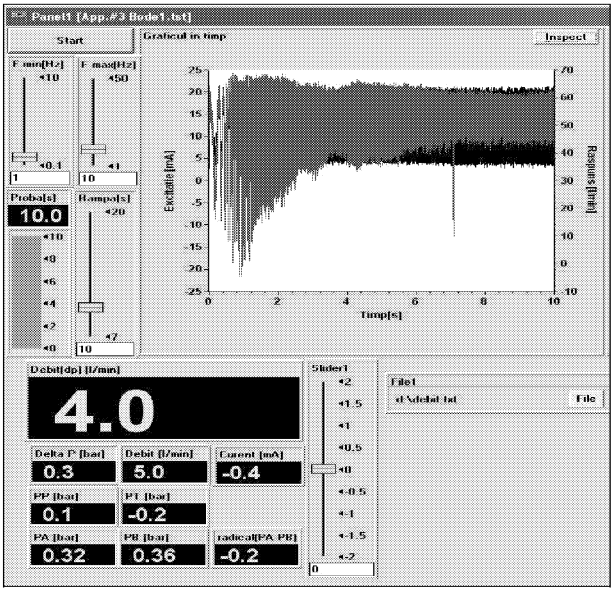


Figure 9: Command current amplitude ± 8 mA

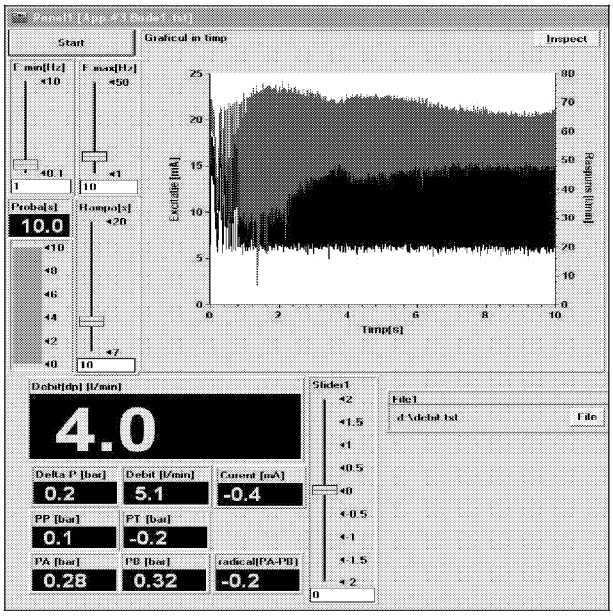


Figure 10: Command current amplitude ± 6 mA

The curves presented above reveal some noises superposed on the main signals, because of the functional aspects existing in the measurement chains of the experimental data. The analyze of these results implies a filtering process, without introducing phase delays, done with the facilities installed in the MATLAB SIMULINK software package (subroutine "filtfilt"). Some representative results are presented below (Figures 11 & 12).

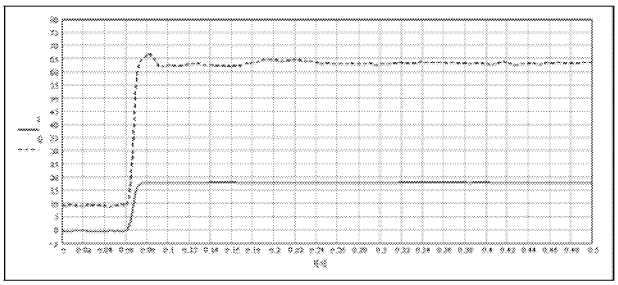


Figure 11: Step response

IDENTIFICATION OF THE TRANSFER FUNCTION

The equivalent transfer function of the servo valve can be written as below:

$$H_{SV}(s) = \frac{\Delta z(s)}{\Delta i(s)} = \frac{K_{iz}}{\frac{s^2}{\omega_n^2} + \frac{2\zeta}{\omega_n} s + 1} \qquad (2)$$

where: K_{iz} = 8,25 m/A, is the gain current – spool position of the servovalve, z; ω_n = 650 rad/s, is the natural frequency of the servovalve; ζ = 1, is the dynamic damping coefficient. These two characteristic dynamic have typical values, very close to the manufacturer catalogue: gain - 7 dB, for a delay phase of 90°. The above mentioned values were determined by implementing a simple qualitative procedure of calibration the Bode diagrams – the ones corresponding to the transfer function to the others determined experimentally (Schwarzenbach 2002, Rashidy, 2003).

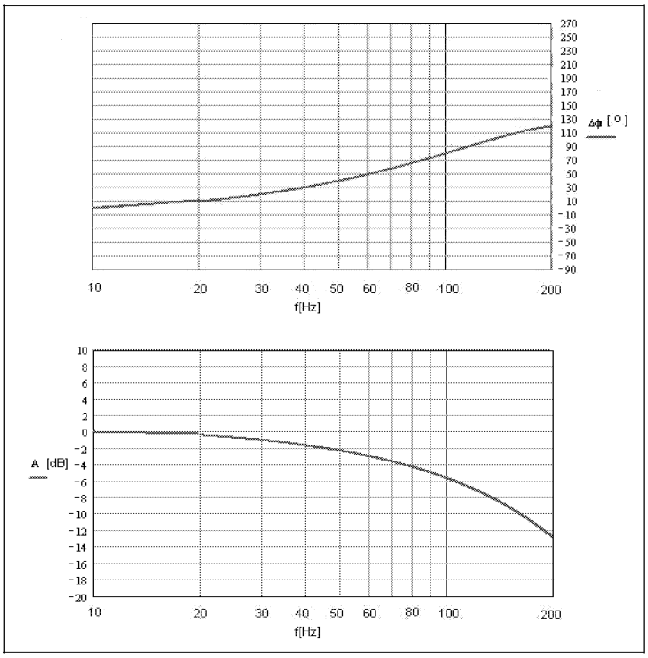


Figure 12: Bode diagrams

CONCLUSIONS

The experimental identification of the transfer function of the servo valve has proved some specific observations, relied to the difficulties to introduce in analytical models of the equipments some uncertain parameters - operational and environmental variables that produce significant differences in the actual dynamic response - such as: variations of the supply pressure, input signal level, hydraulic fluid temperature, ambient temperature, valve loading and so forth. Another difficulty in assigning simplified, linear transfer functions to represent servo valve response is that these valves are highly complex devices that exhibit high-order, nonlinear responses. If a first, second, or even third-order transfer function is selected to represent servo valve dynamics, still only an approximation to actual response is possible. Fortunately, for most physical systems, the servo valve is not the primary dynamic element, so it is only necessary to represent valve response throughout a relatively low frequency spectrum.

The informatics platform performed well the experimental tests done as mentioned above, the specific graphical dependencies and dynamic equivalent parameters correspond to the data mentioned in the manufacturer catalogue. As a conclusion, the developed experimental procedure is valid and able to be integrated in a more complex technology for identification the steady state and dynamic performances of two stage modern servo valves.

REFERENCES

D'Amore, M. and G.Pellegrinetti. 2001. "Dissecting High-Performance Electrohydraulic Valves, Machine Design", (April).

DeRose, D. 2003. "Proportional and Servo Valve Technology", Fluid Power Journal (March/April).

Jelali, M. and A.Kroll. 2003. "Hydraulic Servo Systems - Modeling, Identification & Control", Springer.

Jones, J.C. 1997. "Developments in Design of Electrohydraulic Control Valves", Moog Technical Paper, (November).

Neal, T. P. 1974. "Performance Estimation for Electrohydraulic Control Systems", Moog Technical Bulletin 126, (November).

Rashidy, H. Rezeka, S. Saafan, A. Awad, T. "A hierarchical neuro-fuzzy system for identification of simultaneous faults in hydraulic servovalves", American Control Conference, 2003.

Schwarzenbach, J. and K.F.Gill. 1992. "System Modelling & Control", Edward Arnold.

Electrohydraulic Valves A Technical Look, Moog Technical Paper.

Moog 760 Series Servovalves, product datasheet.

Signal Conditioning an LVDT using a TMS320F2812 DSP (SPRA946).

TMS320F2812 Data Manual (SPRS174).

http://www.measurementcomputing.com/testpoint.html

BIOGRAPHY

Ph. D. Petrin Drumea – specialist in fluid power systems, focused on experimental and theoretical research on mechatronic systems, industrial robots, lubrication systems, flexible hydraulic tools at 700 bar, solid waste processing machines, recovery devices in automotive applications, hydraulic agricultural machinery.

OPTIMIZATION OF ELECTRO-HYDRAULIC SYSTEMS USING MECHATRONIC CONCEPTS

Alexandru Marin
University POLITEHNICA from Bucharest
313 SPLAIUL INDEPENDENTEI, 060042, Bucharest
Romania
E-mail: marin@hydrop.pub.ro

Petrin Drumea and Adrian Mirea
Hydraulics and Pneumatics Institute from Bucharest
14, Cutitul de Argint, Bucharest
Romania
E-mail: ihp@fluidas.ro

KEYWORDS
Control systems, Industrial engineering, Filtering, Dynamic modeling, Continuous simulation

ABSTRACT

General objective of the research activities presented in the paper is relied to the modernization of electro-hydraulic equipment for increasing their performances at the level of similar products in Europe and globally, by implementing mechatronic concepts in all stages of design, development, testing, operation and maintenance activities. The paper extensively presents the experimental identification of the equivalent mathematical model for an electro - hydraulic servomechanism, with MOOG servovalve. We focused on integrating the existing and new knowledge and technologies in industrial applications, as well as implementing some innovative aspects.

INTRODUCTION

Mechatronic systems are electromechanical systems where an integrated design procedure is used with respect to mechanical and electrical aspects (De Silva 2004). These systems possess computational intelligence, concepts of sensing and information acquisition being crucial for mechatronic systems in all stages of design, development, testing, operation and maintenance. A realistic characterization of the dynamic performances, done by experimental tests, followed by the calibration of a mathematical model is one essential step in implementing the modern mechatronic concepts in the study of electro-hydraulic servomechanisms (Alexandrescu 2001 & 2002).
All important international companies, such as Bosch-Rexroth, Parker, MOOG supported constantly researches for improving performances of their electro-hydraulic equipment and systems and are actually implicated in important development projects, in collaboration with universities and research centers from Europe, USA, Japan etc. In our country several important companies (HIDROSIB Sibiu, HESPER Bucharest, HIDRAULICA Plopeni, HERVIL Rm. Valcea), having tradition in fluid power equipment, have cooperated with specialized research institutes and universities to modernize their products.
INOE 2000 – Hydraulics and Pneumatics Institute from Bucharest (http://www.fluidas.ro/ihp/) has realized several theoretical and experimental studies for electro-hydraulic components, respective command and control systems, i.e. implementing for series production proportional valves (for 210 bar and 315 bar), servo-valves (frequency 30 Hz, flow rate 240 l/min, pressure 200 bar), servo-systems with electronic modules included (structural and in design).
For illustrating the above mentioned aspects, we propose to extensively present the experimental identification of the mathematical model for an electro-hydraulic servomechanism, with MOOG servo valve (Pachter 1997).
In Figure 1 is presented the experimental set-up, which contains the servo valve SV, the supply pressure being ensured with a pump P and a pressure valve S; load on the output stage of the servo valve is simulated with a variable hydraulic resistance R. The command signal is done by a function generator GF and the measurement equipment contains four pressure transducers TP (one max. 400 bar supply pressure P, 2 max. 160 bar on A and B ports and one max. 25 bar on tank port T), connected to all ports of the servo valve and a flow transducer Q on the input supply chain (with max. 67 l/min). All electrical signals are gathered with a TEST POINT acquisition device SA, having USB interface, 16 analogical inputs (sample rate 100 ks/s), 2 analogical outputs (analogical resolution 12 bytes), 8 digital input/output ports, 2 counters on 24 bytes.
For dynamic measurements of the servo valve exit flow we used the diaphragm method, respectively an equation containing the pressure drop on the variable hydraulic resistance R:

$$Q = k * sign\,(P_A - P_B) * \sqrt{|P_A - P_B|} \qquad (1)$$

where: Q is the exit flow of the servo valve; k – calibration constant determined in steady state conditions with the flow transducer Q; PA and PB are the pressure values in the ports A and B of the servo valve.

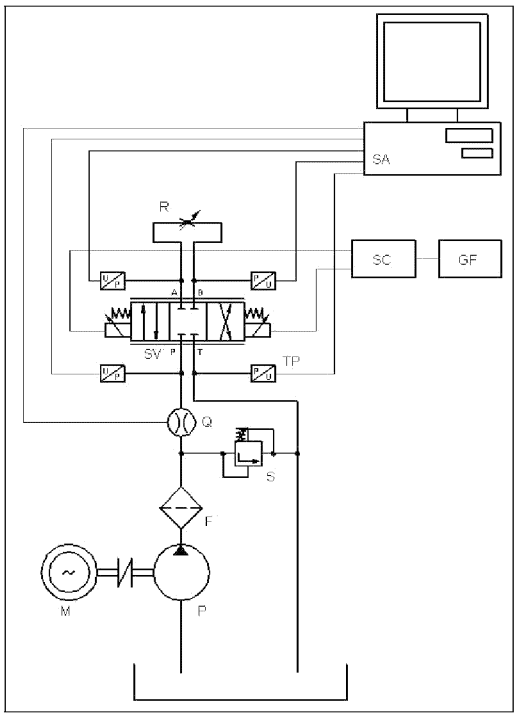


Figure 1: The experimental set-up

For the servo valve, in the paper "Informatics platform for determining the dynamic performances of a two stages servo valve", elaborated by the same scientific group from INOE 2000 – Hydraulics and Pneumatics Institute from Bucharest, the authors experimentally identified the equivalent transfer function, to be used in the model of the studied servomechanism (Marin ISC 2007).

MATHEMATICAL MODEL

The equations specific for the dynamic response of the servomechanism are:
- transfer function of the servovalve (Marin ISC 2007):

$$H_{SV}(s) = \frac{\Delta z(s)}{\Delta i(s)} = \frac{K_{iz}}{\frac{s^2}{\omega_n^2} + \frac{2\zeta}{\omega_n} s + 1} \qquad (2)$$

where: K_{iz} = 8,25 m/A, is the gain current – spool position of the servovalve, z; ω_n = 650 rad/s, is the natural frequency of the servovalve; ζ = 1, is the dynamic damping coefficient.
- continuity equation on the flow domain servovalve-hydraulic cylinder of the servomechanism:

$$A_p \dot{y} + k_{lp} P + \frac{A_p^2}{R_h} \dot{P} = \begin{cases} b_d c_{dd} (|z| - z_0) \operatorname{sign} z \sqrt{\frac{p_S - |P|}{\rho}} & ; \quad |z| > z_0 \\ 0 & ; \quad |z| \le z_0 \end{cases} \qquad (3)$$

where: $b_d = \pi d_{SV} \alpha_d$ is the flow perimeter in the servovalve; d_{SV} = 6 mm is the diameter of the valve spool; α_d = 0,4 – ratio usefulness of entire circular flow perimeter; c_{dd} = 0,6 – flow coefficient in the valve; A_p = 15x10^{-4} m^2 – active surface of the cylinder piston; k_{lp} is the leakage coefficient in the domain servovalve-hydraulic cylinder of the servomechanism; $y, \dot{y}, \ddot{y}$ - position, speed, respective acceleration of the cylinder piston; P – pressure drop across cylinder; R_h – equivalent stiffness of the domain servovalve-hydraulic cylinder of the servomechanism: $R_h = 2\frac{\varepsilon_l}{\vartheta_0} A_p^2$; $\dot{P}$ - time gradient of the pressure drop across cylinder P; ε_l = 10000 bar is the equivalent elasticity of the liquid; ϑ_0 = 0,785x10^{-5} m^3 – oil volume submitted to pressure variations, between servovalve and cylinder; z_0 = 0 – overlap of the valve, in neutral position; P_S – pump stabilized pressure; ρ = 900 kg/m^3 – density of the liquid.

A nonlinearity of the mathematical model is associated to the saturation of the available pump flow, at the value: 63 l/min.
- dynamic equilibrium equation, on the mobile part of the hydraulic cylinder of the servomechanism:

$$m_p \ddot{y} = P A_p - F_f - F_{load} \qquad (4)$$

where: $m_p \cong 4$ kg, is the mass of the mobile part of the hydraulic cylinder; $F_f = F_{fs} + F_{fv}$ is the friction force on the hydraulic cylinder; $F_{fs} = F_{fs0} \operatorname{sign} \dot{y}$ is the Coulomb force; F_{fs0} – absolute value of the Coulomb friction force; $F_{fv} = k_{fv} \dot{y}$ is the viscous friction force $F_{fs} << F_{fv}$, neglecting the Coulomb forces; F_{load} is the load acting on the rod piston of the hydraulic cylinder.
The simulation network corresponding to the mathematical model presented above is presented in Figure 2. To the parameters from equations (2), (3) and (4) we add the specific values of the blocks describing functional chains of the studied servomechanism: position transducer gain $K_{pozitie}$ = 1 m/m; variable constants of the supplementary speed and acceleration reactions K_{viteza} and $K_{acceleratie}$; proportional gain of the PID controller, K_P; conversion factor, specific to the chain command of the servovalve K_{ui} = 0,015 A/V.

EXPERIMENTAL IDENTIFICATION

On the basis of the experimental frequency response of the servomechanism, simulated with the scheme from Figure 2 and applying PEM algorithm implemented in MATLAB-SIMULINK, it was identified the equivalent close loop transfer function (see Figure 3). For the prescribed signal - "Prescris", with frequency of 10 Hz and for time sampling of 1 ms, corresponding to the experimental data acquisition time frame, the expression of the identified transfer function is given below:

$$H(z^{-1}) = \frac{-0.7621 z^6 + 1.381 z^5 - 0.6228 z^4}{z^6 - 2.05 z^5 + 1.144 z^4 + 0.2209 z^3 - 0.4994 z^2 + 0.1219 z + 0.07596} \qquad (5)$$

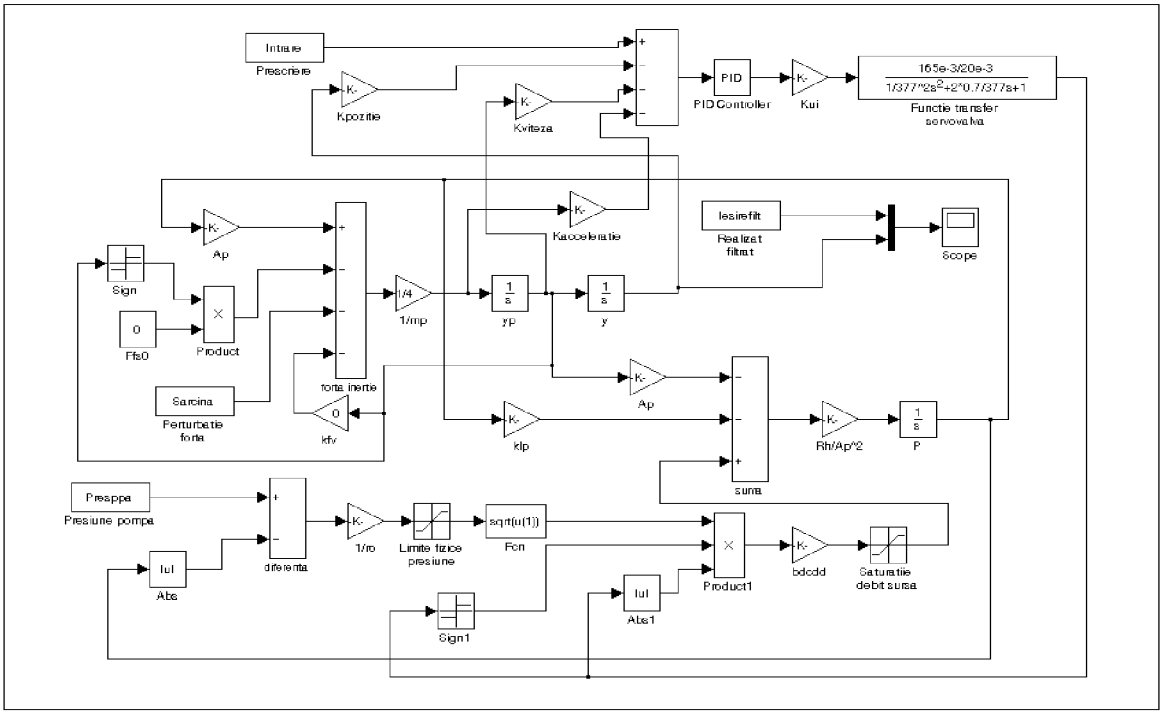


Fig. 2 Simulation network for the studied servomechanism

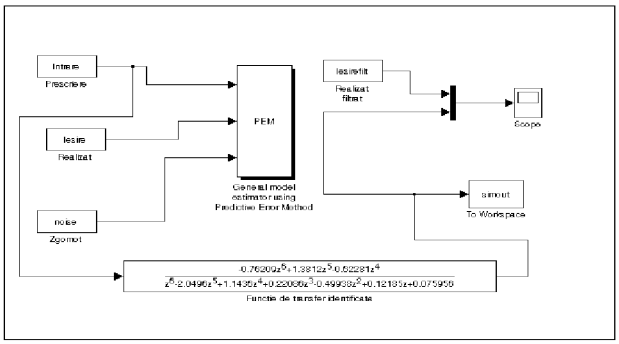


Figure 3: PEM identification method setup

The comparison between experimental and simulated corresponding data is given in Figure 4, the graphics being very close.

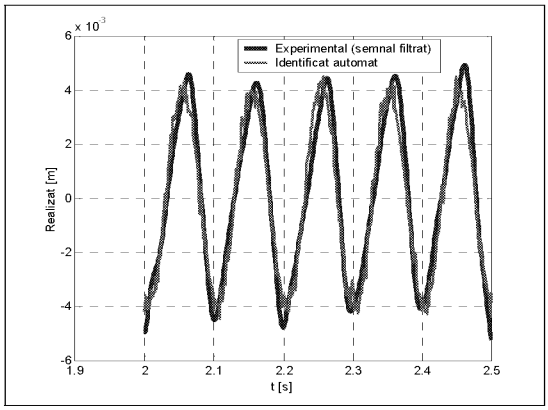


Figure 4: Comparison between experiment and simulation

With the transfer function given in equation (5), by already built commands, it were determined the step response (see Figure 5) and Bode diagrams (Figure 6).

This curves show the following response quality factors:

- transient regime is oscillatory damped, with an important overshoot (48%) and a significant time response (200 ms);
- amplitude margin is Ma_{dB} = - 1,0058, at a frequency of 778,5 rad/s and the phase margin is $M\varphi$ = 4,2535°, at a frequency of 674,8 rad/s;
- bandwidth is given by the frequency ω_B = 32,6 rad/s, corresponding to the absolute value of the transfer function M = 0,707 (M_{dB} = -6,93).

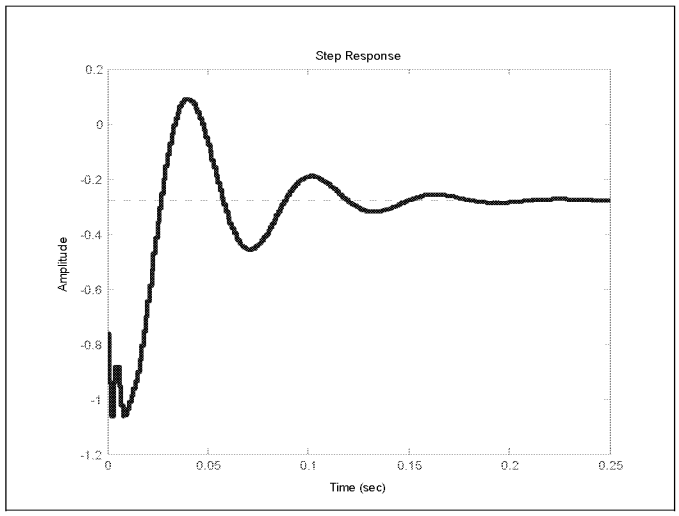


Figure 5: Step response of the servomechanism

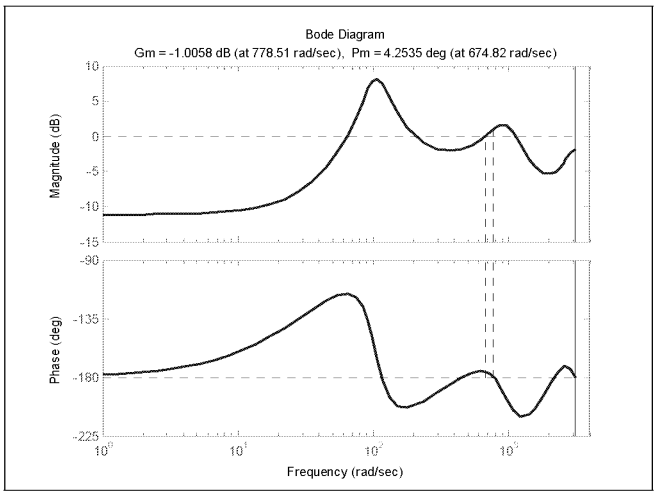


Figure 6: Bode diagrams

CONCLUSIONS

In this stage of the study, we decided upon the set of equipments and performances to be studied, respectively the dynamic performances of an electro - hydraulic servomechanism, with MOOG servovalve. The experimental data served to calibrate an equivalent mathematical model, integrated in a simulation network, more details about the data acquisition process being mentioned in the paper Marin ISC 2007: "Informatics platform for determining the dynamic performances of a two stages servo valve" – ENGIN-03. The perspectives of the present research are relied to the integration of the determined model in the design, development, testing, operation and maintenance procedures of an intelligent displacement unit with hydraulic actuation, specific for robots industry.

ACKNOWLEDGMENTS

The activities presented in this paper were financed by the project, code "CeEx PC-D04-PT13-602", financed by the Excellence Research Program, of the Romanian Ministry of Education and Research.

REFERENCES

Alexandrescu, N. "Mechatronic Positioning System with Hydraulic Actuation", International Conference on Multidisciplinary Design in Engineering, Montreal, Canada, 2001.

Alexandrescu, N. "Control of Hydraulic Unit with Accuracy Actuating on Integrated Mechatronic System", 47 Internationales Wissenschaftliches Kolloquim, Ilmenau, Germany 2002.

De Silva, C.W. "Sensory Information Acquisition for Monitoring and Control of Intelligent Mechatronic Systems", International Journal of Information Acquisition, Vol. 1, No. 1 (2004) 89-99.

Fok, S. C., W. Xiang and F. F. Yap. 2001. "Internet-enabled Virtual Prototyping Workbench for Fluid Power Systems", the Proceedings of ICFP'2001, China, (April).

Habibi, S., A. Goldenberg. "A Mechatronics Approach for the Design of a New High- Performance Electrohydraulic Actuator", International Off-Highway & Powerplant Congress & Exposition, September 1999, Indianapolis, IN, USA.

Marin, Al., P. Drumea, M. Blejan and I. Ilie. 2007. "Informatics platform for determining the dynamic performances of a two stages servo valve", ISC 2007, ENGIN-03.

Martin, D. J. and C.R. Burrows. 1976. "The Dynamic Characteristics of an Electrohydraulic Servovalve," ASME J. Dyn. Syst., Meas., Control, 98, pp. 395–406.

Pachter, M., C. H. Houpis and K. Kang. 1997. "Modeling and Control of an Electro-Hydrostatic Actuator," Int. J. Robust Nonlinear Control, 7, pp. 591–608.

Vilenius, M., A. Ellman, T. Nykänen, M. Pietola, J. Kajaste and H. Kauranne. 2002. "The effect of parameter uncertainty on the reliability of virtual testing of fluid power systems". In: Hakalehto, K. et al. (edited) Proceedings: Mechanical engineering research, new possibilities by co-operation. Tampere: Tampere university of technology, Seminar 3 January 2002. Tampere 2002, p. 115 – 123, 178 p..

Xiang, W., S.C. Fok & F.F. Yap. 2000. "Computational Tools for Fluid Power system Design: Towards distributed artificial intelligence and virtual reality, Journal of Computer Applications in Technology", Vol. 13(6), 2000, pp. 295-301.

BIOGRAPHY

Ph. D. Alexandru Marin – specialist in applied hydraulics, with contributions in experimental and theoretical research on mechatronic systems, automatic electro-hydraulic systems, non-Newtonian automotive hydraulic dampers, lubrication systems, flexible hydraulic tools at 700 bar, solid waste processing machines.

ERROR DETECTION SIMULATION

ON-LINE ERRORS IDENTIFICATION IN MULTI-AXIS MACHINES FOR ADAPTIVE COMPENSATION

Abderrazak El Ouafi and Noureddine Barka
Mathematics, Computer and Engineering Department,
University of Quebec at Rimouski, Rimouski (Quebec)
Canada, G5L 3A1

KEYWORDS

Multi-axis machine accuracy, metrology, calibration, error compensation, laser interferometer.

ABSTRACT

In order to improve multi-axis machine accuracy, error compensation techniques have been widely applied. However, the lack of reliable methods for direct and global measurements implies that all compensation techniques are based on off-line sequential error components measurement. These measurements provide static results, and cannot reflect the actual measurement conditions. Thus, these results are not representative of the real working conditions because of disturbances from thermal distortions and dynamic perturbations. This paper presents an on-line error identification approach for multi-axis machines. Based on the simultaneous measurement of error components, the proposed identification scheme is built to ensure volumetric errors prediction for adaptive error compensation. Implemented on a moving bridge type CMM, the approach led to a significant improvement of the three-dimensional measurement accuracy.

1. INTRODUCTION

The increasing demand for higher dimensional accuracies and the strong trend toward shop floor automation induce the need to develop cost-effective methods to improve the performance of multi-axis machines. The dimensional accuracy is one of the most important factors in determining this performance. Errors affecting a multi-axis machine are defined as the difference between the real and measured position and orientation of the end effectors. To maintain high quality machine performance, these errors have to be detected and eliminated.

In typical machine tools, quasistatic errors are responsible for a very large proportion of the observed total deviation. They account for 70% of the volumetric error (Ragunath 1985). Quasistatic errors vary slowly with time and are related to stationary geometric errors resulting from manufacturing defects, misalignments due to assembly and installation, and thermal errors associated with thermal distortion of machine components due to internal and external heat sources (Okafor and Ertekin 2000). Other errors include dynamic errors resulting from spindle motion, cutting force variations and deflections induced by inertial forces. Obviously, these errors also have an important influence on the machine accuracy. Moreover, whereas some error sources affect machine accuracy directly, others are interrelated with each other and their combined effects cause operational errors. By considering all these diverse and variable error sources, one can understand the difficulties involved in improving multi-axis machine accuracy.

The traditional procedure to provide a relatively error-free multi-axis machine structure is to improve the structural components design and manufacture. This approach is usually uneconomical since it leads to a rapid exponential increase in manufacturing costs. An alternative approach using the concept of reducing the errors by compensation has proven to be more effective in upgrading machine accuracy (Ramesh et al. 2000). Several studies have focused on enhancing the accuracy through on-line monitoring and computer error prediction and correction (Zhang et al. 1985; Donmez et al. 1986).

Various approaches based on quasistatic error analysis, modelling and compensation have been proposed over the last three decades. Some address the modelling problems of the final observed volumetric error in the multi-axis machine workspace by using empirical or analytical models (Belforte et al. 1987; Ferreira and Liu 1986). Earlier studies have applied analytic geometry (Portman 1980; Love and Scarr 1973), vectorial (Schultschick 1977; Hocken et al. 1977) and matrix error representation (Zhang et al. 1985; Donmez et al. 1986; Belforte et al. 1987; Ferreira and Liu 1986). More recent reports describe the use of rigid-body kinematics with homogeneous transformation matrices to model geometrical errors (Ferreira and Liu 1986). Unfortunately, the extensive experimental and analytical efforts required to build an accurate compensation procedure based on the conventional off-line software compensation methods still hinder their commercial implementation.

In the conventional off-line error compensation, the establishment of the volumetric accuracy of a multi-axis machine is usually carried out by evaluating sequentially 21 individual error components and by combining these errors by means of a mathematical model (Fan 2002). This approach is time consuming and requires expensive equipment. Based on the simultaneous measurement of error components, the proposed on-line identification approach is built to ensure error prediction for an adaptive error compensation system that can compensate for not only the geometric errors but also the time-variant errors.

10 of the 21 error components associated with a 3-axis machine are measured on-line (El Ouafi and Barka 2003). The remaining error components are measured by available off-line measurement systems. To illustrate the identification approach, two error models have been developed to synthesize the total error in a three dimensional error form. One model corresponds to the conventional off-line error compensation and the other to the on-line error compensation. The errors predicted by these models are then used for real-time error compensation. Implemented on a moving bridge type CMM, the approach led to a significant improvement of the measurement accuracy.

2. THE PROPOSED IDENTIFICATION APPROACH

In general, a multi-axis machine consists of three mutually orthogonal axes. Each axis of motion is accompanied by six degrees of freedom geometric errors, three of which are translational and three rotational. If the squareness errors between axes are included, there are 21 error components. To compensate for these errors, they need to be characterized. The traditional off-line software compensation technique uses commercially available measurement system as means of calibration. Among these systems, laser interferometer is widely used. However, for a laser interferometer system, it is necessary to set up the optics and align laser path before every measurement. As only one error component can be measured at one setup, much time and effort are also required. In order to measure with less time and effort, diagonal test and circular test such as ball-bar system, have been proposed. For an efficient on-line error compensation system, such measurement approach are no longer applicable because at least the majority of the error components need to be measured in a very short time so that adaptive compensation can be performed. Therefore, new measurement approach is needed, which can measure multiple error components simultaneously.

2.1. Error measurement system

A schematic diagram of the measurement system is shown in figure 1. The system consists of two parts: a moving part fixed on the moving slide to be measured and a stationary part mounted on a reference base where the motion is described. To ensure high measurement accuracy, a high performance He-Ne laser source (LS) is selected with a good beam pointing stability. The laser beam is split into three beams by means of three beam splitters (BS1, BS2, and BS3). The first and third beams are used for straightness and roll measurements. The straightness errors in the vertical and the horizontal planes are measured by the position detector PD2 .The roll error is obtained from the combination of the displacement signals of the PD1 and PD2 and the distance between the two reflecting mirrors (MR1 and MR2). The second beam is used to measure pitch and yaw, using the principle of laser autocollimation.

To illustrate the on-line error measurement, two axes were chosen in the investigation tests. Experimental tests are performed on a moving bridge coordinate measuring machine type. The reference part of the x-axis measurement system is mounted on the granite base of the CMM while the moving part is mounted on the base of the bridge, which moves along the x-axis. The reference part of the y-axis measurement system is mounted on the top of the bridge and the moving part is mounted on the horizontal carrier, which moves along the y-axis. The moving errors in the z-axis are measured using an off-line measurement system. These tests are sufficient to demonstrate the feasibility of the on-line identification for multi-axis machine error compensation.

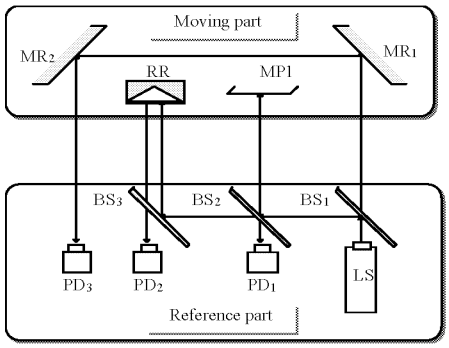


Figure 1 : Layout of the measurement system

2.2. Error components measurement

In order to test the accuracy of the proposed identification approach, on-line measurement systems (OLMS) and a Renishaw laser calibration system (RLCS) are used to measure the error components of the CMM at the same time for comparison. Graphs illustrated in figure 2 show the results of the x-axis measurements. The two systems show good agreement for straightness measurement (less than 1.5 μm) but for angular errors, the difference is approximately 2 arcsec. The divergences between the two systems come from various error sources. The reference positions for the two systems are different and the relatively long travel distance introduces extra noise and drift due to air turbulence on the laser beam. The results for the y-axis illustrated in figure 3 are similar to those for the x-axis. The differences in the y-axis measurements between the two systems can be similarly explained. The results for the z-axis are illustrated in figure 4.

3. THE PROPOSED COMPENSATION SYSTEM

In the proposed error compensation system, three personal computers are used. The CMM is controlled by the first computer P1. At each position in the working zone, the P1 computer collects the current coordinates from the CMM controller and then sends them to the P2 computer. Based on these coordinates, the P2 computer then refers to a database, which includes all the precalibrated error components, in order to retrieve the geometric errors at this position. With these error components, the off-line error model is used to calculate the errors at the probe stylus tip. In the meantime, the P2 computer also collects data from the on-line measurement system for the ten error components via a data acquisition system. Based on these measured error components and the remaining error components retrieved from the database, the on-line error model is used to calculate the errors at the probe stylus tip

again. The errors obtained from these two models are then used to correct the coordinates. The RLCS is controlled by the P3 computer.

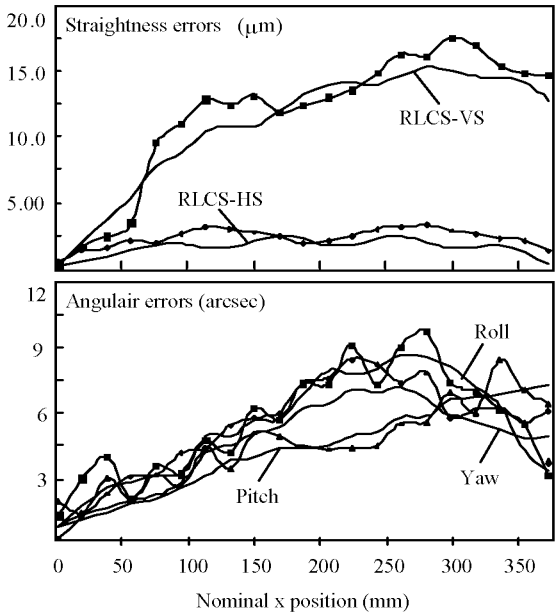


Figure 2 : Results of the X-axis error measurement

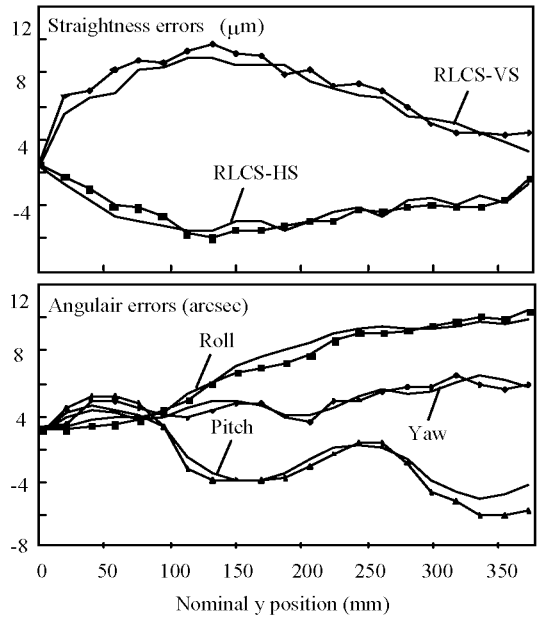


Figure 3 : Results of the Y-axis error measurement

To achieve the correction, two models must be developed to synthesize the measured errors components and thus estimate the final position errors at a concerned position such as a probe stylus tip. Assuming the machine as a rigid body and using the homogeneous coordinate transformation method (HCTM), the well-known 21 error components, which include six error components for each axis and three squareness errors between axes, can be used to represent the spatial geometry of the machine.

To find the error model for the off-line compensation, we have to define a series of coordinate frames. Applying the HCTM to the CMM under the assumptions of small angular motions, we obtain a series of matrices, each describing one coordinate frame with respect to another. The HTM of the probe stylus tip with respect to the reference frame is given by the multiplication of this series of matrices. From the position vector of the probe, we can obtain the coordinates of the probe stylus tip and by subtracting the nominal coordinates of the probe tip, we can have the coordinate errors.

For the on-line compensation system, we need to take into account the fact that the measurement systems are mounted at positions different from those at which the machine calibration was performed. This positional offset will not cause any effect for rotational errors under the rigid-body assumption but it will cause differences for translational errors, which are proportional to the rotational errors and the offset. Using extra coordinate frames, the HCTM can be used again to find the coordinates of the probe stylus tip and the coordinate errors.

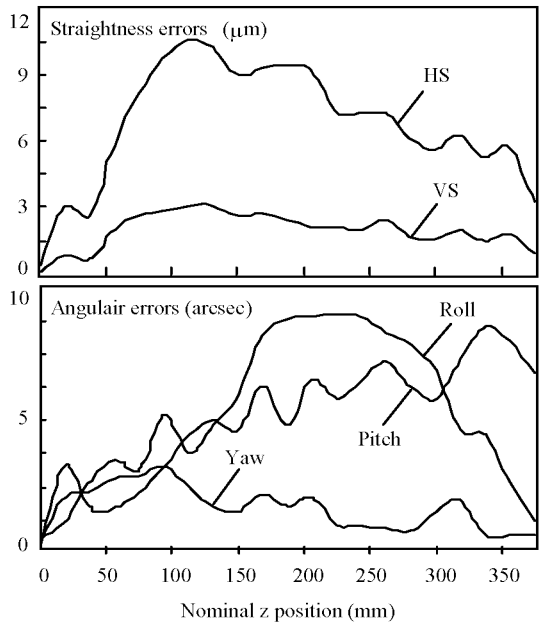


Figure 4 : Results of the Z-axis error measurement

In order to evaluate the performance of the adaptive error compensation system, diagonal tests are designed using the RLCS as the reference. In each test, the CMM is controlled to move in specific plan along the diagonal. The RLCS system is set to measure the travel distance of the machine along the diagonal. The total travel distance along each diagonal is 500 mm. The CMM is controlled to move by 20 constant steps. At each step position, the P1 computer collects the current coordinates from the CMM controller, the P2 computer collects data from the on-line measurement system and the P3 computer collects the coordinates and the distance reading from the RLCS, and then sends them to the P2 computer where the errors are calculated using the off-line and the on-line models. These errors are used to correct the travel distance. The off-line and the on-line corrected distances are then compared with the reading from the RLCS. The P1 computer then moves the CMM to the next position along the diagonal and the same procedure is repeated until the final position is reached. The residual errors of the compensation system are represented by the differences between the corrected distances and the reading from the RLCS.

Figures 5, 6, and 7 present typical results of the diagonal tests for the xy-plan, the yz-plan and xz-plan respectively. Each figure shows the uncompensated result, the compensated result using the off-line error model, and the compensated result using the on-line error model. As can be seen from the figures, the maximum volumetric residual errors for the off-line model are approximately 2.50, 2.95 and 1.75 μm for the xy-plan, the yz-plan and xz-plan respectively. In contrast, the maximum residual errors for the on-line model are 1.30, 2.00 and 1.60μm. These results indicate significant improvement of the three-dimensional measurement accuracy.

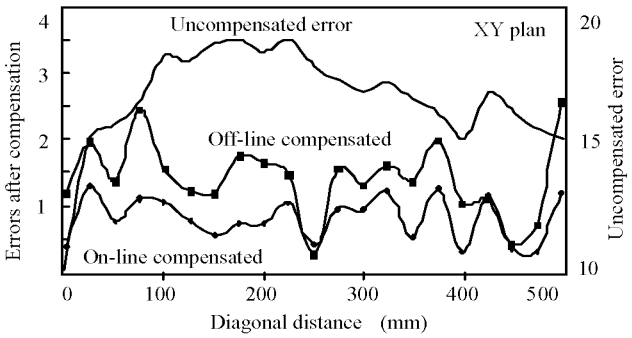


Figure 5 : Results of the compensation test in the xy-plan

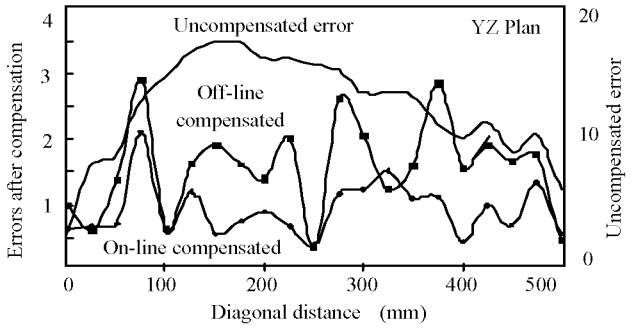


Figure 6 : Results of the compensation test in the yz-plan.

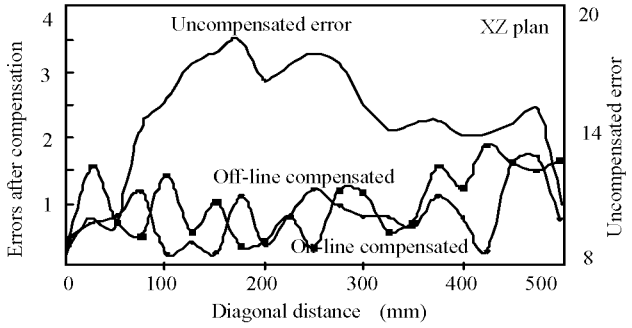


Figure 7 : Results of the compensation test in the xz-plan

4. CONCLUSION

This paper presents an on-line identification approach of volumetric errors in multi-axis machines for error compensation. Based on the measurement of multiple error components simultaneously, the proposed identification approach is built to ensure errors prediction for an adaptive error compensation system that can compensate for not only the geometric errors but also the time-variant errors. The measurement system evaluates 10 of the 21 error components of the machine and the remaining error components are measured by off-line methods. Two mathematical models were developed to synthesize the error components. Implemented and tested on a moving bridge type CMM, the diagonal test results demonstrated that by using the on-line identification approach, the performance of the compensation system presents significant improvement as compared with using the conventional off-line method.

5. REFERENCES

Belforte G., Bona B., Canuto E., Donati F., Ferraris F., Gorini I., Moreti S., Peisino M. and Sartori S., 1987, "Coordinate measuring machines and machine tools selfcalibration and error correction", *Ann. CIRP, 36 (1).*

Donmez M. A., Blomquist D. S., Hocken R. J., Liu C. R. and Barash M. M., 1986, "A General Methodology for Machine Tool Accuracy Enhancement by Error Compensation", *Precision Engineering, 8 (4).*

El Ouafi A. and Barka N., 2003, "Systeme de mesure en temps reel des erreurs dans les machines-outils pour une compensation active", *4th International Conference on Industrial Automation, Montreal, Canada.*

Fan J. W., Guan J. L., Wang W. C., Luo Q., Zhang X. L. and Wang L. Y., 2002, "A universal modeling method for enhancement the volumetric accuracy of CNC machine tools", *Journal of Materials Processing Technology, 129 (1-3).*

Ferreira P.M. and Liu R.C., 1986, "A Contribution to the Analysis and Compensation of the Geometric Error of Machining Center", *Ann. CIRP, 35 (1).*

Hocken R. J., Simpson J. A., Borchardt B., Lazar J., Reeve C. and Stein P., 1977, "Three Dimensional Metrology" *Ann. CIRP, 26 (2).*

Love W. J. and Scarr A. J., 1973, "The Determination of the Volumetric Accuracy of Multi-Axis Machines", *MTDR Proceeding Conference, Vol. 14.*

Okafor A. C. and Ertekin Y. M., 2000, "Vertical machining center accuracy characterization using laser interferometer : Part I - Linear errors" and Part II - Angular errors", *Journal of Materials Processing Technology, 105 (3).*

Portman V. T., 1980, "Error Summation in the Analytical Calculation of Lathe Accuracy", *Machines Tooling, 51 (1).*

Ragunath V., 1985, "Thermal Effects on the Accuracy of Numerically controlled Machine Tools" *Ph. D dissertation, Perdue University, West La fayette, Indiana.*

Ramesh R., Mannan M. A. and Poo A. N., 2000, "Error compensation in machine tools - a review : Part I - geometric, cutting-force induced and fixture-dependent errors, and Part II - thermal errors", *International Journal of Machine Tools and Manufacture, 40 (9).*

Schultschick R., 1977, "The components of volumetric Accuracy", *Ann. CIRP, 25 (1).*

Zhang G., Veale R., Charlton T., Borchardt B. and Hocken R., 1985, "Error Compensation Of Co-ordinate Measuring Machines", *Ann. CIRP, 34 (1).*

ON-LINE SIMULATION BASED FAULT DETECTION FOR A HYDRAULIC DRIVEN INDUSTRIAL ROBOT

C. Angeli
Department of Mathematics and Computer Science
Technological Institute of Piraeus
Konstantinoupoleos 38, N. Smirni
GR-171 21 Athens, Greece
E-mail: c_angeli@otenet.gr or
c.angeli@computer.org

A. Chatzinikolaou
S. Patsi 62
GR-118 55 Athens, Greece
E-mail: achatzi@otenet.gr

KEYWORDS

Modelling, Simulation, Hydraulic systems, Fault detection

ABSTRACT

The development of on-line simulation based methods for drive and control systems are of main importance for the modern production technology. This paper presents a fault detection scheme that has been applied to a hydraulic driven three axes industrial robot. For this purpose a mathematical model for the drive system of the robot is developed. The simulation results and the measured quantities from the actual system are compared on-line in a suitable environment. The final fault diagnosis process is achieved by interaction of modelling information, on-line measured values of the system variables and stored knowledge.

INTRODUCTION

Model-based fault detection methods of industrial processes have received a lot of theoretical and practical attention in recent years due to their importance in the automation processes. These methods could increase the operational efficiency and reduce the maintenance costs of important technical systems. Several papers have been published for the implementation of model-based methods in automated industrial processes such as (Patton et al. 2000, Basseville 2003, Gertler 1998, Isermann and Balle 1996, Mangoubi and Edelmayer 2000).

Hydraulic systems are used as drive and control systems in a wide range of industrial applications from precision control systems for machine tools to heavy industrial production machines and processes, due to their compactness and the ability to transmit and manage high forces. The implementation of model based fault detection methods to hydraulic systems improves the reliability and performance of these systems. Common faults in hydraulic drive systems include worn pumps and actuators. These faults are often reflected in leakage and increased temperature of the hydraulic medium, loss of efficiency, noise and environmental contamination.

In previous research work for the fault detection process of hydraulic components the pressure signal analysis (Angeli and Atherton 2001, Randall 2002, Mitchell 1991), the vibration signal analysis (Liu and Huang 1991, Gao et al. 2003, Edwards et al. 1998) or the measurement of the components leakage (Chen and Lu 1995, Angeli and Chatzinikolaou 2005) have been used as information source depending on the nature of the specific hydraulic components. In this work the signals of the pressure in specific parts of the circuit, the angular velocity signal of the hydraulic motor as well as digital input signals that indicate the state of the components have been used cooperatively.

The hydraulic motor-mass-pipes drive system of the robot axes was modelled using mathematical equations. Real-time data are acquired from the actual system and compared with the relevant variable values of the simulation process referring to the pressure and the angular velocity signals. The parameter values of the equations were obtained from data given by the manufacturer of the components, after laboratory testing, so that the model represents the behaviour of the motor as accurately as possible. The DASYLab data acquisition and control software was used for the data acquisition process as well as for the implementation of the comparison process between measured and simulated data and the abnormality detection in first stage, before the final diagnostic conclusion by the expert system.

MODELLING OF THE DRIVE SYSTEM

The drive system of robot consists of three hydraulic motors, a power supply unit and a control block. The task of the hydraulic system is to move the hydraulic motors and the attached loads. The hydraulic motors used are fixed displacement low speed motors with attached rotary encoders. Figure 1 shows one of these motors.

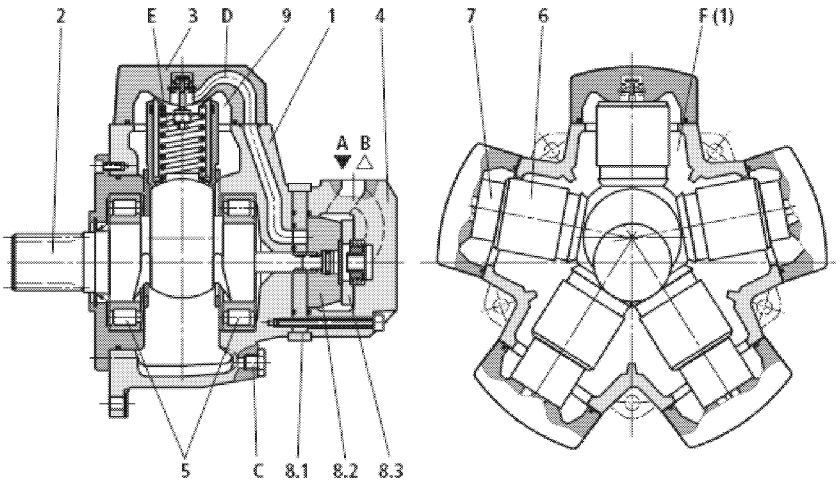


Figure 1. Hydraulic motor (Rexroth)

The function diagram represent the operating curves of the motor for various values of the pressure drop Δp. From these curves the pressure drop in bar for a given load in kW as well as the actual flow in l/min for a demanded rotation speed in min^{-1} can be found. The flow curves, Figure 2, can be used for the estimation or validation of the volumetric efficiency of the hydraulic motor.

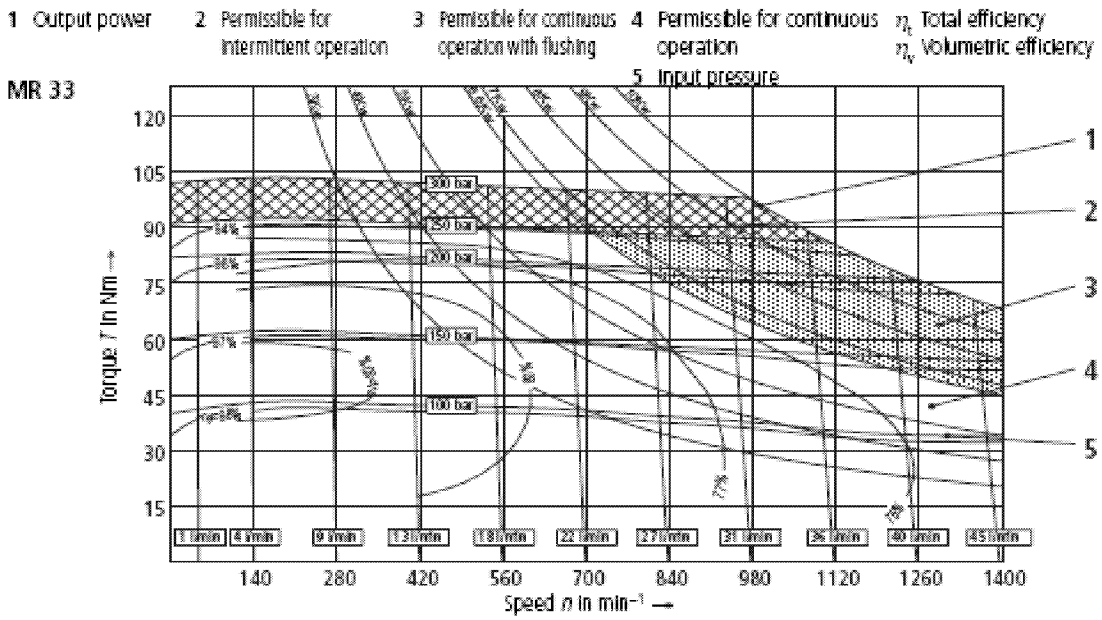


Figure 2. The function curves of the hydraulic motor

Figure 3 shows a typical hydraulic scheme of the hydraulic motor (2) of a robot axis, controlled by a proportional 4-way valve (1). The hydraulic motor must be able to accelerate and rotate the attached mass with a given speed. The rotation speed of the hydraulic motor depends on the actual flow rate through the motor and on its displacement volume. The pressure valves 3 and 4 serve as safety valves in the case of overload or emergency stop.

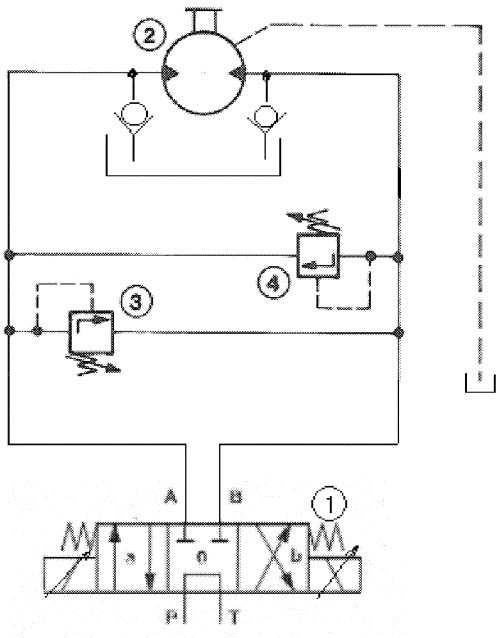


Figure 3. The Hydraulic motor in the hydraulic system.

For the fault detection process a model of the hydraulic motor-mass system was developed. In this model physical relations and relations derived from the technical specifications provided by the manufacturer were used.

For the hydraulic motor with the volumetric efficiency η_v the inflow and outflow are given by:

$$q_{ma} = \left(\frac{C_m}{\eta_v}\right) \cdot \dot{\varphi}$$

$$q_{mb} = C_m \cdot \dot{\varphi}$$

where:

$$C_m = \frac{V_m}{2 \cdot \pi}$$

Each hydraulic motor has to move and accelerate the mass of the corresponding axis. The application of the Newton law gives the following relationship:

$$\frac{d^2\varphi}{dt^2} = \left[\left[\frac{V_m \cdot (p_a - p_b)}{2 \cdot \pi}\right] - M_r\right] \cdot \frac{1}{J_m}$$

where: V_m Displacement volume of the hydraulic motor

M_r Mechanical friction torque

J_m Moment of inertia of the moving parts of the machine related to the motor axis

Pa, pb Pressure at the motor ports

The volumetric efficiency is a parameter that indicates the loss of the motor speed in relation to the theoretical motor speed that corresponds to the oil flow to the motor. This loss is caused by the natural internal leakage of the hydraulic motor. The volumetric efficiency is the result from the division of the actual motor speed by the theoretical motor speed and is given by the relationship:

$$\eta_v = \frac{q_{mb}}{q_{ma}} = \frac{n_{actaul}}{n_{theoretical}}$$

where $$n_{theoretical} = \frac{q_{ma} \cdot 1000}{V_m}$$

and n_{actual} the actual motor speed in r/min

with q_{ma} the flow to the inlet of the hydraulic motor in [l/min]

and V_m the motor displacement in [cm^3]

The volumetric efficiency can be derived from the function curves of the hydraulic motor or from the measurements of the actual speed and the flow through the hydraulic motor for various Δp values near the operating pressure (validation).

For the validation process, the actual hydraulic system of the robot was connected through sensors with a data acquisition system and the acquired data were compared

with the simulation results. The measured quantities are the pressures at critical points of the hydraulic system and the angular velocity.

Values from the simulation and from measurements that refer to the pressure at the port A of the hydraulic motor are illustrated in Figure 4 where it can be seen that the experimental results and the simulation results have a high degree of consistency.

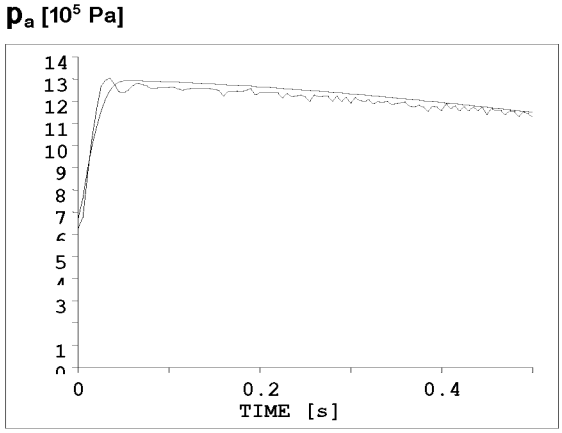


Figure 4. Comparison of measured and calculated results.

FAULT DETECTION

The fault detection process is performed on-line after comparison of measured and calculated data by the expert system. In this process are involved measurement information, modelling information and stored knowledge derived from the expert. In the case where the fault can be detected after a comparison of the measured values with the simulation results, the comparison result is translated to symbolic information which is used by the knowledge base of the system. This information is passed in the structure of the knowledge representation scheme and can activate specific sets of rules for the final decision.

The input to the actual system is the flow to the hydraulic motor that corresponds to a voltage signal U to the proportional valve and the outputs which are fed to the expert system are the angular velocity ω, the pressures p_a, p_b, at the ports of the hydraulic motor and the state signals from the devices of the power unit.

The comparison process of measured and calculated values of the angular velocity is presented in the DASYLab diagram of Figure 5.

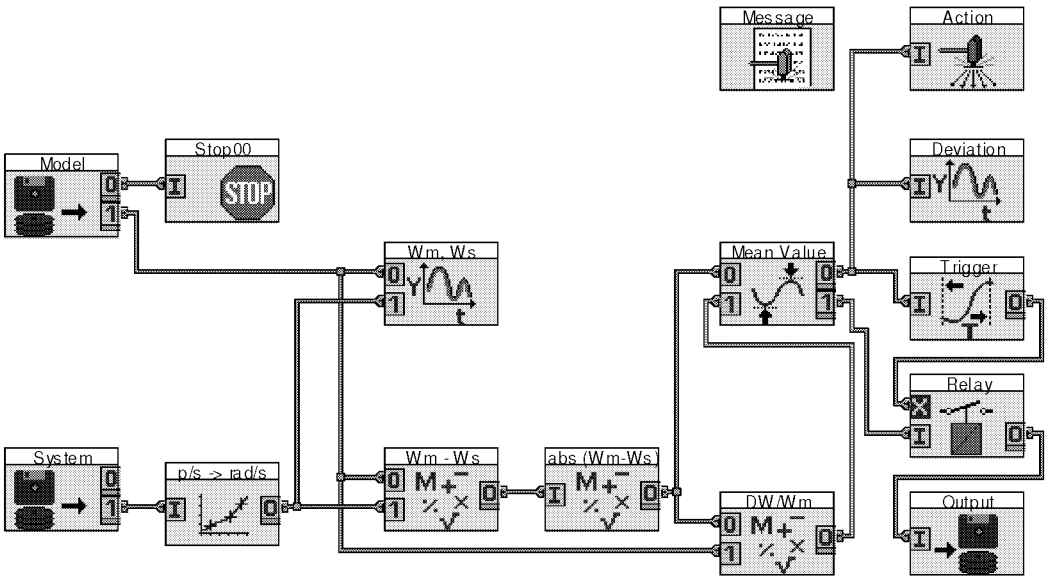


Figure 5. DASYLap diagram for the comparison of measured and calculated data

In Figure 5 the module "Model" contains the numerical data from the simulation process while the module "System" contains the measured values produced by the data acquisition. The module "p/s -> rad/s" transforms pulses per second to rad per second. The module "Wm - Ws" calculates the difference of the angular velocity ω from model and system and the module "abs(Wm-Ws)" calculates the absolute value of the differences. The module "DW/Wm" calculates the values Δωs/ωm. The module "Mean Value" calculates the mean value that corresponds to the steady state of Δωs/ωm. The module "Trigger" switches ON the module "Relay" if activated. The module "Relay" lets the data stream coming from the "Mean Value" module pass to the "Output" file module that writes the mean value of the last block of the Δωs/ωm values to the output file. The "Action" module initiates an action when the maximum error of the absolute difference of the angular velocity from model and system in the steady state exceeds a specific threshold. The "Message" module is used to display output messages received from the "Action" module as "ok", "med", "high", referring to the difference of the angular velocity ω, and to write them to a file for further processing by the expert system.

CONCLUSION

The development of on-line model based fault detection procedures for technical systems improve the operational reliability of the system and offer a higher degree of safety in industrial processes. In this paper, modelling information was used in comparison with on line measurements referring to the pressure signals as well as to the angular velocity signal for the detection of abnormal behaviour of the hydraulic drive system of an industrial three axes robot. The comparison results were evaluated by an expert system for the final diagnostic conclusion. This method enables quick and effective detection of process abnormalities and triggers fault diagnostic activities. The experimentation results show that the diagnosis process is reliable and the method is applicable to real world systems.

REFERENCES

Angeli, C. and D. Atherton. 2001. "A model-based method for an online diagnostic knowledge-based Systems." *Expert Systems,* Vol. 18, Nr. 3, 150-158.

Angeli, C. and A. Chatzinikolaou. 2005. "Model-based Fault Detection for Hydraulic Motors." *In Proceedings of the 24th IASTED International Conference Modelling, Identification and Control MIC 2005*, February 10-13, Innsbruck, Austria, 333-339.

Basseville, M. 2003. "Model-based statistical Signal Processing and Decision Theoretic Approaches to Monitoring." *In Proceedings Safeprocess 03*, Washington, U.S.A. 1- 12.

Chen, Z. and Lu Y. 1995. "State Monitoring and Fault Diagnosis for Hydraulic systems." *Pneumatics and Hydraulics*, 2, 3-7.

Edwards S., A. Less and M. Friswell. 1998. "Fault diagnosis of rotating machinery, Shock and Vibration." Digest, 30, 4-13.

Frank, P.1996. "Analytical and qualitative model-based fault diagnosis: A survey and some new results." *European Journal of control*, Vol. 2, 6-28.

Gao, Y.; O. Zhang and X. Kong. 2003. "Wavelet-based Pressure Analysis for Hydraulic Pump Health Diagnosis." *Transaction of ASAE, Power and Machine Division*, Vol 46(4), 969-976.

Gertler, J. 1998. *Fault Detection and Diagnosis in Engineering Systems*, New York, Marcel Dekker.

Isermann, R. and P. Balle 1996, "Trends in the Application of Model Based Fault Detection and Diagnosis of Technical Processes." *IFAC World Congress 1996*, San Francisco.

Liu, S. and C. Huang. 1991. "Pulsating parameter method for fault diagnosis for a hydraulic pump." *In Proceedings IEEE International Conference on Control*, Piscataway, N. J. 2145-2150.

Mangoubi, R.S. and A. M. Edelmaye.r 2000. "Model based fault detection: the optimal past, the robust present and a few thoughts on the on the future." *In Proceedings Safeprocess 00*, Budapest.

Mitchell, J. 1991. "Research into a sensor-based diagnostic maintenance expert system for the hydraulic subsystem of a continuous mining machine." *IEEE Transactions on Industry Application,* 5, 1192-1199.

Patton, R.; P. Frank and R.Clark. 2000. *Issues in Fault Diagnosis For Dynamic Systems,* Springer-Verlag

Randall, R. 2002. "State of the art in monitoring rotating machinery." *In Proceedings ISMA 2002*, 4, 1457-1477.

ROBOT NAVIGATION SIMULATION

HYBRID NEURO-FUZZY SYSTEM FOR MOBILE ROBOT REACTIVE NAVIGATION

Munaf S. N. Al-Din
Deaprtment of Electrical Engineering
Tafila Technical University
Tafila - Jordan
msnaldin2001@yahoo.com

Ahmed A. Abdullah
Deaprtment of Electrical Engineering
Tafila Technical University
Tafila - Jordan
ahmed_samarai6@yahoo.com

KEYWORDS
Moble Robot, Neuro_fuzzy, Navigation.

ABSTRACT

This paper addresses the problem of mobile robot autonomous navigation in a non structured environment. The objective is to make the robot move along a collision free trajectory until it reaches its target. The approach taken here utilizes a hybrid neuro-fuzzy technique where the inference engine of a classical fuzzy system is replaced by a collection of five parallel neural networks in order to reduce computational time for real-time applications. The five neural networks were trained using data sets randomly selected from the original fuzzy decision matrix. Simulation results were conducted to test the performance of the developed system and the results proved that the proposed approach to be practical for real time applications. Finally, the developed neuro-fuzzy controller was tested on a prototype mobile robot which was designed and constructed as part of this research project.

INTRODUCTION

Among all the soft-computing methods suggested for mobile robot reactive navigation, fuzzy logic systems have been found to be the most attractive. They are tolerant to noise and error in the sense of information coming from the sensory system, and most importantly they are factual reflection of the behavior of human expertise. In general, there are two approaches to the application of fuzzy logic in mobile robot navigation, namely, behavior-based approach (Aguirre E. and González A 2000, Lee C. C. 1990 and Saffiotti A. 1997) and classical fuzzy rule-base approach (Proychev T. and Petrov M. 1996, Ross T.J. 1998 and Zadeh L. A. 1996). The design of fuzzy logic rules is often reliant on heuristic experience and it lacks systematic methodology. Therefore these rules might not be correct and consistent, do not possess a complete domain knowledge, and/or could have a proportion of redundant rules. Furthermore, when a better precision is needed the number of input variables and their fuzzy values need to be increased, for example, when using four input variables each mapped by seven fuzzy values besides 2401 if–then rules maybe required to define the rule-base of the inference system. Such huge expansion in a multi-dimensional fuzzy rule-based system adds further ad hoc to the design of the system (Tanaka K. 1996).
Several successful reactive navigation approaches based on neural networks have been suggested in the literature (Callan R. 1999, Carpenter G. et al 1992 and Fausett L. 1994). In spite of various suggested network topologies and learning methods, neural reactive navigators still perceive their knowledge and skills from demonstrating actions. Therefore, they suffer from a very slow convergence, lack of generalization due to limited patterns to represent complicated environments, and finally information encapsulated within the network can not be interpreted into physical knowledge (Hagras H. and Sobh T. 1998). Consequently, the utilization of neural networks in reactive mobile robot navigation is limited when compared to fuzzy logic. However, the role of neural networks has been found to be very useful and effective when integrated with fuzzy systems (Cao J., Liao X. and Hall E. 2000 and Yung N. H. C. and Ye C. 1999). The birth of this integration between these two soft-computing paradigms is the neuro-fuzzy systems. Neuro-fuzzy systems provide an urgent synergy that can be found between the two paradigms, especially the capability to mimic human experts as in fuzzy logic, and learning from previous experience capability as in neural networks.
In general, neuro-fuzzy systems can be classified into three categories, neurally adaptive fuzzy inference system, neurally performed FIS, and combinatorial, or hybrid, neuro-fuzzy systems. The neurally adaptive fuzzy inference system is the most widely used neuro-fuzzy systems, and they are designed to combine the learning capabilities of neural networks and reasoning properties of fuzzy logic (Medsker A. R. 1995).
In this paper, a new approach is proposed to design a simple hybrid neuro-fuzzy navigation system. The proposed system has two apparent advantages in structure that simplify and reduce the processing time and improve the performance. First, the if-then rule base is replaced by a set of simple neural networks. However, the second one is the five parallel simple neural networks are utilized to replace the fuzzy inference system acquired by the robot's sensory system. With such technique, the required time needed to infer the decision for the robot movement is greatly reduced.

THE PROPOSED NEURO-FUZZY NAVIGATION SYSTEM

The mobile robot is required to explore several paths in a maze, of a pattern of successive combinations of left and right turns. Its task is to reach a desired position at the end of one channel. The mobile robot uses a kind process, sequentially adopting cyclic pattern of the left and right turns. Eventually, it ends up with the desired position, at which time a signal is injected, causing the robot to record the correct pattern. The mobile robot is assumed to be

equipped with three physical ultrasonic sensors and one virtual sensor as shown in Figure 1.
The physical sensors are used to detect obstacles in front of the robot, the right side, and the left side, respectively. The maximum distance that can be sensed by these sensors is assumed to be 6 meters. The virtual sensor is used to guide the robot towards the target. This sensor is especially needed when the target direction of movement is totally blocked by an obstacle. The virtual sensor will guide the robot back towards the target once the obstacle is avoided.

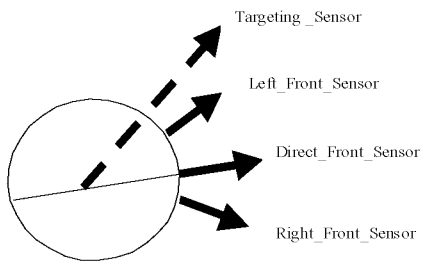


Figure 1: Mobile Robot with Frontal Sensors

Henceforth, the robot travels quickly and accurately along the track to accomplish any job that has been assigned. It is assumed that the robot will not face any traps (or get into a situation where it is required to backtrack or turn around). Such a problem is out of this paper scope.
The four sensors provide the path planning system (in our case a fuzzy logic system) of the robot with three distances front (dc), right (dr), left (dl), and target orientation (theta), respectively. From these inputs, the fuzzy logic controller will make up a decision in which direction should the robot move in order to reach the target. The fuzzy logic controller should pass through three stages, i.e., fuzzification, inference, and defuzzification as shown in Figure2.

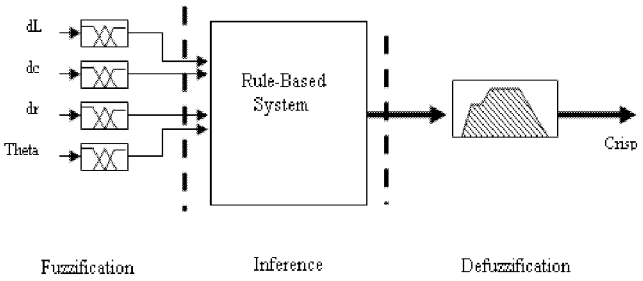


Figure 2: Fuzzy Logic Controller Stages

Fourty Rules Fuzzy Navigator system

The fuzzy logic controller (FLC40) was analyzed and tested for different cases based on the same parameters and rules used by (Xu W.L., Tso S.K. 1996). The robot motion results have been considered with relation to different cases. Problems were recorded and investigated and the reasons behind the failure of this robot, in these cases, were related to the limited number of the sets used (FAR, NEAR), and the limited angle of orientation (turning angle), which are five sets. Due to this limitation, the robot touches the obstacles slightly in all cases considered as shown in Figure 3. To avoid these problems, a relaxation of the rules was done by increasing the number of sets for the input distances from two to five sets, accordingly, the number of rules was increased to 625 activation rules which will be discussed in the next section.

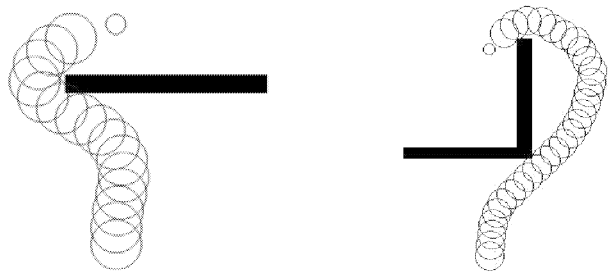


Figure 3: FLC40, simulated motion

Development of the improved Fuzzy Navigator system

As it has been already noted, that the FLC40 is not capable to avoid collision with the edges of the obstacles in all cases. The main reason behind that failure is the low resolution due to two fuzzy sets, i.e., FAR and NEAR. An improvement to the system can be easily made by increasing the number of fuzzy sets in order to achieve better resolution. In this paper, it is proposed to increase the fuzzy sets to five linguistic labels (VL, L, M, S, VS) as shown in Figure 4, 5 and 6. The fuzzy sets in this case become shorter than before, so the accuracy and the performance of the controller are improved. As the number of sets is increased the fuzzy rules are increased as well up to 625 activation rules ($5 \times 5 \times 5 \times 5 = 625$ activation rules).

As an example, a sample is presented where the activation rules are:

IF dr is VL and dc is VL and dl is VL and tr is LB THEN Sa is TLB
IF dr is M and dc is L and dl is VL and tr is RB THEN Sa is TRS
IF dr is S and dc is VL and dl is S and tr is RS THEN Sa is TZ

The results obtained from this improved fuzzy logic controller have been improved. The robot avoids collision with the obstacles as shown in Figure 7, but the main problem in using that improved controller is the processing time. It is very long, since the number of rules is high and requires more time to create a decision and this will affect the response time of the robot.

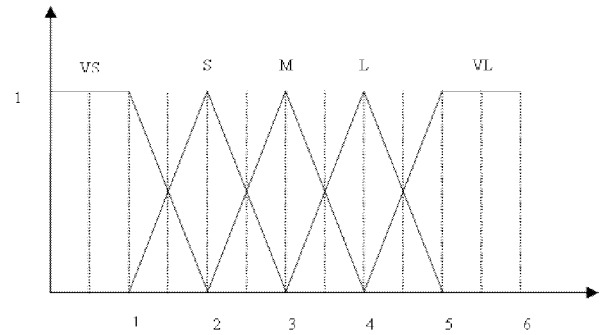


Figure 4: Fuzzy Set Definition for the Four Input Distance Variables
(VL: Very Large,L: Large, M: Medium, S: Small, VS: Very Small)

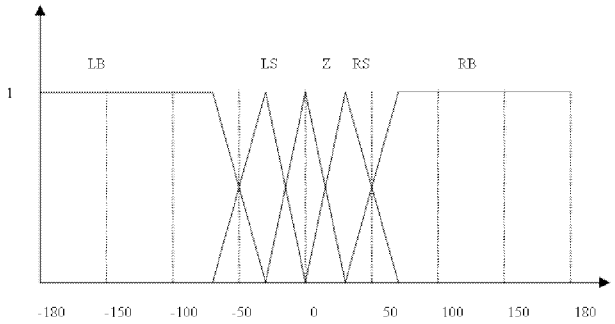


Figure 5: Fuzzy Set Definition for the Input Target Orientation Variable
(LB: target_ Left Big,LS: target_ Left Small, Z: target_Zero, RS: target_ Right Small, RB: target_ Right Big)

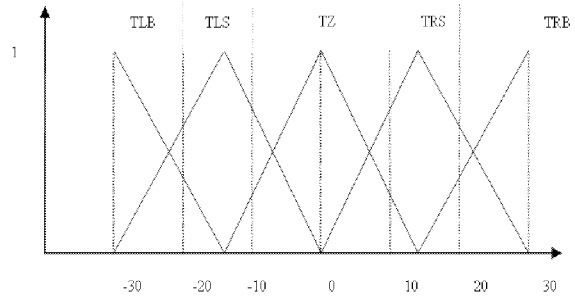


Figure 6: Fuzzy Set Definition for the Output Turning Angle Variable

(TLB: Turn_ Left _Big, TLS: Turn_ Left _Small, TZ: Turn_ Zero, TRS: Turn_ Right _Small, TRB: Turn_ Right _Big)

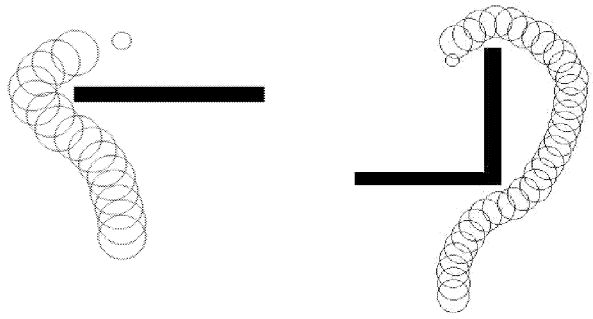


Figure 7: Improvement FLC625 success cases

DEVELOPMENT OF THE NEURO-FUZZY NAVIGATION SYSTEM

The main problem in the fuzzy logic controller is the inference block, which consists of a large number of rules that need a long processing time. To solve this problem of processing time, the inference engine was replaced with a neural network. The system is investigated by considering the results of the integration between both systems (Fuzzy logic and neural networks) as shown in Figure 8.
The outcome has been efficient and accurate but it requires training, which was introduced for that system. To perform the training process, the sample turns out to be very large (83521 × 20) and the system faces problems. To overcome that huge sample, the NN was structured as five parallel networks where each one of this network has 20 fuzzy input nodes and one fuzzy output node as shown in Figure 8. This new structure has improved the performance and the response time. The main idea in this neuro-fuzzy system is the replacement of inference engine by neural networks where it has a fuzzy inputs and a fuzzy output.

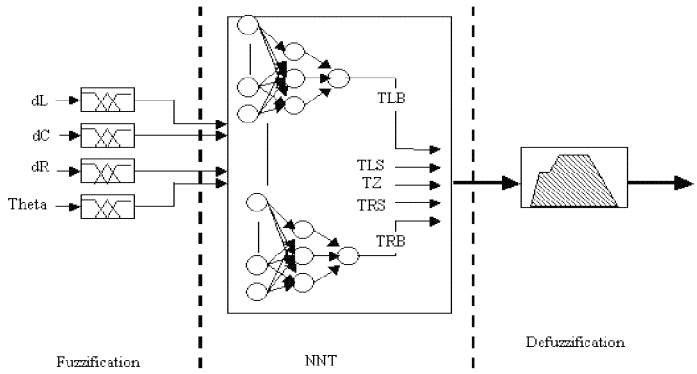


Figure 8: Hybrid Neuro- Fuzzy Architecture

The performance of the Hybrid neuro- fuzzy controller is the same as the improved fuzzy logic controller. The neural network in this controller is trained to do the same action as the inference engine in the improved fuzzy logic controller as shown Figure 9.

The main advantage gained by utilizing the hybrid neuro-fuzzy controller is the reduction in the inference time to about one fourth of the time required when using classic FLC, which increases the response of the controller and improves the performance of the robot. Practically, simulation-using PC doesn't show the differences in the CPU time for the three controllers since the PC is very fast and the response of the hardware is slow. The CPU time for the three controllers is noticed when using micro-controller chip to control the robot motion and download the program to the implemented robot. In the FLC40, the controller response time will be faster than both controllers, but the performance is limited. On the other hand, the FLC625 worked out well but with low response, which introduced a deficiency in the robot motion (create a dead point in the robot controller). The neural network which programmed in

five chips, the data of the main micro-controller entered to the five parallel NNT and this increased the response of the whole controller and improves the performance of the robot motion.

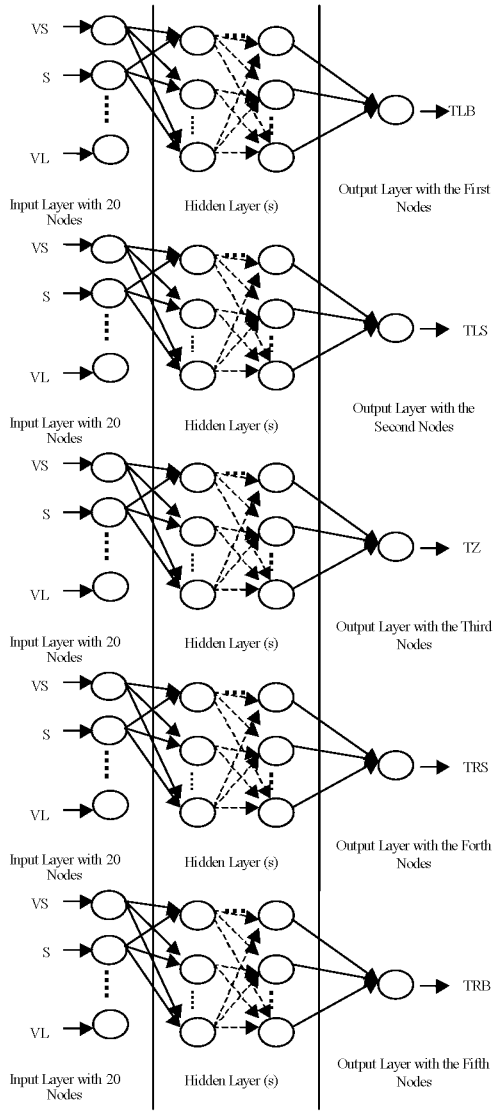


Figure 8: Five Neural Network with Parallel Inputs and Five Separated Outputs

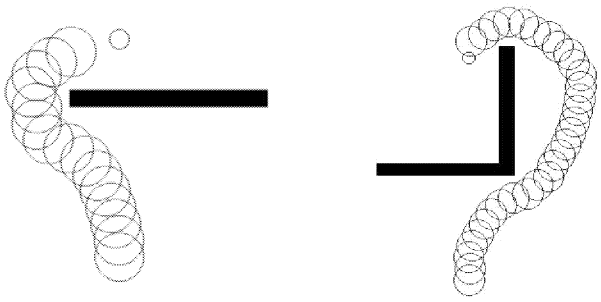


Figure 9: NFC Success Cases

Experimental Prototype

A mobile robot, constructed at the authors' universities, is used to conduct practical experiments. It has a cylindrical shape with a radius of 30 cm, and travels at a maximum speed of 8m/min. The robot has two independent wheels, driven by geared PM DC motors, located at the ends of an axis near to one of the ends of the circular base, and one free caster at the at the other end of the base. Such arrangement provides a simple and effective differential-velocity steering control by varying the applied voltage to the motors. The motion control of the two PM DC motors is accomplished by simple motion control board designed by the authors, which consists a full bridge chopper circuit, and PIC16f877 micro-controller. The advantage of using this micro-controller is that it accepts velocity commands from the onboard computer and to control two DC motors independently.

Three ultrasonic sensors are mounted at the front, and at the two ends of the central axis of the robot, where the right and the left sensors are directed at 45° from the central axis as shown in figure 10. The target can be detected by another ultrasonic sensor, which is allowed to rotate, using a small stepper motor. The reason in mounting the ultrasonic sensors in such arrangement will be discussed in details in a future work. The error eliminating rapid ultrasonic firing (EERUF) method is used to minimize the error in distance measurements due to the noise that affect the ultrasonic sensors, and the crosstalk problem was eliminated by using alternating delays method. A number of simple experimental tests were performed on this mobile robot to test the validity of the proposed strategy, Figure 10 and 11 show the schematic diagram and a life photo for the mobile robot.

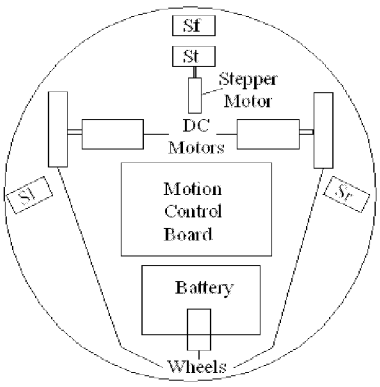


Figure 10: Schematic of the Moble Robot with sensor locations.

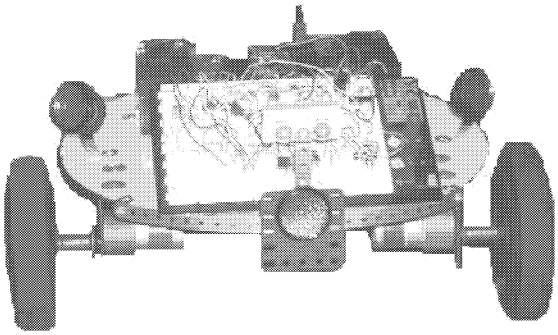


Figure 11: The Moble Robot Prototype.

Conclusion:

The performance of the FLC625 is good and slightly improved the performance of the robot compared to the FLC40 since the robot doesn't touch any obstacle and the robot avoids collision with any obstacles as shown in the above cases. But the inference time is much more than the FLC40. However, the proposed approach that based on using neuro-fuzzy system instead of the inference engine is reduced the processing time and increased the performance. The response of the implemented robot has shown an excellent reduction with respect to the response time.

Reference:

Aguirre E. and González A. 2000. "Fuzzy behaviors for mobile robot navigation: design, coordination and fusion", *International Journal of Approximate Reasoning*, 25(3): 255-289.

Lee C. C. 1990. "Fuzzy logic in control systems. Fuzzy logic controller- part 1 and part 2", *IEEE Trans. on systems, Man and Cybernetics,* 20 (2): 404-435.

Proychev T. and Petrov M. 1996. "Fuzzy Navigator of Mobile Robots in Unstructured Environment", *IEEE Transaction on Systems, Man, and Cybernetics-Part B: Application and Reviews*, 26(5).

Ross T.J. 1998. "*Fuzzy Logic in Engineering Applications"*, McGraw-Hill Inc, New York.

Tanaka K. 1996. "*An Introduction to Fuzzy Logic for Practical Application.* 1st Edition", Spring-Verlag, New York.

Zadeh L. A. 1996. "The roles of fuzzy logic and soft computing in the conception, design and deployment of intelligent systems", *BT Technology,* 14(4): 32-36.

Callan R. 1999. "*The Essence of Neural Networks"*, Prentice-Hall, Europe.

Carpenter G., Grossberg S., Markuzon N., Reynold H. J. and Rosen D. B. 1992. "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analogue Multi-dimensional Maps", *IEEE Tran on Neural Networks,* 3(5): 698-713.

Saffiotti A. 1997. "Fuzzy Logic in Autonomous Robotics: behavior coordination", *IEEE International Conference of Fuzzy Systems*: 573-578.

Fausett L. 1994. "*Fundamentals of Neural Networks"*, Prentice-Hall, Europe.

Hagras H. and Sobh T. 1998 ."Intelligent Learning and Control of Autonomous Robotic Agents Operating in Unstructured Environments", *IEEE Transaction on Systems, Man, and Cybernetics-Part C: Application and Reviews,* 26(4).

Cao J., Liao X. and Hall E. 2000. "Reactive Navigation for Autonomous Guided Vehicle Using the Neuro-fuzzy Techniques", *Engineering Application of artificial intelligent,* 6(4): 320- 329.

Medsker A. R. 1995. "*Hybrid Intelligent System"*,. Kluwer Academic Publishers, Europe.

Yung N. H. C. and Ye C. 1999. "An Intelligent Mobile Vehicle Navigator Based on Fuzzy Logic and Reinforcement Learning", *IEEE Transaction on Systems, Man, and Cybernetics-Part C: Application and Reviews*, 29(2).

Xu W.L., Tso S.K. 1996. "Real time Self–Reaction of a Mobile Robot in Unstructured Environment Using Fuzzy Reasoning", *Engineering Application of artificial intelligent,* 9(5): 475- 485.

BIOGRAPHY

Munaf S. N. Al-Din was born in Baghdad, Iraq and went to the University of Technology in baghdad, where he studied Electronics Engineering and obtained his B.Sc. degree in 1986 and his Ph.D. 1997 from the same university. He worked for six years at Al-Balqa' Applied University in Jordan, then he worked as Assistant Professor at Informatic Institute for Postgraduated Studies at the University of technology in Baghdad from 2005-2006. He is now an Assistant Professor in the department of Electrical Engineering, Tafila Technical University, Jordan.

Ahmed A. Abdullah was born in Baghdad, Iraq and went to the University of Technology in baghdad, where he studied Electronics Engineering and obtained his B.Sc. degree in 1986, M.Sc. 1993 and his Ph.D. 2004 from the same university. He worked for two years lecturer at Informatic Institute for Postgraduated Studies at the University of technology in Baghdad. He is now an Assistant Professor in the department of Electrical Engineering, Tafila Technical University, Jordan.

PARTICLE FILTER-BASED LOCALIZATION OF SIMULATED MINI-ROVERS WITH LOW-COST IR SENSORS

Fabrizio Abrate, Basilio Bona and Marina Indri
Dipartimento di Automatica e Informatica, Politecnico di Torino,
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
E-mail: fabrizio.abrate@polito.it, basilio.bona@polito.it, marina.indri@polito.it

KEYWORDS

Monte Carlo Localization, IR sensors, grid mapping.

ABSTRACT

Localization is one of the most important and investigated problems in mobile robotics. A great number of solutions have been described in many papers and many different theoretical approaches have been presented. Among the sampling-based approaches **M**onte **C**arlo **L**ocalization (MCL) exploits the Monte Carlo techniques to solve the localization problem. The MCL approach has been intensively tested on problems involving robots equipped with range sensors such as sonar or laser sensors leading to very good performances. In this paper the application of MCL techniques to the localization of mobile mini-robots, equipped with only low-cost Infra-Red (IR) sensors, is investigated and discussed in simulation.

INTRODUCTION

In the recent years, a great interest has been shown by several researchers to the problem of *mobile robot localization*, i.e. the problem of determining the pose of a robot relative to a given or a previously learned map of the environment. The real hardness and the characteristics of the localization problems can be deeply different: a taxonomy of such problems exists, classifying them according to the knowledge that is available initially and run-time, to the environment properties, to the adopted strategy and to the number of robots involved. Hence we can have *local* or *global* localization, *static* or *dynamic* environments, a *passive* or *active* approach and one or more robots involved (*single* or *multi* robot localization), as it can be seen in (Thrun *et al.* 2005).

From the theoretical point of view, a great number of different approaches has been developed. Kalman-filter based techniques have proven to be robust and accurate for keeping track of the robot's position, as shown in (Ivanjko *et al.* 2005). Regrettably Kalman-filter can do filtering and prediction operations only assuming *linear Gaussian systems*. We can overcome this last limitation with the *Extended Kalman Filter* (EKF), in which the state transition probability and the measurement probabilities are governed by nonlinear functions, which are then linearized via *Taylor expansion*, but problems remain to globally re-localize a robot in case of a localization failure. Markov localization is the straightforward application of Bayes filter to localization problem and in its grid-based form it can represent arbitrarily complex probability densities. Unfortunately this approach suffers from high computational and memory requirements. A common method to obtain interesting results is to mix different approaches as done in (Guttman 2002), where the *Markov-Kalman localization* (ML-EKF) is presented. It mixes the robustness of grid-based Markov localization and the accuracy of Kalman filtering.

One recent but very popular and interesting approach is the Monte Carlo Localization method (MCL), where the uncertainty is represented in a different way: instead of trying to describe explicitly the density function itself, this density is approximated by a set of *samples* randomly drawn from it, called *particles*. To maintain updated the density over time, Monte-Carlo methods are used (Handschin 1970). Monte Carlo Localization seems to show several key advantages with respect to other approaches, for example with respect to the ones described above. There are several papers that show these advantages; for example (Dellaert *et al.* 1999) presents some localization performances of MCL when the robots used are equipped with *sonar* or *laser* range finders, and in (Wolf *et al.* 2005) a vision-based approach to mobile robot localization that integrates an image-retrieval system with MCL is developed.

A current important challenge in robotics is to reduce the dimensions of the devices up to the frontier of *microrobotics* and, at the same time, to use such small robots within teams of cooperative agents to more efficiently perform localization, mapping and exploration tasks. The team can be heterogeneous, involving robots of different size and equipped with different sensor types, e.g. in tasks where small robots are used in conjunction with bigger ones, to reach and explore narrow places. Sonar and laser range finders, successfully used in the implementation of the most common localization techniques, are bulky devices for such small robots, which must be equipped with lower dimension, light sensors, such as the Infra-Red (IR) ones, and only sometimes by tiny, ad-hoc cameras. In a heterogeneous team of cooperative robots, it is then realistic to assume that some of them are equipped *only* by the low-cost IR sensors, and hence it becomes very important to have self-localization techniques working well also in this case. In (del R. Millàn and Arleo 1997) and (Huang and Beevers 2004) two different approaches to the mapping problem with IR are presented, but no relevant works on localization using low cost IR sensors are available in literature, at the best or our knowledge.

In this paper we investigate the possibility of successfully

applying an testing the MCL approach in simulation to solve a *global* localization problem in a *static* environment, with a *passive* approach, considering a mini rover equipped with low cost IR sensors.

The paper is organized as follows. Section "THE LOCALIZATION APPROACH" briefly illustrates the basic characteristics of the localization approach we applied. Section "THE FRAMEWORK AND THE SENSORS" describes the simulated robot, the set-up used in the tests and the characteristics of the IR sensors, discussing the problem of their modelling. In Section "THE ARENAS FOR THE TEST" the arena for the simulation test is described, while in Section "SIMULATION RESULTS" the results of the performed tests are reported and discussed. Section "CONCLUSIONS" finally draws the conclusions.

THE LOCALIZATION APPROACH

This section briefly introduces the main theoretical elements of the considered localization approach: the use of an *occupancy grid map* of the environment given a priori, and the Monte Carlo Localization technique, based on particle filters.

Grid Maps

The mapping process, in general, consists in acquiring a map of the environment in which the rover moves. Many different ways of organizing and storing the information exist. The method implemented in this work is the *occupancy grid mapping* (Elfes 1989); the basic idea of such an approach is to represent a map as a field of random variables, arranged in an evenly spaced grid. Each random variable is binary and corresponds to the occupancy of the location it covers. In this particular case the map is a priori given.

Particle Filters And Monte Carlo Localization

In the considered localization problem, $X = \{x_t, t = 1, ..., N\}$ is the state vector, where $x_t = (x, y, \theta)$ is the robot pose, $Z = \{z_t, t = 1, ..., N\}$ is the set of all measures and $U = \{u_t, t = 1, ..., N\}$ is the set of all controls, where t represents the current discrete time step, and N is the number of time iterations. The posterior density over the state space is $p(x_t|Z)$, where the set Z contains all the measures up to the current time. In Monte Carlo Localization this posterior density can be represented by a set of M samples called *particles* $P_t = \{p_t^i, i = 1, ..., M\}$ drawn from it, so that the set approximately describes the density. The formal filtering problem that exists when dealing with localization can be divided into two different phases, the *Prediction Phase* and the *Update Phase*. This problem is well explained in (Dellaert *et al.* 1999), so our goal here is only to briefly point out what happens during these two phases.

Prediction Phase

The key idea of this phase is to apply the motion model to each particle p_{t-1}^i belonging to the set of particles computed in the previous iteration. This is accomplished by sampling from the density $p(x_t|u_t, x_{t-1})$.
A new set of particles P_t' is thus generated, whereas the new measures are still to be included.

Update Phase

In this phase the *importance factor* of each particle is computed, and the *importance sampling* (the so-called *resampling*) is performed on the basis of the measures Z. The importance factor is the probability of the measurement z_t under the particle p_t^i, given by $\omega_t^i = p(z_t|x_t^i)$. If we interpret ω_t^i as the *weight* of a particle, the set of weighted particles approximates the Bayes filter posterior probability distribution. The resampling draws with replacement M particles from this temporary weighted set, and the probability of drawing each particle is given by its importance factor. This process does not change the number of the particles in the set but it changes their distribution. At the end we obtain a new set of particles P_k that approximates the density $p(x_t|Z)$.

THE FRAMEWORK AND THE SENSORS

The analysis of the possibility of performing global localization using a robot equipped with low-cost IR sensors has been carried out in the case of a simulated Khepera II© robot. Localization performances have been investigated both in simulation using KiKS (Storm 2001*a*). In the following, we introduce the simulation tool and the main features of Khepera II© robot simulated in the KiKS environment, and we illustrate the sensors model adopted to implement the *Update Phase* described in Section "THE LOCALIZATION APPROACH".

KiKS

KiKS (the acronym of ***Kiks Is a Khepera Simulator***) is a Matlab© based Khepera II© simulator, realized by Theodor Storm.It uses the Khepera© serial toolbox (Piguet and Legon 1998) to communicate with the robot. This simulator allows the creation of bidimensional environments that can be populated with walls, balls and cylinders; a snapshot of the simulator can be seen in Figure 1. It is also possible to insert light sources in the environment. The simulated robot is a virtual clone of either Khepera© robot or Khepera II© robot, and it is important to notice that only one instance of the robot can be simulated in an instance of KiKS. This is due to the single-thread structure of Matlab©. Experiments that involve multiple robot instances can be performed using KiKSnet (Storm 2001*a*).
The most important feature of KiKS is that the simulation engine does not distinguish between real or simulated robot. In particular this means that the function calls are forwarded to an engine that answers with model-based values of sensor readings and odometry if the robot is virtual or, otherwise, they are forwarded to the real device. This feature allows to use KiKS as a framework also for experimental tests and localization tasks with the real Khepera II©. In the simulation tests the values of sensor readings are model-based, and in particular they are generated by an identification process employing a neural network. More details on how this neural network has been trained and other information can be found in (Storm 2001*b*).

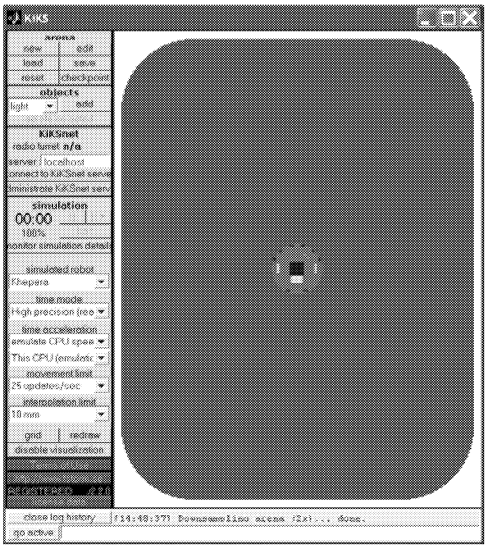


Figure 1: Kiks front-end example

Khepera II©

The Khepera II© robot is a non-holonomic mobile robot. It embeds a 25 MHz Motorola© 68331 processor with 512 kBytes RAM memory and 512 kBytes programmable Flash memory. The robot is moved by 2 DC brushed servo motors with incremental encoders that can drive the robot from a minimum velocity of 0.02 m/s up to a maximum of 0.6 m/s. The robot is quite small and lightweight, since its diameter is 0.07 m, its axle length is 0.053 m, its height is 0.03 m and it weights only 0.8 kg. It is equipped with 8 IR proximity and ambient light sensors with up to 0.1 m range; the set-up of the sensors can be seen in Figure 2.

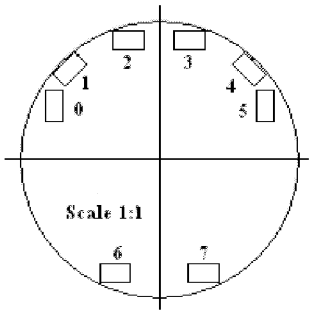


Figure 2: The position of the 8 Infra-Red sensors

IR Sensors: Characteristics And Models

In general IR sensors are small devices that can emit an infrared light ray and sense the presence of an object by using the IR-beam reflected from the object itself. IR sensors are often used in *optoelectronic scanning* and *switching devices*, i.e., index sensing and coded disk scanning. Considering the robotics field and especially mobile robotics, IR sensors have often been used as *proximity sensors*, with the task of detecting obstacles. In this work these sensors are used to measure distances, after they have been properly modelled.

In our opinion there are fundamentally two approaches to build a perceptive model of a mobile robot equipped with IR sensors. The first approach considers the whole set of sensor readings at time t to determine only one distance and one direction, like in a polar coordinate system, where ρ and θ are respectively the minimum estimated distance and the estimated angle from an object. The second approach computes a distance value from every sensor, in order to have at time t as many distance estimates as the number of embedded IR sensors; then an appropriate probabilistic model is selected and an estimation of the intrinsic model parameters is applied.

We have primarily considered the first approach and we have developed a sensor model starting from the theoretical aspects examined in (Garcia and Solanas 2004) and (Novotny and Ferrier 1999), where it is shown how the *Phong Illumination Model* can be exploited to measure distances using IR sensors. Before being able to measure distances, the reflective properties of the material are to be determined. The sensor model thus obtained shows a good distance estimation behavior when it returns a distance value, displaying low mean and standard deviation estimation error. Unfortunately, an insufficient number of sensors in a non homogeneous configuration surrounds the considered robot (as shown in Figure 2), so that the model is too often unable to return a distance value, making it not applicable to our localization problem. However, this model could be implemented when dealing with robots equipped with IR sensors all around in homogeneous configuration.

The second approach consists of three consecutive steps:

1. finding the characteristic of the sensors;
2. selecting the appropriate probabilistic model of the measurement process;
3. applying an algorithm to estimate the intrinsic parameters.

In Figure 3 the characteristic of sensor 3 (referring to Figure 2) is shown. A linear fit of data can be adopted, thus obtaining a linearity interval large enough (50 mm) to allow the robot to sense the obstacles, avoiding undesirable contacts. The probabilistic model we have selected incorporates three types of measurement errors: small measurement noise, errors due to failures to detect objects and finally random unexplained noise. This model is a variant of the *beam perception model* proposed in (Thrun *et al.* 2005). Finally we adopted a *Maximum Likelihood estimator* similar to the one proposed in (Thrun *et al.* 2005) that iteratively estimates the intrinsic parameters of the probabilistic perception model considered. This *Maximum Likelihood estimator* is an instance of the *expectation maximization* algorithm. The probabilistic model and the estimator have been applied in simulation, showing that this way of modelling the IR sensors behavior can lead to good results.

THE ARENAS FOR THE TEST

To perform the localization experiments, an arena have been prepared in KiKS environment for simulation purposes. This map is 1 m x 0.6 m.

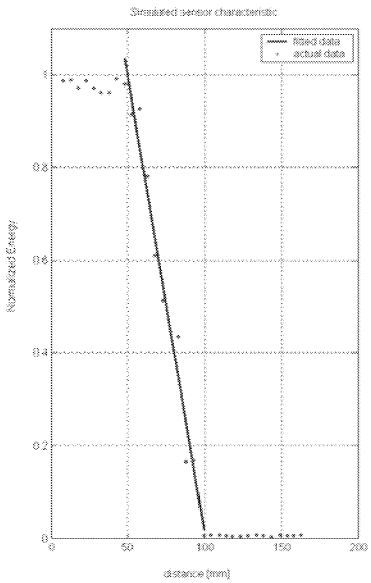


Figure 3: The sensors characteristics

Simulation

The simulated arena can be seen in Figure 4, where it is shown in a KiKS snapshot. The walls in this arena have been generated using a KiKS's feature that allows the user to simulate a wall with a specified color. We have chosen to simulate white walls.
The robot travels inside the arena for 100 s, using a follow-type control that allows the robot to reach walls and follow them.

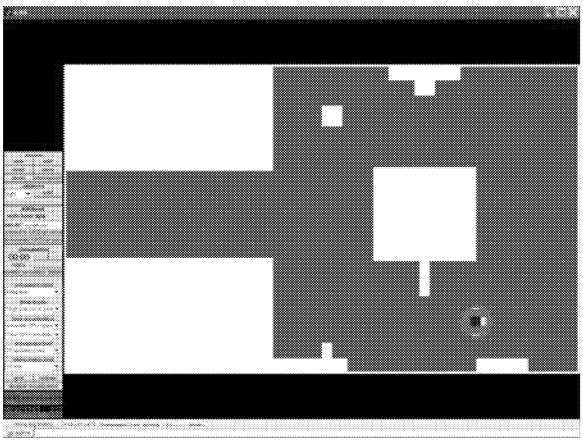


Figure 4: The map built with KiKS simulation tool

SIMULATION RESULTS

We have tested the performance of the implemented MCL algorithm by considering three different situations concerning the uncertainty on the initial robot heading, but assuming that the x and y Cartesian robot's initial position coordinates are totally unknown. In the first situation we have analyzed the algorithm behavior assuming $\pm 45^\circ$ uncertainty on the robot's initial position and spreading 20.000 particles with uniform density distribution all over the free space of the arena. In the second situation we have considered $\pm 90^\circ$ of heading uncertainty and 40.000 particles, in the third and last one we have considered a totally unknown heading and 80.000 particles. In substance we have increased the number of particles linearly with the growing heading uncertainty. A key concept that has to be noticed is that the number of particles reported above (20.000, 40.000 and 80.000) has been used only in the first iteration of the algorithm, and then reducing the number of particles within an interval between 100 and 1000.
The graphs we are going to show pertain to the third situation mentioned above, which is the most critical one; in the other two situations even better results have been obtained. The first plot of Figure 5 shows the average pose estimation error depending on the simulation time, while the second plot shows the number of particles included inside the area occupied by the robot. In Table 1 the percentage of time in

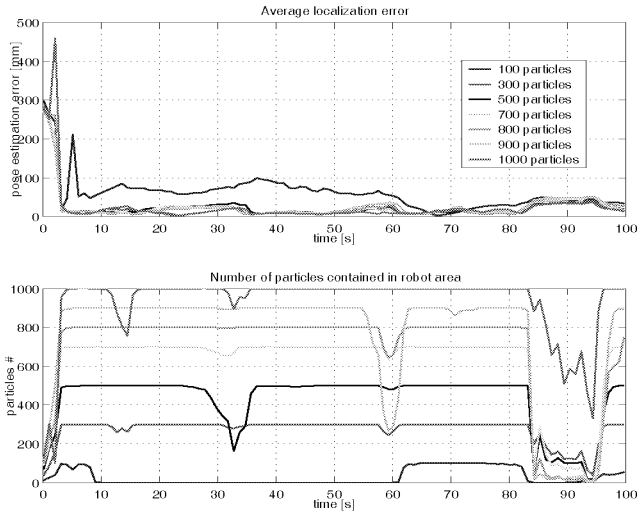


Figure 5: Simulation statistics

which the robot is correctly localized is shown. Here more experiments with respect to Figure 5 are considered. Correct localization is assumed when all the particles are contained inside the area covered by the robot. Except from the experiment carried out with 100 particles, which shows a very low localization percentage, we can state that generally the robot localizes itself quickly and performs good position tracking all over the path. Looking at Figure 5, we can see that at approximately 85 s of simulation the particles are no longer contained in the area occupied by the robot. This happens when the robot is in front of the narrowest corridor on the left (see Figure 4) and the sensors cannot perceive obstacles and walls anymore since they are too far for the sensor range, thus letting the particles spread out. Afterwards, the robot motion allows the sensors to collect new data, and the particles become compact around the correct position estimated value, so that the localization can occur again.

Table 1: Percentage localization error in simulation

Particles #	Localization percentage
100	27 %
200	66 %
300	92 %
400	85 %
500	85 %
600	87 %
700	87 %
800	85 %
900	82 %
1000	95 %

CONCLUSIONS

In this paper we have presented an approach to solve the localization problem using an MCL algorithm and grid maps, when dealing with simulated robots equipped only with low-cost IR sensors. We have first theoretically described the algorithm and we have then identified a perceptive model of the sensors suitable for the implementation of the Update Phase of the algorithm. Finally we have investigated the performance of the approach with a simulated Khepera II© robot, showing in particular that satisfactory results can be achieved by using a suitable number of particles, and that the correct localization of the robot can be also quickly recovered after its motion across too wide areas for the sensor range. Current work is devoted to experimental localization tests carried out with a real mini-rover. The first obtained results seem to be in good agreement with those of the simulation test.

References

del R. Millàn, Josè and A. Arleo (1997). Neural network learning of variable grid-based maps for the autonomous navigation of robots. In: *Proceedings of 1999 IEEE International Conference on Robotics and Automation*. pp. 40–45.

Dellaert, F., D. Fox, W. Burgard and S. Thrun (1999). Monte carlo localization for mobile robots. In: *Proceedings of 1999 IEEE International Conference on Robotics and Automation*. pp. 1322–1328.

Elfes, A. (1989). Using occupancy grids for mobile robot perception and navigation. *Computer* **22**, 46–57.

Garcia, M.A. and A. Solanas (2004). Estimation of distance to planar surfaces and type of material with infrared sensors. In: *Proceedings of the 17th International Conference on Pattern Recognition (ICPR04)*. pp. 745–748.

Guttman, J. (2002). Markov-kalman localization for mobile robots. In: *Proceedings of the 16th International Conference on Pattern Recognition*. pp. 601–604.

Handschin, J. E. (1970). Monte carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica* (6), 555–563.

Huang, Wesley H. and Kristopher R. Beevers (2004). Topological mapping with sensing-limited robots. In: *Workshop on the Algorithmic Foundations of Robotics (WAFR)*. 367-382.

Ivanjko, E., M. Vaŝak and I. Petrovič (2005). Kalman filter theory based mobile robot pose tracking using occupancy grid maps. In: *International Conference on Control and Automation (ICCA2005)*. pp. 869–874.

Novotny, P.M. and N.J. Ferrier (1999). Using infrared sensors and the phong illumination model to measure distances. In: *1999 IEEE Internation Conference on Robotics and Automation*. pp. 1644–1649.

Piguet, Y. and S. Legon (1998). kmatlab package. http://www.k-team.com/download/khepera.html.

Storm, T. (2001*a*). *KiKS User's Guide*.

Storm, T. (2001*b*). *Master Thesis*.

Thrun, S., W.Burgard and D. Fox (2005). *Probabilistic Robotics*. MIT Press.

Wolf, J., W. Burgard and H. Burkhardt (2005). Robust vision-based localization by combining an image-retrieval system with monte carlo localization. *IEEE Transactions on Robotics* **21**(2), 208–216.

BIOGRAPHY

FABRIZIO ABRATE received his bachelor in Computer Science in 2004 and his M.S. degree in Electronic Engineering in 2006 from Politecnico di Torino, and he is now a Phd student in the same institution. His current research interests involve SLAM algorithms in mobile robotics and the distributed estimation problems with teams of cooperative robotic platforms.

BASILIO BONA received the Laurea Degree in Electrical Engineering from Politecnico di Torino, Italy, in 1971. He is Full Professor of Industrial Robotics at Politecnico di Torino, where he also chairs the Center for Systems Prototyping. He is the In-Charge of the Robotics Research Group, Dipartimento di Automatica e Informatica, Politecnico di Torino. His research interests include real-time architectures for rapid prototyping of robotic control systems, vision and force control and mobile robots.

MARINA INDRI received the Laurea Degree in Electronic Engineering in 1991 and the Ph.D. degree in Systems Engineering in 1995, both from Politecnico di Torino, Italy. She is an Associate Professor of Automatic Control at Politecnico di Torino. Her research interests are in the industrial and mobile robotics field, including rapid prototyping of robot controllers, interaction and impact control, friction identification and compensation, vision control. Since 2001 she is Associate Editor of the Conference Editorial Board of the IEEE Control System Society.

VALIDATION AND ANALYSIS IN ROBOTICS

MODELLING AND VALIDATION OF A 1-DOF ARM POWERED BY PNEUMATIC MUSCLES

Aron Pujana-Arrese
Javier Arenas
Leire Maruri
Ana Martinez-Esnaola
Joseba Landaluze

IKERLAN Research Centre
E-20500 Arrasate/Mondragon. Spain
E-mail: {APujana|JArenas|LMaruri|AIMartinez|JLandaluze}@ikerlan.es

KEYWORDS
Modelica, pneumatic muscle, model, robotic arm.

ABSTRACT

Pneumatic artificial muscles – or McKibben muscles – are low-impedance actuators and their characteristics make them very interesting for developing robotic applications that feature low impedance in terms of interaction with the environment, as for instance prothoses, orthoses or some wearable robots. In order to research the applicability of these actuators in industrial applications and in the development of orthoses and wearable robots, a one-degree-of-freedom arm driven by pneumatic muscles manufactured by Festo has been designed and built at the Ikerlan research centre. This paper focuses on the modelling of the experimental set-up, and on the model validation with experimental data. As a basis for that, a pneumatic muscle model has been developed in Modelica. After a description of the set-up constructed, the modelling of the whole system in Dymola/Modelica is explained. Finally, model validation is presented, where experimental results are compared to those obtained in simulation.

INTRODUCTION

Most robots use actuators and control systems which feature high mechanical impedance. In contrast, there is evidence in the natural world that natural impedances in animals are quite low. For instance, the human arm has an impedance of around 50 Nm/rad when trying to follow a path, and vertical stiffness during running is on the order of 10 kN/m (Pratt 2003). These values are very low compared to typical position-controlled robots.

Nowadays many robotic applications require low impedance in terms of interaction with the environment. Protheses, orthoses and wearable robots are some of such cases. The observation of nature and the use of biologically-inspired components (sensors and actuators) open up new ways in the design of biomimetic robotic devices.

Although electrical actuators are not biomimetic, they are still the most-used ones nowadays in biomimetic robots, owing to the fact that they are easily integrated, are high-powered, low cost and have easy interface. However, major advances are being made in the development of artificial muscles in order to replicate the biological qualities of the animal muscle. Among the new technologies existing, electroactive polymers (EAPs) should be mentioned, although their applicability remains far from industrial applications. Pneumatic artificial muscles – or McKibben muscles – are a very interesting alternative.

The pneumatic muscle – the biologically-inspired artificial muscle – was invented in the 1950s by the physician Joseph L. McKibben, to actuate an arm orthosis so as to help control a disabled hand (Tondu and Lopez 2000). Although this type of actuator subsequently fell into oblivion, interest in it was revived at the end of the 20th century due to its high performance actuating features, good power/weight ratio and lightness, which enables low impedance robotic applications in terms of their interaction with the environment to be easily designed. The manufacturers Shadow and Festo market industrial versions of the McKibben muscle.

Unfortunately, pneumatic artificial muscles evidence certain very non-linear force-length characteristics, like animal muscles, and controlling them and obtaining the performance features demanded by some applications are no easy issues. Modelling of pneumatic muscles and controller design have been active research issues in recent years (Tondu and Lopez 2000; Colbrunn et al. 2001; Petrovic 2002; Schröder et al. 2003).

Nowadays, Ikerlan is working on Intelligent Assist Devices (IADs), a special class of body amplifiers. They are biomechatronic devices whose aim is to help the user to carry out daily tasks that require a certain effort. For the purpose of researching the applicability of pneumatic muscles in those industrial applications and especially in the development of orthoses and wearable robots, an experimental one-DoF arm powered by pneumatic muscles manufactured by Festo has been designed and constructed. The initial work developed has been geared towards obtaining a model of the complete system in the Dymola/Modelica package, the modelling and simulation

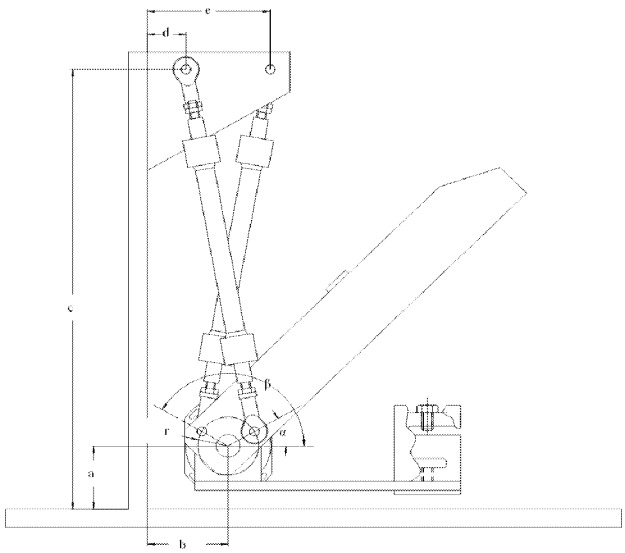


Figure 1: Geometric model of the one-DoF robotic arm

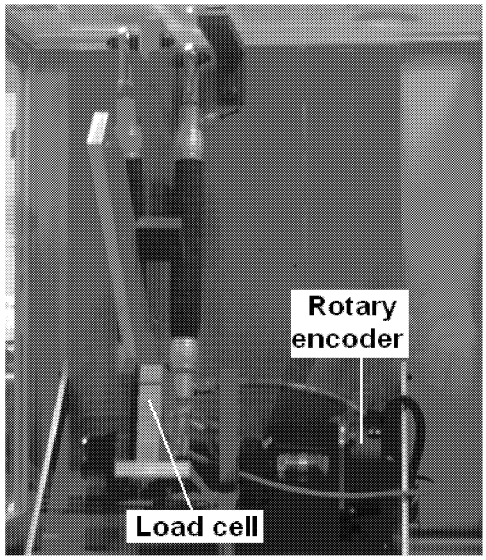


Figure 2: Picture of the experimental set-up

tool used as standard in Ikerlan, for which purpose a model of a pneumatic muscle has been made in the Modelica modelling language.

This paper presents the modelling of the one-DoF arm, as well as the model validation with experimental data. First of all, the paper briefly describes the experimental set-up. It then goes on to explain the modelling of a pneumatic muscle and of the whole system in Dymola/Modelica. The paper concludes by presenting some experimental and in-simulation results, obtained both in open- and closed-loop mode, as a way to validate the models and the modelling procedure.

DESCRIPTION OF THE 1-DOF ARM

A human arm orthosis-type application has been taken into consideration when designing the set-up. To this end and albeit with a single degree of freedom, it was considered that it should allow for the greatest angular displacement possible, and that it should be able to transport the greatest mass possible at the tip (emulating a weight borne by the hand). On the other hand, however, it needed to be confined to the length of the pneumatic muscles. In seeking a compromise between all the specifications, a displacement of around 60° and a maximum mass to be moved at the tip of 8 kg were set. By trying to minimize the length of the muscle required, the design focused on the mechanism that would enable the arm and inertias to rotate with good dynamics by means of the two muscles.

The pneumatic muscle that was chosen was the DMSP-20-200N manufactured by Festo, and the resultant mechanism is shown in Figure 1. The parameter values that define the mechanism are:

a=5 mm; b=85 mm; c=491 mm; d=40.6 mm
e=129.4 mm; α=0°-60°; β=120°-180°; r=32 mm

From these values the distance L (mm) between the ends (joining points of the mechanism) of the pneumatic muscles is:

$$L = \sqrt{175059 + 2841.6 \cdot \cos\alpha - 26624 \cdot \sin\beta}$$

When the muscles are without pressure, the distance L is of 423 mm, with the length of the muscle fibre being 200 mm. The centre of the arm mass with regard to the centre of rotation is at a height of 17.6 mm and at a horizontal distance of 205 mm, considering that the arm is in the horizontal position. The arm mass is 0.987 kg. The centre of the additional masses placed on the end of the arm (up to a maximum 8 kg) would be at a height of –24 mm and at a horizontal length of 367 mm with regard to the centre of rotation, always bearing in mind that the arm is in the horizontal position. Figure 2 shows a picture of the prototype constructed. The set-up may be rotated so that the arm moves in a horizontal plane and the effects of gravity are therefore cancelled out.

Two Festo MPYE-5-1/8HF pneumatic servovalves are initially used for actuation, with working pressure being 6 bar. Festo SDE-D10 pressure sensors are used to establish the entry pressure of each muscle. The model includes a FAGOR S-D90 encoder which supplies 180000 pulses per turn, so as to measure the rotation angle of the arm accurately. A load cell is also included on the lower stop of the model (Figure 2), so that the force exerted by the arm against this stop may be measured. The stop may be fastened at different angles, whereby the force exerted by the muscles may be calculated at different lengths and with different supply pressures.

PIP8 hardware made by the MPL company, which is similar to The Mathworks xPCTargetBox, is used to control the system. Control algorithms are implemented in Simulink and code generated and downloaded in the aforementioned hardware by means of The Mathworks RTW.

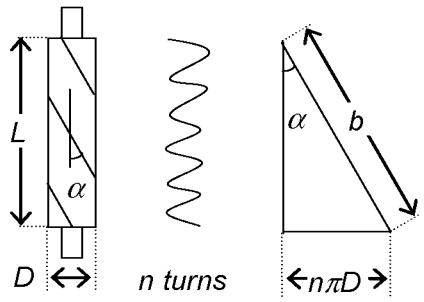


Figure 3: Geometric model of a pneumatic muscle

MODELLING OF THE EXPERIMENTAL SET-UP

Model of a pneumatic muscle in Modelica

As has been previously mentioned, pneumatic muscle actuators are highly non-linear systems. Many authors have worked on the idea of obtaining a model that represents the behaviour of these devices, with the most important ones being shown below. The objective of the model is to find a relation between the pressure and length of the pneumatic muscle with the force it exerts along its entire axis. Figure 3 shows the well-known outline of the constitution of the pneumatic muscle. L is the length of the cylinder and D the diameter. Assuming inextensibility of the mesh material, the geometric constants of the system are the thread length b and the number of turns n for a single thread. α is the angle between the thread and the long axis of the cylinder. The angle changes as the length of the muscle changes.

From these premises, Tondu and Lopez (2000) have carried out a mathematical development based on the application of theorem of virtual work, which has led them to obtain the following equation:

$$F(\varepsilon, P) = (\pi r_0^2)P[a(1-\varepsilon)^2 - b] \qquad (1)$$

where:

$$\varepsilon = \frac{(l_0 - l)}{l_0} \qquad 0 \le \varepsilon \le \varepsilon_{max} \qquad a = \frac{3}{\tan^2(\alpha_0)}, \qquad b = \frac{1}{\sin^2(\alpha_0)}$$

r_0 being the nominal inner radius, l the length of the muscle, l_0 the initial nominal length, P the pressure and α_0 the initial angle between the membrane fibres and the muscle axis.

In this initial equation two main approximations are considered. The first one is that the thickness of the muscle walls is very small, and the secondly, it may also be assumed that pressure is transmitted evenly throughout the membrane. However, there is a major phenomenon which Tondu and Lopez (2000) consider needs improving (1): the fact that the shape of the muscle is not cylindrical on the ends, but rather is flattened. To represent this, they incorporate a correction factor k, with equation (1) taking the following form:

$$F(\varepsilon, P) = (\pi r_0^2)P[a(1-k\varepsilon)^2 - b] \qquad (2)$$

Tondu and Lopez (2000) establish two ways for obtaining parameter k: On the one hand, they propose considering a constant value for k which may vary depending on the material that the muscle is made of and, on the other, they suggest making the parameter k depend on the pressure in the muscle at any given time. According to their studies, this last-mentioned option is the one that provides the most complete model.

Other analytical expressions for calculating the force of a pneumatic muscle may be found in technical literature. Thus, in (Colbrunn et al. 2001), (Petrovic 2002) and (Umetani et al. 1999), formulas equivalent to (2) are put forward based on the same principles as Tondu and Lopez (2002), and may be readily deduced from each other. In all of them is reflected the fact that the force exerted by the muscle has a linear relation with the pressure inside it and non-linear with the contraction.

For the model put forward in this work, the equation (3) proposed in (Petrovic 2002) has initially been chosen due to the simplicity of 2nd order polynomial adjustment.

$$F = (D_1 + D_2 \cdot q + D_3 \cdot q^2)P \qquad (3)$$

with q being the displacement $q = l_0 - l$.

However, in the experimental trials carried out, it was noted that to identify the coefficients of equation (3), it had to incorporate a new term, independent from the pressure and dependent on contraction. Therefore, the ratio between force, pressure and contraction in the muscles takes the following form:

$$F = (D_1 + D_2 \cdot q + D_3 \cdot q^2)P + \varphi(q) \qquad (4)$$

This correction term $\varphi(q)$ introduced into the model will be determined in each case from the experiences needed to identify the parameters of the equation (4) taken into consideration.

The model described before has been implemented by using the good qualities of the modelling language of physical systems known as Modelica. Modelica is a freely available, object-oriented language for modelling of large, complex, and heterogeneous physical systems. It is suited for multi-domain modelling, for example and in this case, mechatronic models in robotics, involving pneumatic, mechanical and control subsystems. Dymola/Modelica is the standard modelling and simulation tool used in the Control Engineering department of Ikerlan.

The muscle model has been developed ensuring its connectivity with other objects described in Modelica, in order to construct more complex mechanical-pneumatic systems. For this reason the model contains two mechanical interfaces that correspond to the muscle anchorage points

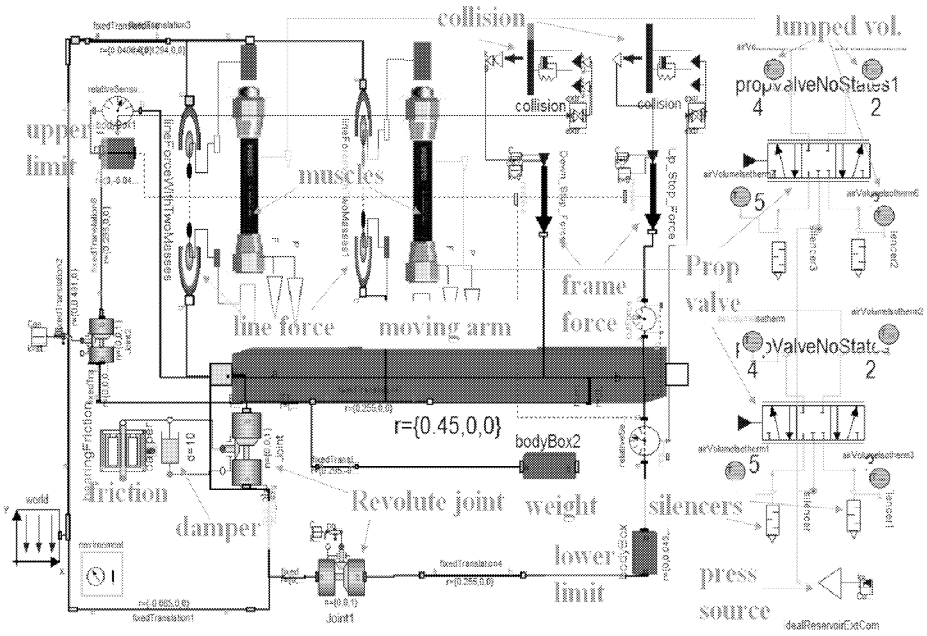


Figure 4: Model of the whole system in Dymola/Modelica

and a pneumatic interface that makes it possible to connect to an external pneumatic circuit.

Three main parameters have been taken into consideration that characterize the physical properties of this type of muscle: nominal length, nominal diameter and the initial angle between the membrane fibres and the muscle axis. Other magnitudes that may be considered as parameters are, for instance, the heat transfer coefficient or the maximum pressure beyond which the system delivers a warning message.

In order to characterize the physical behaviour of the muscle a set of equation have been implemented in the Modelica code. The first one corresponds to the previously analyzed equation (4) that describes the mechanical force exerted by the muscle.

Other equation used is the state equation of ideal gas:

$$P \cdot V_{eff} = m \cdot R \cdot T_{muscle} \qquad (5)$$

where the air mass (m) and the temperature in the inside (T_{muscle}) are considered as state variables of the model. V_{eff} is the effective volume of the muscle, where even the flattening experienced on the ends of the muscle when it is contracted has been taking into consideration for its calculation.

The last of the equations taking into account corresponds to the thermodynamic equation:

$$m \cdot \frac{dT_{muscle}}{dt} + T_{muscle} \cdot m_{dot} = \gamma \cdot T_{inout} \cdot m_{dot} + \frac{dV_{eff}}{dt} \cdot P / Cv - h \cdot s \cdot (T_{muscle} - T_{sou}) / Cv \qquad (6)$$

with m_{dot} being air mass flow rate, γ the ratio of specific heat capacities, T_{inout} the temperature of air entering or leaving muscle, Cv specific heat capacity at constant volume, s the heat transfer surface area, h heat transfer coefficient and T_{sou} the temperature of the surroundings.

A more complete description of the Modelica code implemented can be found in (Pujana-Arrese et al. 2007).

Model of the whole system in Modelica

For composition of the whole system Dymola, a modelling and simulation tool based on Modelica, was chosen due to the properties of its graphic interface. Dymola supports hierarchical model composition, and libraries of reusable components are available in many engineering domains.

Figure 4 corresponds to the graphic representation of the model implemented in Dymola. Its most significant components are explained in detail below.

In constructing the model, elements which have been expressly developed for this application (such as the pneumatic muscle model) coexist with components belonging to commercial libraries. Given the multidisciplinary nature of the model, there exist components of different domains which are related to each other by special objects that carry out the connection work between the mechanical part and the pneumatic part.

To model the mechanical structure, which has been explained in a previous section, objects from the *Multibody Library* and *Pneulib* libraries have been mainly used.

The element that represents the metallic arm is an object called *bodyBox* which models a rectangular-shaped rigid

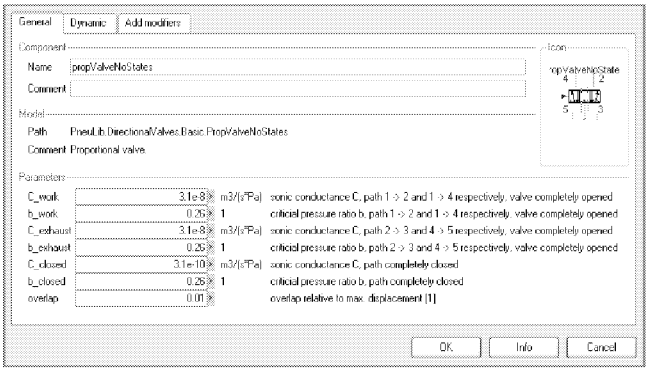


Figure 5: Parameters definition for the model of servovalve Festo

solid. The mass of the latter is determined via specified parameters such as density, length, thickness and height. An actuated revolute joint object is used to define the rotation axis of the body. These objects have an additional flange connection. Thus, different elements of the Rotational Library can be attached in order to change the behaviour of the movement (Pujana-Arrese et al. 2006). Specifically in this case, a *damper* and a non-linear element known as *bearingFriction* – which includes Coulomb friction due to the joint between the axis and the arm – have been connected. As has been described in the previous section, the possibility exists of coupling up to 8 plates of 1 kg each on one end of the arm. The model also features this option, for which purpose a *bodyBox* of the same size as the plates has been placed into the model. Furthermore, the amount of weights may be configured using a global parameter.

The stops against which the arm impacts when it reaches the limits have also been modelled using *bodyBox* elements. Moreover, taking into account that the position of the limits may be configured in the real model (at least the lower limit), the *bodyBoxes* have been connected to two *revolute joints*. The rotation angle of these joints is set parametrically prior to carrying out the simulation, thus emulating the configuration of the real model.

A special class named *Collision* has been developed in order to model the impact between the arm and the stops mentioned in the previous paragraph, of which two instances have been included in the model – one for each point of contact. The input for these objects is the relative distance between the point of impact of the arm and the mechanical stop. When the distance is detected as becoming negative – in other words, when there is a penetration of one body into another – the force of the impact is calculated based on a spring and damper model. The forces calculated are applied to the points of contact on the arm using force input elements known as *frameForce*. The normal direction is always considered to be that towards the surface of the impact.

All the rest of the components of the model, which are going to be analyzed now, correspond to the pneumatic part of the system. The objects that make it possible for the pneumatic components to interact with the mechanical ones

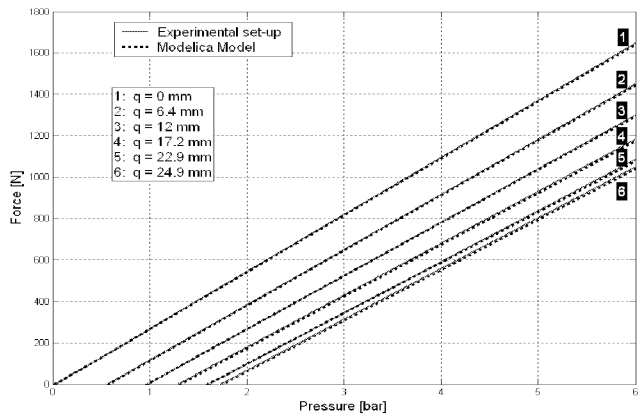


Figure 6: Force-pressure results for different lehgths

are connected to the anchorage points of each muscle. These objects (*lineForce*) describe a line of force between their two mechanical ports. The magnitude of this force depends on the actuator element connected to the ports. Obviously, in this case the actuators will be two objects of the class that models the pneumatic muscle which has been previously described in detail.

The Festo MPYE-5-1/8HF servovalve model used is worthy of a comment. Modelling and identification of the pneumatic servovalves is a none-too trivial matter which has been subject to match activity lately. In the case being dealt with here, it has been made from the basic *propValveNoStates* model of the *Pneulib library,* which describes a proportional valve, with second order spool dynamics and no lumped volumes at the ports in such a way that the position of the rail is almost proportional to the signal that commands it. It uses a linear interpolation for *SonicConductance* and *CriticalPressureRatio*. Both parameters are used according to ISO 6358. The values of these parameters have been taken as they are supplied by Festo. Figure 5 shows a screen with the parameters used.

Each pneumatic muscle is connected to an out port of the valves ensuring opposite dynamics. The remaining outputs of the valves have been fitted to silencers that function as exhausts. The element that provides the system with the necessary pressure is the externally commanded source which is connected to the entry port of the two valves. When a valve is inactive, it attempts to close its five ports, letting a flow of leaks escape that is also modelled. To model the circuit pressure dynamics, this has been completed by connecting a *lumpedVolume* object to each port of the valves. All the elements mentioned in the last three paragraphs form part of the *Pneulib* commercial library.

MUSCLE MODEL IDENTIFICATION

As stated before equation (4) expresses the basic equation of the muscle model made. In every case its coefficients should be identified, analytically or by means of experimental trials. Depending on the experimental set-up trials suitable to be performed are different. In the present

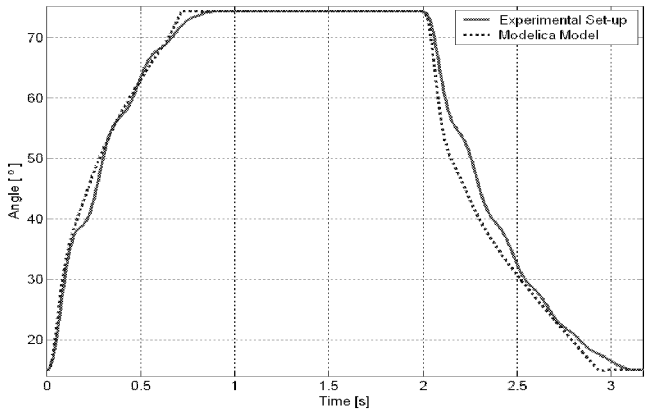


Figure 7: Results in open-loop without load

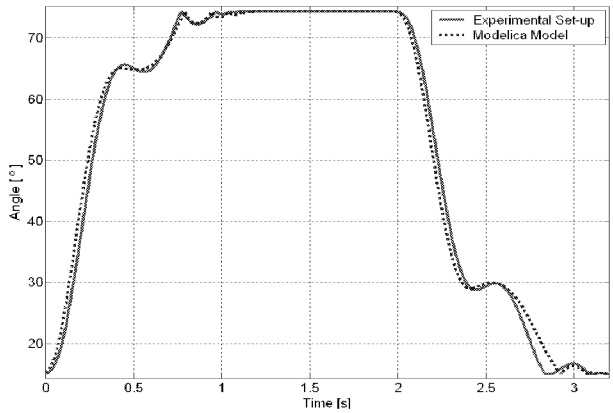


Figure 8: Results in open loop being the load 3 kg

case it was feasible to keep the muscle in a state of constant contraction and to increase slowly the pressure, recording the force exerted on a load cell. In this way, a family of straight lines was obtained with different contractions that relate force to pressure.

To carry out the trials, the muscle was firstly set at a determined, known length, thanks to the possibility of varying the lower stop of the prototype. Furthermore, the servovalves were removed so as to prevent undesired dynamic influences in such a way that pressurized air was introduced directly into the muscle. The experience involved blowing air into the muscle in order to make the arm descend until it was positioned on the lower stop, where the length of the muscle remained constant. The air pressure was slowly varied in this position within the range of 0 – 6 bar and the force exerted by the arm on the lower stop was measured by the load cell. The real force exerted by the muscle was estimated using that data, together with geometric ratios.

The results obtained are shown in Figure 6. The different straight lines correspond to different constant contractions of the muscle. The upper straight line refers to the nominal length and the lower one to the maximum contraction that may be experienced by the muscle in this model. It can be seen that the greater the length of the muscle, the greater the force it is capable of developing, as is gathered from the equation (4). It is also noted that the greater the contraction, the further the cut of the straight lines with the abscissa axis is from the origin, an effect which gave rise to the introduction of the correction term $\varphi(q)$ in equation (4).

In the case described here and based on the experimental results shown in Figure 6, the term $\varphi(q)$ was identified by means of a 4th order polynomial adjustment. Consequently, the values obtained for the different parameters of the model are:

$$D_1 = 280 \qquad D_2 = -2.2 \qquad D_3 = 0.033$$

$$\varphi(q) = 0.00046 \cdot q^4 - 0.053 \cdot q^3 + 2.02 \cdot q^2 - 43.69 \cdot q + 61.62$$

The same Figure 6 shows the data obtained in simulation with the complete model of the prototype for the same experience. As is noted, the results in simulation coincide very well with the experimental data, which to a certain extent is normal, given that this experimental data is the basis used for identifying the parameters of the muscle model.

OPEN-LOOP VALIDATION OF THE MODEL

In order to validate the whole model previously explained, some experimental and in-simulation trials have been carried out in open-loop. The prototype has been rotated in order to prevent the effects of gravity in such a way that the arm moves on a horizontal plane. Trials have been carried out using different weights placed on the end of the arm. The control elements used are the servovalves, which accept an input signal within the range of 0 – 10 V. An impulse train of ±0.9 V added to a central value of +5 V, over a period of 4 seconds, was used as input signal. In this way the arm covers the whole range of displacement. Figures 7 and 8 show an example of the results obtained both experimentally and in simulation. Figure 7 shows the results in the case there is no weight on the arm tip. Figure 8 shows the results when there is a 3 kg weight on the end of the arm. The angle reflected in the ordinate axis of the figures corresponds to the angle measured from the vertical (supposing the prototype had not been rotated).

As may be observed in the figures, especially in Figure 8, both in the model and in the prototype there appears a vibration which originates from displacements in both directions in the intermediate area of movement, changing the dynamics of the system and making it slower. These vibrations are due to the prototype's own configuration: when the arm moves, two opposed phenomena take place in the performing muscle. On the one side, it tightens and diminishes its strength. On the other side, the pressure boosts and so does the strength. Both effects are inversely reflected in the other muscle, too, and thus there comes a moment in which the strength generated by both muscles evens. As a consequence of this, the arm vibrates. This behaviour is also reflected in the results of the model, the key being the change in volume and the flattening at the ends during contraction in the model of the pneumatic muscle. Although the above-mentioned vibration is not

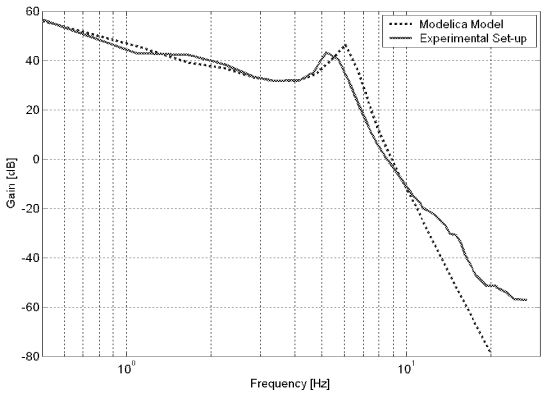


Figure 9: Experimental and in-simulation Bode diagram

evidenced in the results shown in many bibliographical references, it does appear in Tondu and Lopez (2000).

In addition to these trials in open-loop, a Bode diagram was also carried out, both on the model and on the experimental set-up. Although the system is very non-linear, a sinusoid of amplitude 0.5 V was applied to the valves and the arm amplitude measured after some cycles. For the case no weight is on the arm tip the Bode diagrams obtained are shown in Figure 9. A resonance of approximately 5.5 Hz can be clearly seen, which is also reflected in the model.

CLOSED-LOOP VALIDATION OF THE MODEL

After the tests in open-loop were carried out, some closed-loop trials were performed. A basic PI controller was used. The parameters were tuned up with a position leap of 10 degrees and slope of 20 degree per second carried out in the central part of the arm displacement area. Figure 10 shows the results for this experience when the position leap has been accomplished in that area without any weight on the arm. The proportional and integral gains of the controller were K_p=0.15 and K_i=0.05. As it was observed in open-loop trials, a vibration appears in the intermediate area of movement. After the vibration has taken place, system dynamics change and the arm moves more slowly until it reaches the stationary set-point.

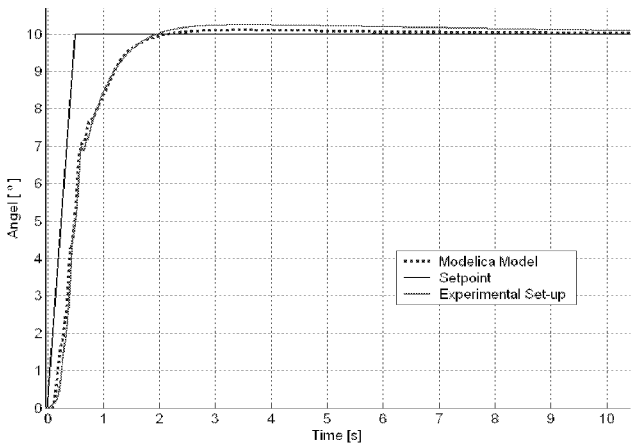


Figure 10: Results in closed-loop without load. Step in central arm displacement zone

As it has been mentioned before, the prototype does not present a linear behaviour along its movement range and thus there are notable differences depending on the area the position leap is performed. For the same PI controller the vibration is more obvious the nearer the arm is to the bottom stop. Both the moment the vibration appears and its intensity depend on the displacement area. For instance, Figure 11 shows the experimental and in-simulation responses to the position step when carried out in the first part of the arm displacement area. It can be observed that the model reflects well the response dynamics, being the rise time more or less the same, but near the steady-state position some differences appear: in simulation the response presents smaller overshoot. Figure 12 shows the results obtained when the position step is carried out in the last part of the arm displacement area. As mentioned before, the vibration appears clearly in the transient response and it is reflected by the simulation result, which follows very well the experimental one.

As an example of the results obtained with different masses on the arm tip, Figure 13 shows the closed-loop response in the central arm displacement zone when the load is 3 kg. In this case the proportional and integral gains of the controller

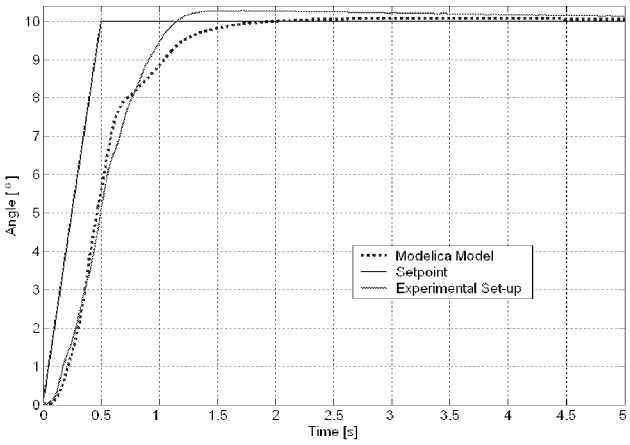


Figure 11: Results in closed-loop without load. Step in first arm displacement zone

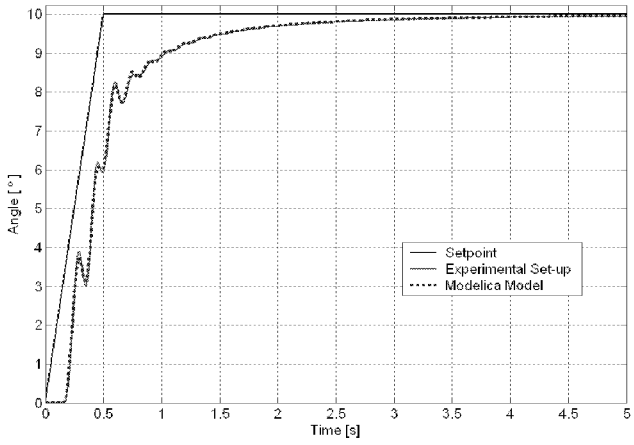


Figure 12: Results in closed-loop without load. Step in last arm displacement zone

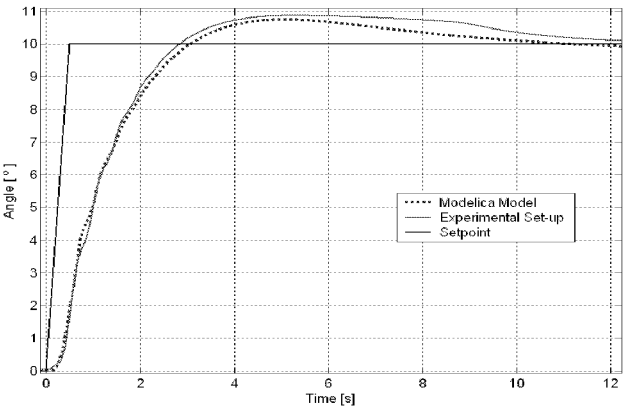


Figure 13: Results in closed-loop being the load 3 kg. Step in central arm displacement zone

were K_p=0.06 and K_i=0.03.

As mentioned before, the model does not present such notable differences in relation to the displacement area. Therefore, the model tuned up in a way the simulation results fitted the experimental ones in the central displacement area has been taken as the nominal one (Figure 10 and Figure 13). Minor changes were performed in the nominal model to adapt it to the responses in other displacement areas, as shown in Figures 11 and 12. As a conclusion, due to the non-linear characteristics of the experimental set-up built, it is very difficult to obtain models valid for all displacement zones. Variations in the system behaviour depending on the displacement area will be considered as system uncertainties for designing other more complex controllers.

CONCLUSIONS

Pneumatic artificial muscles, or McKibben muscles, are very interesting actuators for the development of robotic applications that feature low impedance in terms of their interaction with the environment, as may be the case with orthoses or certain wearable robots. In order to research about the applicability of such actuators, a one-DoF arm powered by two pneumatic muscles has been designed and built in Ikerlan. A model of the whole experimental set-up has been developed in Dymola/Modelica, based on the model of a pneumatic muscle. The model has been validated by open-loop and closed-loop experimental trials, underlining that it is very difficult to obtain a unique non-linear model valid for all displacement zones.

ACKNOWLEDGMENT

The material used in this paper was partly supported by the Spanish Ministry of Education and Science and European FEDER Fund (research project DPI2006-14928-C02-01).

REFERENCES

Colbrunn, R.W.; G.M. Nelson; and R.D. Quinn. 2001. "Modeling of Braided Pneumatic Actuators for Robotic Control". *International Conference on Intelligent Robots and Systems (IROS'01*), Maui, HI.

Petrovic, P.B. 2002. "Modeling and Control of an Artificial Muscle". *X Conference on Mechanical Vibrations.* Timosoara, Belgrade.

Pratt, G.A. 2003. "Biologically Inspiraded Components of Robots – Sensors, Actuators and Power Supplies". In *Biologically Inspired Intelligent Robots*, Y. Bar-Cohen and C. Breazeal. ISBN 0-8194-4872-9, pp. 188-192.

Pujana-Arrese, A.; A. Martinez-Esnaola; F. Martinez; C.F. Nicolás; and J. Landaluze. 2006. "Construction machinery dynamic models for Real-Time training simulators". *Industrial Simulation Conference ISC'2006.* 5-7 June, Palermo, Italy.

Pujana-Arrese, A.; J. Arenas; I. Retolaza; A. Martinez-Esnaola; and J. Landaluze. 2007. "Modelling in Modelica of a pneumatic muscle: Application to model an experimental set-up". *21st European Conference on Modelling and Simulation ECMS 2007.* 4-6 June, Prague, Czech Republic.

Schröder, J.; D. Erol; K. Kawamura; and R. Dillmann. 2003. "Dynamic Pneumatic Actuator Model for a Model-Based Torque Controller". *IEEE Int Symp. On Computational Intelligence in Robotics and Automation.*

Tondu, B. and P. Lopez. 2000. "Modeling and Control of McKibben Artificial Muscle Robot Actuators". *IEEE Control Systems Magazine,* April, pp.15-38.

Umetani, Y.; Y. Yamada; T. Morizono; T. Yoshida; and S. Aoki. 1999. "Skil Mate, wearable exoskelton robot". *16th IAARC/IFAC/IEEE International Symposium on Automation and Robotics in Construction (ISARC'99).* September 22-24, Madrid, Spain.

AUTHOR BIOGRAPHIES

ARON PUJANA-ARRESE, **JAVIER ARENAS**, **LEIRE MARURI, ANA MARTINEZ-ESNAOLA** and **JOSEBA LANDALUZE** are with the Control Engineering department of Ikerlan, a technological research centre located in the Basque Country (Spain). Their current research interests include design, modelling and control of biomimetic mechatronic systems, virtual prototyping and linear and non-linear robust control techniques.

Modeling, Stress Analysis and Design of an Underwater Remotely Operated Vehicle (ROV)

Ahmad Bagheri
Shahed Aliakbar
Salar Basiri

Mechanical Eng. Department
Guilan University, Rasht, Iran
E-mail: bagheri@guilan.ac.ir

Keywords
ROV, Finite Elements Analysis, Underwater Robot

Abstract

In this paper, firstly the modeling of the robot has been performed and then the analysis of stress and displacement has been carried out using CATIA finite element techniques. The body of the robot is under pressure due to the amount of water above it. The design and construction of an underwater has been considered for transferring the information about alive organisms under the sea.
The purpose of the construction of this ROV is recording visual data from the seabed. The camera has been installed on the frame connected on the base. The base and the engine are connected to the ROV's main body. Inside of the main body consists of electrical equipment and a vessel that balances the buoyant force.
According to the analysis increasing the thickness of the body is recommended.

Introduction

The progress made in the field of underwater vehicles, either autonomous (AUV) or teleoperated (ROV), has been very important in the last years [1,2].

This progress, based mainly on the use of new sensors or improved environment perception systems in these vehicles, has also carried with it an outstanding increase in their cost. This cost can restrict in some cases their applicability, hence, some researchers have been involved in the design and development of low-cost underwater robots [3,4].

Due to cost restrictions and with the aim to provide a biologists' team with a system able to perform long observations over the flora and some determined undersea fauna, we were entrusted to study and develop a low-cost prototype.

The underwater robot that has been constructed in Giulan University is for study on behave of fishes and discover not useful organism in Caspian Sea so it is an explorer ROV.

The situation of the body of the robot poses significant problems in the accurate fabrication of ROV because the stresses that perform in the body may lead to fracturing body or damaging electrical equipment are into the vessel. Therefore, estimating the magnitude of stresses and displacement and characterizing the effect of pressure of water on the body are deemed necessary.

In this work the amount of the von misses stresses and displacement of the body have been achieved using CATIA finite element techniques.

Design and Construction of Underwater Robot

This robot includes a vessel to produce the power and also there are the electrical equipments inside. The body of underwater robot combines of two half part of thin wall vessel. It is made from steel.
There are five engines in the bottom part of robot. Two of them for the vertical movements of robot and two of them are being used for horizontal movements of robot. Under the robot there is an engine to produce the vibrating of robot to take place on the Flore of the sea[5].

The engines has been connected with the pins that has been used on the handles of engine. This robot includes of three projectors and one camera in closed circle form on the part connected to the base of the robot for taking picture which the pictures transfer to the operator in the ship.

The height of robot is 80 cm and the diameter is 30 cm. also the diameter of base ring is 60 cm. The Top of the robot is connected to a cable which, is connected to the ship. In conclusion, according to the analysis the thickness of the body is recommended.

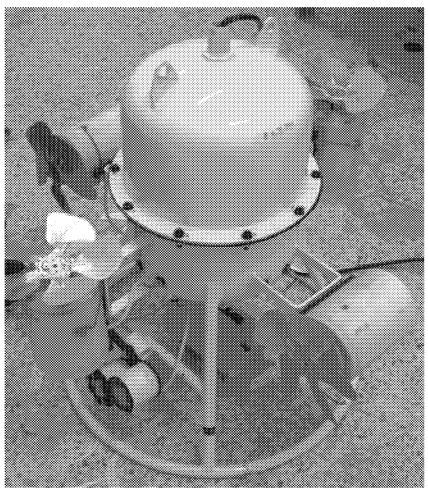


Figure 1: Underwater remotely operated vehicle constructed in Guilan University

Theoretical consideration

Von mises stress and displacement is calculated by CATIA finite element techniques, that for both component of the stress tensor within the part cylindrical axis uses:

$$\sigma_{rr} = P \frac{a^2}{(b^2 - a^2)} \left[1 - \frac{b^2}{r^2}\right]$$

$$\sigma_{\theta\theta} = P \frac{a^2}{(b^2 - a^2)} \left[1 + \frac{b^2}{r^2}\right]$$

Where a, b are the inner and outer radiuses.

And for the component of the displacement using the same axis uses:

$$u_r = \frac{p}{E} \frac{a^2}{b^2 - a^2} r \left[(1 - v) + (1 + v)\frac{b^2}{r^2}\right]$$

Where υ is the Poisson ratio and E is the Young's-module[6,7]. p is the pressure of hydrostatic power that is the amount of weight of water above the robot that according this formula:

$$P = P_0 + \int_0^h \rho\, g\, dh = P_0 + \rho\, g\, h$$

Where P_0 is the atmosphere pressure, ρ is density of water and h is the height of water above the robot.

Finite Elements Analysis of Underwater Robot

the cylindrical body and the frames around it have been designed like sheet by 1mm thickness to choose shell element for them, and it cause to possibility for testing different thickness for the body and finally select the best thickness.

It has been tried that analyze occur in critical condition. The speed of this robot is so low because for best condition of recording film, (the speed is about 0.5 m/s).

It cause to dynamic forces like drag and tension force of the rope above the robot be negligible, also the effect of the pressure of the water is more than these dynamic forces in this speed. So analysis occur in the static conditions.

Finite Element Analysis consist of three steps:
1- Preprocessor 2- Solving 3-Postprocessor

At first step (preprocessor), these procedures will be done:
1- Selecting the kind of elements
2- Adjusting the property of the elements
3- Appling the material and etc.

In this process of construction the model, the thickness of body is equal 1mm, because we want to choose Shell element for the body, to changing in the property of the shell element and we can test different thickness for the body until find the best thickness for it[8].

It has been tried to model all the connections and analyze all the details. However this is the advantage of CATIA software and it is almost impossible in other soft wares so it is the reason of using this software in this work[9].

In this way the meshing of the model is prepared and now we can solve the problem whit the determining of body conditions(Figure2).

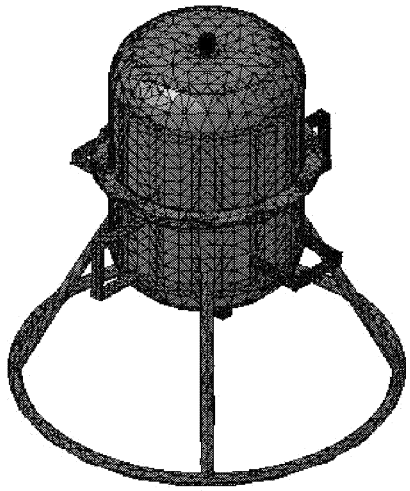


Figure 2 : Meshing the model

After meshing, this is time for Solution the problem which in this level loading the model, the time adjusting and solving problem has happened.

The first body condition of the pressure of hydrostatic power is the amount of water above the robot that according this formula:

$$P = P_0 + \int_0^h \rho\, g\, dh = P_0 + \rho\, g\, h$$

Where p_0 is the amount pressure of atmosphere of above the surface of water and ρ is the density of sea water and g is the gravity and finally h is the height water on the top section of Robot.

Since this critical height is most important condition in the robot, the depth that has been considered is 300 meters. And we can find out the atmosphere pressure in the above formula.

Then we model the reflective power from the weight of the engines in case theses are effective.
For solving this problem, in theses kind of software, we should consider a place on the model to be fixed that we chose the circular base of robot for it.

Result and discussion

Here shown the exit results such as Von Mises stress and displacement that these have been got from the CATIA:
As it seen the amount of stress is from 25MPa until 3750MPa, according to the yield strength for the safety factor it has been recommended that be 1.And also according to some places on the body that the von mises stress are more than ultimate strength so the thickness of body should be increase.

Figure 3 is shown critical zone that it can be found out, from the changing of the color.

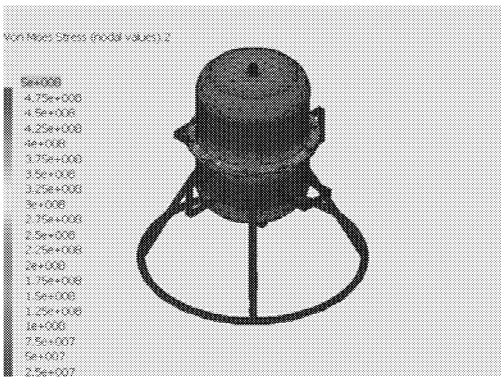


Figure 3 : The graph of the von mises stress

Another part of model have been brought in below figures to recognize the effect of pressure in each zone.

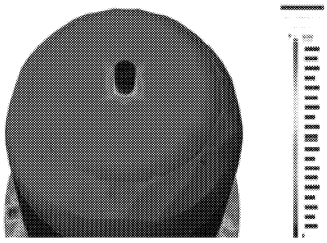


Figure 4 : Top of the robot (caps)

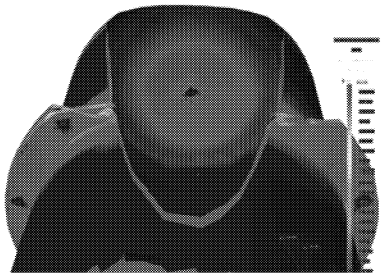


Figure 5-a : The cut view of the body is shown at the center of cylinder

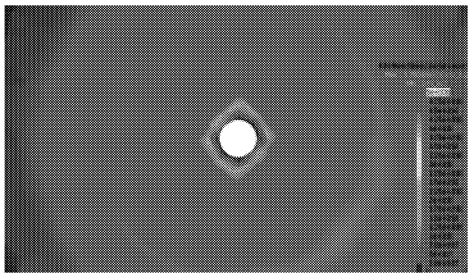


Figure 5-b : The cut view of the body is shown at the center of cylinder

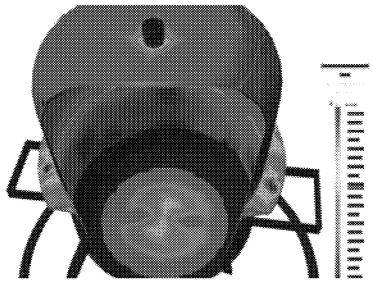


Figure 6 : View of the center of bottom half body

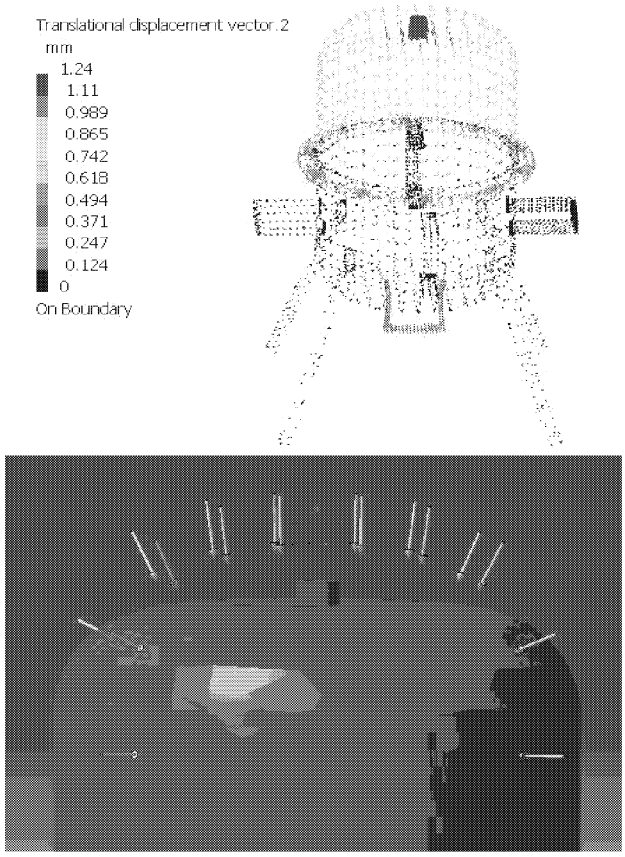


Figure 7 : Graph of displacement

From this figures it will be clear that where zone should be repaired.

Conclusion

Based on the above result of analyzing and also according to the achieved data of stress & displacement of the robot's body, the design should be optimized.

The yield stress is 250MPa, never less the maximum stress is on the frame around the body and is equal 3750MPa, and the safety factor has achieved 0.066 which needs to be more than 1, so the design should be changed until the safety factor be more than 1. At first the thickness of frame and also the circle of bolts can be changed.

So after testing different thickness and different amount for bolt's circle finally thickness is better than be 10mm and the circles be 8 mm and also the distance between center of bolts and center of the cylindrical body is better than be 163mm.

But now this time the maximum stress is on the top center of the body and is equal 656MPa. Now according to this formula $\sigma_l = \frac{pd_i}{4t}$ for thin wall vessel by increasing the thickness of the body the amount of the stress decreases.

So the thickness of the body have been increased, the first thickness that cause to the maximum stress be lees than yield strength is 8mm, that the maximum stress achieved to 237Mpa and again it has occurred on the frame.

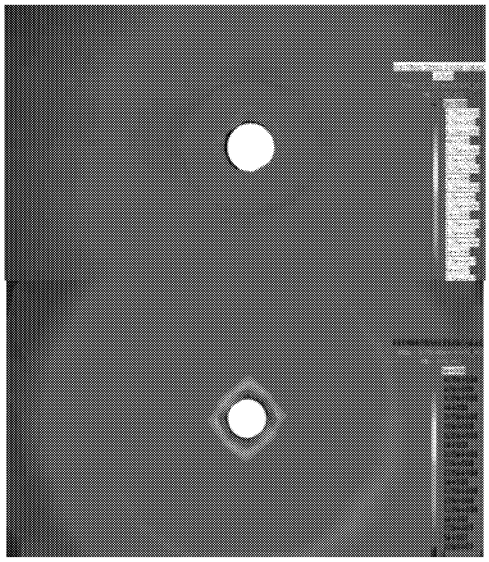


Figure 8 : Stress at the center of cylinder before and after modification

This time it is less than yield strength and the safety factor is equal 1.05 that is reasonable. If the thickness of body be equal about 1/20 of its circle it can be called thin wall vessel, and this body obey from this low by 8mm thickness.

But if instead of steel ASTM-A36 use from steel ASTM-514 that is plated baked because its yield strength is 690MPa it causes to there is more limit for maximum stress until the thicknesses and bolt's circles don't be increased.

By using of this steel the thickness of frame and the bolt's circle can be both 6mm, and the thickness of the body can rest 4mm.

Figures 8 to 10 compare robot before modification and after.

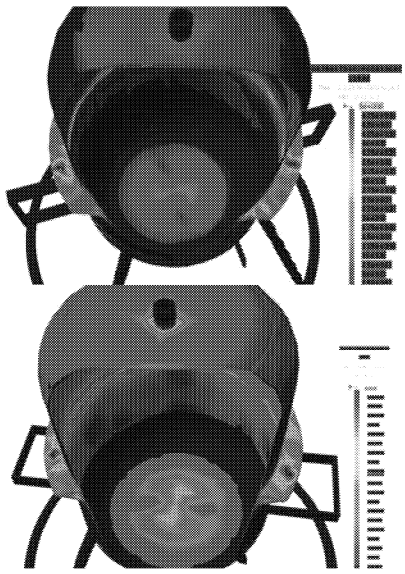


Figure 9 : Stress at the center of cylinder before and after modification in View of the center of bottom half body

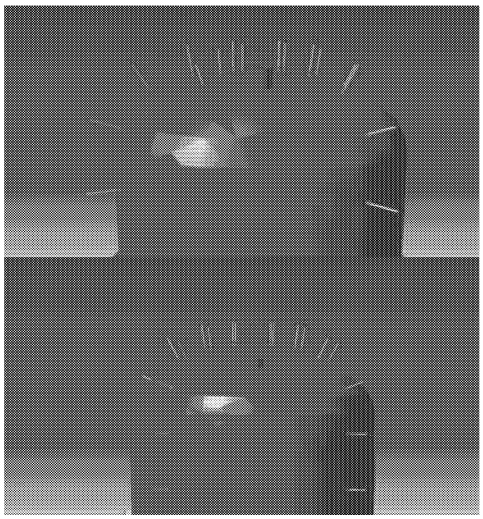


Figure 10 : Graph of displacement before and after modification

REFERENCES

[1] V. Rigaud, Robotics state of the art versus subsea intervention, Report of the UNION project D1.2, ESPRIT-BRA, 1994.

[2] I.D. Bonnon, Rov intervention for effective cable burial-an overview of recent achievements, Oceans 94, vol. 2, Brest, France, 1994, pp.515–520.

[3] J. Amat, et al., Garbı´: The UPC low cost underwater vehicle, Undersea Robotics & Intelligent Control, A joint US/Portugal Workshop, 1995, Lisbon, Portugal, pp. 19–24.

[4] S.M. Smith, Modular low-cost AUV systems design, Undersea Robotics & Intelligent Control, A joint US/Portugal Workshop, Lisbon, Portugal, 1995, pp. 153–157.

[5] J. F. Lots, D. M. Lane and E. Trucco. Application of 2 ½ D visual servoing to underwater vehicle station keeping. Proc. of the IEEE Oceans Conference, Providence, USA, September 2000.

[6] Beer. F.P, Johnston. E.R, 1998, "Mechanics of Material", Mc Graw-Hill

[7] Shigley. J.E, 1998, "Mechanical Engineering Design", Mc Graw-Hill

[8] Saeed Moaveni, 2007, " Finite Element Analysis Theory and Application with ANSYS", Prentice Hall, 3 edition

[9] Zamani. N.G, 2005, " CATIA V5 FEA Tutorials, Release 14", Schroff Development Corporation

AUTOMOTIVE SIMULATION

A CONCEPT OF AN INTEGRATED SUSPENSION CONTROL LOGIC ARCHITECTURE AND ITS TESTING

Grzegorz Ślaski
Instytut Maszyn Roboczych i Pojazdów Samochodowych
Politechnika Poznańska
ul. Piotrowo 3
60-965 Poznań
Poland
E-mail: Grzegorz.Slaski@put.poznan.pl

KEYWORDS

Motor vehicles, controlled suspension, dynamic simulation, adjustable damper

ABSTRACT

The article presents a concept of integrated suspension control software with individual modules and blocks in the software structure presented in detail. Some results of simulation testing of this software are described. Later on some information about prototype installation of the whole control system in a bus is also presented. The article finishes with a description of the most important results obtained during a field test of prototype hardware installation and prototype software implemented on special hardware and software platform dedicated to control software prototyping in the automotive industry.

INTRODUCTION

One way to improve ride quality and safety level is to optimize suspension parameters to the weight of car and car load, and also the type of road (in effect the type of road excitation). There are two elements of suspension having influence on car suspension performance – a spring element and a damper. In passive suspension the designer must assume some most frequent conditions of driving and find the best suspension stiffness and damping for those conditions.
Using spring elements of variable stiffness and damping elements of variable damping ratio we can get suspension able to adopt to various driving conditions.
As in passive suspension the designer has to find an optimal parameter set for the spring and damper to reach the compromise between comfort and safety; in the suspension of variable stiffness and damping he can find an area of compromise (not single values of stiffness and damping).
The only problem then is to find an algorithm to find current best suspension parameters set. Of course after finding an algorithm we will also find additional problems with putting this algorithm into practice.
The abilities of today's industry provide us with various types of dampers of variable damping ratio. The only problem is to build an appropriate control system to control suspension with such a damper.
This article presents some concepts as to how to organize control software for suspension of variable damping ratio.

CONTROL STRATEGIES AND INPUT DATA

Significant improvements with new and cheaper electronic technologies in the area of sensors, actuators, control units give a new perspective on the selection of control strategy.
We can find such fields for improvement in car dynamics using controlled suspensions:

- improvement in ride comfort – the driver is less tired, active safety and fatigue strength of car construction are improved,
- preventing vehicle's body pitch - possible during sudden braking and accelerating, fewer changes in suspension travel, smaller changes in wheel geometry, better vehicle handling,
- preventing vehicle's body from rolling – possible in sudden steering maneuvers, fewer changes in suspension travel and the rise of gravity center, smaller changes in wheel geometry, better vehicle handling,
- minimizing dynamic wheel loads – a better use of road grip, better stability and braking performance, better handling,
- assistance with other active safety control systems optimizing suspension performance for temporary needs of other safety systems.

The most difficult aspect of designing a suspension controller is to find a way to get a trade-off among conflicting goals – better ride comfort or better handling.
There are some comfort strategies used to control suspension with adequately fast adjustable shock absorbers – for example the skyhook strategy or groundhook strategy to get better ride comfort. Sometimes safety is more important than better ride comfort, and we have to find other strategies to increase vehicle active safety. A special algorithm is always needed to find each time which strategy is more important.
To execute this algorithm and algorithms of the comfort and safety strategies we need to have information about vehicle dynamics at a given moment. It can be the data about body mass and wheel mass accelerations and velocities, brake pedal movement or steering wheel angle and rotational velocity. Some of those data cannot be measured directly. An estimation algorithm is necessary in order to find that; it should use other measured data. It concerns for example body mass and wheel mass velocity data. This problem is one of the most difficult and important tasks of controller's logic.

CONCEPT OF THE SUSPENSION CONTROLLER ARCHITECTURE

Construction of the suspension controller with aforementioned functionality requires designing multilayer and hierarchical structure shown in figure 1.

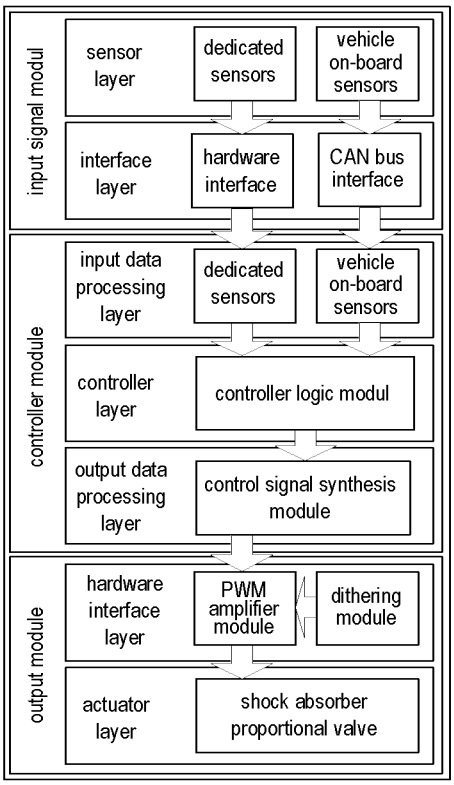


Figure 1. Suspension controller architecture diagram

Controller architecture shown in figure 1 assumes three principal modules:

1. input signal module – hardware signal conditioning module, and an interface to software part on input side of the system,
2. controller module – the module responsible for executing control algorithms; it consists of software signal conditioning layer, a control logic layer and software control signal synthesis layer,
3. output module – a second part of the system hardware, a software-hardware interface for output signals with a power stage (PWM amplifiers) for shock absorbers proportional by-pass valves.

Input signal module consist of sensor layer and interface layer – both for special dedicated sensors for controlled suspension and for on-board vehicle sensors used for the other control systems in today's cars. On-board sensors provide their information using CAN communication, so our controller needs to have special CAN bus interface. An investigated prototype of a bus wit controlled suspension had such information provided by an on-board sensor via CAN communication: vehicle speed, front and rear axle operational loads and brake pedal position.
Other dedicated sensors have their own hardware interface.

The controller module is the main module for the functionality of the suspension, It is responsible for this suspension's available functions.
Input data processing layer is a software interface to hardware sensor interface. Its main task is to change the electric signal value to measured physical parameter value. It can also consist of digital filter or detrend procedures.
The most important layer is the controller layer responsible for executing control algorithms of the controlled suspension. It has to execute complicated strategies of ride comfort and handling.
Next layer is the one of shaping output data – control signals on the basis of the results of logic operation. The functioning of this layer depends on the type of actuator, which imposes type and shape of control signal.
The last, output stage module of suspension control system consists of a hardware interface layer and actuator layer.
The hardware interface layer is hardware part we can call a power stage – it has to be an interface to the software output data processing layer and simultaneously it has to be able to produce bigger power signals. So it consists of some amplifier elements.

CONCEPT OF CONTROLLER LAYER

As the most important from point of view of executing control algorithm, the controller layer is first of all responsible for good control of car suspension. Its overall form is shown in figure 2.
The core of the system is an arbiter block, which coordinates the control and chooses a more important strategy at the moment. It can choose between comfort and safety strategies, so two other very important blocks are just a comfort strategy block and a safety strategy one.
Both these two blocks constantly make calculations and decisions. Only the arbiter block decides which of the two control signals should be used at given moment and uses one of them.
The last block is control signal synthesis and adjusting block shaping and also adjusting control signals to specific shock absorbers.

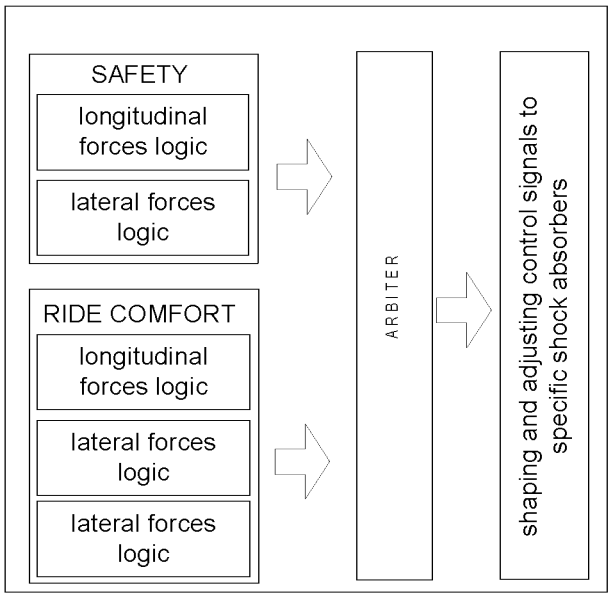


Figure 2. Suspension controller's logic architecture diagram

Safety block strategy is shown in figure 4. This is a block executing control algorithm responsible for good vehicle handling. It is assumed that unsafe situations occur during sudden driver's operations with the brake, accelerator pedal or steering wheel movement. For that reason two blocks inside "safety" block are "steer_logic" and "brake_logic".

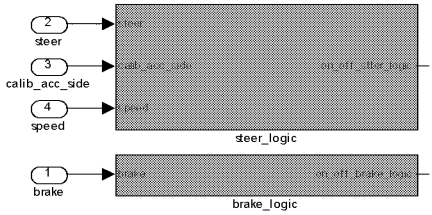


Figure 3. Suspension controller's logic architecture diagram

"Steer_logic" block detects sudden and fast steering wheel operations and bigger than trigger value of side accelerations. Taking into account also vehicle velocity this block takes decision about increasing damping ratio. An example of input and output signal of this block is shown in figure 4. The output signal is a control signal and its domain is logic signal domain - 0 or 1 value.

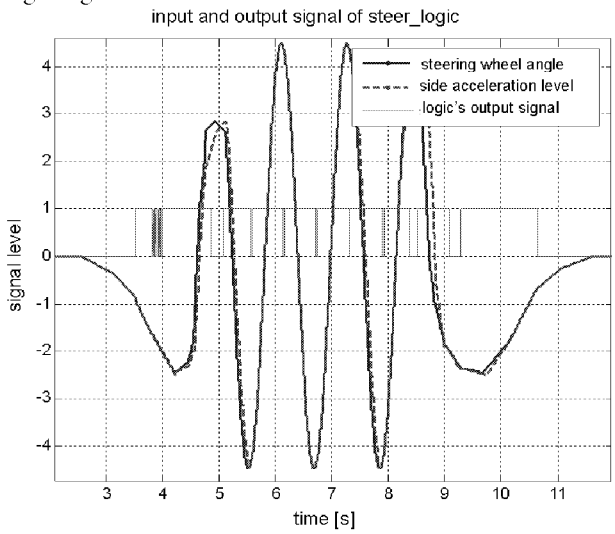


Figure 4. Input and output signals of "steer_logic" block

The second block is built in a similar way, but it detects sudden and very fast movement of the brake pedal. It is a "brake_logic" block. It has to detect not only very big brake pedal travel but also big travel velocity to recognize an unsafe situation.

Next module of control logic is comfort block. The task of these blocks is to control damping in such a way that the RMS terms of body acceleration are minimized. There can be skyhook policy ("skyhook_logic" block) used for pure semiactive suspension or adjustment of damping ratio adequately to load changes in slower adjusted suspensions ("load_logic" block).

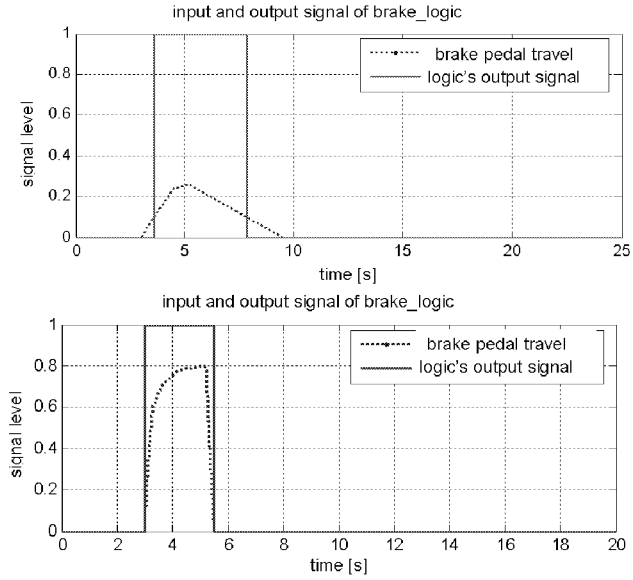


Figure 5. Input and output signals of "brake_logic" block

A prevention of excessive wheel travel ("hop_logic") is also needed. The configuration of this block is shown in figure 6.

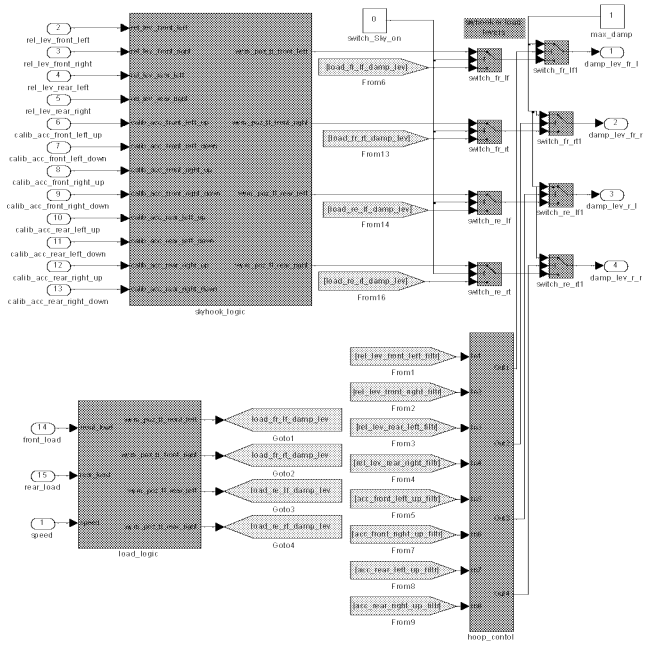


Figure 6. Structure of "comfort_logic" block

The superior module of the controller is the arbiter module. Its task is to choose which strategy – comfort or safety – is optimal to use at a given moment and adequately switch control signal synthesis to appropriate logic. The arbiter module works using logic signal value to make switching between blocks. This logic works on an established principle that safety logic is more important and first to use in case of a sudden manoeuvre made by the driver. The internal structure of arbiter block is shown in figure 7.

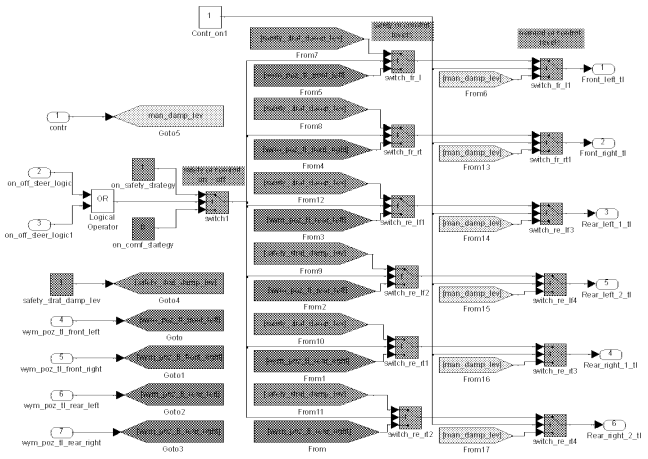


Figure 7. Structure of "arbiter" block

An example showing the functioning of this block and switching between comfort and safety logic is shown in figure 8.

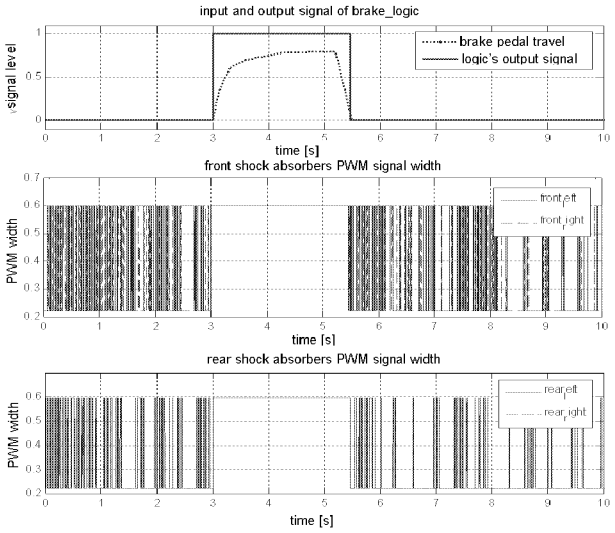


Figure 8. Results of a simulation made for testing "arbiter" block showing priority of safety logic.

TESTING OF CONTROLLER CONCEPT

One of the most difficult things during designing suspension controller is not only to find rules and algorithms but also to find algorithm parameter trigger levels switching function and strategy of control. It can be done during testing the controller.

It is possible to build a controller and make field tests of a vehicle equipped with such a controller or to make some simulation tests with controller software or even with a real controller. The second case concerns hardware in the loop simulation.

During tests simulation and field tests were used. Simulation for examining control strategy was first carried out to examine separate control strategies – separately for "steer_logic", brake logic, etc. It was done with only the input signal needed for a given strategy and using logical and physical value output signal. Examples of results of these tests are presented in figures 4,5 and 7.

First artificial input signals were built, later, after some field test in real vehicle, signals recorded during field tests were used.

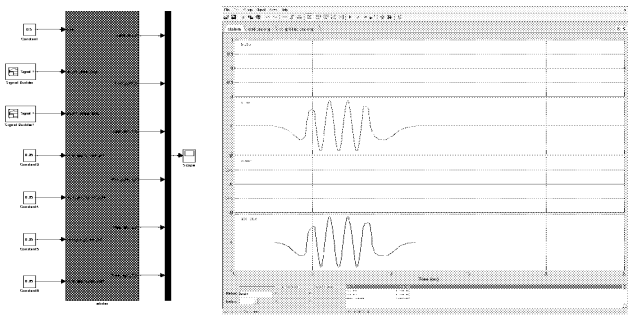


Figure 9. Simulation configuration of model and test input signals used for these simulations.

Tests made using simulations allowed to select correct tentative trigger values for activate safety values for "brake_logic" or "steer_logic". These values were calibrated later in field tests.

After using simulations, a prototype of the whole controlled bus suspension was prepared. Hardware structure of this system is shown in figure 10.

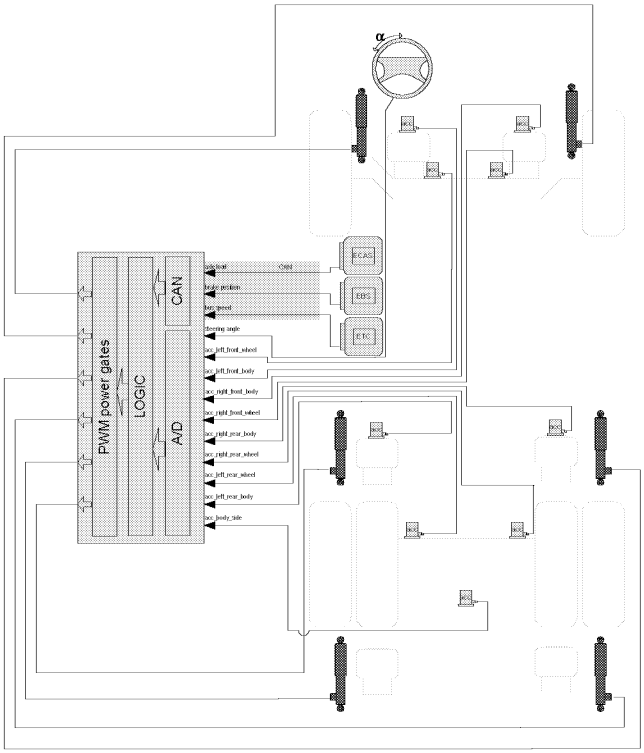


Figure10. Structure of hardware installation of prototype controlled suspension in a bus.

Front shock absorbers are controlled separately and rear ones are controlled in pairs - two left together and two right together.

This system was controlled by hardware and software platform used for prototyping control systems in automotive industry with software prepared during simulation tests. This system also allowed observing and recording data of control process for later off-line analyses.

Electronic hardware for prototyping control systems is shown in figure 11 as it was placed in a bus during field experiments.

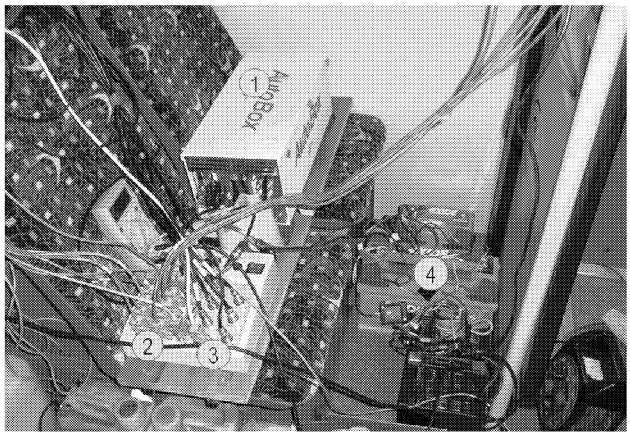


Figure 11. Structure of "comfort_logic" block: 1 - AutoBox - controller prototyping device, 2 – input signal conditioning module, 3 – output signal conditioning module (PWM), 4 – system power supply,

During field test control software was tested for:

- braking logic in medium and hard braking tests,
- safety logic in ISO double lane change test,
- comfort logic during tests on different roads.

Results for "steer_logic" have shown that during the time of double lane change this logic gave signal for using safety damping settings. The problem was with not really sharp borders between states for which safety settings were needed. It is the topic of further research.

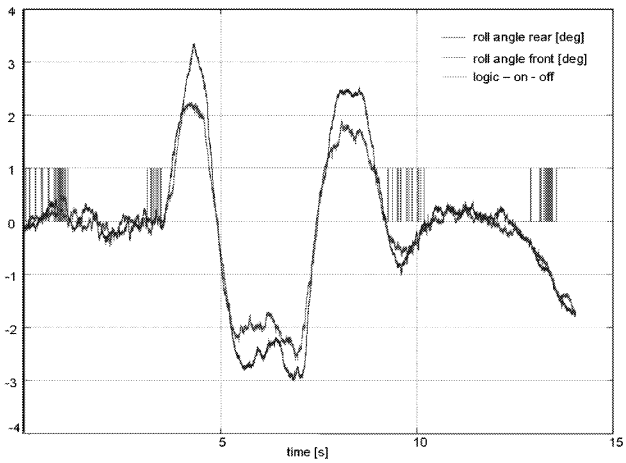


Figure 12. Results of "steer_logic" field test – safety settings are switched on during double lane change.

There was also possibility to use constant but various levels of damping ratio for bus shock absorbers, There were three settings used, called respectively "comfort" "road;' and "sport" for each time harder settings of shock absorbers. The "road" level was comparable with passive safety. These settings were compared with active comfort logic control. Figure 13 shows that there was some improvement in ride comfort using active control.

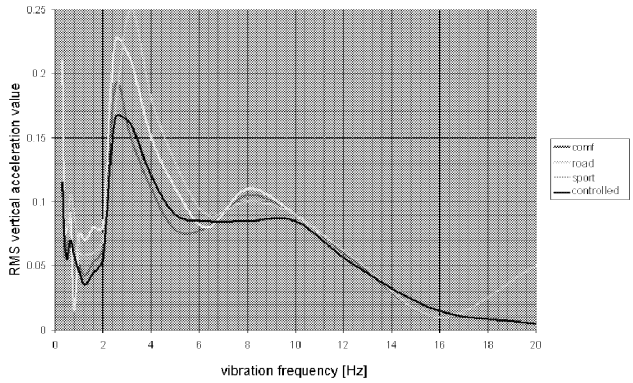


Figure 13. Results of comparing comfort ride for various damping levels and active comfort control of suspension

CONCLUSIONS

Off-line simulations and field tests of prototype control software proved that assumed concept of integrated suspension control logic is effective for improving ride comfort and handling properties of the car. It is especially effective with transient manoeuvres.

Controlled suspension gives possibilities to use comfort damping ratio during normal, smooth drive and always to have access to and a possibility to activate fast such damping ratios that will improve handling properties. It is especially important when driving at higher velocities – so more important in coaches than in for example city-buses. But with the city-buses there is a big operation load gap between a crowded bus during rush hours and an almost empty bus during for example late evening hours or weekend days.

Connecting variable and adjustable damping ratio with variable stiffness of pneumatic suspension it is possible to get fully adaptive suspension, appropriate for the type of road, driving style and load level.

REFERENCES

Ślaski G. 2006: Simulation of control system of semiactive suspension In process of controller building with Simulink use. Zeszyty Naukowe Wydziału Mechanicznego Politechniki Koszalińskiej, nr 38, p. 202-209,

Williams R. A.1997: Automotive active suspensions Part 1: basic principles., Engineering Centre, Jaguar Cars, Coventry Proc. Instn Mech. Engrs, Part D, p. 415-426,

Williams R. A.1997: Automotive active suspensions Part 2: practical considerations., Engineering Centre, Jaguar Cars, Coventry Proc. Instn Mech. Engrs, Part D,1997,p. 427-444,

AUTHOR'S BIOGRAPHY

GRZEGORZ ŚLASKI is a lecturer an researcher at the Institute of Working Machines Motor Vehicles at Poznań University of Technology. His research cover two domains, both connected with vehicle dynamics: field testing of car dynamics and car dynamics control system simulation and prototyping.

NUMERICAL SIMULATION OF VEHICLE PLATOON AERODYNAMICS

Marek Maciejewski
Ewa Narbuntowicz
Instytut Maszyn Roboczych i Pojazdów Samochodowych
Politechnika Poznańska
ul. Piotrowo 3
60-965 Poznań
Poland
E-mail: marekm@sol.put.poznan.pl

KEYWORDS

Road vehicles, vehicle platoon, aerodynamic interference

ABSTRACT

Fundamental purpose of realized simulations was the estimation of the size and importance of aerodynamic interference between vehicles moving in a platoon. In the work, it has been mainly focused on the aerodynamic drag and the analysis of its formation for particular vehicles. Main indicators of interference were the changes of aerodynamic drag coefficients of particular vehicles according to the configuration of a platoon and the distance between particular vehicles. Quite a large number of simulation results have made the base for realization of extensive comparisons and analyses.

INTRODUCTION

One of efficient ways to cope with increasing problems of road congestion, environment pollution and excessive energy consumption is introduction of an advanced, intelligent and automatic control of road traffic. Such control has little connection with the traditional traffic light control, where vehicles are grouped in order to pass in organised manner by an intersection; but it consists in automatically controlled, unphysical coupling of the vehicles in groups moving at common speed and with little gaps between particular vehicles. This road traffic technique is defined as platooning, and leads not only to considerable increase of road capacity, but it also improves road traffic safety. Not trivial is also substantial decrease of aerodynamic drag for the automatically coupled group of vehicles, and as consequence of this, considerable energetic savings and reduction of exhaust gases emission.

Traditionally, aerodynamic analyses are usually focused on consideration of individual vehicles with the purpose of determining the drags and drag coefficients, or else defining physical phenomena concomitant with air flow. Till now it is carried out only few experimental tests and numerical simulations of aerodynamic phenomena occurred between vehicles coupled automatically and moving in the direct proximity. In such situation, it is advisable to complete the typical studies of individual, isolated vehicles with more complicated and advanced studies, i.e. the investigation of the mutual interaction of coupled vehicles in aspect of their aerodynamics.

Presentation of results has been complemented with the description of numerical method, detailed definition of the research range and simulation conditions, discussion of examined objects and discretizations of computational domains. Selected results have been documented through suitable visualisations helpful in formulation of conclusions.

METHOD OF THE NUMERICAL SIMULATION

The numerical simulation of the aerodynamic phenomena around the vehicles moving in close order is based on the Navier–Stokes equations for incompressible fluid. A disadvantage of such flow description is the lack of a time evolution term in the continuity equation. It causes that this equation can be only treated as a divergence-free constraint for velocity. Simultaneously, the pressure occurring only in the momentum equations can be interpreted as a specific Lagrange multiplier, which the task is to force such form of velocity field that ensures the fulfilment of divergence-free constraint. One of solutions, which allows to circumvent the above constraint, is suitable relaxation of the problem. In our simulations, it is realised by introduction into the continuity equation an additional term containing the time derivative of pressure, or otherwise, the introduction of the artificial compressibility (Drikakis and Rider 2005, Nithiarasu 2003), which ensures sufficient and explicit coupling the continuity and momentum equations.

The solution of the problem for assumed initial and boundary conditions has been realised with the use of spatial finite volume discretization. In this method, the search for solution in computational domain is carried out separately in all cell-centred finite volumes. In every cell all variables assume constant, volume-averaged values, and their variation results entirely from the flux equilibrium conditions on cell faces, calculated from the approximate solution of the Riemann problems for reconstructed fields of primitive variables. For solving the above problem, it has been used an upwind scheme referring to the signs of local eigenvalues. The determination of convective and viscous fluxes across mesh edges (faces) has been given a base for the time integration and computation of variable values in particular cells. The length of time steps has been chosen automatically with regard to stability conditions. In general,

for considered unsteady problem it would be advisable to use the dual time integration, but owing to the adaptive mesh refinement and thus very small real time steps, it has been kept the standard time integration.

Taking into account the essential influence of mesh arrangement onto accuracy of results, and the fact that the optimal mesh shape can vary and it is a priori unknown, the program for numerical flow simulation has been prepared in an adaptive version. It means that the program performs a suitable standard analysis, determines the subdomains of the mesh where the solution is not sufficiently accurate, and improves the discretization in those regions until the assumed accuracy criterion has been achieved. Thus the adaptive algorithm, in comparison with the standard ones, contains two new elements: an *a posteriori* error estimator and a mesh refinement procedure. This adaptive procedure changes the mesh density through the bisection of the longest cell faces, and thus also finite volumes, in indicated regions. Just as the fundamental CFD program, the adaptive procedures have been fully developed by present writers.

RESEARCH RANGE AND CONDITIONS

The domain of considerations has been restricted exclusively to the vehicles moving though independently still in organised sets characterised by small clearance gaps between particular vehicles. In the research has been tried to imitate typical road traffic conditions, i.e. all vehicles have moved in space unlimited by physical or artificial boundaries influenced the aerodynamic phenomena. Moreover, realised simulations in no wise tried to imitate aerodynamic situation from wind tunnels, and thus have avoided primary sources of error in wind tunnel testing, i.e.:

- scale (Reynolds number) effect,
- air flow in relation to pavement and generating the spurious boundary layer,
- flow blockage.

All simulations have been carried out with care for reconstruction of real road conditions. Vehicle speed was constant, equal to 20 m/s (72 km/h). Physical conditions (the density and viscosity) have corresponded with data characteristic for the air. Numerical studies have been carried out in relation to one typical, though geometrically simplified car model (hatchback). Assumption of such idealised model was of no importance for numerical simulations, but it has been let to observe some physical phenomena in flow domain more distinctly than in the case of the more streamlined model.

Though the full aerodynamic simulation of the flow around vehicles requires three-dimensional analyses, nevertheless, in order to carry out the simulations much quickly and less labour-consuming, it has been decided on the two-dimensional simulations. It is naturally conditioned by the fulfilment of certain geometric requirements regarding the form and symmetry of vehicle bodies.

The creation of simplified vehicle models, and then their placement in a properly geometrically specified outer boundaries of flow domain, it has been defined configurations of computational domains. In order to ensure that no correction of drags and drag coefficients is needed, it has been assumed that the height of computational domains ("pseudo-wind-tunnels") is equal 12 meters, the length is equal to 50 meters, and the placement of the first vehicle is always identical.

The base for the discretization was unique defining the edges of computational domains through the nodal points placed on the lines of vehicle profile and outer outline of flow domain. Exemplary form of the initial discretization has been presented in Figure 1, and corresponded with it, the final, post-adaptive computational mesh has been inserted in Figure 2.

The boundary conditions for each specific computational domain have been defined in accordance with following scheme:

- on the vehicle profile: motionless solid boundary,
- on the left- and right-side of the "tunnel": inflow and outflow boundaries, respectively,
- at the top of flow domains: symmetric boundary conditions,
- at the bottom of "wind-tunnel": moving (with inflow speed) solid boundary.

For description of configurations of computational domains, it has been introduced the uniform notation according to the pattern:

ABCD–

where particular symbols (A, B, C and D) denote:

A the profile type (h – the hatchback),
B first gap (from 1 to 4 meters, every 1 meter), between the first and second vehicle,
C second gap (in meters, from 1 to 4), between the second and third vehicle,
D third gap (in meters, from 1 to 4), between the third and fourth vehicle.

In some figures (in legends) it has been additionally specified the position of each vehicle in the platoon according to the scheme

ABCD–#E

where the additional symbol denotes:

E the consecutive vehicle number in a platoon.

The comparative configuration, composed of only one vehicle, has been denoted by h000–.

RESULTS OF AERODYNAMIC SIMULATIONS

When analysing the simulation results it has been first of all focused on comparisons of aerodynamic drag coefficients (oblique simulation results), and not on presentation and discussion of direct results (pressure and velocity component distributions). Presentation of this later (also at different time moments) has no reason in search of general rules of the aerodynamic drag formation, though it can have great weight for studying certain detailed solutions. Moreover, the presentation of aerodynamic features expressed by oblique results can be simultaneously and dynamically (in time) referred to many various forms of vehicle platoon. Focusing on aerodynamic drag coefficient comparisons of vehicles in different platoon configurations has been prompted to conclude about consequences of aerodynamic interactions, i.e. the aerodynamic interference between vehicles in a platoon. In particular, basing oneself on 65

cases of vehicle configurations, it has been considered the following general groups of problems:

- interdependence of aerodynamic drag coefficients of vehicles in different platoon configurations, with reference to aerodynamic features of the comparative case, i.e. the individual vehicle,
- effect of the only one gap change on the aerodynamics of particular vehicles in platoons,
- effect of proportional change of two gaps on the aerodynamics of particular platoon vehicles,
- effect of proportional changes of all gaps between vehicles on the aerodynamics of particular vehicles.

In this publication, it has been illustrated only some, demonstrative examples of aerodynamic interference of vehicles. First example (Figure 3) refers to the aerodynamic drag formation of all vehicles in a platoon configuration, together with reference to single vehicle drag. All remaining examples (Figures 4–7) concern the effect of change in vehicle locations in platoons on the aerodynamic drag of particular vehicles, with restriction only to the last problem from their groups specified above. Namely, all gaps between vehicles in a platoon are equal, but different (1, 2, 3 or 4 meters) for various platoons.

Note that in all figures the first steps (seconds) of time analysis are connected with large fluctuations of drag coefficients. It results from unrealistic definition of initial conditions, but with time (after few seconds) these coefficients achieve the proper level of accuracy. The whole records of aerodynamic simulation results can be found in the report (Maciejewski et al. 2003).

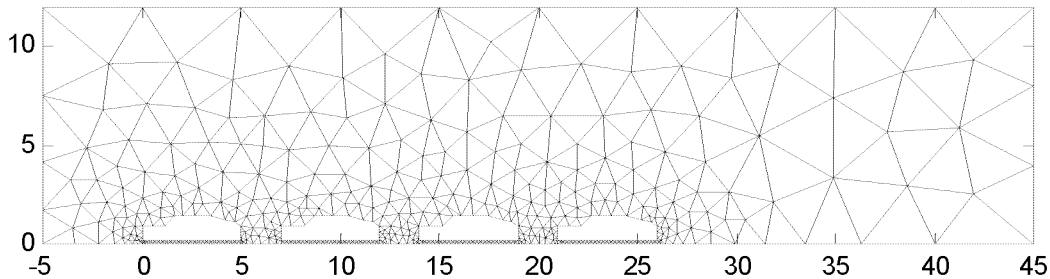


Figure 1: Exemplary Initial Mesh for the Computational Model h222– (all dimensions in meters)

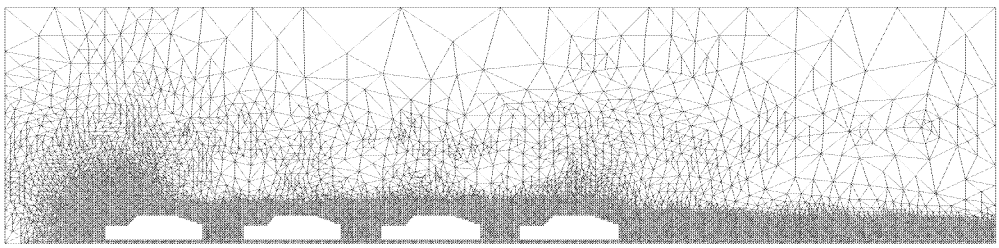


Figure 2: Exemplary Post-adaptive Mesh for the Model h222–

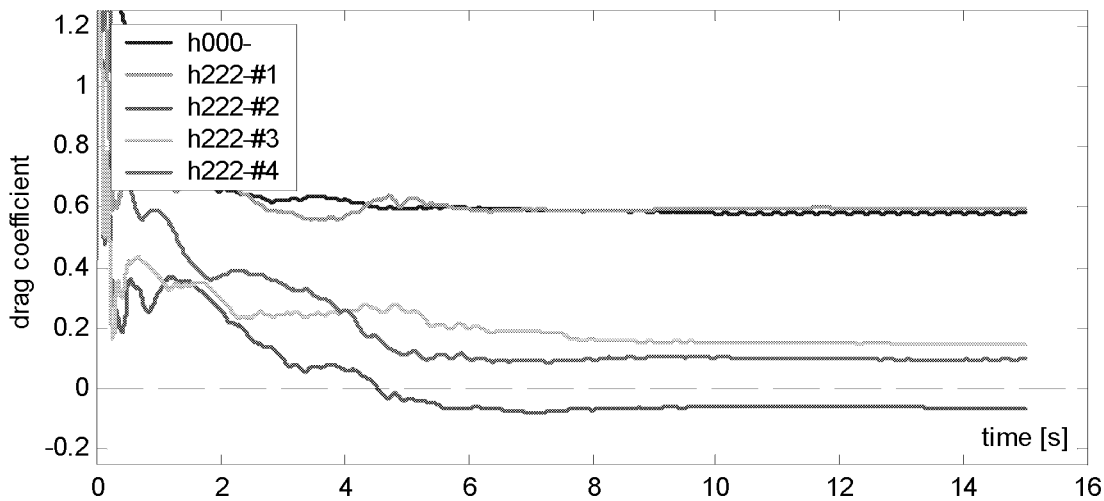


Figure 3: Aerodynamic Drags of Vehicles in the h222– Configuration and in Comparative Configuration (h000–)

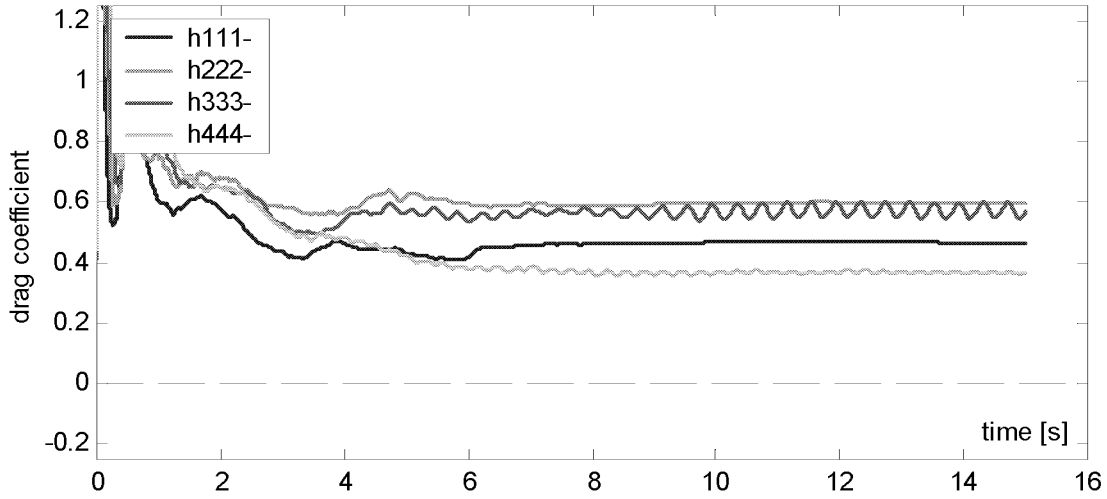


Figure 4: First (#1) Vehicle Drags for Constant Gaps (1, 2, 3 or 4 meters) in Particular Platoons

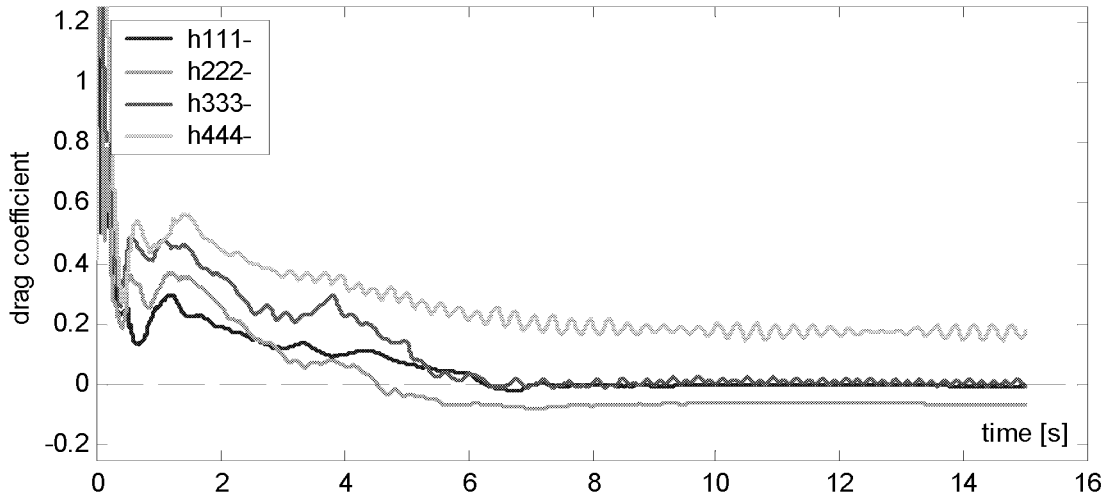


Figure 5: Second (#2) Vehicle Drags for Constant Gaps (1, 2, 3 or 4 meters) in Particular Platoons

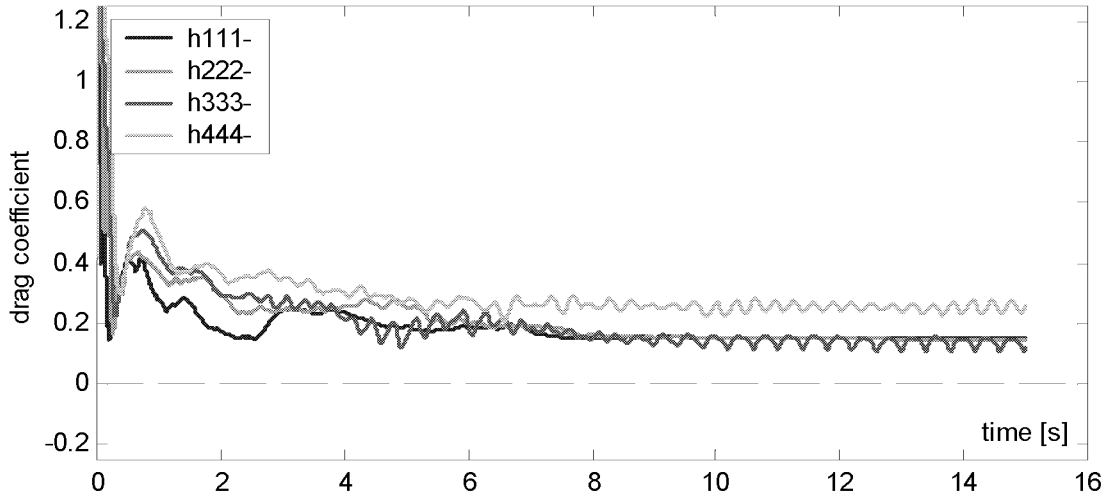


Figure 6: Third (#3) Vehicle Drags for Constant Gaps (1, 2, 3 or 4 meters) in Particular Platoons

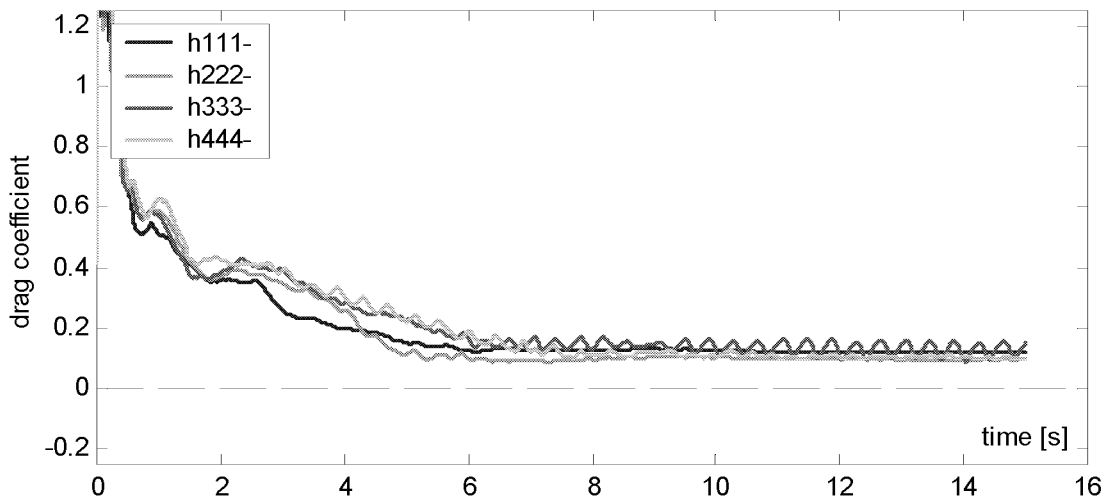


Figure 7: Fourth (#4) Vehicle Drags for Constant Gaps (1, 2, 3 or 4 meters) in Particular Platoons

CONCLUSION

Global and concise analysis of oblique results, and first of all the course of aerodynamic drag coefficients in time, has given cause for doing a number of observations and for deriving several recapitulative conclusions regarding the aerodynamic aspects of vehicle movement in platoons. The most general and most important of them have been placed below:

1. Movement of a vehicle platoon coupled automatically decreases global aerodynamic drag of all vehicles.
2. In general, the lower sum of gaps between particular vehicles, the lower global aerodynamic drag.
3. Gap decrease between two successive vehicles not always results in decreasing the aerodynamic drag coefficient of the second vehicle, and the first vehicle drag can increase or decrease.
4. The smaller gaps between vehicles (in examined range), the lesser tendency to fluctuations in the drag of particular vehicles.
5. The highest aerodynamic drag of the first vehicle (exceeding distinctly comparative level characteristic for single isolated vehicle) occurs for the gap from second vehicle equal to 2 meters, regardless of the size of remaining gaps. At that time the drag of second vehicle is always decidedly lower.
6. In vehicle platoon, in general, the largest is the aerodynamic drag of the first vehicle, and the least – the second vehicle drag, nevertheless there are happened deviations from this rule.
7. General rules forecasting the course and size of the aerodynamic drag are possible to define, but certain specific platoon forms can be not subject to them.

Many other, more specific observations in the range of influence of inter-vehicle gap changes on aerodynamic characteristics of particular vehicles in a platoon can find in the full description of this problem (Maciejewski et al. 2003). One should remember that some aerodynamic relationships defined in dependence on inter-vehicle gaps can change together with changes in vehicle shapes.

REFERENCES

Drikakis D. and Rider W. 2005. *High-resolution Methods for Incompressible and Low-speed Flows.* Springer-Verlag, Berlin, Heidelberg.

Maciejewski M. et al. 2003. „Symulacja interferencji aerodynamicznej przy ruchu pojazdów sterowanym automatycznie". Research Report 51-975/2003 BW, PP IMRiPS, Poznań (Nov), 1–76.

Nithiarasu P. 2003. "An efficient artificial compressibility (AC) scheme based on the characteristic based split (CBS) method for incompressible flows." *International Journal for Numerical Methods in Engineering* 56, No.13 (Apr), 1815–1845.

AUTHOR BIOGRAPHY

MAREK MACIEJEWSKI is engaged in two research domains. Both are connected with numerical methods in engineering. The first is geometrically nonlinear theory of thin-walled beams and its finite element realisation, and the second – the solution of viscous, incompressible flows on adaptive unstructured finite volume meshes, with applying to the road vehicle aerodynamics.

MATERIALS HANDLING, LOGISTICS AND TRAFFIC SIMULATION

PRODUCTION SCHEDULING AND MATERIALS HANDLING

IMPLEMENTING AN EVOLUTIONARY ALGORITHM FOR PRODUCTION SCHEDULING IN A CARDBOARD CORRUGATED FABRICATION PLANT

Fabian Vargas-Nieto
Departamento de Ingeniería Industrial, Universidad del Norte, Km 5 via Puerto Colombia, Barranquilla, Colombia

Jairo R. Montoya-Torres
Escuela de Ciencias Económicas y Administrativas, Universidad de La Sabana, autopista norte de Bogotá, Chía, Colombia

German Velásquez
Empaques Industriales Colombianos
Autopista frenta antiguo Aeropuerto, Soledad, Colombia

KEYWORDS
Manufacturing, scheduling, evolutionary algorithm, AI-based optimization.

ABSTRACT
This paper considers a production scheduling problem in a cardboard corrugator manufacturing plant. In the literature, this problem is known as the two-dimensional cutting-stock problem, and most of the optimisation techniques only focus on minimising either trim waste or pattern changes. This paper proposes an evolutionary algorithm to optimise the service level constrained to some preferences of production planning. Conducted experiments using real data are compared with the current scheduling method used in the plant.

1. INTRODUCTION

The manufacturing of corrugated cardboard boxes consists of the stages of pattern layout and finishing (i.e. printed, folded and glued) according to specifications that may vary between product styles. Because of its complexity for production management, the most important part of this manufacturing process is the pattern-layout stage. In the literature, the problem of pattern layout optimisation is known as the two-dimensional cutting-stock problem. Solution procedures are traditionally based on linear programming models or heuristic algorithms. In real-life practice, however, some plants still schedule corrugator manually. The major reason is that analytical methods and good heuristics do not fully capture the problem complexity. In effect, a pattern layout that is optimal or near-optimal in terms of trim waste may lead to bottlenecks at the finishing stage or to sub-optimal scheduling solutions for the whole plant. In addition, these troubles on production also concern the delivery of final product, affecting in this way due-date related performance indexes.

In this paper, we are interested in proposing a solution approach for the corrugator scheduling problem based on evolutionary algorithms in order to optimise the weighted order completion rate. In fact, due to the complexity of parameters to take into account and to the relation between these parameters, the corrugator scheduling problem is a combinatorial optimisation problem. Since an evolutionary algorithm (EA) is an intelligent computer-based optimisation technique that has provided very good results when applied to solve other combinatorial and engineering optimisation problems (Fogel, 1995; Back, 1995), it seems thus interesting to apply EA to solve the corrugator scheduling problem.

2. PROBLEM STATEMENT

In the literature, researchers have traditionally referred the corrugator production scheduling problem as the two-dimensional cutting-stock problem (also named the trim or trim-waste problem). In its classical definition, a factory produces a material such as linerboard in long rolls of fixed width; customer orders specify a desired number of sheets of a certain length and width. The widths of the cut sheet rarely allow full utilisation of the roll, resulting in trim waste, which often is not salvageable. The possible configurations of pattern layouts are as shown in Figures 1 to 4.

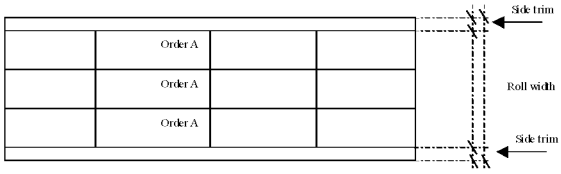


Figure 1. Pattern layout type 1.

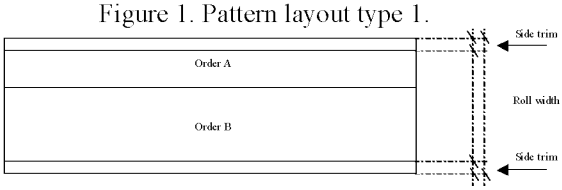


Figure 2. Pattern layout type 2.

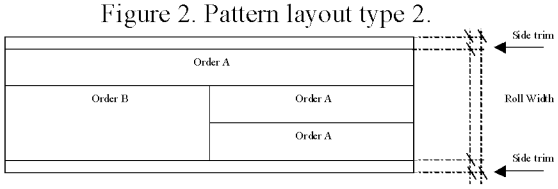


Figure 3. Pattern layout type 3.

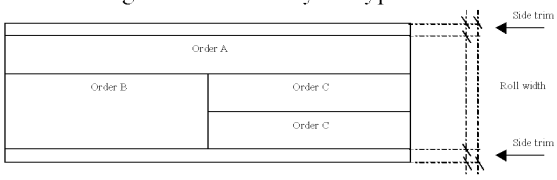


Figure 4. Pattern layout type 4.

The objective is to arrange orders so the sum of sheet widths simultaneously being cut will most closely equal the roll width. This formulation applies to the manufacture of corrugated boxes because the corrugator knives operate independently. Thus, the choice of two orders to be run at the same time is unaffected by the sheet length of either. A feasible solution involves a set of cutting patterns and the number of times each pattern will be used (a cutting pattern is simply a combination of sheet widths whose sum does not exceed that of the roll). Pattern generation can easily become an overwhelming task since corrugator scheduling also requires the sequence in which the chosen patterns will be run.

Formally, the classical cutting stock problem can be formulated as a 0-1 linear program as follows:

$$Minimize \sum_j c_j x_j + \sum_k s_k y_k$$

subject to:

$$\sum_j a_{ij} x_j = 1 \qquad \forall i$$
$$\sum_j u_{jk} x_j \le A_k y_k \qquad \forall k$$
$$x_j \in \{0,1\} \qquad \forall j$$
$$y_k \in \{0,1\} \qquad \forall k$$

Where
x_j is 1 if element j is used and 0 otherwise (j=1,…,n),
y_k is 1 if stock size k is used and 0 otherwise,
a_{ij} is 1 if order i is completed in element j and 0 otherwise,
u_{jk} is the feet of stock size k required by element j,
A_k is the feet of stock size k available in inventory,
c_j is the total cost of using element j exclusive of the cost of changing to the required roll stock size. It includes the cost of corrugator time and paper used plus the cost of pattern changes. This value is adjusted if any order is not produced at the maximum quantity to reflect the cost of producing the whole order,
s_k is the cost of loading stock size k on the machine. It is assumed that if two or more elements use the same stock size, they will be run sequentially so there will only be one set-up for each stock size.

The objective is to arrange orders so the sum of sheet widths simultaneously being cut will most closely equal the roll width. This formulation applies to the manufacture of corrugated boxes because the corrugator knives operate independently. Thus, the choice of two orders to be run at the same time is unaffected by the sheet length of either. A feasible solution involves a set of cutting patterns and the number of times each pattern will be used (a cutting pattern is simply a combination of sheet widths whose sum does not exceed that of the roll). Pattern generation can easily become an overwhelming task since corrugator scheduling also requires the sequence in which the chosen patterns will be run.

Early works on the cutting-stock problem are based on linear programming formulation in order to minimise trim-waste (Gilmore and Gomory, 1963; Wade, 1964; Haessler, 1975; Dyckhoff, 1981; Haessler and Talbot, 1983; Umetani et al., 2003;). Although these formulations are still very popular and are used in various commercial computational packages, they lead to a big solution space and have difficulties in dealing with non-linear problems, which are most common in the real world. As a result, heuristic solution procedures have become increasingly popular in the literature in order to consider more practical situations in which it is needed a balance between the waste objective and customer service, production costs, and machine and workforce utilisation (Haessler, 1975; Haessler and Talbot, 1983; Hinxman, 1980; Bookbinder and Higginson, 1986). The efficiency and effectiveness of heuristic solution procedures, however, heavily depend on the heuristic used. Finding good heuristics is often as difficult as solving the problem itself (Liang et al., 2002). When applied to the corrugator scheduling problem many heuristic approaches to the cutting-stock problem have been unsuccessful either because they attempted to generate patterns sequentially and had trouble with trim loss at the end of a sequential procedure, or because they used linear programming to minimise trim, and performed poorly with regard to number of pattern changes and order congruity (Haessler and Talbot, 1983; Bookbinder and Higginson, 1986). The relationship between customer service and trim waste has been less studied. In this paper, we propose an evolutionary algorithm that optimises the weighted order completion rate.

3. DESCRIPTION OF THE EVOLUTIONARY ALGORITHM

3.1. Background

An evolutionary algorithm (EA) is a problem solving technique that uses the concepts of evolution and hereditary to produce good solutions to complex problems that typically have enormous search spaces and are therefore difficult to solve (see Figure 5). The biggest difference with other meta-heuristics (like Taboo Search (TS) or Simulated Annealing (SA)) is that EA maintains a population of solutions rather than a unique current solution. Solutions are coded as finite-length strings called chromosomes and a measure of their adaptation (the fitness) is computed by an engine. Starting from an existing population, each iteration generates new chromosomes by applying operators (like crossover and mutation) to two randomly chosen parents. The main advantage of EA is its intrinsic parallelism, which allows the exploration of a larger solution space than in TS or SA (Sevaux and Dauzère-Pérès, 2003). A well-designed EA allows for the efficient and effective exploration and exploitation of the problem's search space of feasible solutions in an effort to identify the global optima, or near-optimal, solution to difficult problems.

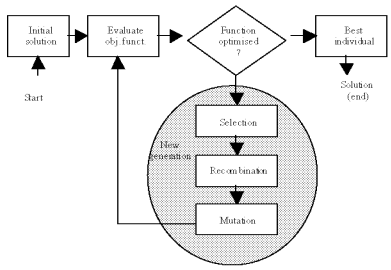


Figure 5. General description of an EA.

Early applications of EA's are found in the literature to solve complex combinatorial optimisation problems (see (Back, 1995; Fogel, 1995)). In a broad way, the evolutionary algorithm presented here seeks to maximise the weighted order completion rate, measured in terms of weighted proportion of units programmed from the total units requested per order. The algorithm works according to the order preferences established by the corrugator's scheduler for the orders available for planning. The algorithm will provide an order combinations set for a particular grade (combination of liner and medium paper types required to provide specific product properties to the cardboard) that will correspond to a corrugator schedule.

3.2. Solution representation

Each order combination in a corrugator schedule consists of several cutting patterns, which define the way one or more orders are going to be cut. Besides, every scheduling solution must determine the following information:
- Number of cutting patterns to run in the corrugator.
- The length of every cutting pattern.

A scheduling solution will be thus represented as a vector of integer number denoting the identification of the selected cutting patterns, and as a vector of numbers ranged between 0

and 1 representing the proportion of the total length to be produced for each selected cutting pattern.

It is to notice that the total length of a particular cutting pattern will be determined by the first of the two orders that completes its particular number of units requested. Hence, we can deduce the following statements:

- A solution will be represented by two chromosomes.
- The chromosomes will not necessarily have the same size because two solutions may be formed by a different number of cutting patterns.

The cutting patterns that create the solutions belong to a previously generated set of cutting patterns. The former set must meet the following conditions:

- The side trim of every cutting pattern must oscillate between the maximum and minimum levels allowed (determined by the corrugator's scheduler).
- Orders longer than 300 meters must be combined within a cutting pattern no shorter than 300 meters.

Even though the algorithm does not control the side trim directly through its fitness function, the selection of cutting patterns does.

3.3. The algorithm

The algorithm is now described in detail.

Step 1: Initialisation

The following heuristic procedure was developed in order to create the initial population of the EA:

1 Define the maximum and minimum side trim allowed.
2 Based on the parameters defined in the previous step, generate the set of acceptable cutting patterns. That is, all possible combinations in pairs of orders to cut whose side trim are between the maximum and the minimum allowed.
3 Calculate the length of the acceptable cutting patterns.
4 Do until the proportion planned of each order to be planned is at least 100%:
   4.1 Choose randomly an acceptable cutting pattern, which has not been chosen before, and its length.
   4.2 Calculate the proportion planned of each order to be planned.

Step 2: Fitness function evaluation

To evaluate fitness function two aspects must be taken into account:

- To plan the greatest number of due-date and on-date orders.
- To plan the least number of anticipated orders and the least number of upgraded orders.

To fulfil these preferences, an order weighting method was developed. This procedure establishes that the scheduler ranks the orders according to urgency level or to the necessity of planning a specific order. Another alternative that allows the scheduler to transmit his preferences is the option of defining a priority status which consists of identifying an order as immediate or optional. Immediate orders are the due-date and on-date orders. Optional orders are the anticipated and upgraded orders. Therefore, fitness will be computed as the difference between the number of units planned and the number of units requested multiplied by its weight. The smaller this difference, the better a specific solution.

The target quantity to be planned for each order will depend directly on the priority status determined for that specific order. If the order was labelled as "immediate", the target quantity to be planned will be the number of units the client requested. On the other hand, if an order priority is "optional", the ideal or target quantity will be 0. That is because it is not desirable to produce anticipated orders or upgraded orders (orders which had an original grade lower than the one they were planned). In general, the weight measures the scheduler tolerance level related to the difference between the total planned units and the total requested units per order. The bigger the weight for an order, the lower the scheduler tolerance level.

Step 3: Selection

The original tournament selection procedure is used to pick k parents randomly and after a few calculations it returns the best of them (Fang, 1994).

The value of k is set as the 1/10th of the original population size.

Step 4: Recombination

Two recombination operators are defined: One point crossover and uniform crossover, i.e. in each crossover the offspring is 4 sons. The crossover operators are described below:

- One point crossover consists on picking a number that is less than the chromosome size. The offspring will be formed with the genes inherited from the parents located before the position determined by the number selected, and the genes located after that position will switch each other.

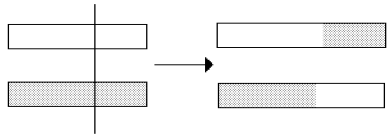


Figure 6. One point crossover

- In uniform crossover, every gene coming from the first parent has a probability of 0.5 of switching position with the corresponding gene in the second parent (Fang, 1994).

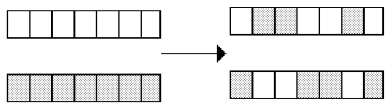


Figure 7. Uniform crossover

Step 5: Mutation

Two mutation operators are determined:

- Cutting pattern addition mutation, which adds a cutting pattern randomly along the chromosome.
- Proportion mutation, which generates a new proportion for an existing cutting pattern.

Step 6: End

The number of generations is the stop criteria for the algorithm.

4. IMPLEMENTATION

The results obtained after implementing the algorithm were compared with the results provided by the Company using the current manual production schedule. 11 different order sets were selected to represent the different schedules for the corrugator. The order sets used cannot be presented in detail because of confidentially reasons, but they consist on a variety of grades, dimensions, units requested, and immediate and optional order combinations. Performance metrics considered for the analysis of solutions provided by the algorithm were

Weighted Order Completion Rate (WOCR), average side trim and additional cost incurred because of order upgrading.

Table 1 presents the average value obtained for each performance metric using the evolutionary algorithm and the manual scheduling method currently used in the Company. The WOCR for the algorithm increases on about 10% by using the proposed algorithm against the current manual scheduling method. Also, we can obtain other improvements such as a reduction in the upgrading cost. This last factor suggests a great level of flexibility, from the production manager perspective.

Table 1. Results

	This paper	Current
WOCR (%)	79.26%	69.09%
Side trim (cm)	6.49	6.27
Upgrading costs (C$)	$76.409	$146.764

5. CONCLUDING REMARKS

This paper proposed a new evolutionary algorithm (EA) to schedule production activities in a corrugator manufacturing plant. The algorithm also provides some insights about the trim lost and costs incurred when upgrading the corrugator. In addition to the improvement in the weighted order completion rate (WOCR), the algorithm is enough flexible to incorporate the empirical knowledge of the production manager when selecting the most appropriate schedule for the factory. Thus, it integrates both computer-based and empirical knowledge. Reduction in upgrading costs is also obtained after the application of the algorithm. These features make our algorithm very attractive for actual implementation in the Company.

For future work, the multiple-objective optimisation problem using a more global approach can be addressed. For instance, optimal solutions for the cutting-stock problem at the first stage of processing cardboard boxes may induce some problems when sequencing production tasks at the finishing stage of the corrugator. A balance between on-time customer delivery and optimisation of production resources has to be achieved in order to maintain the competitiveness in today's global market.

ACKNOWLEDGEMENTS

Part of this work was performed while the second author was Visiting Professor at the Department of Industrial Engineering, Universidad del Norte, Barranquilla, Colombia. Second author would like to thank the financial support of School of Economics and Management Sciences, Universidad de La Sabana, Chía, Colombia.

REFERENCES

Back, T. 1995. *Evolutionary Algorithms in Theory and Practice*, New York, Oxford University Press.

Bookbinder, J.H. and J.K. Higginson. 1986. "Customer Service vs. Trim Waste in Corrugated Box Manufacture". *Journal of the Operational Research Society*, vol. 37, pp. 1061-1071.

Dyckhoff, H. 1981. "A New Linear Programming Approach to the Cutting Stock Problem". *Operations Research*, vol. 29, pp. 1092-1104.

Fang, H.I. 1994. "Genetic Algorithms in Timetabling and Scheduling". PhD thesis, Department of Informatics, The University of Edinburgh.

Fogel, D.B. 1995. *Evolutionary Computation, Towards a New Philosophy of Machine Intelligence*. IEEE Press.

Gilmore, P.C. and R.E. Gomory. 1963. "A Linear Programming Approach to the Cutting-Stock Problem – Part II". *Operations Research*, vol. 11, pp. 863-888.

Haessler, R.W. 1975. "Controlling Cutting Pattern Changes in One-Dimensional Trim Problems". *Operations Research*, vol. 23, pp. 483-493.

Haessler, R.W. and F.B. Talbot. 1983. "A Zero-One Model for Solving the Corrugator Trim Problem". *Management Science*, vol. 29, pp. 200-209.

Hinxman, A.I. 1980. "The Trim-Loss and Assortment Problems: a Survey". *European Journal of Operational Research*, vol. 5, pp. 8-18.

Liang, K.H., X. Yao, C. Newton, and D. Hoffman. 2002. "A New Evolutionary Approach to Cutting Stock Problems with and without Contiguity". *Computers & Operations Research*, vol. 29, pp. 1641-1659.

Sevaux, M. and S. Dauzère-Pérès. 2003. "Genetic algorithms to minimize the weighted number of late jobs on a single machine". *European Journal of Operational Research*, 151, 296-306.

Umetani, S., M. Yagiura and T. Ibaraki. 2003. "One-Dimensional Cutting Stock Problem to Minimize the Number of Different Patterns". *European Journal of Operational Research*, vol. 146, pp. 388-402.

Wade, C.S. 1964. *I.B.M. 1620 Corrugator Trim and Schedule Program*. I.B.M. Corporation, Kingston, N.Y.

BIOGRAPHIES

Fabian Vargas-Nieto is currently working as Thermal Transfer Division Director at Finotex Colombia S.A., in Barranquilla, Colombia. He is industrial engineer from Universidad del Norte, Barranquilla, Colombia, and is currently finishing the Master of Science degree at the same university.

Jairo R. Montoya-Torres is Associated Professor and Director of the undergraduate program on Marketing and International Logistics Management, School of Economics and Management Sciences, La Sabana University, Chia, Colombia. He holds the Ph.D. in industrial engineering from Ecole des Mines de Saint-Etienne and Université Jean Monnet, France. His research interests include scheduling, supply chain, and modelling and analysis of logistic and production systems. His e-mail is jairo.montoya@unisabana.edu.co

Germán Velásquez holds the degrees of Industrial Engineer and Master of Science in Industrial Engineering, both from Universidad del Norte, Barranquilla, Colombia. He is currently working as Production Planner at Empaques Industriales Colombianos S.A., a company within to the cardboard manufacturing industry.

DESIGN AND IMPLEMENTATION OF A BI-OBJECTIVE EA-BASED SCHEDULING STRATEGY: CASE STUDY IN THERMAL-PRINTED LABEL MANUFACTURING

Fabian Vargas-Nieto
Department of Industrial Engineering
Universidad del Norte
Km 5 via Puerto Colombia, Barranquilla, Colombia

Jairo R. Montoya-Torres
School of Economics and Management Sciences
Universidad de La Sabana
Km 21 autopista norte de Bogotá, Chía, Colombia
jairo.montoya@unisabana.edu.co

KEYWORDS
Scheduling, evolutionary algorithm, makespan, tardy jobs, thermal transfer, printed labels manufacturing.

ABSTRACT
This paper considers our work in progress on the problem of scheduling production activities in a real-life manufacturing plant. The problem is modelled as a two-stage flexible flow shop problem with minimization of both the makespan and the number of tardy jobs. We described the design approach of an evolutionary algorithm. Preliminary experiments are performed and results obtained motivate us to actually implement the proposed procedure.

1. INTRODUCTION

Thermal transfer labels industry has experienced an enormous growth in the past decade in Colombia. The Colombian apparel market is demanding every day labels with more variable information, such as size, style, references, descriptions, barcodes, etc. In addition, it is also claiming for a smaller minimum order quantity in production orders that allow the apparel companies request the exact label amount they need. Many times label amount that clients need is not bigger than 500 units per reference, which difficult its manufacturing through traditional label production processes: woven label and flexographic printed label processes. For the company under study in this research work, during the last two years, sales growth caused a 50% lead-time raise, in spite of the sufficient installed production capacity to attend market demand. However, during the last 6 months the lead-time has increased approximately 250%, which has begun to threaten the permanence of the most important client accounts for the company. After reviewing the causes that motivated the former lead-time increase, the company determined that as demand grew, the number of orders to program also grew, making manual scheduling harder to handle. Hence, it is necessary to implement a software application base on a multi-objective scheduling model to execute client orders under the minimization of both the makespan and the number of tardy jobs.

2. PROBLEM DESCRIPTION AND LITERATURE

- This paper presents an evolutionary algorithm-based approach for production scheduling at for thermal-printed label manufacturing. The case study is taken from a real plant located in Colombia. The manufacturing process consists of two stages: Printing and Cutting. Client order is received at printing stage and labels are printed according to the design requested by the client. Because of the nature of the process, each color in the label design needs one pass through the printing machine. This means, for example, that a label with a design of two colors printed on the face and one color printed on the back needs to pass three times through the printing machine. The plant has between 4 or 5 printing machines. At the cutting stage, label rolls transferred from printing stage are cut according to the label length established in the label design given by the client. There are between 3 and 5 cutting machines. Finally, client orders are delivered to the client in carton boxes and the order is finished. Formally, this scheduling problem can be modelled as a Flexible Flow Shop (FFS) with two stages, as shown in figure 2. It consists of a set of k=2 stages in series; at stage s, s=1,...,k, there are m_s identical machines in parallel. There is unlimited intermediate storage between two successive stages. Job j, j=1,...,n has to be processed at each stage on any one machine. The processing times of job j at the various stages are $p_{1j}, p_{2j},...,p_{kj}$. The goal is to find a schedule without preemption that minimizes the maximum completion time of all jobs (makespan).

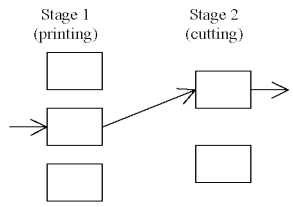


Figure 1. Job routing in a two-stage flexible flow shop

The two-stage FFS is known to be strongly NP-hard (Gupta, 1988). In the literature, works on two-stage FFS problems consist on exact methods, involving mathematical programming formulations and branch-and-bound algorithms, as well as approximation algorithms, dedicated heuristics or even lower bounds (Brah and Hunsucker, 1991; Haouari and M'Hallah, 1997, Soewandi and Elmaghraby, 2001, Allaoui and Artiba, 2006, Azizoglu et al., 2001, Jin et al., 2006). The problem of simultaneously minimizing makespan and the number of tardy jobs has not yet been studied. In this paper, we present the design and preliminary implementation of an evolutionary algorithm to solve the scheduling problem on a two-stage flexible flow shop with these two objectives. As stated previously, data will be obtained from a real-life manufacturing plant.

3. DESIGN OF THE BI-OBJECTIVE SCHEDULING EVOLUTIONARY ALGORITHM

An evolutionary algorithm (EA) is a problem solving technique that uses the concepts of evolution and hereditary to produce good solutions to complex problems that typically have enormous search spaces and are therefore difficult to solve (see Figure 3). The biggest difference with other meta-heuristics is that EA maintains a population of solutions rather than a unique current solution. Solutions are coded as finite-length strings called chromosomes and a measure of their adaptation (the fitness) is computed by an engine. Starting from an existing population, each iteration generates new chromosomes

by applying operators (like crossover and mutation) to two chosen parents. The main advantage of EA is its intrinsic parallelism, which allows the exploration of a larger solution space. Early applications of EA's are found in the literature to solve complex combinatorial optimisation problems (Back, 1995; Fogel, 1995).

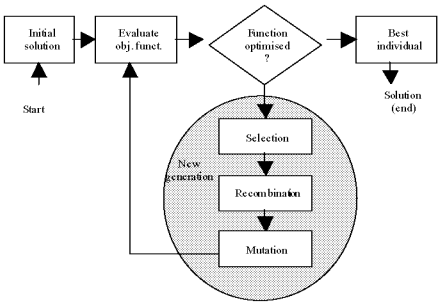


Figure 3. General description of an EA.

3.1. Solution representation

In a broad way, the evolutionary algorithm presented here is a bi-objective optimization procedure that seeks to minimize the total completion time of all jobs that has to be scheduled (or makespan) and the number of tardy jobs. The chromosome is represented as shown in figure 4. The vector contains the job numbering (1,2,…,n) as well as the numbering of printing machines (I_1,I_2, …, I_m). The example presented in figure 4 consists on a chromosome with 3 printing machines and 6 jobs. This vector defines the processing sequence of each job on each machine. The corresponding processing order at this stage will determine the processing sequence at the cutting stage according with the modified First-In-First-Out (M-FIFO) rule considering a workload balancing among the cutting machines.

I_1	1	2	I_2	5	6	4	I_3	3

Figure 4. Chromosome to represent the solution

3.2. Crossover operator

The crossover operator corresponds to an uniform crossover, as illustrated below:

Father 1:

I_1	1	2	I_2	5	6	4	I_3	3

Father 2:

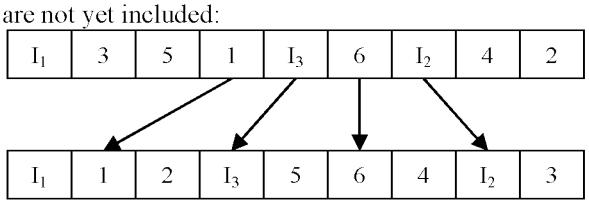
I_1	3	5	1	I_3	6	I_2	4	2

Genes are selected uniformly from Father 1 to Son 1:

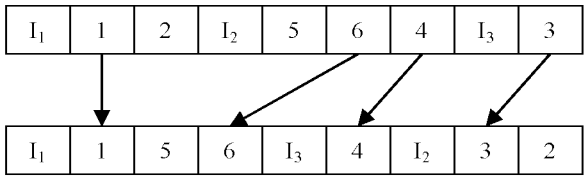
Son 1:

I_1		2		5		4		3

Afterwards, empty spaces correspond to genes from Father 2 that are not yet included:



For the second son, genes from Father 2 are selected uniformly, while genes from Father 1 are selected in the same manner as previously explained for Son 1.



3.3. Mutation

The mutation procedure is based on an interchange strategy. First, a chromosome of any father is selected, lets say:

I_1	1	2	I_2	5	6	4	I_3	3

Then, two genes of this father are randomly selected, lets say genes I_2 and 6. These genes change their positions, resulting in the following son:

I_1	1	2	6	5	I_2	4	I_3	3

3.4. Parameters

In order to implement the evolutionary algorithm, the following parameters have to be defined: crossover and mutation probabilities, initial population and no-dominated population sizes. The actual values of these parameters are defined next in our experiment.

3.5. Algorithm steps

Step 1. Generate initial population: Initial population is generated using an iterative random procedure that assigns the first gene for all individuals from the machine group. Then the rest of genes are selected randomly from both: machine group and jobs group.

Step 2. Calculate printing times: Printing times are calculated for each job in each solution as the time when each job finished printing according to the order established by the solution representation for each individual in the population.

Step 3. Assign jobs to cutting machines: Jobs are assigned according with the modified First-In-First-Out (M-FIFO) rule considering a workload balancing among the cutting machines.

Step 4. Calculate Ending times: Ending times for each job are calculated for each individual as the time when each job finished cutting according to the order established by the solution representation in each individual.

Step 5. Calculate Makespan and Number of Tardy Jobs: Makespan is calculated as the larger ending time for all the jobs in each individual. The number of tardy jobs is calculated by comparing the ending time for each jobs versus the due date for each job in each individual.

Step 6. Calculate fitness: Fitness for each individual is calculated as follows: $fitness_i = n/N$, where n is the number of solutions that individual i dominates among population, and N is the size of population.

Step 7. Select individuals: The individuals with better fitness are selected and copied to the non-dominated archive.

Step 8. Recombine non-dominated individuals: Apply crossover and mutation operators described previously to non-dominated individuals based on the crossover and mutation probabilities.

Step 9. Selection of new individuals: Copy new individuals to the population, and then go to step 6.

Repeat this procedure until the maximum number of generations is reached.

4. SIMULATIONS

4.1. Experimental design

The proposed meta-heuristic scheduling procedure was programmed using Microsoft Excel® macros based on Visual Basic. Experimental tests were performed on a PC AMD® 2400 Plus 2.2 GHz and 512 MB RAM. The partner company of this research provided preliminary test data. In order to analyse the performance of the proposed production scheduling procedure, we initially considered a set of 21 jobs (this is the average quantity of jobs to be scheduled at any plant of the company under study) with total processing times ranging from 2 minutes to 64 hours, and due dates ranging from 16 to 64 hours. The number of colours required to finish the labels was 1 or 2 depending on the type of label produced. The complete information about the processing of jobs is kept confidential. The number of machines at the printing and cutting stages was 5 and 3 machines respectively. The population size was set to 167 individuals while the file size of non-dominated solutions was set to be 20 individuals. The probability of mutation was set to be 0.3. Finally, the probability of crossover was set to be 0.8, as most authors recommend. However, for the purpose of our experiment, was also considered a value of 0.5 for the crossover probability.

4.2. Analysis of results

Figures 5 and 6 present, respectively, the solutions after one generation with crossover probability of 0.5 and 0.8. We can observe that the lower the probability, the closer the points in the figure. Besides, it is possible to observe that with 0.8 of probability of crossover, the Pareto frontier is more visible for the makespan value.

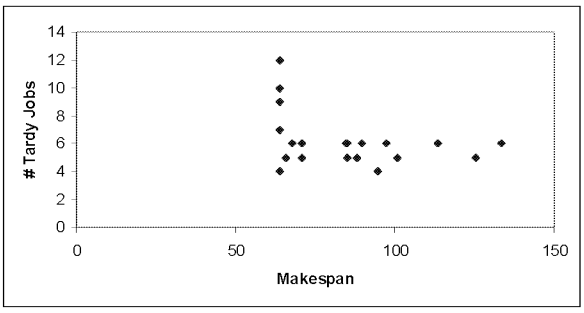


Figure 5. Initial solution with 0.5 of crossover probability

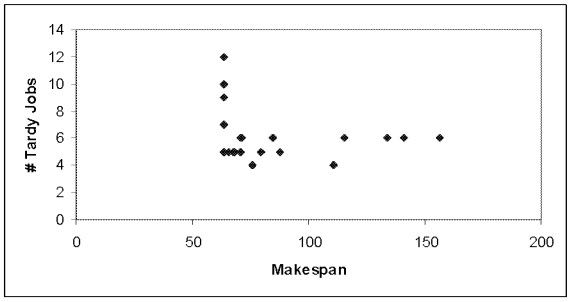


Figure 6. Initial solution with crossover probability of 0.8.

Figures 7 and 8 represent the final solution obtained after 10 iterations with crossover probabilities of 0.5 and 0.8 respectively. It is to recall that the mutation probability was set to be 0.3 and remained fixed for all the experiments. It is possible to see that when using crossover probability of 0.8, the Pareto frontier is already visible, while in the 0.5 crossover probability case, much more points are still not close to the Pareto front. Thus, when arriving to the twentieth iteration, the Pareto frontier is already practically defined, as shown in figure 9. In contrast, when using a lower value for the crossover probably, like 0.5 in this experiment, it is necessary to arrive to a higher number of iterations in order to get defined the Pareto front.

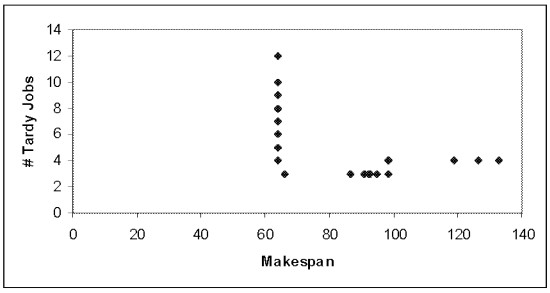


Figure 7. Solution after 10 iterations with 0.5 of crossover probability

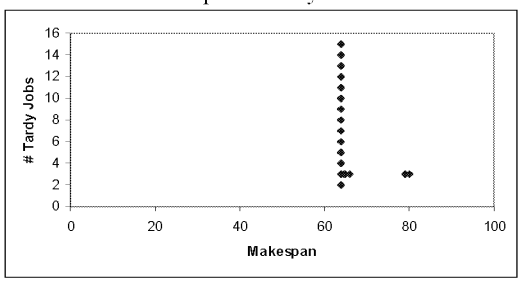


Figure 8. Solution after 10 iterations with 0.8 of crossover probability

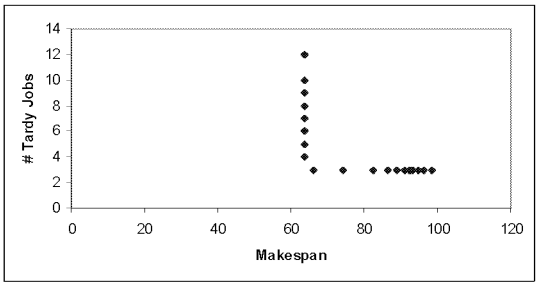


Figure 9. Solution after 20 iterations with 0.5 of crossover probability

When arriving iteration number 100, it is possible to see that when the crossover probability is 0.8 the value for the number of tardy jobs is lower than the one obtained in the other experiments (Number of tardy jobs = 2. In other experiments, 3 and 4). For this real-life case of a production scheduling plant, these results verify the proposed conjectures presented in most of the literature works concerning the parameters definition when implementing evolutionary algorithms.

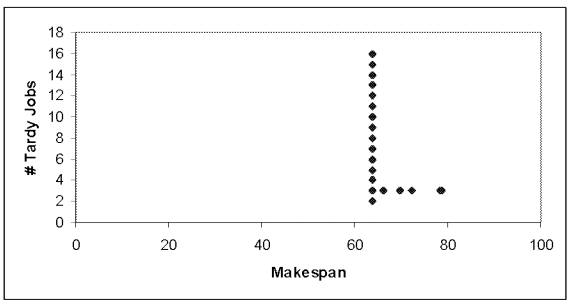


Figure 10. Solution after 20 iterations with 0.8 of crossover probability

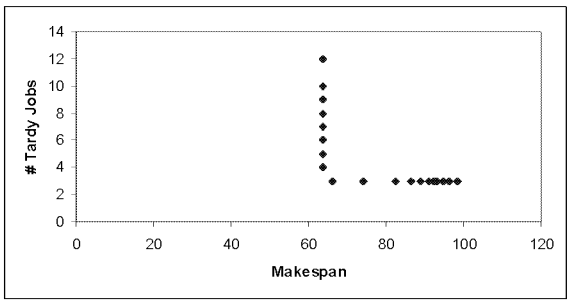


Figure 11. Solution after 100 iterations with 0.5 of crossover probability

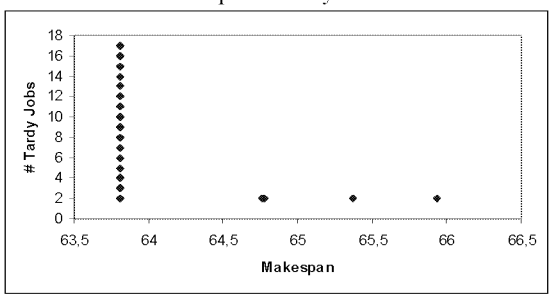


Figure 12. Solution after 100 iterations with 0.8 of crossover probability

4.3. Selection of a single solution from the Pareto front

After implementing the EA-based scheduling strategy developed, the production supervisor would still have to select one solution from the Pareto front to implement it as a real schedule. In order to help the production supervisor to select a solution to be implemented, the software provides a single solution by selecting the solution from the Pareto front with better machine workload balance in both stages: printing and cutting. To perform the former stated, the workload per machine is calculated for each individual belonging to the Pareto front (non-dominated solutions), after the last generation has been performed. The solution with less difference in workload between the most loaded machine and the least loaded machine in both printing stage and cutting stage is selected.

4.4. Comparison with real results

A comparison between the real results for the makespan and number of tardy jobs for the 21 chosen jobs, after a manual scheduling defined by the production supervisor *versus* the solution obtained by the scheduling algorithm. The gap, in percentage, was computed as $\%dev = (Z_p - Z_A)/Z_p$, where Z_p and Z_A are, respectively, the solutions obtained by the production supervisor and by the algorithm. Those results show an improvement of near to 25% on the makespan value (i.e. the time needed to finish all jobs) and a decrease of 75% on the number of tardy jobs. This strongly suggests that the proposed algorithm would probably contribute to achieve better results in real life.

5. CONCLUDING REMARKS

This paper considered the problem of scheduling production activities in a real-life manufacturing plant. The problem was modelled as a two-stage flexible flow shop problem, with makespan and number of tardy jobs minimization. The problem was solved using a bi-objective evolutionary algorithm. Experimental results motivate the extension of the set of experiments in order to actually implement the proposed scheduling procedure at the factory. Algorithm parameters have to be well defined in order to obtain accurate performance in terms of both solution values and computational time. So, as shown in this preliminary study, a well-defined idea about their actual values is obtained by testing several scenarios depending on the analysts' interests.

ACKNOWLEDGEMENTS

Second author would like to thank the financial support of the School of Economics and Management Sciences, La Sabana University, Colombia.

REFERENCES

Allaoui H. and A. Artiba (2006) Scheduling two-stage hybrid flow shop with availability constraints. *Computer & Operation Research* 33, 1399-1419.

Azizoglu M., E. Cakmak, and S. Kondakci (2001) A flexible flowshop problem with total flow time minimization. *European Journal of Operational Research*, 132, 528-538.

Back, T. (1995). *Evolutionary Algorithms in Theory and Practice*, New York, Oxford University Press.

Brah S.A. and J.L. Hunsucker (1991) Branch and Bound algorithm for the flow shop with multiple processors. *European Journal of Operation Research*, 51, 88-99.

Fogel, D.B. (1995). *Evolutionary Computation, Towards a New Philosophy of Machine Intelligence*. IEEE Press.

Gupta, J. (1988). Two stage hybrid flow shop scheduling problem. *Journal of the Operations Research Society* 38, 359-364.

Haouari, M. and R. M'Hallah (1997). Heuristic algorithms for the two-stage hybrid flowshop problem. *Operations Research Letters* 21, 43-53.

Jin, Z., Z. Yang and T. Ito (2006). Metaheuristic algorithms for the multistage hybrid flowshop scheduling problem. *International Journal of Production Economics*, 100, 322-334.

Soewandi, H. and S.E. Elmaghraby (2001). Sequencing three-stage flexible flowshops with identical machines to minimize makespan. *IIE Transactions* 33, 985-993.

BIOGRAPHIES

Fabian Vargas-Nieto holds the Industrial engineering degree from Universidad del Norte, Barranquilla, Colombia, and is currently finishing the Master of Science degree at the same university. He is working as Thermal Transfer Division Director at Finotex Colombia S.A., in Barranquilla, Colombia.

Jairo R. Montoya-Torres is Associated Professor and Director of the Marketing and International Logistics program at the School of Economics and Management Sciences, La Sabana University, Bogotá, Colombia. He holds the Ph.D. in industrial engineering from Ecole des Mines de Saint-Etienne and Université Jean Monnet, Saint-Etienne, France. His research interests include production scheduling, supply chain management, and optimisation of manufacturing and logistic systems.

SIMULATING WIRELESS MONITORING IN AUTOMATED MAINTENANCE OF BELT CONVEYOR SYSTEMS

S. Güngen, A. Lopez de la Cruz, H. P.M. Veeke, G. Lodewijks
Delft University of Technology, faculty 3mE,
e-mail: H.P.M.Veeke@tudelft.nl

KEYWORDS
Resource management, Data acquisition, Maintenance strategies

ABSTRACT

This paper discusses concepts for automated maintenance of belt conveyors by means of wireless temperature monitoring. A powered maintenance trolley that can travel autonomously over the structure of a belt conveyer system is adapted as a platform for the maintenance system. A robot on the trolley performs replacements of the bearings. The wireless concepts are simulated and compared to an earlier maintenance concept, which is based on vibration analyses.
Wireless Sensor Mesh Networks and RFID Systems are two wireless technologies which can potentially be used for condition monitoring of belt conveyors. With a Wireless Sensor Network, condition monitoring can be performed at any time, independent of trolley passage times. In case an RFID system is used, the data reader has to be located on the trolley, which is why condition monitoring is restricted to trolley passage times.
Compared to the Vibration and RFID Strategies, the maintenance strategies, which are based on a Wireless Sensor Mesh Network, have much lower fail rates and higher flexibility.

INTRODUCTION

Belt conveyor systems are generally used for the transportation of commodities over long distances. Some conveyors can reach lengths of more than 10 km. The belt is supported by rollers spread over equal distances. Belt conveyors for the transportation of iron ore contain up to 400 roll bearings per 100 meters. The operational conditions for these bearings are harsh, due to high operational loads and contaminating dust from the transported commodity. Bearing failure can occur frequently and potentially cause damage to the expensive rubber belt. To avoid this, the bearings' condition should be monitored frequently, so that the rolls can be replaced before actual failure occurs. In the current belt conveyor systems, condition monitoring is a manual task, which is generally outsourced to specialists who visually inspect the bearings one-by-one. With 40000 bearings spread over a distance of 10km, this is a time-consuming and expensive task.

Automated maintenance of belt conveyor systems is a promising alternative to outsourcing maintenance, in particular if looking at efficiency, accuracy and costs. A powered maintenance trolley that can travel autonomously over the structure of a belt conveyer system was adapted as a platform for the maintenance system [Lodewijks, 2004] and [Lodewijks, Ottjes,2005]. Data acquisition equipment for vibration analyses was installed on this trolley along with a robot that can replace bearings. The optimum maintenance strategy was determined by means of a logistic simulation model. In this concept, data on a specific bearing can only be collected when the maintenance trolley is present at the location. An alternative to this concept would be the use of a Wireless Monitoring System, where temperature sensors on each bearing actively signal the maintenance system when attention is needed.

This paper discusses concepts for automated maintenance of belt conveyor systems by means of wireless temperature monitoring. Existing Wireless Monitoring Systems that can be used for this will be introduced. Then a concept for the logistic control of an automated maintenance system will be explained and simulated and finally conclusions and future research will be presented.

MAINTENANCE STRATEGIES

Real-world conditions such as improper lubrication, impact loading, vibration, excess

temperature, contamination, excessive loading, and misalignment, will decrease the life expectancy of bearings. If these conditions are severe, they may lead to premature failure of bearings. For belt conveyors, contamination of the lubricant (dust from the commodity transported) and impact loading (too large transported volumes) are the most important reasons for bearing failure. In the final stage of any bearing failure, increased friction between the bearing components causes a steep rise in temperature.

Condition monitoring of bearings, by means of temperature monitoring, is not new to the industry. "Axlebox" bearing units incorporating speed and temperature sensors are already used in high-speed trains, passenger coaches, and locomotives for preventive maintenance [De Man, 2006]. Alternatively, condition monitoring of bearings can be done by vibration monitoring or lubricant analysis, but for this research we restrict ourselves to temperature monitoring.

WIRELESS MONITORING

A typical Wireless Sensor Network is composed of SmartTags (data acquisition points), µNodes (communication waypoints), and Gateways (data collection points). The units are truly "wireless" in the sense that they require neither power nor data wiring. The only wires necessary are between the sensors and the data collecting SmartTags. These autonomous devices self-organize at power-up and quickly re-configure as devices join, leave or move around the network. The system is capable of executing user defined rules, thereby triggering user-defined events such as an increase in the temperature.

The most common layout is the "meshed" network, in which the µNodes form a meshed topology (see figure 1). The SmartTags are the end-points of the network. They can communicate with either one µNode (star topology) or with multiple µNodes (mesh topology). The SmartTags forward their data to the Gateways using redundant paths in the network.

The maintenance trolley would no longer have to be used for condition monitoring of the bearings, as data readings would be sent frequently via the network to the fixed Gateway at either end of the belt conveyor. This would make it possible to create a list of bearings to be replaced (updated frequently), before departure of the trolley. The trolley could be programmed so as to only start a maintenance run, if the repair time of a bearing on the list is due. Suppliers claim that battery life would be 4-8 years at a condition update frequency of twice per day. Although data would have to be collected from a massive 40000 sensors, air bandwidth would not to be a limiting factor.

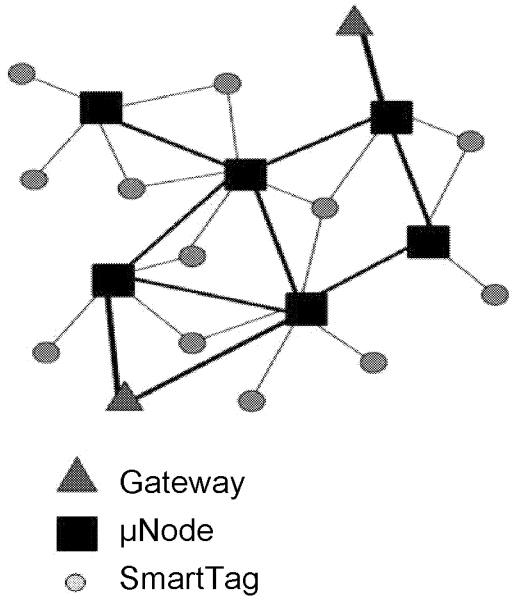


Fig. 1. A Wireless Sensor Mesh Network

A monitoring system for belt conveyors, based on RFID (Radio Frequency Identification), would consist of a Reader/Writer (Data collection Point), which is placed on the maintenance trolley, and of RFID tags (with integrated temperature sensors) placed on each individual bearing. Compared to passive tags, active RFID tags have a longer communication range in harsh environments and are ideal for collecting data from sensors. Therefore the most realistic choice for monitoring belt conveyors would be active RFID tags.

The amount of system components required would be comparable to the Wireless Sensor Mesh Network introduced earlier, except that the µNodes could be left away. Sensor data would be stored in the active RFID-tags and transmitted to the reader on the maintenance trolley when it passes by. The trolley would then have to be detected in time by the Tags, requiring much more listening time (very frequent listening), whether the trolley is coming into the line of sight and therefore requiring much more battery power. The trolley would not have to stop for monitoring of each individual bearing and therefore no time would be lost. However, monitoring would be restricted to trolley passage

times just like in earlier maintenance concepts which work on vibration analyses.

SIMULATION MODEL

A simulation model has been developed in TOMAS [Veeke, Ottjes, 2000], with the intention to compare 3 methods of condition monitoring of the rolls:

1. Vibration analysis as described by Lodewijks [2004].
2. Wireless Sensor Mesh Network (WSMN, Fig. 1)
3. active RFID system.

The model contains the (classes of) elements:

- the belt conveyor
- the rolls
- the automated maintenance robot
- the condition monitoring system
- the estimation of the remaining roll lifetime

The belt conveyor is specified in terms of its length and the number of rolls. Each roll is supported by two bearings. The life length of a specific bearing in a roll is allocated via a tabularized distribution. Under and upper limits can be specified assuming a uniform distribution (minimum and maximum life length as specified by roll and bearing manufacturer). The chance of failure before reaching the minimum life length can be specified, again according to a uniform distribution. All distributions can be changed for the middle and the side rolls of the carrying as well as the return idler sets.

Condition monitoring can be performed either by an active RFID system or by a Wireless Sensor Mesh Network (WSMN). For simplicity, it is assumed in this study that all RFID system- and Wireless Network- components have endless battery life.

The automated maintenance robot consists of a trolley and a replacement robot. It travels back and forth over the structure of the belt conveyor at a constant speed. Replacements are only performed in the forward direction of the robot. It is assumed that the robot is available 24 hours per day. The total replacement time consists of a fixed set-up time (seconds/per frame) for the replacement robot and of a replacement time (seconds/roll), in which the old roll is replaced with a new one.

The total lifetime estimation of an individual roll is based on temperature measurements taken from the sensors on each bearing of the roll.
The lifetime of the roll is defined as the minimum of the lifetimes of its two bearings. Each bearing is equipped with a temperature sensor that communicates with the central data collection point, either through an active RFID system or through the WSMN. For the simulation, the temperature-vs.-time curve of a roll is modeled as shown in figure 2 below. The normal operation temperature of a bearing is assumed constant (T_{Normal}) for the major part of its life. A short period before failure of the roll, the temperature will start to rise abruptly. Depending on the failure mode, failure can be due to damage in the bearing structure or due to contamination of the lubricant. In the final stage of either failure mode, increased friction between the bearing components causes this fast rise in temperature.

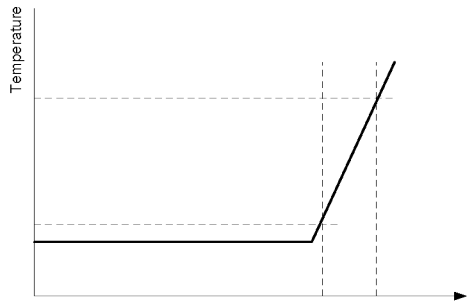


Figure 2. Temperature of the roll

The discrepancy between the roll's actual temperature at a given point in its Signal Interval and the measured temperature is simulated with 2 parameters. The measured temperature is a sample from a normal distribution with as mean the current "actual temperature" of the roll (see Figure 2). The deviation of this distribution determines the accuracy of the measured temperature and is controlled by the first parameter (d). With the second parameter (f), a bias is introduced. The estimator of the measured temperature becomes conservative, biased towards overestimating the temperature.
The measured temperature M is defined by:

$M = A + d*(F–A)*X + f (F– A)$	
F	Fail Temperature of the Roll
A	Current Actual Temperature of the Roll
X	Random variable, sampled from N(0,1)
d	Deviation, as fraction of the residual temperature
f	Safety factor, as fraction of the residual temperature

SIMULATION SETTINGS

Two options are presented for condition monitoring of the rolls: a Wireless Sensor Mesh Network (WSMN) or an active RFID system.

WSMN: Condition monitoring of all the rolls is performed independent of the robot movement, at fixed *Long* Monitoring Intervals, which must be specified. Once a roll's temperature has reached T_{Signal}, the sensor autonomously sends a warning signal to the network. From that point onwards, the WSMN will start monitoring the temperature of that roll at fixed *Short* Monitoring Intervals.

Active RFID system: Condition monitoring is performed via the central RFID reader located on the robot. This means, monitoring can only be performed during the robot cycles. All temperature measurements are used to estimate the residual lifetime of the roll.

By combining the different options presented above, several maintenance strategies are possible. Four strategies are considered in this paper.

S1	Condition monitoring is performed by a WSMN at fixed Monitoring Intervals (long, then short). The robot performs maintenance cycles at fixed Cycle Intervals.
S2	Condition monitoring is performed by a WSMN at fixed Monitoring Intervals (long, then short). The robot performs maintenance cycles at flexible intervals.
S3	Condition monitoring is performed by a WSMN at fixed Monitoring Intervals (long, then short). The robot performs Ad-Hoc Maintenance Runs.
S4	Condition monitoring is performed by an active RFID system. The robot performs maintenance cycles at fixed intervals.

The Safety Temperature (ST) defines a temperature above which a roll is replaced by the robot. The safety temperatures are explained below for each strategy.

Strategies 1, 4:
ST must be chosen so that the remaining Lifetime of the roll is at least larger than the Robot Cycle Interval.

Strategy 2:
- ST must be chosen so that the remaining Lifetime of the Roll is larger than the short monitoring interval + average Robot travel time per cycle.
- A second Safety Temperature (ST2) is needed which defines a temperature at which replacement may take place. ST2 must be lower than ST, otherwise the robot would replace too few rolls per cycle, dedicating a whole cycle to only one or two roll replacements.

Strategy 3:
ST must be chosen so that the remaining Lifetime of the Roll is larger than the short monitoring interval + average Robot travel time from one roll to another (Ad-Hoc).

Experiments were performed with different conveyor lengths, ranging from 100 to 10000 meters, with a frame distance of 2 meters.
The lifetime of a bearing is distributed uniformly between LMin and LMax. A certain rate of the bearings fails between 0 and LMin. This failure is uniformly distributed between 0 and LMin. From these specifications the average lifetime of each type of bearing can be calculated. The lifetime of a roll is the minimum of the lifetime of two bearings.

Bearing Lifetimes (days)	Lmin	Lmax	Average
Upper side roll	1750	2083	1812.4
Upper middle roll	1667	2000	1733.5
Lower side roll	1875	2208	1931.1

The Signal Interval of a bearing is distributed uniformly between 7 and 10 days.
The bearings' Operation Temperature is taken 30 °C and Fail Temperature 80°C.
The figures for the maintenance trolley are given below.

Speed	0.5 m/sec
Monitoring time	0 sec
Robot setup time	30 sec/frame
Replacement time	60 sec/roll
Cycle interval	6 days

The performance of a maintenance strategy is mainly recorded of:

- the number of rolls that has been replaced (too) late (Failure)
- the average time between replacement and end of lifetime of a roll (Waste)

SIMULATION RESULTS

Because there is no actual system to compare with, it is not possible to validate the simulation model in this stage. Therefore the model has only been verified

There are two different versions of the WSMN Flex strategy:

- WSMN Flex1: ST and ST2 are very close in range (± 2 °C)
- WSMN Flex2: ST and ST2 are far in range with ST2 equal to ST of the WSMN Fixed Strategy (± 35 °C).

The five strategies (including Flex2) are simulated several times to find the optimum ST for each individual strategy. This is repeated for 5 different conveyor lengths, ranging from 100 to 10000 m.
WSMN Fixed and RFID Fixed: These two strategies show almost equal results. The travelled distances for these strategies remain within proportion and relatively low for all conveyor lengths. The Average early replace times are the longest at all conveyor lengths, compared to other strategies.

WSMN Flex1: The strategy is outperformed by other strategies for all conveyor lengths. It is only feasible up to a conveyor length of ±150m. For higher conveyor lengths, the travelled distance increases significantly.

WSMN Flex2: This strategy is ideal for conveyor lengths ≥750m. The strategy outperforms all the other strategies in its low travelled distance. However the "price" paid is a higher average early replace-time.

WSMN AD-Hoc: This Strategy proves to have the shortest travelled distance and a low average early replace time until conveyor lengths of ≤ 750m. For short conveyors, this strategy is the ideal solution. Up to ±150m the distance travelled is at least 1:3 compared to other strategies. With higher conveyor lengths, the travelled distance increases significantly.

For the sensitivity of the strategies for the discrepancy between real and measured temperature each strategy is simulated twice: once with a discrepancy of 0% and once 10%.
An analysis is done by comparing the difference in the percentage of late replaced rolls between the two simulations. This experiment is repeated for 3 different conveyor lengths. The results are summarized in the table below.

	100m	1000m	10000m
WSMN Fixed	0.8%	0.5%	0.2%
WSMN Flex1	1.1%	0.0%	0.0%
WSMN Flex2	0.0%	0.0%	0.0%
WSMN AdHoc	1.3%	0.1%	0.0%
RFID Fixed	0.8%	0.6%	0.5%

WSMN Fixed: For all conveyor lengths, this strategy is relatively sensitive to changes in the bearings' temperature. The sensitivity decreases with increasing conveyor length.

WSMN Flex 1: For short conveyor lengths, this strategy is sensitive to changes in the bearings' temperature. For long conveyor lengths the sensitivity decreases to 0%.

WSMN Flex 2: For all conveyor lengths, this strategy has 0% sensitivity to changes in the bearings' temperature. It thereby outperforms all other strategies.

WSMN AD-Hoc: This strategy is most sensitive for simulations with short conveyor lengths. However, the sensitivity decreases to 0% with increasing conveyor length.

RFID Fixed: For long conveyor lengths, this strategy has the highest sensitivity to changes in the bearings' temperature. The sensitivity decreases with increasing conveyor length.

Vibration Monitoring vs. Wireless Temperature Monitoring

Earlier, vibration analyses for condition monitoring of bearings was theoretically compared to temperature monitoring of bearings. In theory, damage or wear to a bearing can be

detected earlier by listening to the vibrations than by measuring the temperature of the bearing.

The simulation parameters of the vibration analyses strategy are different from the parameters for the wireless strategies. In the simulation program for Vibration analyses, the time to replace a roll is indicated as a certain time interval before the lifetime of the bearing. The user can assign a deviation to this lifetime estimation. In the simulation program for the wireless strategies, the time to replace a roll is when the roll has reached a certain temperature. The user can assign a deviation to the measured temperature. Despite these differences (amongst others) in the two programs, the strategies can be compared to each other:

The RFID fixed strategy is the wireless strategy, which is most similar in working principle to the vibration analyses strategy. In both strategies, condition monitoring is done on the trolley and is therefore dependent on the trolley cycle intervals. These two strategies are simulated several times, to find the respective deviation settings, at which both strategies render the same % of late replaced rolls. For best comparison, the remaining wireless strategies are then simulated using the deviation in measured temperature, which was found for the RFID Strategy.

The next experiments are performed with a conveyor length of 10000m. Wireless strategies are restricted to small monitoring- and cycle intervals, because of the short Signal Interval of 7 days, during which a bearing's temperature rises. Therefore, all strategies (including vibration analyses) are simulated with a Cycle Interval of 6 days. For the vibration Analyses, the Safety Time is set equal to the Cycle Interval, and the Inspection Time was set to twice the Cycle Interval, as was recommended in [Lodewijks, 2004].

Calibration
For this experiment, the deviation in the estimated lifetime of the vibration strategy is set to 1%. A simulation with these settings renders 3.3% of late replaced rolls. By means of repetitive simulations with the RFID strategy, it is found that, when the deviation in measured temperature is set to 29%, this simulation also renders 3.3% of late replaced rolls. The remaining wireless strategies can now be compared to the vibration strategy.

Strategy	Cycles	Travel time	Early (%)	Early/Late Time (days)
WSMN fixed	608	0.5443	99.2	4.59/0.44
WSMN Flex1	3646	0.4766	99.8	1.67//0.07
WSMN Flex2	986	0.5132	100	5.50//0.03
WSMN AdHoc	47435	0.0315	99.0	1.52/0.11
RFID Fixed	608	0.5443	96.7	4.32/0.597
Vibr. Analysis	608	0.87	96.7	3.00/3.71

The results for the different strategies are listed in the table above.

The RFID strategy renders similar early replace-times to the Vibration analyses, which justifies the choice of using the RFID strategy for calibration of the wireless strategies. The only advantage of using the RFID strategy above Vibration analyses is the low average late replace-time (0.59 d). Despite the extreme deviation in measured temperature (29%), the WSMN strategies render very low percentages of late replaced rolls (0-1%), compared to 3.3% for the RFID and Vibration strategies. WSMN Flex2 is the only strategy that renders 0% late replacements. However, the price paid for this good performance is a highest average early replace-time of 5.5 days and a relatively higher travelled distance.

CONCLUSIONS AND FUTURE RESEARCH

Several Wireless Temperature Monitoring strategies for belt conveyors have been introduced and simulated in this paper. The performances of these strategies have been compared to each other and to the Vibration analyses concept, which was introduced by [Lodewijks, 2004].
If a wireless sensor network were to be applied to a 10km belt conveyor, the network would need at least 40000 sensors. This represents a massive investment. However, the current growth of this technology's market promises a significant reduction in costs in the nearby future. Amongst the features of Wireless Temperature Monitoring are:

- Possibility to monitor from a distance, away from location (outsourcing easier)

- Provides an overview of the conditions of all the rolls in the conveyor, at user specified intervals.
- Extremely low fail rate, due to the Signal feature
- Low early replacement times
- Is a basis for automation of the maintenance/replace robot

The features listed above promise low operational costs and easy roll stock management for belt conveyors on which this technology will be applied in future.
Compared to the Vibration Analyses and RFID Strategies, the maintenance strategies, which are based on a Wireless Sensor Mesh Network, have much lower fail rates and higher flexibility. For Vibration Analyses and RFID, the inspection intervals depend fully on the robot's fixed cycle intervals. Therefore, information on the rolls' condition is not available before the robot starts a cycle. This is why these two strategies render the highest fail rate at an elevated deviation.
The Wireless Sensor Mesh Network, in combination with a trolley robot that does maintenance cycles at Flexible Intervals (WMSN Flex 2), has proven to be the most feasible wireless maintenance strategy for long belt conveyors. This strategy has on average, long robot cycle intervals and a low sensitivity for changes in roll temperatures. The result is a very flexible maintenance system with the lowest fail rate.
The Wireless Sensor Mesh Network, in combination with a trolley robot that does Ad-Hoc maintenance runs (WSMN Ad-Hoc), has proven to be the most feasible technology for conveyor lengths ≤ 750m. Below this length, the total distance travelled by the robot is relatively very low and the average early replacement time for the rolls is negligible.
An advantage for Vibration Monitoring is that vibrations, due to damage or wear of a roll, can be detected earlier than an increase in temperature. Because of the relatively short time interval, during which a failing bearing's temperature rises, the wireless strategies, based on temperature monitoring, are restricted to small monitoring- and robot cycle intervals. As a result, the Vibration Monitoring trolley can be set to do maintenance at larger cycle intervals than the Wireless Temperature Monitoring strategies, whilst still maintaining relatively low percentages of late replacements.

Research should be done into the possibilities of combining Wireless Sensor Mesh Networks with Vibration Monitoring. Compact sensors capable of detecting vibrations would be the ideal solution for belt conveyors.
A small scale model of the maintenance system should be built for validation of the simulation model. Tests can then be performed to find the exact values for the temperature characteristics (operation temperature, fail temperature, signal interval, deviation) of conveyor bearings and to further validate the results of the simulation.

REFERENCES

Albers, A. (2006), University of Karlsruhe, Institute of Product Development, Germany, "Acoustic Emission Analysis", Practicing oil Analysis Magazine. May 2006

De Man, H. (2006), LogicaCMG, "On Train Axle Bearing Temperature Measurement", www.logicacmg.com/file/3942

Lodewijks, G. (2004), "Strategies for Automated Maintenance of Belt Conveyor Systems", Bulk Solids Handling Vol. 24 (1), pp. 16-22

Lodewijks, G., Ottjes, J.A., (2005), "Belt Conveyor Inspection Tools base don Fuzzy Logic", Bulk Solids Handling 25 (5), pp. 284-289

Veeke, H. P.M., Ottjes,J.A., 2000. "Tomas: Tool for Object-oriented Modelling And Simulation". *Proceedings of Advanced Simulation Technology Conference (ASTC2000), April 16-20, 2000*, Washington, D.C. pp. 76-81, The Society for Computer Simulation International (SCS)

PERFORMANCE EVALUATION OF LOT DISPATCHING AND AUTOMATED MATERIAL HANDLING APPROACHES BY DISCRETE-EVENT SIMULATION

René Drießel, Lars Mönch
Chair of Enterprise-wide Software Systems
Department of Mathematics and Computer Science
University of Hagen
Universitätsstraße 1, D-58097 Hagen, Germany
E-mail: {rene.driessel|lars.moench}@fernuni-hagen.de

KEYWORDS
Semiconductor Manufacturing, Automated Material Handling Systems, Shop-Floor Emulation via Simulation, Performance Evaluation

ABSTRACT

In this paper, we present design principles for an infrastructure that allows simulation-based performance assessment of production control approaches for a complex manufacturing system with Automated Material Handling System (AMHS). We describe the suggested coupling architecture. The systems includes the simulation engine AutoSched™ AP for dispatch decisions, AutoMod™ for simulating the AMHS, and a blackboard-type data layer that is between the production control application and the two simulation engines. We present computational results for simple dispatching rules applied to lot and vehicle dispatching respectively.

INTRODUCTION

In semiconductor manufacturing, integrated circuits are produced on silicon wafers. This type of manufacturing is very capital intensive. Lots are the moving entities in a semiconductor wafer fabrication facility (waferfab). Each lot contains a fixed number of wafers. The process conditions are very complex (Uzsoy et al. 1994, Atherton and Atherton 1995, Schömig and Fowler 2000). We have to deal with parallel machines, different types of processes like batch processes and single wafer processes, sequence dependent setup times, prescribed customer due dates for the lots, and reentrant process flows. Batch machines can process several lots at the same time on the same machine. Very often, we also have to face with an over time changing product mix including a large number of different products.
Especially, in modern 300-mm wafer fabs AMHS are very important because of the increased factory automation within these manufacturing facilities. We refer to Agrawal and Heragu (2006) and Montoya-Torres (2006) for recent survey papers on AMHS in semiconductor manufacturing.
So far, very often scheduling and AMHS decisions are made independently. In (Qu et al. 2004) the authors suggest to incorporate AMHS decisions within a shifting bottleneck type scheduling approach. However, only a few static test instances are considered. In this paper, we extend the simulation framework of Mönch et al. (2002) to the case where additionally the simulation of an AMHS is necessary. In this particular case, we have to add an additional simulator that is responsible for the emulation of the AMHS.
The paper is organized as follows. We describe the overall coupling architecture and its design criteria in the next section. Then we continue with describing the suggested blackboard-type data layer. In the third part of the paper, we present results of a simulation study that uses the suggested coupling architecture.

DESIGN CRITERIA FOR THE COUPLING ARCHITECTURE

The suggested coupling architecture has to support the following two tasks:

- mimic the behavior of the shop-floor that communicates with a shop-floor control system over a message bus, the shop-floor control system has to take AMHS decisions into account,
- provide a generic interface to plug in an arbitrary shop-floor control system.

Note that the support of the AMHS appears as a new requirement compared to the architecture suggested by Mönch et al. (2002).

OVERALL ARCHITECTURE

Involved Subsystems

The architecture consists of four parts that are tightly connected together:

- a shop-floor control system that makes scheduling decisions for the lots of the wafer fab and vehicles of the AMHS,
- a simulation model for emulating the shop-floor (called process model),
- a simulation model for emulating the AMHS (called material handling model),

- a blackboard that is between the two simulation models and the shop-floor control system.

The process model is modeled with AutoSched™ whereas the material handling model is developed using AutoMod™. We consider in this paper an AMHS that consists of an Overhead Hoist Transport (OHT) system that connects several bays. Each bay has an OHT mono rail. One vehicle is used within each bay.
The developed blackboard is an extended version of the blackboard by Mönch et al. (2002) that is refactored and enriched by new classes for the material handling domain. The blackboard is an object model that holds objects for all relevant business objects of the shop-floor. The objects of the blackboard are updated in an event-driven manner.
The shop-floor control system consists for the sake of simplicity of a set of dispatching rules for the lots and for the vehicles. However, because of the design of the architecture an arbitrary shop-floor control approach seems to be possible. We are especially interested in scheduling approaches that make lot scheduling and AMHS scheduling decisions at the same time as suggested for example by Qu et al. (2004). The overall architecture is depicted in Figure 1.

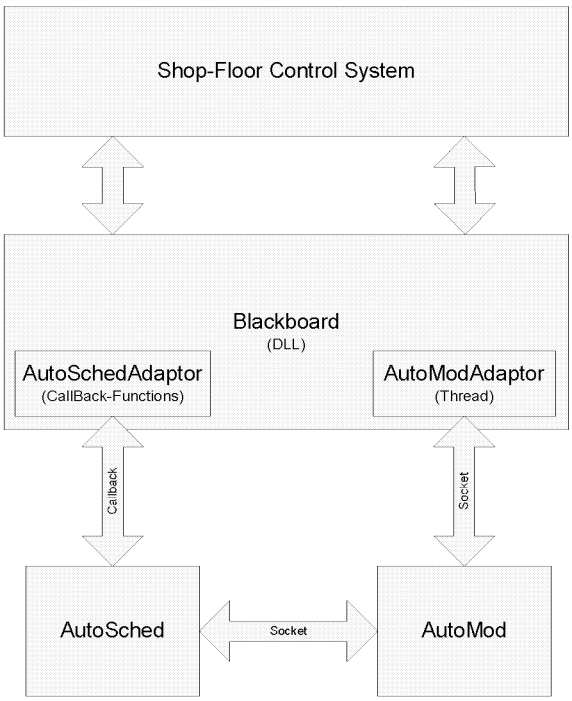


Figure 1: Overall Architecture

Coupling of AutoSched AP and AutoMod

The communication between the two simulation models is ensured through the Model Communication Module (MCM) of Brooks Automation. A move request for a vehicle is sent from AutoSched™ to AutoMod™ for each move from one storage to another storage. The transport is simulated by the used material handling system. When a move is completed a move complete message is sent to AutoSched™. The simulation inside AutoSched™ is stopped until AutoMod™ has sent a corresponding complete message.

Coupling with the Blackboard

The blackboard is implemented as a dynamic link library (DLL) that is loaded by the AutoSched™ AP simulation engine. This DLL is written in the C++ programming language.
All necessary data structures for representing the state of the entire wafer fab are implemented within the library. The coupling with the process model is realized through callback (notification) functions that are called for each subscribed event within the simulation.
The coupling with the material handling model is implemented by a thread (AutoModAdaptor) that implements a small complexity socket server. The AutoMod™ simulation engine sends a socket message to the AutoModAdapator when one of the relevant business objects changes its state.

BLACKBOARD-TYPE DATA MODEL

The data model is divided into three parts. Classes for the process model and the material handling model are the most import content of the blackboard. Furthermore, the blackboard contains classes that are used to represent scheduling entities. The scheduling model consists of the following classes:

- **Resource:** A resource represents a physical resource that is needed to process lots. Resources are given by tools or vehicles.
- **Route:** A route describes the steps that are needed to produce a lot.
- **RouteStep:** This class represents a single route step.
- **Part:** A part represents the produced lot type (product).
- **Lot:** The lot class represents a lot to be processed.
- **Job:** A job is an instance of a route for a specific lot. The job has a release date, a due date, and a list of steps that should be assigned to a corresponding resources.
- **JobStep:** This class describes a single step of a job. Each job step is assigned to a route step in the route of the job.
- **Activity:** The activity class represents a single activity that can be processed by a resource (i.e. the processing of a batch or a movement of an vehicle, a batch is defined as a collection of lots that are processed at the same time on the same machine).
- **Schedule:** A schedule is represented by an ordered list of activities that are assigned to a resource.

The scheduling model is shown in Figure 2 as UML class diagram. The entities of the scheduling model are used to represent the results of the shop-floor control system. These results will be implemented within the two simulation models.
The process model is given by the following classes:

- **Tool:** A tool (machine) is the central resource in the process model. This class is derived from the Resource class. A tool group is defined as a collection of tools with the same functionality.
- **Storage:** The storage class represents a place where lots can wait for their processing.
- **ProductionStep:** A production step is inherited from the generic route step.
- **Batch:** A batch class represents a formed batch that is processed on a certain tool. The batch class is inherited from the activity class.

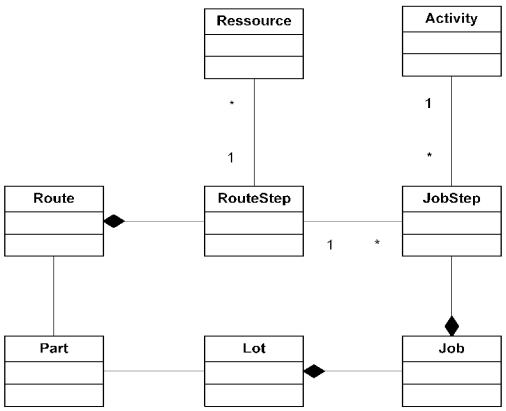


Figure 2: UML Class Diagram for Scheduling Entities

The process model is depicted in Figure 3 as UML class diagram. The process model is used to represent the business objects from the manufacturing process (emulated by the AutoSched™ simulation model).

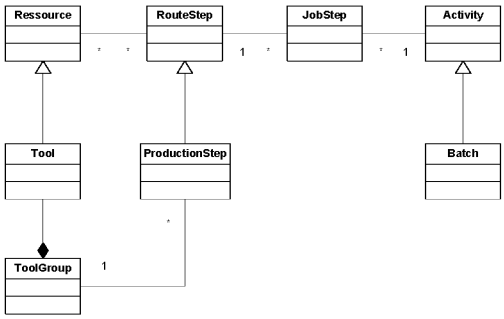


Figure 3: UML Class Diagram for the Process Model

Finally, the material handling model contains the following entities:

- **Vehicle:** A vehicle is used to transport a lot from one location to another.
- **Location:** This class represents a single location in the wafer fab. Typically locations are the position of resources and storages in the wafer fab.
- **Path:** A path links two locations together. A vehicle can only travel on a path.
- **TransportStep:** The transport step represents the travel between two process steps through the factory. When a lot is transported more than one vehicle is involved.
- **Pickup:** This activity represents the pickup time of a vehicle for a specific lot.
- **Movement:** The movement activity represents a vehicle movement.
- **SetDown:** This activity represents the set down time of a vehicle for a specific lot.

The classes of the material handling model are shown as UML class diagram in Figure 4.

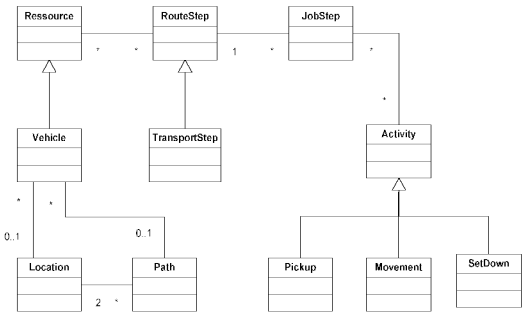


Figure 4: UML Class Diagram for Material Handling

The communication between the two simulation engines and the blackboard is bidirectional. The blackboard acts as a mirror of the manufacturing process. The corresponding business objects in the blackboard are updated whenever the corresponding business objects in the manufacturing system change their state.
Each resource (i.e., tools or vehicles) in the simulation is controlled by the blackboard. A function in the blackboard is called to support the decision-making process when a resource has to make a dispatching decision
Different scheduling and dispatching decisions can be based on the data within the blackboard. This offers benefits for assessing different scheduling and dispatching rules within one simulation testbed.

LOT AND VEHICLE DISPATCHING

The blackboard can be used to implement and evaluate several dispatching strategies for the process and material handling model because all the required data is available in the blackboard. It is hard to implement these rules within AutoMod™ or AutoSched™,
A second motivation for the suggested coupling architecture comes from more sophisticated look-ahead dis-

patching rules that take future lot and vehicle arrivals into account.
Lot dispatching that take positions of future lots and vehicles into account requires information that is only available inside the AutoMod™ simulator. Therefore, it is necessary to represent this information within a blackboard. Scheduling algorithms that consider the AMHS have to be based on integrated data structures too.
We implement a First-In-First-Out (FIFO) rule for lot dispatching. The lot that arrives first in the stocker of a tool group is selected first for processing.
Vehicles are used to transport the nearest lot first. The corresponding rule is called nearest lot first (NLF). The distance between the current vehicle position and the lot position is determined using Dijkstra's shortest path algorithm.
Furthermore, we also implement an earliest due date (EDD) type rule for the vehicles. In this particular case, the vehicles pick first the lot with the smallest due date among the lots that are available for a transport.

SIMULATION EXPERIMENTS

In this section, we describe some simulation experiments to assess the suggested architecture.

Used Simulation Model

We use a simulation model that consists of three bays for our simulation experiments. The machinery is given by the tools of the MiniFab model (El Adl et al. 1996). We add the AMHS described in (Shikalgar, Fronckowiak and MacNair 2002) to our simulation model. There is a bay for each tool group in the fab. The bays are connected by an Overhead Hoist Transport (OHT) system. Each bay has also a OHT mono rail. Only one vehicle is used within a bay. The vehicles are not able to travel between the different bays. The layout of the transportation system is depicted in Figure 5.
The circles represent stockers that are modeled as storages in the blackboard and in the AutoSched™ simulator. All storages are represented as connection points within AutoMod™. The dark squares are the positions of the tools within the simulation model.

Results of Computational Experiments

Two load situations (average load and high load) of the wafer fab are considered. Runs with and without the AMHS are performed for each load situation. Lots of three different product types are considered. The due dates are set using a flow factor *FF* to calculate (expected) waiting times by simply multiplying the pure processing time with the flow factor. The used flow factors and start rates of the products are shown in Table 1.
The tools are dispatched with the FIFO rule. Experiments with NLF and EDD are conducted for vehicle dispatching. We simulate 20 days.
Four different performance measures are considered. Our primary performance measure is total tardiness (TT). TT is given by the sum of the tardiness values of the different lots. The tardiness of lot *j* is defined as

$$T_j := max(0, C_j - d_j). \quad (1)$$

Here, we denote by C_j the completion time of lot *j*. The notation d_j is used for the due date.

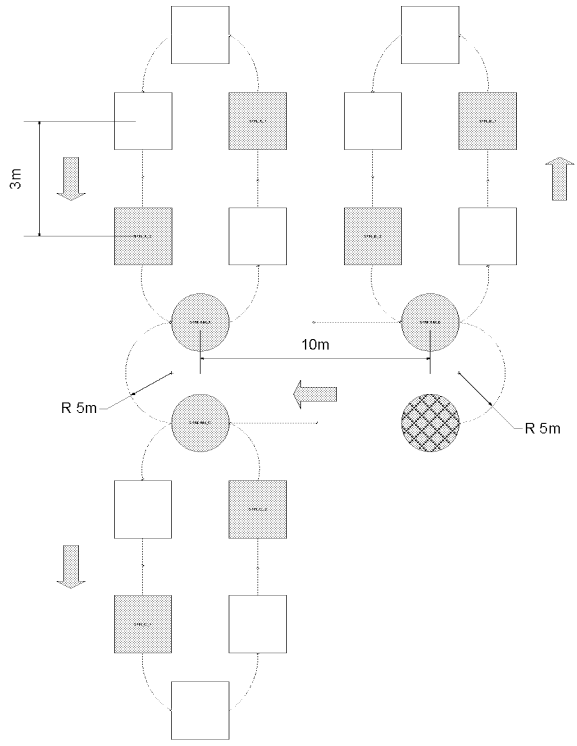


Figure 5: Used Factory Layout for the Experiments

Table 1: Parameters of the Simulation Experiments

Product	Order repeat time for average load (min)	Order repeat time for high load (min)	*FF*
A	160	141	1.54
B	300	285	1.49
C	3000	2946	1.44

The second performance measure of interest is the average cycle time (ACT). ACT is given by

$$ACT = \frac{1}{n}\sum_{j=1}^{n}(C_j - r_j). \quad (2)$$

Here r_j represents the release date for every lot *j* and *n* denotes the number of all completed lots. The number of completed lots (throughput TP) and the number of late lots is also reported in Table 2. Table 2 shows the results of six simulation experiments.

We conclude from the experiments that a consideration of the AMHS has an impact on all performance measures. The tardiness and cycle time is increased in both load situations when the transport system is taken into account.

Table 2: Computational Results

Experiment	TT (h)	ACT (h)	TP	#Late Lots
Average Load				
No transport	79:33	15:12	279	95
Transport, NLF	123:03	15:32	277	119
Transport, EDD	118:32	15:30	277	116
High Load				
No transport	162:53	15:46	305	152
Transport, NLF	186:02	15:56	306	164
Transport, EDD	176:42	15:53	306	162

The results for TT and ACT are slightly improved using the due date oriented EDD dispatching rule for the vehicles. Therefore, it makes senses to look for more advanced strategies for vehicle dispatching (or even scheduling).
The time to conduct a simulation experiment without AMHS is approx. 20 seconds. It increases up to 40 minutes when the AMHS is added. This behavior is mainly caused by the communication between the AutoMod™ simulator and the blackboard.

SUMMARY AND OUTLOOK TO FUTURE RESEARCH

In this paper, we described a coupling architecture that allows for simulation-based performance assessment of simultaneous lot dispatching/scheduling and AMHS decisions. The architecture consists of a simulation engine for emulating a wafer fab and a second simulation engine for emulating the material handling system. The shop-floor control system is coupled with two simulation engines via a blackboard-type data layer. The data layer contains all the necessary data for making scheduling and material handling decisions. The suggested architecture allows for a performance assessment of dispatching rules for lots and vehicles.
There are several directions for future research. First of all it seems to be possible to develop and test more sophisticated look-ahead dispatching rules. A next logical step consists in the development of more global scheduling approaches that take AMHS decisions into account.

REFERENCES

Agrawal, G. K. and S. S. Heragu. 2006. "A Survey of Automated Material Handling Systems in 300-mm Semiconductor Fabs." IEEE Transactions on Semiconductor Manufacturing, 19(1), 112-120.

Atherton, L. F. and R. W. Atherton. 1995. *Wafer Fabrication: Factory Performance and Analysis*. Kluwer Academic Publishers, Boston, Dordrecht, London.

El Adl, M. K., A.A. Rodriguez, and K.S. Tsakalis. 1996. "Hierarchical Modeling and Control of Re-entrant Semiconductor Manufacturing Facilities." In *Proceedings of the 35th Conference on Decision and Control*, Kobe, Japan.

Fowler, J. W. and J. Robinson. 1995. *Measurment and Improvement of Manufacturing Capacities (MIMAC): Final Report*. Technical Report 95062861A-TR, SEMATECH, Austin, TX.

Mönch, L., O. Rose, and R. Sturm. 2002. "Framework for Performance Assessment of Shop-Floor Control Systems." In *Proceedings of the 2002 Modeling and Analysis of Semiconductor Manufacturing Conference (MASM 2002)*, ed. J. W. Fowler, J. K. Cochran, 95-100.

Montoya-Torres, J. R. 2006. "A Literature Survey on the Design Approaches and Operational Issues of Automated Wafer-transport Systems for Wafer Fabs." *Production Plannng & Control*, 17(6), 648-663.

Qu, P., B. Steinmiller, and S. J. Mason. 2004. "Incorporating Automated Material Handling Systems Into a Disjunctive Graph." In *Proceedings of the 2004 Industrial Engineering Research Conference*, Houston.

Shikalgar, S. T., D. Fronckowiak, and E. A. MacNair. 2002. 300mm "Wafer Fabrication Line Simulation Model." In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C.-H. Chen, J.L. Charnes, 1365-1368.

Schömig, A. and J. W. Fowler. 2000. "Modelling Semiconductor Manufacturing Operations." In *Proceedings of the 9th ASIM Dedicated Conference Simulation in Production and Logistics*, ed. K. Mertins and M. Rabe, 55-64.

Uzsoy, R., C.-Y. Lee, and L. A. Martin-Vega. 1992. "A Review of Production Planning and Scheduling Models in the Semiconductor Industry, Part I: System Characteristics, Performance Evaluation and Production Planning." *IIE Transactions on Scheduling and Logistics*, 24: 47-61.

MASM Test Data Sets. ftp://ftp.eas.asu.edu/pub/-centers/masmlab/factory_datasets/.

AUTHOR BIOGRAPHIES

RENE DRIESSEL is a Ph.D. student in the Department of Mathematics and Computer Science at Fern-Universität in Hagen. He received a master's degree information systems from the Technical University of Ilmenau, Germany. After receiving his master's degree he worked for two years as an IT-consultant in several

SAP projects in the logistics domain. His research interests are in production control of semiconductor wafer fabrication facilities, automated material handling systems, multi-agent-systems, simulation, and ERP software.

LARS MÖNCH is a Professor for Enterprise-wide Software Systems in the Department of Mathematics and Computer Science at FernUniversität in Hagen. He received a master's degree in applied mathematics, a Ph.D. in the same subject from the University of Göttingen, Germany, and a Habilitation degree in Information Systems from the Technical University of Ilmenau. After receiving his Ph.D. he worked for two years for Softlab GmbH in Munich in the area of software development. His current research interests are in information systems for manufacturing, simulation-based production control of semiconductor wafer fabrication facilities, applied optimization, and artificial intelligence applications in manufacturing. He is a member of GI (German Chapter of the ACM), GOR (German Operations Research Society), SCS and INFORMS.

PRODUCTION EFFICIENCY IMPROVEMENT

MODEL AND SIMULATION OF A GAS COOKERS ASSEMBLY LINE: PERFORMANCE ANALYSIS AND SUGGESTIONS TO INCREASE EFFICIENCY

Giovanni Andrioli and Giorgio Romanin-Jacur
Department of Management and Engineering
University of Padova,
Stradella San Nicola 3
36100 Vicenza
Italy
E-mail: romjac@dei.unipd.it

KEYWORDS
Manufacturing, discrete simulation, decision support system.

ABSTRACT

In a factory producing high class cookers and ovens, one of the three production departments includes an assembly shop. There an assembly line with 23 on-line and 9 off-line manual and semi-automatic stations is devoted to assemble final products. A study of its behaviour, aimed at increasing its efficiency without imposing dramatic structural and functional changes, was performed. After a qualitative and quantitative analysis on the field, a discrete stochastic simulation model was built up and validated. Based on the current model outputs, two different modification of work organization were suggested and simulated, with positive results on the assembly system performance.

ENVIRONMENT

The environment is a factory with 110 employees which produces high class cookers and ovens, destined to both Italian and foreign markets. In such a factory raw materials and components are bought from external suppliers, works-in-process are partly manufactured inside and partly outsourced, final products are assembled inside; the factory includes three production departments, where all final products are assembled and some works-in-process are manufactured, and a logistic department, where final products are temporary stored and then dispatched to customers.
We consider one of the production departments, which includes: an office; two warehouses where final products, works-in-process and components are stored; three workrooms dedicated to work-in-progress manufacturing; an assembly shop, containing a serial assembly line with 23 on-line working stations and 9 off-line working stations, which perform either test operations or pre-assembly operations on components and works-in-process, and feed the assembly line.

STUDY OBJECT AND DEVELOPMENT

The assembly shop in its whole is the object of the present paper; at the beginning of the study, the line had been repeatedly adjusted, adapted and modified, in correspondence with new products design and new technological solutions; unfortunately, such changes concerning processed parts, performed operations and employed machines had not been supported by the necessary actions on line management, with special attention to operating speed and distribution of single operations among workers. Therefore the study had the scope of analysing performed operations and related interactions, of modelling and simulating the whole shop and finally of looking for possible improvements; two peremptory condition had been imposed: improvements had to be obtained with zero investment and minimal structural revolution. The study was limited to one of the products, by considering that the developed model may be easily adapted to other products by suitably changing operating parameters; moreover we think the adopted method may be extended to different serial assembly lines.
The work performed included in the order: analysis of all information about assembly cycles at disposition; check on its correspondence with reality; after a negative result of the check, rebuilding a new valid information, by taking all assembling times and generally all other operating times; building up and implementing a simulation model on the computer; simulation experiment based on current times; model validation based on current behaviour; critical analysis of current performance parameters; new experiments based on different cycle times and related analysis of resulting performance parameters; new experiments based on both different cycle times and different operations scheduling, and related analysis of performance parameters.

ASSEMBLY SHOP

As shown in Figure 1, the serial assembly line we studied has U shape and initially it was able to process 24 items per hour (2.5 minutes per item); the throughput time was 145 minutes. The line has 23 working station, each one served by one operator, dedicated to on-line assembling operations; 9 more stations, always served by one operator, are dedicated to off-line operations, which consist of the assembling of subsets and to some test of components and subsets: they are adjacent and feed the assembling line. At three on-line working stations (numbers 12, 18 and 20) a test on partially assembled items is performed: items which do not comply with specifications are eliminated; at two off-line stations (10D and 10E) a test on some components (burners) is effected, before assembling them.

The line may be divided into two parts: the first one, including the first seven stations, works without imposed rate, and items, i.e., partially assembled final products, are manually shifted on free rollers; between station 7 and station 8 items enter the synchronized section of the line, ruled by a device which imposes a fixed rate, and they further move at constant speed on drawing rollers as far as the end of the line, after station 23; at the beginning of the study, the interval between two consecutive item entrances was 2.5 minutes, and the distance between two consecutive items was 1.5 metres; the line advanced at the speed of 0.01 metres per second, then the item movement was so slow that it created no problems to operators; the space around items was sufficient to reach the item sides. In general operators perform a manual assembly work on items moving on the line; some of them are equipped with automatic machines devoted either to item positioning before an assembling operation, or to item (or part) testing.

SYSTEM ANALYSIS AND MODELING

The first part of the study was dedicated to line qualitative and quantitative description, as, due to the reasons discussed above, no exact information was available. All operations were accurately described, for what concerns both interested mechanical parts and manual movements on the moving items. Operation sequence for every worker was pointed out and all operation durations estimated; in order to get a good estimation, each operation duration was repeatedly taken for all operators, at different times of the day; eventually a rectangular time distribution was adopted as it resulted the best fitting one.

Once the single operations had been described, the total line flow diagram, considering also employed resources, was outlined.

Now a simulation model, based on the flow diagram, was built up by means of package MicroSaint in its version 3.2. The Task Network, reported in Figure 2, included 100 tasks, plus 22 MOVE tasks used only to describe a movement time between successive stations, 111 integer real variables, 74 real variables, 3 integer arrays and 4 real arrays of dimension 1500.

8 tasks are independent arrival generators, one to start simulation and 7 to simulate arrivals of components to off-line stations. 27 tasks describe off-line stations with related entering queues; each station is described by tasks connected in a specific way, according to performed operations. 14 tasks describe stations which are not obliged to follow a fixed rate; every station is described by two tasks, one to represent the operations performed and the other to represent the manually effected movement towards either the next off-line station or the assembly line, as soon as the related buffer is idle. 1 task describes the rating device. 64 tasks describe stations working at fixed rate; every station is described by four tasks, the first one representing the entry of the current item into the operating area, the second one the operations performed, the third one the movement on rollers inside the area, the fourth one the movement towards the area of the following station. The remaining tasks represent finite products and refused ones which go out.

MODEL EXPERIMENTS AND APPLICATIONS

The first simulation experiment, based on current position at the beginning of the study, was first used to validate the model; the differences in time between real and simulated times, for what concerns both single stations and the whole line, were absolutely small, which means that the model was correct. An analysis of performance parameters gave a low saturation of resources, i.e., of operators (< 80% for all, with lower minima), and a low line efficiency (59.90%).

Then a second experiment, based only on a decrease of cycle time from 150 seconds (2.5 minutes) per item to 130 seconds per item could obtain a lot of favourable results: more precisely higher operators' saturation, higher production (+ 15.31%) and higher line efficiency (+ 9.46%). The decrease of cycle time and consequently the increase of the item speed and the reduction of the space around the moving items did not cause any difficulty.

A third experiment was effected, based on a further decrease of cycle time to 125 seconds per item and on light modifications on 4 operations sequence and allocation to stations, together with some modification on the related components' feeding. The results included maximum operators' saturation at about 90%, a production increase of 20% and an increase of line efficiency of 12.05%, if compared with initial conditions.

Note that all improvements were suggested by the simulation model, and could be applied at zero investment with very small changes in the operations planning.

CONCLUSION

The paper shows an interesting application of simulation on an actual industrial department, with good results obtained at low cost and without dramatic changes.

BIBLIOGRAPHY

Various Authors 1998. "MICRO SAINT 3.0: User's Manual." Micro Analysis and Design Simulation Inc., Boulder, Colorado, U.S.A.

BIOGRAPHY

GIOVANNI ANDREOLI went to the School of Engineering of the University of Padova in Vicenza, where he obtained the Laurea Magistrale degree (equivalent to Master of Science) in Management Engineering in 2006.

GIORGIO ROMANIN-JACUR went to the School of Engineering of the University of Padova, where he obtained the Laurea degree (equivalent to Master of Science) in Electrical Engineering in 1970. He was Fellow and then Research Associate of the Italian Council of Researches until 1980; he was Lecturer, then Associate Professor and eventually Full Professor of Operations Research at the School of Engineering of the University of Padova, where he currently teaches.

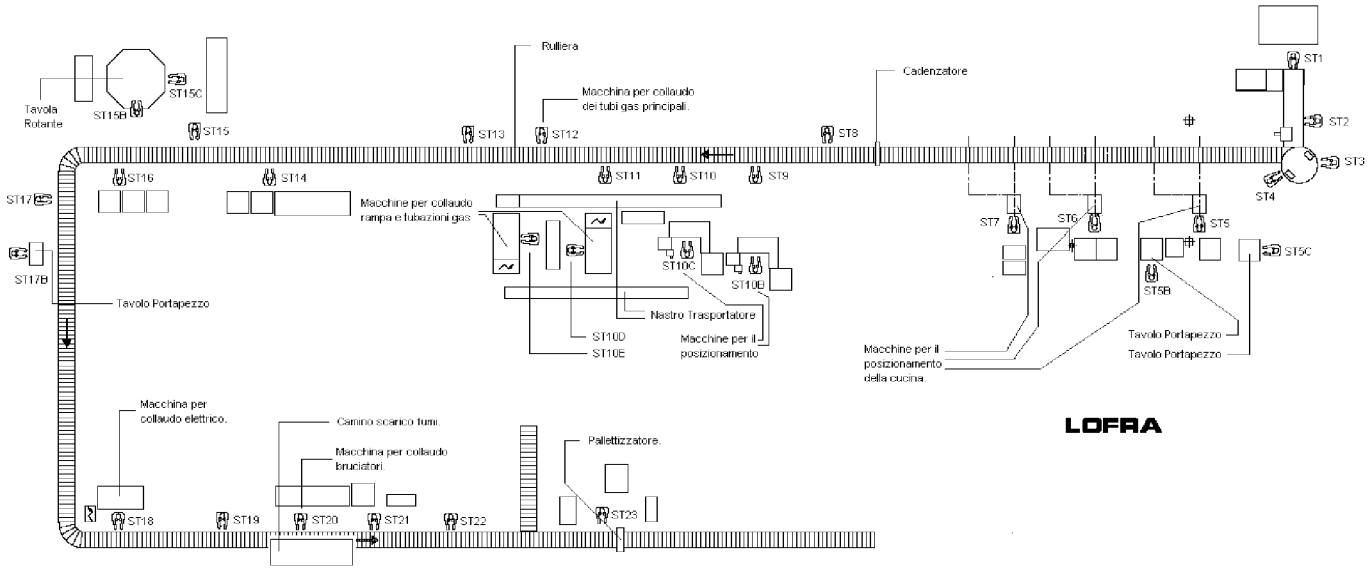


Figure 1: Assembly shop layout

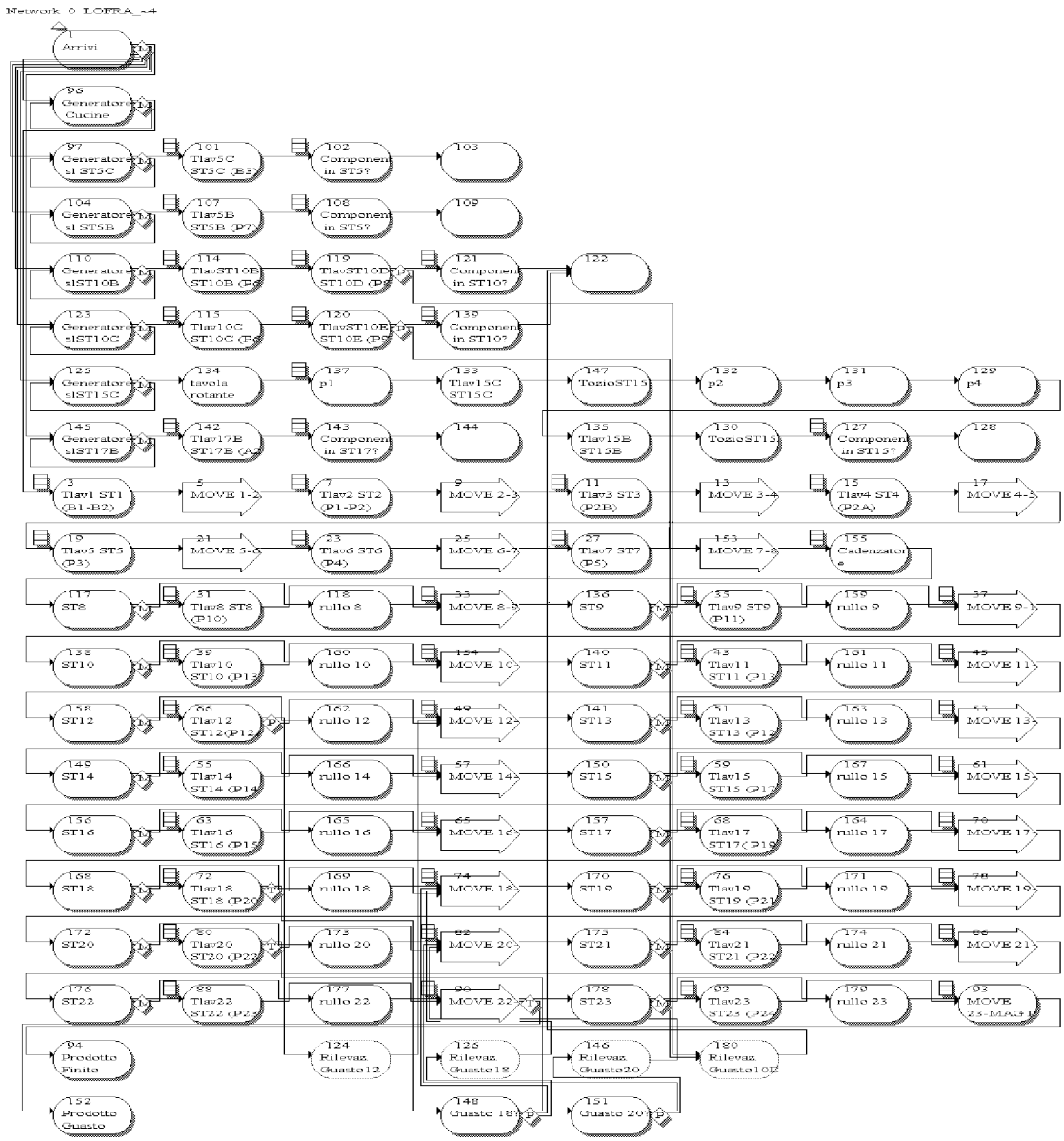


Figure 2: Micro Saint task network model

SIMULATION BASED PLANNING OF CAPACITY INCREASE FOR THE PRODUCTION OF LITHIUM POLYMER BATTERIES

Claus-Burkard Boehnlein [1]
Christian Sahr [2]

[1] Department of Business Information Integration, University of Wuerzburg
Neubaustrasse 66, 97070 Wuerzburg, Germany
boehnlein@wiinf.uni-wuerzburg.de

[2] Bullith Batteries AG
Fraunhoferstrasse 1b, 25524 Itzehoe, Germany
cs@sollith.com

KEYWORDS

Industrial Process, Manufacturing, Decision Support, Discrete Simulation, Petri Nets

ABSTRACT

Since 2004 the Bullith Batteries AG (later on BBA), a Spin Off of the Fraunhofer Institute for Silicon Technology (FhG ISiT) in Itzehoe/Germany, has been producing high performance Lithium-Polymer batteries at its site in Itzehoe. In the spring of 2006 an existing site in Munich was closed and the machines were transferred to Itzehoe to concentrate research and production in one place.

In this paper a concept and measures for a gradual increase in capacity up to the medium-term target value of 7,000 batteries per week are developed based on a material flow simulation. The goal is, to give the company a tool that supports resource protecting growth and gives the management the opportunity to flexibly react to future market trends.

PRODUCT AND PROZESS STRUCTURE

Unique characteristics of BBAs products are a high energy density, grade of safety and life time as well as thermal and mechanical stress resistance. Another characteristic is the very high shape flexibility as BBAs batteries abstain from rigid metal cases. Therefore the management expects promising markets in health care, identification technology and in the production of functional clothing.

By using two different material systems for the anode foil of the batteries, two product families with different electrical characteristics can be produced on the same machines and with the same process structure.

Under the initial conditions the existing production capacity could only be assessed, because the transfer of machines from Munich to Itzehoe had not been finished yet. Based on internal forecasts the capacity was calculated with 1,200 batteries per week and one shift each day.

The production system is separated in ten production plants with different functionalities and 23 machines and manual work centers in total. The equipment is arranged by the principle of a line production, which means all products have to pass all machines in the same order (cp. Figure 1).

The output rate of the machines and manual work centers differ from each other. This is characteristic for the organization type of the existing production system. Therefore buffers are necessary in the production process (Fries 1995; Bloech 2004).

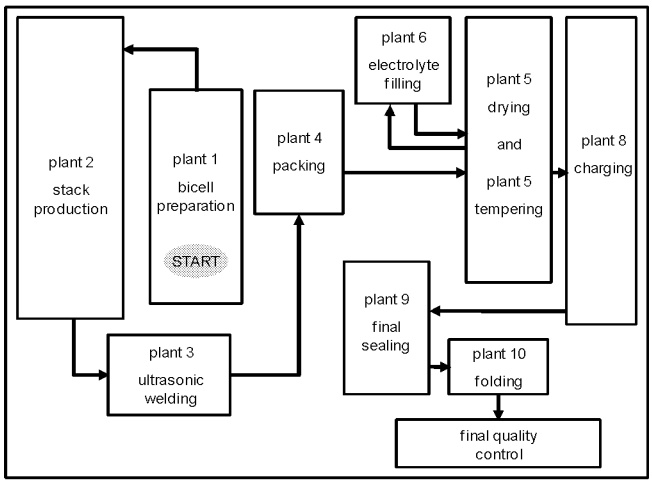


Figure 1: Production process at the site in Itzehoe

A battery consists of a stack of bicells, with the same geometry and electrical voltage. In plant 1 the bicells are made of two cathodes covering two separators and one anode in between. After that in several steps a varying number of bicells are stacked in plant 2. The products of a family only

differ in the number of stacked bicells and thus in their electrical characteristics. In plant 3 the battery bodies are tested for short circuits and then they are contacted by means of ultrasonic welding. In plant 4 the battery bodies get boxed and after that they get dried in an oven with a protective gas atmosphere in plant 5. The drying prevents undesirable chemical reactions in plant 6 when the boxed battery bodies are filled with liquid electrolyte. Back again in plant 5 the battery bodies are tempered to achieve a homogeneous dispersion of the liquid electrolyte. In plant 8 the battery bodies get charged for the first time and after that they get sealed in a vacuum in plant 9. The finished batteries are being folded to achieve the defined dimensions in plant 10 and in the end there is a final quality control.

In the phase of pre-production and sample delivery small lots, up to 500 units, were produced for customers. This phase is finished by now. For the medium-term it is planned to produce batteries in high volume at the site in Itzehoe.

This paper focuses on the analysis of the initial production system and on the development of measures for an efficient and gradual increase up to the target value of 7,000 batteries per week.

The operation time at the different machines and workplaces depends on the number of bicells needed for a battery type. The following analysis is based on the assumption of three different battery types in order to find an appropriate configuration of the production system that will help achieve the output target for each of the products (cp. Table 1).

Table 1: Characteristics of selected products

product	number of bicells	material system
A	4	1
B	8	1
C	7	2

It is planned to produce batteries of material system 1 from 4 up to 8 bicells in the medium-term. The products A and B represent the boundaries of this range. Product C with 7 bicells was selected, because of the expected high sales volume for material system 2.

METHOD AND APPROACH

With the prevailing conditions the total output of the production system depends on the output of the bottleneck in the production process, which moves with the produced product type. GOLDRATT – with his Optimized Production Technology (OPT) – provides an iterative approach to improve the total output in production systems with bottlenecks. The main steps of this approach are (Goldratt 1988):

- Identify the bottleneck of the system,
- maximize the use of the bottleneck,
- increase the capacity of the bottleneck,
- evaluation of the target achievement and
- start the next iteration, if the bottleneck has moved and the target is not yet achieved.

In this paper a material flow simulation is executed to find the total output of the initial production system and to determine the impact of individual measures.

So the following steps need to be taken: At first the structure of the existing production system has to be analyzed and relevant data have to be gathered. Based on this information a simulation model of the existing production system can be elaborated. The gathered data are used for the validation of the simulation model. The model is valid, if it creates the same output compared to the existing production system regarding the relevant system aspects. Once this is done successfully, reliable conclusions based on analyses of different scenarios and simulation runs can be drawn.

Further on the production system under the initial conditions has to be determined. The following aspects are of particular interest: The overall output per week of the production system, stocks within the production process and the rate of capacity utilization of each plant. Based on these results bottlenecks in the production process can be identified and avoided regarding to the iterative approach mentioned above.

Next different measures to increase the output are systematically tested and analyzed. The performance of combinations of those measures will be especially investigated. Therefore several scenarios, based on the model of the initial production system, are necessary to be extended in stages by additional measures. Each scenarios stage of development has to be validated in terms of efficiency and effectiveness. The result is a catalog of measures, which provides BBA with efficient strategies for a stepwise output increase regarding to the target value.

DATA COLLECTION AND MODELLING

Most of the necessary simulation data were gathered in the production process. The processing time at any manual work center was gathered and averaged over a period of three month. Some data had to be guessed by the machine operators and validated by the production manager because of low plant loading rates within the survey period.

Some processing periods, like the drying period of packed battery bodies, the tempering period or the cycle time of charging and discharging, are specified process parameters. Improvements of the production process and the used material systems in the past led to a stable failure rate of the semi finished items.

The plants are operated in shifts of eight hours with the exception of unsupervised automated process steps. Some plants have to be operated by different workers at the same time. Maintenance work is done at the end of a shift.

The simulation runs are executed without initial inventory in the model. Each simulation run simulates ten production weeks. The lead time of a lithium polymer battery in the production system takes from eight up to ten days, depending on the number of stacked bicells and the material system. Therefore only the simulation data of the last eight of ten production weeks are analyzed.

The organization and structure of the production system allows different shift models in each plant. To enable this option the parameter "plant available" is used. Plants or manual work centers, that are only temporarily available, are modeled with this additional resource. Thus each single plant or manual work center can be activated or deactivated depending on different shift plans.

Because of low plant loading rates in the past there is a lack of realistic operating data. In this situation the model validation reduces to a plausibility check.

The simulation model was build with PACE 5.0, a tool that allows hierarchical modeling and event driven simulation of higher Petri Nets (Eichenauer 2002). This type of Petri Nets is appropriate to model and analyze resource conflicts and parallel processes (Oberweis 1996). PACE uses Smalltalk instructions to model time and additional condition and action code. Figure 2 shows the top level of the production model. The different plants are modeled in detail in subnets.

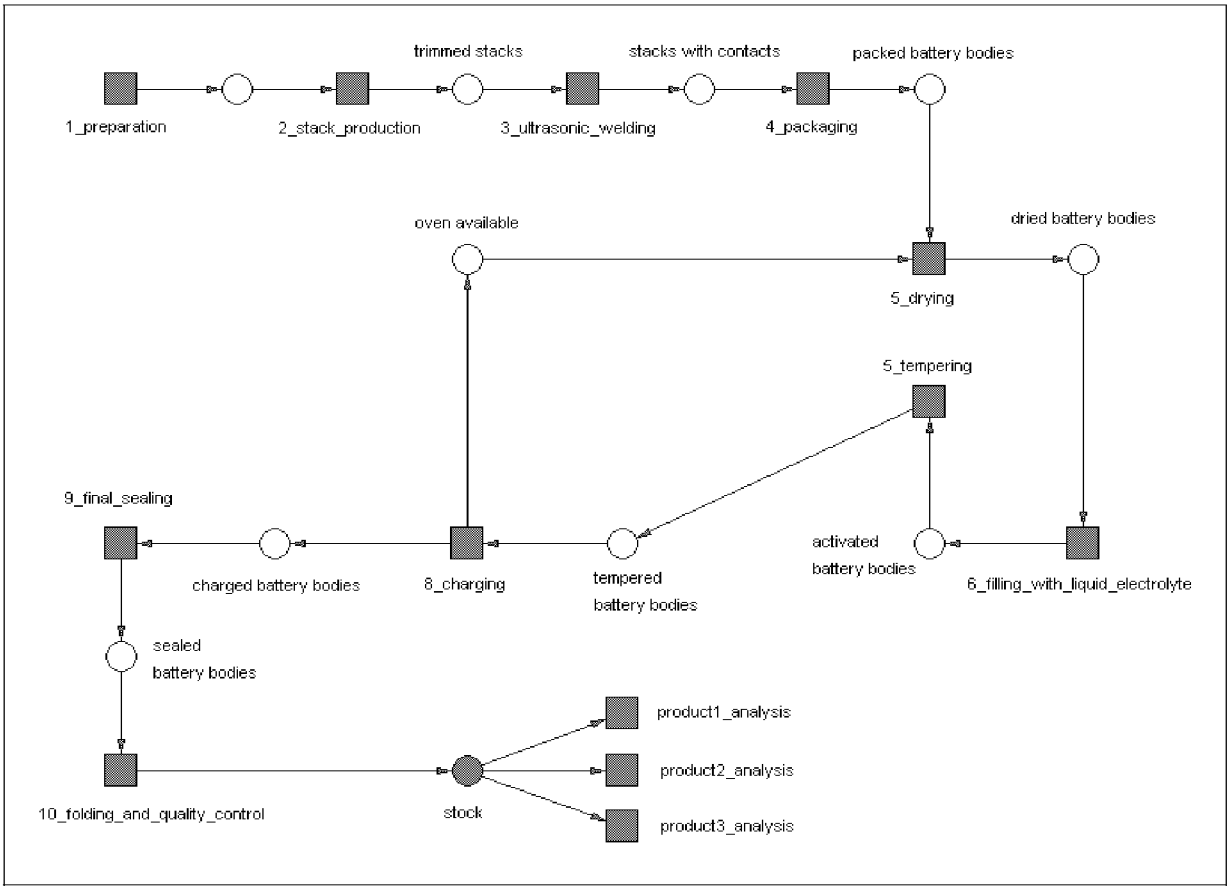


Figure 2: Petri Net model of the production system

ANALYSIS OF THE PRODUCTION SYSTEM

The relevant target value in terms of the given scope is the feasible output depending on the type of product. Table 2 gives an overview of the determined quantities. Under the initial conditions the required output of 7,000 batteries per week is clearly missed by all of the three selected products.

The fluctuation margin of the weekly output is striking. The analysis of the present production resulted, that the fluctuation margin can be traced back to the batching of 10 magazines apiece 72 battery bodies for the drying in plant 5.

Table 2: Output under the initial conditions

product	number of bicells	average output per week	fluctuation margin of the weekly output
A	4	2,336	2,157 - 2,741
B	8	1,240	1,155 - 1,693
C	7	1,459	1,160 - 1,761

The determined output data for the three selected products are applied as reference values for the evaluation of the model scenarios analyzed later on and the improvement measures derived from them.

In the simulation model the capacity of buffers is modeled as unlimited. Through this decoupling of the single plants in the production process, a backlog is prevented. Thus a simple bottleneck-analysis is provided. In the case that the output of two sequential plants temporarily differs, but equals in the long run, cyclical fluctuation at the intermediate buffer is the consequence. The dimensioning of the real buffer derives then from the maximum buffer level in the model.

If, however, the buffer between two plants or manual work centers increases over the time, then this indicates a bottleneck at the following plant or manual work center.

Under the initial conditions, all of the plants in the considered production are run in a single shift operation. The simulation of ten production weeks for product B shows that the stock of punched anodes, prior to the stack production (plant 2), increases by 4,500 units every production week. Only slight stock variations can be detected in the other buffer locations. Thus, plant 2 is the production's bottleneck and needs to be analyzed more closely for an expansion of product B's total production performance.

Plant 2 is also the bottleneck for the production of product A and C. And yet another bottleneck becomes evident for product A during the ultrasonic welding as the buffer prior to plant 3 continually increases over time.

The analysis' results of the present production give important indications for further analyses and the development of model scenarios.

SCENARIOS AND SIMULATION RESULTS

A bottleneck's utilization should be improved in order to increase its performance. This can be achieved by reducing the operation time at the bottleneck and by preventing wastefulness.

Qualified personnel are deployed in the production. Refittings are rarely needed in the given production system. That is why a further decrease of the operation time at manual work centers due to learning curve effects is not expected in this survey. The operation time at the fully automated plant components are preset by the machine's cycle times.

Table 3: Measures for a stepwise output increase for product A

	plant 2 stack production	plant 3 ultrasonic welding	plant 5 drying	plant 6 liquid electrolyte filling	plant 5 tempering	plant 8 charging	plant 9 final sealing	units per week
initial situation	one shift	one maschine, one shift	one shift	one shift	one shift	one shift	one shift	2336
enhancement 1	one shift	two maschines, one shift						2400
enhancement 2	two shifts	two maschines, one shift						4592
enhancement 3	two shifts	two maschines, two shifts						4753
enhancement 4	three shifts	two maschines, two shifts						5371
enhancement 5	three shifts	two maschines, two shifts	two shifts	two shifts	two shifts			5371
enhancement 6	three shifts	two maschines, two shifts	two shifts	two shifts	two shifts	three shifts		6869
enhancement 7	three shifts	two maschines, two shifts	two shifts	two shifts	two shifts	three shifts	< 3 extra hours per week	7110

Likewise, some of the processes such as the drying of the battery bodies cannot be changed due to technological reasons. Hence, a further reduction of the operating times is not seen as a starting point for performance enhancements. The level of utilization cannot be increased by relieving the bottleneck as all of the products have to be handled at the bottleneck plant.

Thus the bottleneck has to be eliminated by an addition to capacity. This can be done by

1. increasing the availability at the bottleneck due to extra hours or additional shifts,
2. outsourcing production orders,
3. automating the manual work steps,
4. introducing quality enhancing measures at the bottleneck and at the following plants and
5. by expanding the number of machines or manual work centers at the bottleneck.

In this case it is not possible to outsource production orders as important know-how would be lost and as third-party suppliers aren't able to offer the specific production process needed to produce this product.

The automation of existing manual work steps is very expensive and is therefore not considered to be a reasonable measure. Quality enhancing measures intended to reduce wastage at the bottleneck and at the following work stations is very time-consuming and considered as an ongoing measure that will not be pursued any further in this analysis.

Hence, extra hours or additional shifts and extra capacity via more machines or manual work centers at the bottleneck are the only reasonable measures left.

The simulation model of the production of the three products A, B and C that had been developed in the beginning of this survey was gradually supplemented with capacity enlarging elements followed by simulation runs. By proceeding in this manner, it became evident that the bottleneck started to move and that the production output increases little by little to the target value of 7,000 batteries per week. The result of the simulation analysis of product A is shown in table 3. This product's production output can be increased to 7,110 batteries per week from 2,336 units per week by following seven clearly defined, successive individual measures. The grey elements in the table represent the moving bottleneck position in the production process.

CONCLUSION

By simulating the gradually expanded simulation model it was demonstrated that the total output of all three of the BBA products could be increased to over 7,000 batteries per week. This was achieved by proceeding demand-oriented and in several stages from a single shift operation to a multi shift operation. 10,500 units per week were achieved in some of the scenarios; thereby clearly surpassing the target. The necessary measures affect seven of ten plants.

The overall lead time could be shortened in spite of the increased performance and the capital commitment in terms of the current assets was reduced. Shortened lead times, higher capacity utilization and an increased performance are target figures that are often seen as incompatible (Corsten 1999). However this survey demonstrates that on the basis of a simulation analysis an increased synchronisation of resources and material flow can be achieved which at the same time enables a simultaneous improvement of the listed parameters. The adverse the performance coordination of individual plants and manual work centers was in the initial production system the better it is succeeded.

REFERENCES

Bloech, J. 2004. "Einführung in die Produktion", Springer, Berlin.

Corsten, H. 1999 „Produktionswirtschaft - Einführung in das industrielle Produktionsmanagement". Oldenbourg, Munich.

Eichenauer, B. 2002. „Optimizing Business Processes using Attributed Petri Nets". In *Simulation als betriebliche Entscheidungshilfe,* Biethahn, J. (Ed.) Göttingen, 15-30.

Fries, H. 1995. „Betriebswirtschaftslehre des Industriebetriebs". Oldenbourg, Munich.

Goldratt, E. 1988. "Computerized Shop Floor Scheduling". In *International Journal of Production Research* Vol. 26 (1988) No. 2, 443-455.

Oberweis, A. 1996. „Modellierung und Ausführung von Workflows mit Petri-Netzen". Teubner, Stuttgart.

BIBLIOGRAPHY

Claus-Burkard Boehnlein studied electrical engineering at the Technical University of Munich as well as Business Administration at the University of Munich and the University of Wuerzburg. After he finished his PhD at the University of Wuerzburg in 2001 his main research interest is Supply Chain Management and Simulation with Petri Nets. Since 2002 Claus-Burkard Boehnlein is the leader of the working group "Simulation in Business Administration" in the research Organization ASIM (German Simulation Society).

Christian Sahr studied Business Administration at the University of Wuerzburg. Since 2006 he is the production manager at the Bullith Batteries AG in Itzehoe/Germany.

Accelerating SoC Design using SysML and SystemC

Waseem Raslan
Mentor Graphics Egypt
51st Beirut St., Heliopolis
Cairo, Egypt
Email:waseem_raslan@mentor.com

Ahmed Sameh
The American University in Cairo
Computer Science Department
Down Town, Cairo, Egypt
Email:sameh@aucegypt.edu

KEYWORDS
Electronic System Level, ESL, SysML, SystemC, Code Generation, System on Chip, SoC, System Level Design.

ABSTRACT

Increasing design productivity is one of the main challenges facing the semiconductor design roadmap for >= 32 nm technology as stated by the International Technology Roadmap for Semiconductors, ITRS, report. Increasing the level of abstraction and increasing the design flow automation level are two of the methodologies suggested by the ITRS to address the design productivity challenge. Producing executable specifications is another methodology to increase the productivity.
This research suggests that the automatic mapping of SysML the system level UML based notations adopted by the Object Management Group, OMG, to the IEEE standardized SystemC, can raise the level of design abstraction in an automated environment that produces an executable specification of the design.

INTRODUCTION

As the size of the systems that can fit on a single chip (System on Chip or SoC) is getting larger while the technology is getting more complicated, designers face a huge challenge trying to validate the whole system before going on actual hardware fabrication and software full implementation. Raising the level of design abstraction is one key design methodology that has been suggested by the ITRS to overcome this design productivity challenge.
Electronic Design Automation, EDA, tools are very powerful at the register transfer level of abstraction and the abstraction levels below. Being detailed and slow by nature, these levels of abstraction are not suitable for validating a whole system that could be further implemented in software or hardware. In today's tight schedules and tough time to market competition, the ability to validate the whole system at a higher level of abstraction is becoming a mandatory for designers to avoid unpleasant surprises at critical phases of the system integration. Different levels of abstractions for software and hardware are shown in Figure 1.
Emerging System Level languages, like SystemC and SystemVerilog, have erupted to raise the level of abstraction level above the RTL one. Yet those languages still have upper limitations in terms of visual descriptions and ease of use at the system level. In addition, Martin shows in (Martin 2003) that those new languages can not, by themselves, be used to close the gap between hardware and software designers. Being very useful in describing software, several research works have suggested using UML notations in describing hardware as well.

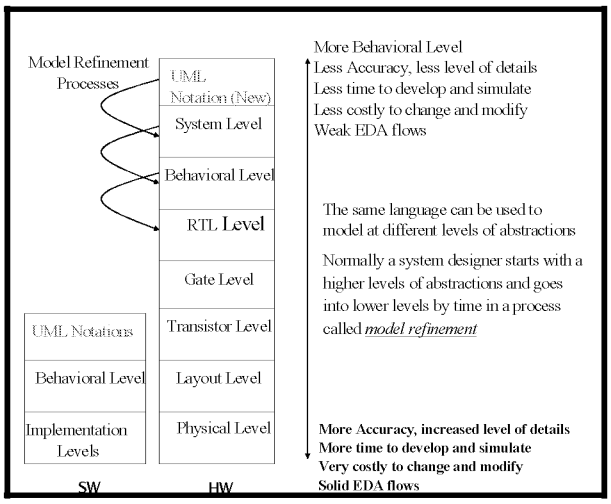


Figure 1 HW/SW Levels of Abstraction

Since SystemC is considered the language that could be best used for describing the system at a very high level of abstraction, and since it is a C++ based library, SystemC is considered the natural language to follow the UML in the abstraction level hierarchy. In addition, SystemC models produces an executable that can be run to simulate the system it models. If SystemC is modeling a system at a very high level of abstraction, this executable can be considered the executable specification of the system, which is one of the methodologies that are suggested by the ITRS to address the design productivity challenge.
Raising the automation level is also one of the methodologies suggested by the ITRS to overcome the productivity challenge, thus automated translation from UML to SystemC has been a hot of subject of research in the last four years. Several system level UML profiles have proposed in the context of those research works.
Realizing the importance to standardize these system level notation profiles, a request for proposal, RFP, have been presented to the OMG group to standardize a system level UML profile that can be best used to model a system. SysML has been proposed for the OMG group an d it has been adopted to be the standard system level UML profile. The goal of this research is to propose and suggest the SystemC modeling techniques that should result in modeling both the structure and behavioral SysML diagrams to

produce a single executable that represents the system behavior. A prototype for a translation tool, a SysML model compiler, has been implemented using a UML editing tool that supports SysML. Un-timed TLM SystemC models have been generated to model the expected behavior of the SysML design.

RELATED WORK

As pointed out earlier, several research works have been devoted to automate the translation of UML data model to SystemC through creating a dedicate UML profile that is used to model the whole system.

Thiagarajan et al. have proposed a system in (Nguyen et al. 2004) that parses the UML design produced from Rhapsody tool in its XML format to produce an abstract tree of the design that is fed to the Velocity template based code generation engine. The proposed system is fed with a set of SystemC templates to generate the SystemC code based on them. Their work was mainly targeting the class and state chart diagrams.

In a later research, (Tan et al. 2004) have proposed a system which they have called RT2SystemC translator. In this research, they have used Rational Rose Real Time UML editing tool to model the whole system. To generate the SystemC code, the C++ generated code from Rational Rose was converted into XML using the GCC_XML utility. The XML output is then preprocessed and parsed using JDOM, then translated into SystemC code.

Prevostini et al. have proposed in (Basu et al. 2005) the ACES design flow that is a complete design flow that starts from the UML level. ACES design flow grabs the system specifications through the class, state machine, activity, sequence and use case diagrams edited by the Rhapsody tool. Then ACES extracts the design information by traversing Rhapsody data model, accessed through Rhapsody APIs, to generate the SystemC models. ACES design flow takes the SystemC generated code down the way through the software, hardware and interface synthesis procedure.

Other research works have been conducted targeting SystemC code generation focusing mainly on the class and state machine diagrams as in (Mellor and Balcer 2003), (Mellor et al. 2005), (Zhu et al. 2002), (Rosenstiel et al. 2004), (Riccobene et al. 2005) and others.

All the previous translators were focusing on a subset of the UML diagrams, mainly class diagrams and state machine diagrams, using a proposed UML profile to match system level modeling. The current research is targeting SysML profile proposed by OMG to be the system level standard. The current research is not focusing on specific diagrams, on the contrary this research is attempting to benefit from most of the SysML diagrams during the SystemC transformation phase to have a matched SystemC model of the SysML design. This paper is suggesting using the redundant information present in SysML diagrams in validating the whole design

SysML OVERVIEW

All of the previously related research works have created their own UML profile to be suitable for system level modeling. Some researchers have gone the extra-mile and created a SystemC specific UML profile. Realizing the importance of standardization, Thales and others have submitted, in March 2003, a RFP "Request For Proposal" to the OMG "Object Management Group" to standardize a UML specific profile for system modeling. A working group has been formed and created a Systems Modeling Language Forum called SysML. The forum has proposed a SysML notation, based on UML 2.0 version, to address the issues raised in the SE RFP. Like UML, SysML became the OMG standard for system level modeling notation.

SysML is designed to be effective in specifying requirements, system structure, functional behavior, and allocations during specification and design phases of system engineering. SysML uses some UML 2.0 diagrams as is, extends some other UML 2.0 diagrams, and introduces new diagrams to achieve its goals as it follows:

1. <u>UML 2.0 diagrams that are reused without changes:</u> Use Case diagram, Sequence diagram, and State Machine diagram.
2. <u>UML 2.0 diagrams that are reused and extended:</u> Activity diagram (extends UML 2.0 Activity diagram), Block Definition diagram (extends UML 2.0 Class diagram), Internal Block diagram (extends UML 2.0 Composite Structure diagram), and Package diagram (extends UML 2.0 Package diagram).
3. <u>New diagrams that have been introduced are:</u> Parametric Constraint diagram, Allocation diagram/Allocation Trace ability Table, and Requirements diagram.

SystemC OVERVIEW

SystemC, as defined by OSCI, is a library of classes and a simulation kernel that extends the standard C++. SystemC enables modeling of systems by providing hardware constructs within the context of standard C++. Being built on standard C++, SystemC closes the gap between SW and HW designers in addition to providing a stable platform for development of system-level tools. Many design companies began to use SystemC in their design flows. SystemC Language Reference Manual (LRM) has been standardized as IEEE 1666 stndard by the IEEE.

SystemC has gained the fame of being superior at architectural design and architectural verification & HW/SW co-verification as suggested by Arnout et al. in their SystemC-SystemVerilog comparison in (Arnout and Brophy 2004). This superiority at architecture level, in addition to being C++ based language, recommended SystemC to be the next level design language to the UML level.

SystemC design building blocks, as defined by its LRM and illustrated in (Grother et al. 2002), are mainly: modules, ports, interfaces, channels and processes.

1. **<u>Modules:</u>** Modules are the main building blocks of the SystemC design. Basically any module is a C++ class derived from SC_MODULE SystemC class. Any module should contain a set of ports to communicate with other modules, a set of processes that describe the functionality of each module. Internal data and channels for holding the model state and communication among internal processes. In addition, modules can contain other modules to model the design hierarchy.

2. **Interfaces:** Interfaces are used to define the set of operations, parameters types and return values without specifying how these operations are actually implemented. These interfaces are used to allow for easy exploration of different implementations - given that they have the same interface - without having to do any change to the computational models.
3. **Ports:** Ports are the means for a module to interact with other modules through well-defined boundaries.
4. **Channels:** Channels are the implementation of the SystemC interfaces, they define how the channels behave and how they are implemented.
5. **Processes:** Defined as member functions in the modules, processes are the basic unit of functionality in SystemC. SC_MODULE methods must be registered, via SC_METHOD or SC_THREAD call, in the scheduler to be considered SystemC processes.

PROPOSED SysML TO SystemC CODE GENERATION

As part of this research, a dedicated survey has been conducted on both commercial and open source UML/SysML editing tools to select the appropriate tool for implementing the prototype. ARTiSAN Studio, Rhapsody and Telelogic TAU G2 were considered to be used in this research. According to the survey results, and given that Rhapsody has been used in a couple of strong similar research works in (Nguyen et al. 2004) and (Basu et al. 2005), Rhapsody 6.1 tool of I-Logix matched our selection criteria.

The proposed SystemC generation approach depends on accessing Rhapsody 6.1 data model information through its COM API interface. Rhapsody provides various ways to access its APIs: through Visual Basic, VB scripts, VC++, and Java. Accessing the APIs through VC++ has been selected. Having accessed Rhapsody SysML design information through its API's, a visitor based mechanism, (Czarnecki and Helsen 2003), has been selected to traverse the SysML design data model, through Rhapsody COM API interface, and generate the appropriate SystemC model for each component.

The implemented prototype also creates a makefile as well as a Microsoft Visual Studio 7.1 project for the generated SystemC code to be able to compile and link the generated SystemC code with the SystemC libraries to produce an executable without the need of the designer manual intervention. In other words the output of the system is a SystemC executable that can be run to validate the whole system, not just a set of SystemC generated files. The automatic generation of this makefile and the Microsoft project files adopts a template based code generation, (Czarnecki and Helsen 2003), methodology.

Semantically correct transformation between SysML design and SystemC generated models will be validated by testing the SystemC executable and ensuring that it behaves as expected for sample SysML designs.

Mapping SysML to SystemC

The prototype that has been implemented is applied to one of Rhapsody provided SysML design examples, the Mobile HandSet example. This example has three main blocks: ConnectionManagementBlock, MobilityManagementBlock and DataLinkBlock, in addition to two actors: the Mobile unit or the MMI actor, and the Network actor. Figure 2 shows the high level architecture of the Mobile Handset SysML design with its high level blocks and actors.

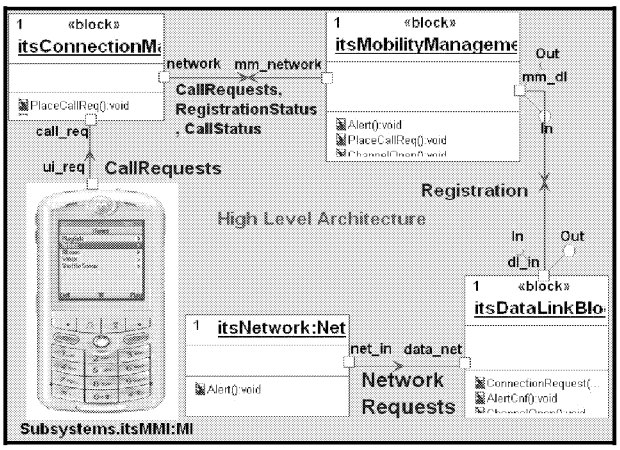


Figure 2 High Level Architecture of Mobile HandSet SysML Example provided by Rhapsody tool

The prototype is able to generate an executable that models both the structure and behavioral diagrams of the SysML design. The generated SystemC executable was run and the sequence of events method calls were compared to the expected sequence of events and methods calls listed in the provided sequence diagrams to find them matching. The prototype was also run on a SysML Radar system design that is another SysML design provided by Rhapsody. The prototype produced a successful SystemC specification of the design.

The following code snippet shows the generated SystemC code that represents the high level architecture design of the Mobile Handset design, where SysML blocks are directly mapped to SystemC modules and SysML ports are mapped to SystemC ports.

```
Subsystems::Subsystems(sc_module_name nm) : sc_module(nm) {
  // SubModule Allocations ...
  // Channel/Submodule allocations ...
  _itsConnectionManagementBlock = new
ConnectionManagementBlock("_itsConnectionManagementBlockInst");
  _itsMobilityManagementBlock = new
MobilityManagementBlock("_itsMobilityManagementBlockInst");
  _itsDataLinkBlock = new DataLinkBlock("_itsDataLinkBlockInst");
  _itsMMI = new MMI("_itsMMIInst");
  _itsNetwork = new Network("_itsNetworkInst");
  // Connectivity ...
  // Process Registration ...
}
Architecture::Architecture(sc_module_name nm) : sc_module(nm) {
  // SubModule Allocations ...
  // Connectivity ...
 _Subsystems = new Subsystems("_SubsystemsInst");
 _Subsystems->_itsConnectionManagementBlock->
call_req(itsConnectionManagementBlock_itsMMI);
 _Subsystems->_itsMMI->
ui_req(itsConnectionManagementBlock_itsMMI);
…
  // Rest of connectivity statements via signals …
  // Process Registration ...
```

}

The connectivity is done in SystemC via creating internal SystemC channels like itsConnectionManagementBlock_itsMMI and connecting the SystemC ports like call_req through statements like: _Subsystems-> _itsConnectionManagementBlock->call_req (itsConnectionManagementBlock_itsMMI); in the above code snippet.

The hierarchical block diagrams are traversed and each sub block is generated as a SystemC module while the SystemC module instantiation and connectivity are taken into consideration. Figure 3 shows the diagram of the ConnectionManagementBlock module.

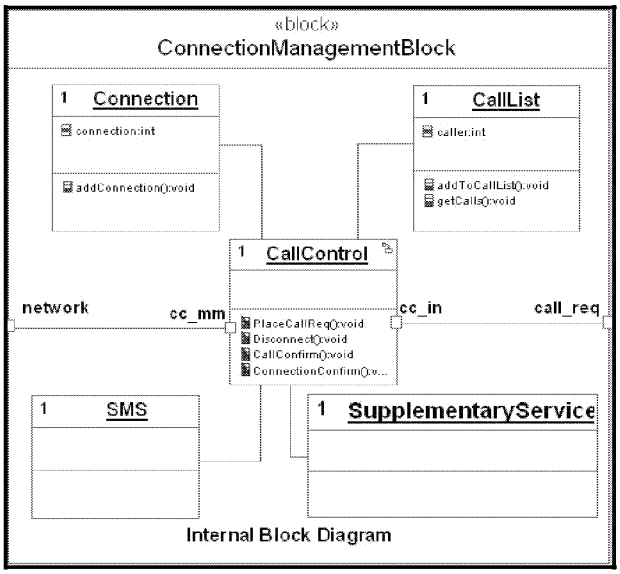


Figure 3 Connection Management Block Diagram

The SystemC generated code instantiates the different modules and makes the appropriate connectivity and association needed. The connectivity of ports is done through SystemC ports through code like: _CallControl->cc_mm(network); _CallControl->cc_in(call_req); While each sub-module has setter methods to represent the associations between CallControl module and the rest of the modules.

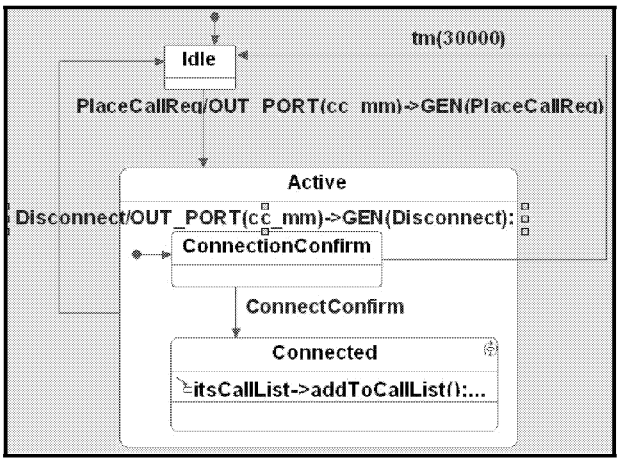


Figure 4 State Machine Diagram of CallControl Block

The dynamic behavior of the SysML can be represented in various diagrams: StateMachine, Activity, Sequence, and Use Case diagrams. Figure 4 represents the state machine diagram that specifies the dynamic behavior of the CallControl SysML block.

The following code snippet represents the main method that is generated to model the behavior described in the SysML state machine diagram shown above. The transition and state action code in the above figure is written in the SysML action language as specified by Rhapsody.

This action language is translated by the prototype to its equivalent SystemC code. For example the SysML action statement OUT_PORT(cc_mm)->GEN(Disconnect); is translated to cc_mm.**write**("Disconnect"); while each SystemC module has a dedicated code to intercept the incoming and fire an internal event that SystemC block behavior code will act upon.

```
void CallControl::processStatechartOfCallControloutputs() {
  switch(_currentStatechartOfCallControlstate) {
  case Connected:
  // currently in a voice or data call
  if (ConnectionManagement::_eventDisconnect.event() == true) {
  _nextStatechartOfCallControlstate.write(Idle);
  //Action that will be translated
  cc_mm.write("Disconnect");
  next_trigger(SC_ZERO_TIME);
  next_trigger();
  break;
  }
  // State Connected Action
  _itsCallList->operationaddToCallList();
  next_trigger(SC_ZERO_TIME);
  _itsConnection->operationaddConnection();
  next_trigger(SC_ZERO_TIME);
  cc_in.write("ConfirmIndication");
  next_trigger(SC_ZERO_TIME);
  next_trigger(
ConnectionManagement::_eventDisconnect.default_event());
   break;
  case Idle:
  // Indicates no call in progress
  if (ConnectionManagement::_eventPlaceCallReq.event() == true) {
    _nextStatechartOfCallControlstate.write(Active);
    //Actions
    cc_mm.write("PlaceCallReq");
    next_trigger(SC_ZERO_TIME);
    next_trigger();
    break;
  }
  // State Idle Action
  next_trigger(
ConnectionManagement::_eventPlaceCallReq.default_event());
  break;
  }
  // Rest if State triggers and actions …
};
```

Another example of modeling the SysML behavior diagrams is modeling the Activity diagrams. The MobilityManagementBlock has a sub-block called MMCallControl. The behavior of MMCallControl is modeled in SysML as Figure 5 shows.

In modeling the SysML Activity diagram, each activity is modeled as a SystemC process while the action language that triggers each activity is translated into the proper SystemC events that control process activation.

Usually the behavior description of the design is modeled in either the state machine or the activity diagrams. The information that reside in the sequence diagrams is considered redundant and represent another view of the sequence of events and operations that the design should go through when a specific event is triggered.

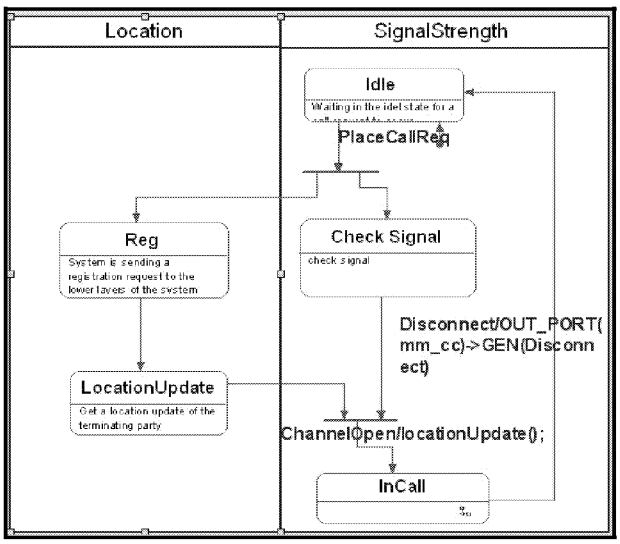


Figure 5 Activity Diagram of MMCallControl

This redundant information can be used to make sure the overall design is intact and goes through the specified sequence of operations as specified in the set of sequence diagrams. The following figure shows the sequence of events and messages that are generated when placing a call is requested.

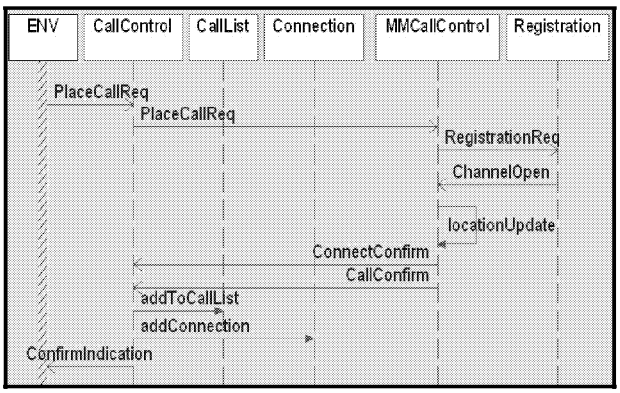


Figure 6 Sequence diagram of placing a call

The generated SystemC code is simply storing the events and operations called throughout the SystemC design execution in a proper data structure.
After the execution, the sequence diagrams events and messages are compared with the sequence of events and operations produced by running the design. If both sequences map to each other, the SysML/SystemC design is validated to be behaving as specified in the set of sequence diagrams.

CONCLUSIONS AND FUTURE WORK

SystemC represents the natural next level of system abstraction after UML/SysML. OMG is fostering SysML as a dedicated system level UML 2.0 profile. In this research, a SysML/SystemC model transformation prototype has been created with very promising results. As far as we know this research is the first dedicated SysML/SystemC model transformation. A study of the SysML/SystemC mapping and a model transformation prototype has been done with promising results.
Future work could be summarized as it follows:

1. Breaking the high dependency on the UML/SysML vendor. Use of open source tools should be considered.
2. Automatic creation of test benches that can be used at various levels of design refinements.
3. Linking this work to the complete flow beyond system level modeling and design.
4. Enabling SystemC visualization and code understanding through round trip engineering support.

REFERENCES

Arnout G. and Brophy D. 2004 "SystemC and SystemVerilog Designing SoC Together" Open SystemC Technology Symposium. San Diego, CA. DAC.

Basu A. S., Lajolo M., Prevostini M. 2005 "A Methodology for Bridging the Gap between UML and Codesign", UML for SoC Design Book by Kluwer/Springer.

Czarnecki K. and Helsen S. 2003 "Classification of Model Transformation Approaches" OOPSLA'03 Workshop on Generative Techniques in the Context of Model-Driven Architecture.

Grother T., Martin G., Liao S., Swan S. 2002 "System Design with SystemC". Kluwer Academic Publishers.

Martin G. 2003 "SystemC and the Future of Design Languages: Opportunities for Users and Research" Proceedings of the 16th Symposium on Integrated Circuits and Systems Design (SBCCI '03), © IEEE 2003.

Mellor S., Balcer M. 2003 "Executable UML: A Foundation for Model-Driven Architecture", Addison-Wesley.

Mellor S, Wolfe J., McCausland C. 2005 "Why Systems-on-Chip needs More UML like a Hole in the Head", Design, Automation, and Test Conference in Europe, DAC.

Nguyen K. D., Sun Z., Thiagarajan P.S., Wong W. 2004 "Model-driven SoC Design Via Executable UML to SystemC", RTSS'04.

Riccobene E., Scandura P., Rosti A., Bocchio S. 2005 "A SoC Design Methodology Involving a UML 2.0 Profile for SystemC", 5th ACM international conference on Embedded software.

Rosenstiel W., Kuhn T., Schweizer T., Winterholer M., Schulz-Key C. 2004 "Object-Oriented Modeling and Synthesis of SystemC Specifications", © IEEE.

Tan W. H., Thiagarajan P.S., Wong W. F., Zhu Y., Pilakkat S.K. 2004 "Synthesizable SystemC Code from UML Models", RTSS'04.

Zhu Q., Matsuda A., Kuwamura S., Nakata T., Shoji. M. 2002 "An Object-Oriented Design Process for System-on-Chip using UML". In The 15th Int'l Symp. on System Synthesis. © ACM.

SCHEDULING OPTIMIZATION AND ANALYSIS

SIMULATION-BASED OPTIMIZATION OF AGENT SCHEDULING IN MULTISKILL CALL CENTERS

Athanassios N. Avramidis, Michel Gendreau, Pierre L'Ecuyer

Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, C.P. 6128, Succ. Centre-Ville
Montréal (Québec), H3C 3J7, CANADA

Ornella Pisacane

Dipartimento di Elettronica Informatica e Sistemistica
Università della Calabria, Via P. Bucci, 41C, Arcavacata di Rende (CS), ITALY

ABSTRACT

We examine and compare simulation-based algorithms for solving the agent scheduling problem in a multiskill call center. This problem consists in minimizing the total costs of agents under constraints on the expected service level per call type, per period, and aggregated. We propose a solution approach that combines simulation with integer or linear programming, with cut generation. In our numerical experiments with realistic problem instances, this approach performs better than all other methods proposed previously for this problem. We also show that the two-step approach, which is the standard method for solving this problem, sometimes yield solutions that are highly suboptimal and inferior to those obtained by our proposed method.

INTRODUCTION

The telephone call center industry employs millions of people around the world and is fast growing. A few percent saving in workforce salaries easily means several million dollars. Call centers often handle several types of calls distinguished by the required skills for delivering service. Training all agents to handle all call types is not cost-effective. Each agent has a selected number of skills and the agents are distinguished by the set of call types they can handle (also called their *skill set*). When such skill constraints exist, we speak of a *multiskill* call center. *Skill-based routing* (SBR), or simply *routing*, refers to the rules that control the call-to-agent and agent-to-call assignments. Most modern call centers perform skill-based routing (Koole and Mandelbaum 2002, Gans et al. 2003). In a typical call center, inbound calls arrive at random according to some complicated stochastic processes, call durations are also random, waiting calls may abandon after a random patience time, some agents may fail to show up to work for any reason, and so on. Based on forecasts of call volumes, call center managers must decide (among other things) how many agents of each type (i.e., skill set) to have in the center at each time of the day, must construct working schedules for the available agents, and must decide on the call routing rules. These decisions are made under a high level of uncertainty. The goal is typically to provide the required quality of service at minimal cost. The most common measure of quality of service is the *service level* (SL), defined as the long-term fraction of calls whose time in queue is no larger than a given threshold. Frequently, multiple measures of SL are of interest: for a given time period of the day, for a given call type, for a given combination of call type and period, aggregated over the whole day and all call types, and so on. For certain call centers that provide public services, SL constraints are imposed by external authorities, and violations may result in stiff penalties (CRTC 2000).

In this paper, we assume that we have a detailed stochastic model of the dynamics of the call center for one day of operation. This model specifies the stochastic processes for the call arrivals (these processes are usually non-stationary and doubly stochastic), the distributions of service times and patience times for calls, the call routing rules, the periods of unavailability of agents between calls (e.g., to fill out forms, or to go to the restroom, etc.), and so forth. We formulate a stochastic optimization problem where the objective is to minimize the total cost of agents, under various SL constraints. This could be used in long-term planning, to decide how many agents to hire and for what skills to train them, or for short-term planning, to decide which agents to call for work on a given day and what would be their work schedule. The problem is difficult because for any given fixed staffing of agents (the staffing determines how many agents of each type are available in each time period), no reliable formulas or quick numerical algorithms are available

to estimate the SL; it can be estimated accurately only by long (stochastic) simulations. Scheduling problems in general are difficult (they are NP-hard) even in deterministic settings where each solution can be evaluated quickly and exactly. When this evaluation requires costly and noisy simulations, as is the case here, solving the problem exactly is even more difficult and we must settle with methods that are partly heuristic.

Staffing in the *single-skill* case (i.e., single call type and single agent type) has received much attention in the call center literature. Typically, the workload varies considerably during the day (Gans et al. 2003, Avramidis et al. 2004, Brown et al. 2005), and the planned staffing can change only at a few discrete points in time (e.g., at the half hours). It is common to divide the day into several periods during which the staffing is held constant and the arrival rate does not vary much. If the system can be assumed to reach steady-state quickly (relative to the length of the periods), then steady-state queueing models are likely to provide a reasonably good staffing recommendation for each period. For instance, in the presence of abandonments, one can use an Erlang-A formula to determine the minimal number of agents for the required SL in each period (Gans et al. 2003). When that number is large, it is often approximated by the *square root safety staffing formula*, based on the Halfin-Whitt heavy-traffic regime, and which says roughly that the capacity of the system should be equal to the workload plus some safety staffing which is proportional to the square root of the workload (Halfin and Whitt 1981, Gans et al. 2003). Scheduling problems are often solved in two separate steps (Mehrotra 1997): After an appropriate staffing has been determined for each period in the first step, a minimum-cost set of shifts that covers this staffing requirement can be computed in the second step by solving a linear integer program. However, the constraints on admissible working shifts often force the second step solution to overstaff in some of the periods. This drawback of the two-step approach has been pointed out by several authors, who also proposed alternatives (Keith 1979, Thompson 1997, Henderson and Mason 1998, Ingolfsson et al. 2003, Atlason et al. 2004). For example, the SL constraint is often only for the time-aggregated (average) SL over the entire day; in that case, one may often obtain a lower-cost scheduling solution by reducing the minimal staffing in one period and increasing it in another period. Atlason et al. (2004) developed a *simulation-based* methodology to optimize agent's scheduling in the presence of uncertainty and general SL constraints, based on simulation and cutting-plane ideas. Linear inequalities (cuts) are added to an integer program until its optimal solution satisfies the required SL constraints. The SL and the cuts are estimated by simulation.

In the *multiskill case*, the staffing and scheduling problems are more challenging, because the workload can be covered by several possible combinations of skill sets, and the routing rules also have a strong impact on the performance. Staffing a single period in steady-state is already difficult; the Erlang formulas and their approximations (for the SL) no longer apply. Simulation seems to be the only reliable tool to estimate the SL. Cezik and L'Ecuyer (2007) adapt the simulation-based methodology of Atlason et al. (2004) to the *optimal staffing* of a multiskill call center for a *single period*. They point out difficulties that arise with this methodology and develop heuristics to handle them. Avramidis et al. (2006) solve the same problem by using neighborhood search methods combined with an analytical approximation of SLs, with local improvement via simulation at the end. Pot et al. (2007) impose a constraint only on the aggregate SL (across all call types); they solve Lagrangean relaxations using search methods and analytical approximations.

Some authors have developed queueing approximations for the case of two call types, via Markov chains and under simplifying assumptions; see Stolletz and Helber (2004) for example. But here we are thinking of 20 to 50 call types or more, which is common in modern call centers, and for which computation via these types of Markov chain models is clearly impractical.

For the *multiskill scheduling problem*, Bhulai et al. (2007) propose a two-step approach in which the first step determines a staffing of each agent type for each period, and the second step computes a schedule by solving an IP in which this staffing is the right-hand side in key constraints. A key feature of the IP model is that the staff-coverage constraints allow *downgrading* an agent into any alternative agent type with smaller skill set, separately for each period. Bhulai et al. (2007) recognize that their two-step approach is generally suboptimal.

In this paper, we propose a simulation-based algorithm for solving the multiskill scheduling problem, and compare it to the approach of Bhulai et al. (2007). This algorithm extends the method of Cezik and L'Ecuyer (2007), which solves a single-period staffing problem. In contrast with the two-step approach, our method optimizes the staffing and the scheduling simultaneously. Our numerical experiments show that our algorithm provides approximate solutions to large-scale realistic problem instances in reasonable time (a few hours). These solutions are typically better, sometimes by a large margin (depending on the problem), than the best solutions from the two-step approach. We are aware of no competitive faster method.

MODEL FORMULATION

We now provide definitions of the multiskill staffing and scheduling problems. We assume that we have a stochastic model of the call center, under which the mathematical expectations used below are well defined, and that we can simulate the dynamics of the center under this model. Our

problem formulations here do not depend on the details of this model.

There are K call types, labeled from 1 to K, and I agent types, labeled from 1 to I. Agent type i has the skill set $S_i \subseteq \{1,\ldots,K\}$. The day is divided in P period, labeled from 1 to P. The *staffing vector* is $\mathbf{y} = (y_{1,1},\ldots,y_{1,P},\ldots,y_{I,1},\ldots,y_{I,P})^{\mathrm{t}}$ where $y_{i,p}$ is the number of agents of type i available in period p. Given $\mathbf{y}$, the *service level* (SL) in period p is defined as

$$g_{k,p}(\mathbf{y}) = \mathbb{E}[S_{\mathrm{g},k,p}]/\mathbb{E}[S_{k,p} + A_{k,p}],$$

where $S_{k,p}$ is the number of type-k calls that arrive in period p, $S_{\mathrm{g},k,p}$ is the number of those calls that get served after waiting at most $\tau_{k,p}$ (a constant called the *acceptable waiting time*), and $A_{k,p}$ is the number of those calls that abandon in period p after waiting at least $\tau_{k,p}$. Aggregate SLs, per call type, per period, and globally, are defined analogously. Given acceptable waiting times τ_p, τ_k, and τ, the aggregate SLs are denoted by $g_p(\mathbf{y})$, $g_k(\mathbf{y})$ and $g(\mathbf{y})$ for period p, call type k, and overall, respectively.

A *shift* is defined by specifying the time periods in which an agent is available to handle calls. Let $\{1,\ldots,Q\}$ be the set of all admissible shifts. We assume that this set is the same for all agent types. The admissible shifts are specified via a $P \times Q$ matrix $\mathbf{A}_0$ whose element (p,q) is $a_{p,q} = 1$ if an agent with shift q works in period p, and 0 otherwise. A vector $\mathbf{x} = (x_{1,1},\ldots,x_{1,Q},\ldots,x_{I,1},\ldots,x_{I,Q})^{\mathrm{t}}$, where $x_{i,q}$ is the number of agents of type i working shift q, is a *schedule*. The *cost vector* is $\mathbf{c} = (c_{1,1},\ldots,c_{1,Q},\ldots,c_{I,1},\ldots,c_{I,Q})^{\mathrm{t}}$, where $c_{i,q}$ is the cost of an agent of type i with shift q. To any given shift vector $\mathbf{x}$, there corresponds the staffing vector $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a block-diagonal matrix with I identical blocks $\mathbf{A}_0$, if we assume that each agent of type i works as a type-i agent for her entire shift.

However, following Bhulai et al. (2007), we also allow an agent of type l to be downgraded to an agent with smaller skill set, i.e., of type i_p where $S_{i_p} \subset S_l$, in any time period p of her shift. Define $\mathscr{S}_i^+ = \{j : S_j \supset S_i\}$, $\mathscr{S}_i^- = \{j : S_j \subset S_i\}$, and let $z_{l,i,p}$ be the number of type-l agents that are downgraded to type i during period p. These are the *skill transfer* variables. A schedule $\mathbf{x} = (x_{1,1},\ldots,x_{1,Q},\ldots,x_{I,1},\ldots,x_{I,Q})^{\mathrm{t}}$ is said to *cover the staffing* $\mathbf{y} = (y_{1,1},\ldots,y_{1,P},\ldots,y_{I,1},\ldots,y_{I,P})^{\mathrm{t}}$ if for $i = 1,\ldots,I$ and $p = 1,\ldots,P$, there are nonnegative integers $z_{l,i,p}$ for $l \in \mathscr{S}_i^+$ and $z_{i,l,p}$ for $l \in \mathscr{S}_i^-$, such that

$$\sum_{q=1}^{Q} a_{p,q} x_{i,q} + \sum_{l \in \mathscr{S}_i^+} z_{l,i,p} - \sum_{l \in \mathscr{S}_i^-} z_{i,l,p} \;\geq\; y_{i,p}. \qquad (1)$$

These inequalities can be written in matrix form as $\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} \geq \mathbf{y}$, where $\mathbf{z}$ is a column vector whose elements are the $z_{l,i,p}$ variables and $\mathbf{B}$ is a matrix whose entries are in the set $\{-1,0,1\}$. With this notation, the *scheduling problem* can be formulated as

$$\begin{array}{ll}
\text{(P0)}: & \text{[Scheduling problem]} \\
\min & \mathbf{c}^{\mathrm{t}}\mathbf{x} = \sum_{i=1}^{I}\sum_{q=1}^{Q} c_{i,q}x_{i,q} \\
\text{s.t.} & \\
& \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} \geq \mathbf{y} \\
& g_{k,p}(\mathbf{y}) \geq l_{k,p} \text{ for } 1 \leq k \leq K \text{ and } 1 \leq p \leq P \\
& g_p(\mathbf{y}) \geq l_p \quad \text{for } 1 \leq p \leq P \\
& g_k(\mathbf{y}) \geq l_k \quad \text{for } 1 \leq k \leq K \\
& g(\mathbf{y}) \geq l \\
& \mathbf{x} \geq 0,\ \mathbf{z} \geq 0,\ \mathbf{y} \geq 0 \text{ and integer}
\end{array}$$

where $l_{k,p}$, l_p, l_k and l are given constants.

In practice, a given agent often works more efficiently (faster) when handling a smaller number of calls (i.e., if his skill set is artificially reduced). The possibility of downgrading agents to a smaller skill set for some periods can sometimes be exploited to take advantage of this increased efficiency. In case where the agent's speed for a given call type (in the model) does not depend on its skill set, one might think intuitively that downgrading cannot help, because it only limits the flexibility of the routing. This would be true if we had an optimal dynamic routing of calls. But in practice, an optimal dynamic routing is too complicated to compute and simpler routing rules are used instead. These simple rules are often static. Then, downgrading may sometimes help by effectively changing the routing rules. Clearly, the presence of skill transfer variables in (P0) cannot increase the optimal cost, it can only reduce it.

Suppose we consider a single period, say period p, and we replace $g_{k,p}(\mathbf{y})$ and $g_p(\mathbf{y})$ by approximations that depend on the staffing of period p only, say $\tilde{g}_{k,p}(y_{1,p},\ldots,y_{I,p})$ and $\tilde{g}_p(y_{1,p},\ldots,y_{I,p})$, respectively. If all system parameters are assumed constant over period p, then natural approximations are obtained by assuming that the system is in steady-state over this period. The single-period multiskill staffing problems can then be written as

$$\begin{array}{ll}
\text{(P1)}: & \text{[Staffing problem]} \\
\min & \sum_{i=1}^{I} c_i y_i \\
\text{s.t.} & \\
& \tilde{g}_k(y_1,\ldots,y_I) \geq l_k \quad \text{for } 1 \leq k \leq K \\
& \tilde{g}(y_1,\ldots,y_I) \geq l \\
& y_i \geq 0 \quad \text{and integer for all } i
\end{array}$$

where c_i is the cost of agent type i (for a single period), and the period index was dropped throughout. Simulation-

based solution methods for this problem are proposed in Cezik and L'Ecuyer (2007) and Avramidis et al. (2006). Pot et al. (2007) address a restricted version of it, with a single constraint on the aggregate SL over the period (i.e., they assume $l_k = 0$ for all k).

In the approach of Bhulai et al. (2007), the first step is to determine an appropriate staffing, $\hat{\mathbf{y}} = (\hat{y}_{1,1}, \ldots, \hat{y}_{1,P}, \ldots, \hat{y}_{I,1}, \ldots, \hat{y}_{I,P})^{\mathrm{t}}$. For this, they look at each period p in isolation and solve a version of (P1) with a single constraint on the aggregate SL; this gives $\hat{y}_{1,p}, \ldots, \hat{y}_{I,p}$ for each p. In their second step, they find a schedule that covers this staffing by solving:

$$
\begin{array}{rl}
\text{(P2)}: & \text{[Two-stage approach]} \\
\min & \mathbf{c}^{\mathrm{t}}\mathbf{x} \\
\text{s.t.} & \\
& \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} \geq \hat{\mathbf{y}} \\
& \mathbf{x} \geq 0, \mathbf{z} \geq 0 \quad \text{and integer}
\end{array}
$$

The presence of skill-transfer variables generally reduces the optimal cost in (P2) by adding flexibility, compared with the case where no downgrading is allowed. However, there sometimes remains a significant gap between the optimal solution of (P0) and the best solution found for the same problem by the two-step approach. The following simplified example illustrates this.

Example 1 Let $K = I = P = 3$, and $Q = 1$. The single type of shift covers the three periods. The skill sets are $S_1 = \{1,2\}$, $S_2 = \{1,3\}$, and $S_3 = \{2,3\}$. All agents have the same shift and the same cost. Suppose that the total arrival process is stationary Poisson with mean 100. This incoming load is equally distributed between call types $\{1,2\}$ in period 1, $\{1,3\}$ in period 2, $\{2,3\}$ in period 3. Any agent can be downgraded to a specialist that can handle a single call type (that belongs to his skill set), in any period. In the presence of such specialists, an incoming call goes first to its corresponding specialist if there is one available, otherwise it goes to a generalist that can handle another call type as well. When an agent becomes available he serves the call that has waited the longest among those in the queue (if any). The service times are exponential with mean 1, there are no abandonments, and the SL constraints specify that 80% of all calls must be served within 20 seconds, in each time period, on average over an infinite number of days.

If we assume that the system operates in steady-state in period 1, then the optimal staffing for that period is 104 agents of type 1. Since all agents can serve all calls, we have in this case an $M/M/s$ queue with $s = 104$, and the global SL is 83.4%, as can be computed by the Erlang-C formula. By symmetry, the optimal staffing solutions for the other periods are obviously the same: 104 agents of type 2 in period 2 and 104 agents of type 3 in period 3. Then, the two-step approach gives a solution to (P2) with 104 agents of each type, for a total of 312 agents.

If we solve (P0) directly instead (e.g., using the simulation-based algorithm described in the next section), assuming again (as an approximation) that the system is in steady-state in each of the three periods, we find a feasible solution with 35 agents of type 1, 35 agents of type 2, and 34 agents of type 3, for a total of 104 agents. With this solution, during period 1, the agents of types 2 and 3 are downgraded to specialists who handle only call types 1 and 2, respectively, and the agents of type 1 act as generalists. A similar arrangement applies to the other periods, mutatis mutandis. Note that this solution of (P0) remains valid even if we remove the skill transfer variables from the formulation of (P0), if we assume that the routing rules do not change; i.e., if calls are always routed first to agents that can handle only this call type among the calls that can arrive during the current period.

Suppose now that we add the additional skill sets $S_4 = \{1\}$, $S_5 = \{2\}$, $S_6 = \{3\}$, and that these new specialists cost 6 each, whereas the agents with two skills cost 7. In this case it becomes attractive to use specialists to handle a large fraction of the load, because they are less expensive, and to keep a few generalists in each period to obtain a "resource sharing" effect. It turns out that an optimal staffing solution for period 1 is 2 generalists (type 1) and 52 specialists of each of the types 4 and 5. An analogous solution holds for each period. With these numbers, if downgrading is not possible, the two-step approach gives a solution with 6 generalists (2 of each type) and 156 specialists (52 of each type), for a total cost of 978. If downgrading is allowed, then the two-step approach finds the following much better solution: 2 agents of type 1 and 52 of each of the types 2 and 3, for a total cost of 742. The reader can easily verify that by appropriate downgrading in each period, this solution can cover the optimal staffing in each period. If we solve (P0) directly with these additional skill sets, we get the same solution as without them; i.e., 104 agents with two skills each, for a total cost of 728. This is again better than with the two-step approach, but the gap is much smaller than what we had with only three skill sets.

Example 2 Observe that in the previous example, if all the load was from a single call type, there would be a single agent type and the two-step approach would provide exactly the same solution as the optimal solution of (P0). The example illustrates a suboptimality gap due to a variation in the *type* of load.

Another potential source of suboptimality (this one can occur even in the case of a single call type) is the time variation of the total load from period to period. If there is only a global SL constraint over the entire day, then the optimal solution may allow a lower SL during one (or more) peak period(s) and recover an acceptable global SL

by catching up in the other periods. To account for this, Bhulai et al. (2007), Section 5.4, propose a heuristic based on the solution obtained by their basic two-step approach. Although this appeared to work well in their examples, the effectiveness of this heuristic for general problems is not clear.

Yet another (important) type of limitation that can significantly increase the total cost is the restriction on the set of available shifts. Suppose for example that there is a single call type, that the day has 10 periods, and that all shifts must cover 8 periods, with 7 periods of work and a single period of lunch break after 3 or 4 periods of work. Thus a shift can start in period 1, 2, or 3, and there are six shift types in total. Suppose we need 100 agents available in each period. For this we clearly need 200 agents, each one working for 7 periods, for a total of 1400 agent-periods. If there were no constraints on the duration and shape of shifts, on the other hand, then 1000 agent-periods would suffice.

OPTIMIZATION BY SIMULATION AND CUTTING PLANES

We summarize the proposed simulation-based optimization algorithm. The general idea is to replace the problem (P0) by a *sample* version of it, (SP0$_n$), and then replace the nonlinear SL constraints by a small set of linear constraints, in a way that the optimal solution of the resulting *relaxed* sample problem is close to that of (P0). The relaxed sample problem is solved by linear or integer programming.

We first describe how the relaxation works when applied directly to (P0); its works the same way when applied to the sample problem. Consider a version of (P0) in which the SL constraints have been replaced by a small set of linear constraints that do not cut out the optimal solution. Let $\bar{\mathbf{y}}$ be the optimal solution of this (current) relaxed problem. If $\bar{\mathbf{y}}$ satisfies all SL constraints of (P0), then it is an optimal solution of (P0) and we are done. Otherwise, take a violated constraint of (P0), say $g(\bar{\mathbf{y}}) < l$, suppose that g is (jointly) concave in $\mathbf{y}$ for $\mathbf{y} \geq \bar{\mathbf{y}}$, and that $\bar{\mathbf{q}}$ is a *subgradient* of g at $\bar{\mathbf{y}}$. Then

$$g(\mathbf{y}) \leq g(\bar{\mathbf{y}}) + \bar{\mathbf{q}}^{\mathsf{t}}(\mathbf{y} - \bar{\mathbf{y}})$$

for all $\mathbf{y} \geq \bar{\mathbf{y}}$. We want $g(\mathbf{y}) \geq l$, so we must have

$$l \leq g(\mathbf{y}) \leq g(\bar{\mathbf{y}}) + \bar{\mathbf{q}}^{\mathsf{t}}(\mathbf{y} - \bar{\mathbf{y}}),$$

i.e.,

$$\bar{\mathbf{q}}^{\mathsf{t}}\mathbf{y} \geq \bar{\mathbf{q}}^{\mathsf{t}}\bar{\mathbf{y}} + l - g(\bar{\mathbf{y}}). \tag{2}$$

Adding this linear *cut inequality* to the constraints removes $\bar{\mathbf{y}}$ from the current set of feasible solutions of the relaxed problem without removing any feasible solution of (P0). On the other hand, in case $\bar{\mathbf{q}}$ is not really a subgradient (which may happens in practice), then we may cut out feasible solutions of (P0), including the optimal one. We will return to this.

Since we cannot evaluate the functions g exactly, we replace them by a sample average over n independent days, obtained by simulation. Let ω represent the sequence of independent uniform random numbers that drives the simulation for those n days. When simulating the call center for different values of $\mathbf{y}$, we assume that the same uniform random numbers are used for the same purpose for all values of $\mathbf{y}$, for each day. That is, we use the same ω for all $\mathbf{y}$. Proper synchronization of these *common random numbers* is implemented by using a random number package with multiple streams and substreams (L'Ecuyer et al. 2002, L'Ecuyer 2004).

The *empirical SL* over these n simulated days is a function of the staffing $\mathbf{y}$ and of ω. We denote it by $\hat{g}_{n,k,p}(\mathbf{y},\omega)$ for call type k in period p; $\hat{g}_{n,p}(\mathbf{y},\omega)$ aggregated over period p; $\hat{g}_{n,k}(\mathbf{y},\omega)$ aggregated for call type k; and $\hat{g}_n(\mathbf{y},\omega)$ aggregated overall. For a *fixed* ω, these are all deterministic functions of $\mathbf{y}$. Instead of solving directly (P0), we solve its *sample-average approximation* (SP0$_n$) obtained by replacing the functions g in (P0) by their sample counterparts $\hat{g}$ (here, $\hat{g}$ stands for any of the empirical SL functions, and similarly for g).

We know that $\hat{g}_{n,k,p}(\mathbf{y})$ converges to $g_{k,p}(\mathbf{y})$ with probability 1 for each (k,p) and each $\mathbf{y}$ when $n \to \infty$. In this sense, (SP0$_n$) converges to (P0) when $n \to \infty$. Suppose that we eliminate a priori all but a *finite* number of solutions for (P0). This can easily be achieved by eliminating all solutions for which the total number of agents is unreasonably large. Let $\mathscr{Y}^*$ be the set of optimal solutions of (P0) and suppose that no SL constraint is satisfied exactly for these solutions. Let $\mathscr{Y}_n^*$ be the set of optimal solutions of (SP0$_n$). Then, the following theorem implies that for n large enough, an optimal solution to the sample problem is also optimal for the original problem. It can be proved by a direct adaptation of the results of Vogel (1994) and Atlason et al. (2004); see also Cezik and L'Ecuyer (2007).

Theorem 1 *With probability 1, there is an integer $N_0 < \infty$ such that for all $n \geq N_0$, $\mathscr{Y}_n^* = \mathscr{Y}^*$. Moreover, under mild assumptions on the arrival processes, see Cezik and L'Ecuyer (2007), there are positive real numbers α and β such that for all n,*

$$P[\mathscr{Y}_n^* = \mathscr{Y}^*] \geq 1 - \alpha e^{-\beta n}.$$

We solve (SP0$_n$) by the cutting plane method described earlier, with the functions g replaced by their empirical counterparts. The major practical difficulty is to obtain the subgradients $\bar{\mathbf{q}}$. In fact, the functions $\hat{g}$ in the empirical problem (computed by simulation) are not necessarily concave for finite n, even in the areas where the functions g of

(P0) are concave. To obtain a (tentative) subgradient $\bar{\mathbf{q}}$ of a function $\hat{g}$ at $\bar{\mathbf{y}}$, we use forward finite differences as follows. For $j = 1, \ldots, IP$, we choose an integer $d_j \geq 0$, we compute the function $\hat{g}$ at $\bar{\mathbf{y}}$ and at $\bar{\mathbf{y}} + d_j \mathbf{e}_j$ for $j = 1, \ldots, IP$, where $\mathbf{e}_j$ is the jth unit vector, and we define $\bar{\mathbf{q}}$ as the IP-dimensional vector whose jth component is

$$\bar{q}_j = [\hat{g}(\bar{\mathbf{y}} + d_j \mathbf{e}_j) - \hat{g}(\bar{\mathbf{y}})]/d_j. \tag{3}$$

In our experiments, we used the same heuristic as in Cezik and L'Ecuyer (2007) to select the d_j's: We took $d_j = 3$ when the service level corresponding to the considered cut was less than 0.5, $d_j = 2$ when it was between 0.5 and 0.65, and $d_j = 1$ when it was greater than 0.65. When we need a subgradient for a period-specific empirical SL ($\hat{g}_p$ or $\hat{g}_{k,p}$), the finite difference is formed only for those components of $\mathbf{y}$ corresponding to the given period; the other elements of $\bar{\mathbf{q}}$ are set to zero. This heuristic introduces inaccuracies, because $\hat{g}_p$ and $\hat{g}_{k,p}$ depend in general on the staffing of all periods up to p or even $p+1$, but it reduces the work significantly.

Computing $\bar{\mathbf{q}}$ via (3) requires $IP + 1$ simulations of n days each. This is by far the most time-consuming part of the algorithm. Even for medium-size problems, these simulations can easily require an excessive amount of time. For this reason, we use yet another important short-cut: We generally use a smaller value of n for estimating the subgradients than for checking feasibility. (The latter requires a single n-day simulation experiment.) That is, we compute each $\hat{g}(\bar{\mathbf{y}} + d_j \mathbf{e}_j)$ in (3) using $n_0 < n$ days of simulation, instead of n days. In most of our experiments (including those reported in this paper), we have used $n_0 \approx n/10$.

With all these approximations and the simulation noise, we recognize that the vector $\bar{\mathbf{q}}$ thus obtained is only a *heuristic guess* for a subgradient. It may fail to be a subgradient. In that case the cut (2) may remove feasible staffing solutions including the optimal one, and this may lead our algorithm to a suboptimal schedule; Atlason et al. (2004) and Cezik and L'Ecuyer (2007) give examples of this. For this reason, it is a good idea to run the algorithm more than once with different streams of random numbers and/or slightly different parameters, and retain the best solution found.

At each step of the algorithm, after adding new linear cuts, we solve a relaxation of (SP0$_n$) in which the SL constraints have been replaced by a set of linear constraints. This is an integer programming (IP) problem. But when the number of integer variables is large, we just solve it as a linear program (LP) instead, because solving the IP becomes too slow. To recover an integer solution, we select a threshold τ between 0 and 1; then we round up (to the next integer) the real numbers whose fractional part is larger than τ and we truncate (round down) the other ones. We memorize the cumulated amount of truncation and whenever it exceeds 1, we reset it to 0 and add one agent of the currently considered type. These two versions of the CP algorithm are denoted CP-IP and CP-LP.

When we add new cuts, we give priority to the cuts associated with the global SL constraints, followed by aggregate ones specific to a call type, followed by aggregate ones specific to a period, followed by the remaining ones. This is motivated by the intuitive observation that the more aggregation we have, the smoother is the empirical SL function, because it involves a larger number of calls. So its gradient is less likely to oscillate and the vector $\mathbf{q}$ defined earlier is more likely to be a subgradient. Moreover, in the presence of abandonments, the SL functions tend to be non-concave in the areas where the SL is very small, and very small SL values tend to occur less often for the aggregated measures than for the more detailed ones that were averaged. Adding cuts that strengthen the aggregate SL often helps to increase the small SL values associated with specific periods and call types.

After adding enough linear cuts, we eventually end up with a feasible solution for (SP0$_n$). This solution may be infeasible for (P0) (because of random noise, especially if n is small) or may be feasible but suboptimal for (P0) (because one of the cuts may have removed the optimal solution of (P0) from the feasible set of (SP0$_n$)). To try improving our solution to (SP0$_n$), we do a local search around it, still using the same n and the same random numbers. This local search has two phases. In phase 1, we attempt to reduce the cost by iteratively removing one shift at a time, until either none of the possibilities is feasible or a time limit is reached. For the CP-LP version, we first round the solution to integer by using a threshold τ as explained earlier. We start with $\tau = 0.5$ and decrease the value of τ by 0.01 successively until we get a feasible solution. In phase 2, we attempt to reduce the cost by iteratively considering a *switch move* in which we try to replace an agent/shift pair by another one with smaller cost. The candidates for the switch are drawn at random, at each step. Phase 2 terminates when a time limit is reached, or when a maximum number of consecutive moves without improvement is reached.

If we start the cutting plane algorithm with a full relaxation of (SP0$_n$) (no constraint at all), the optimal solution of this relaxation is $\mathbf{y} = \mathbf{0}$. The functions $\hat{g}$ are not concave at $\mathbf{0}$, and we cannot get subgradients at that point, so we cannot start the algorithm from there. As a heuristic to quickly remove this area where the staffing is too small and the SL is non-concave, we restrict the set of admissible solutions a priori by imposing (extra) initial constraints. To do that, we impose that for each period p, the skill supply of the available agents covers at least α_k times the total load for each call type k (defined as the arrival rate of that call type divided by its service rate), where each α_k is a constant, usually close to 1. Finding the corresponding linear constraints is easily achieved by solving a max flow

problem in a graph. See Ceşik and L'Ecuyer (2007) for the details.

A NUMERICAL ILLUSTRATION

We consider a call center with $K = 20$ call types and $I = 35$ agent types, whose skill sets are shown in Table 1. There are 52 time periods of 15 minutes each, so the center operates for 13 hours each day. Arrivals are assumed to obey a Poisson process stationary over each period, for each call type, and independent across call types. The arrival rates for each period and call type can be found in an extended version of this paper, available from the authors; they vary from 5 to 27 calls per minute. The rates increase gradually over the first 10 to 12 periods, then they decrease slowly for the rest of the day. All service times are exponential with mean 8 minutes and patience times have a mixture distribution: the patience is 0 with probability 0.001, and with probability 0.999, it is exponential with rate 0.1 per minute. We consider 123 different shifts, all lasting 7.5 hours and including one 30-minute lunch break near the middle and two 15-minute coffee breaks (one pre-lunch and one post-lunch). A description of these shifts can be found in the extended version of the paper. The cost of an agent with s skills is $0.9 + s/10$. The SL constraints are that for each period, at least 80% of the calls (aggregated over all types) must be answered within 20 seconds, on average over many days. That is, $\tau_p = 20$ seconds and $l_p = 0.8$ for each p. This implies that the global constraint with $\tau = 20$ seconds and $l = 0.8$ must also be satisfied. There are no other constraint. All these numbers are inspired from observations in real-life call centers at Bell Canada. In particular, we point out the presence of specialists (agents with a single skill) in the available skill sets for all call types.

We solved this problem using (1) CP using LP and rounding up at each stage (CP-LP), and (2) the two-step approach in which the staffing is first optimized separately for each period using steady-state approximation via simulation with batch means (TS). (The CP-IP is not practical for this problem instance, because the IP is too large to be solved exactly at each step with the given CPU time budget.) Each method was given a CPU time budget of 5 hours and was applied 8 times, with independent random numbers. The 8 solutions thus obtained were then simulated for $n_* = 50000$ days as an additional (more stringent) feasibility test, and each solution was declared feasible or not according to the result of this test, i.e., according to the feasibility of $(\mathrm{SP0}_{n_*})$.

The results appear in Table 2. In this table, n is the number of simulated days for checking feasibility at each step and for the local search at the end of the algorithm (for TS, these n days are split into batches to apply the batch-means method); n_0 is the number of simulated days used for generating the cuts; "min cost" and "median cost" are the minimum and median costs of all solutions (feasible or not) obtained by this method over the 8 independent trials; P^* is the percentage of trials that returned a feasible solution for $(\mathrm{SP0}_{n_*})$; and P_1^* is the percentage that returned a feasible solution with cost within 1% of the best known feasible solution (the lowest-cost feasible solution for $(\mathrm{SP0}_{n_*})$ generated by either algorithm, over all replications and CPU time budgets).

Skill	Agent type
0	0,2,4,6,8,15
1	0,10,12,14,16
2	1,3,5,7,9,17
3	3,11,13,18
4	0,1,4,10,19
5	2,6,7,9,20
6	4,8,11,12,21
7	4,5,9,11,13,14,22
8	1,3,4,5,9,23
9	4,5,8,12,13,24
10	0,4,7,9,11,25
11	3,8,10,13,14,26
12	1,4,6,9,14,27
13	7,8,12,14,28
14	1,5,6,13,29
15	0,4,9,11,30
16	1,5,10,31
17	2,3,12,13,32
18	1,7,11,14,33
19	2,5,7,11,12,13,34

Table 1: Each line of the table lists the agents types that can handle a given call type. The skill set of each agent type can be easily inferred. The numbering in the table is started from 0 instead of 1, for compatibility with the simulation and optimization software.

Algo	n	n_0	min cost	median cost	P_1^*	P^*
CP-LP	300	20	136.2	137.5	50	50
TS	1500		156.1	156.1	0	100

Table 2: Empirical results for the example

Among the 8 solutions found by CP-LP, 4 were declared feasible by the 50000-day simulation, and all of them have a cost of 138.8. For the 4 infeasible solutions, the constraints were violated only by a very small margin: The worst SL in any given period for these 4 solutions was 0.797 (for period 31), 0.797 (for period 31), 0.798 (for period 30), and 0.799 (for period 28), respectively. The cheapest

solution, whose cost is 136.2, was one of the two with worst-case SL of 0.797. In practice, a manager might be willing to use this almost-feasible solution, considering the fact that the center will always experience stochastic variation in the arrival process and the SL in any case. For this reason, it could be useful to report slightly infeasible solutions in general, and not only the feasible ones. All the solutions returned by TS were declared feasible, but they are significantly more expensive, with a cost of 156.1. This shows that large suboptimality gaps with the TS method do occur in realistic call center settings, and not only in artificial examples. We repeated this experiment with a 10-hour CPU budget (n_0 was doubled and n was increased to 400 for CP-LP), and none of the two algorithms found a better solution than with the 5-hour budget.

Table 3 summarizes the empirical optima found by CP-LP and by TS. The agent types are regrouped by cost (number of skills). The table gives the total number of agents of each group (each cost) in the solution. We see that CP-LP selects a larger number of agents than TS, but less expensive ones, whence the lower cost.

agent type	cost	CP-LP	TS
4	1.8	0	1
1,5,9,11,13	1.6	3	32
7,12,14	1.5	15	18
0,2,3,8	1.4	19	33
6,10	1.3	23	23
15,...,34	1.0	55	0
total number		115	107
total cost		138.8	156.1

Table 3: A summary of the best feasible solutions found by CP-LP and by TS

We also made experiments with other variants of this problem, e.g., with a larger variety of shifts or with fewer periods, and also with other (smaller) problems, and the results were similar. For the smaller problems, the gap between CP and TS was generally smaller (this should depend mostly on the structure of the problem more than its size), but TS was always dominated by CP. We also implemented a metaheuristic method based on neighborhood search combined with queueing approximation, along the lines of Avramidis et al. (2006), but we were unable to make it competitive with CP for solving (P0).

CONCLUSION

We have proposed in this paper a simulation-based methodology to optimize agent's scheduling over one day in a multiskill call center. Even though the use of common random numbers reduces the simulation noise (or variance) significantly, there is still randomness in the solution provided by the algorithm, mainly due to the fact that the simulation lengths must be kept short (because the estimation of each subgradient requires simulations at up to thousands of different parameter values). Yet, our approach finds better solutions than with any other method that we know. In practice, one may run the algorithm a few times (e.g., overnight) and retain the best solution found. Future research on this problem include the search for faster ways of estimating the subgradients, refining the algorithm to further reduce the noise in the returned solution, e.g., by improving the way we round non-integer solutions, and extending the technique to simultaneously optimize the scheduling and the routing of calls (via dynamic rules).

ACKNOWLEDGMENTS

This research has been supported by Grants OGP-0110050, OGP38816-05, and CRDPJ-320308 from NSERC-Canada, and a grant from Bell Canada via the Bell University Laboratories, to the second and third authors, and a Canada Research Chair to the third author. The fourth author benefited from the support of the University of Calabria and the Department of Electronics, Informatics and Systems (DEIS), and her thanks also go to professors Pasquale Legato and Roberto Musmanno. The paper was written while the third author was at IRISA, in Rennes, France.

REFERENCES

Atlason J.; Epelman M.A.; and Henderson S.G., 2004. *Call center staffing with simulation and cutting plane methods. Annals of Operations Research*, 127, 333–358.

Avramidis A.N.; Chan W.; and L'Ecuyer P., 2006. *Staffing multi-skill call centers via search methods and a performance approximation.* Submitted.

Avramidis A.N.; Deslauriers A.; and L'Ecuyer P., 2004. *Modeling Daily Arrivals to a Telephone Call Center. Management Science*, 50, no. 7, 896–908.

Bhulai S.; Koole G.; and Pot A., 2007. *Simple methods for shift scheduling in multi-skill call centers.* Manuscript, available at `http://www.cs.vu.nl/~koole/research`.

Brown L.; Gans N.; Mandelbaum A.; Sakov A.; Shen H.; Zeltyn S.; and Zhao L., 2005. *Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. Journal of the American Statistical Association*, 100, 36–50.

Cezik M.T. and L'Ecuyer P., 2007. *Staffing Multiskill Call Centers via Linear Programming and Simulation. Management Science*, 53. To appear.

CRTC, 2000. *Final standards for quality of service indicators for use in telephone company regulation and other related matters.* Canadian Radio-Television and Telecommunications Commission, De-

cision CRTC 2000-24. See http://www.crtc.gc.ca/archive/ENG/Decisions/2000/DT2000-24.htm.

Gans N.; Koole G.; and Mandelbaum A., 2003. *Telephone Call Centers: Tutorial, Review, and Research Prospects. Manufacturing and Service Operations Management*, 5, 79–141.

Halfin S. and Whitt W., 1981. *Heavy-traffic limits for queues with many exponential servers. Operations Research*, 29, 567–588.

Henderson S. and Mason A., 1998. *Rostering by iterating integer programming and simulation.* In *Proceedings of the 1998 Winter Simulation Conference*. vol. 1, 677–683.

Ingolfsson A.; Cabral E.; and Wu X., 2003. *Combining Integer Programming and the Randomization Method to Schedule Employees.* Tech. rep., School of Business, University of Alberta, Edmonton, Alberta, Canada. Preprint.

Keith E.G., 1979. *Operator scheduling. AIIE Transactions*, 11, no. 1, 37–41.

Koole G. and Mandelbaum A., 2002. *Queueing models of call centers: An introduction. Annals of Operations Research*, 113, 41–59.

L'Ecuyer P., 2004. *SSJ: A Java Library for Stochastic Simulation.* Software user's guide, Available at http://www.iro.umontreal.ca/~lecuyer.

L'Ecuyer P.; Simard R.; Chen E.J.; and Kelton W.D., 2002. *An Object-Oriented Random-Number Package with Many Long Streams and Substreams. Operations Research*, 50, no. 6, 1073–1075.

Mehrotra V., 1997. *Ringing Up Big Business. ORMS Today*, 24, no. 4, 18–24.

Pot A.; Bhulai S.; and Koole G., 2007. *A simple staffing method for multi-skill call centers.* Manuscript, available at http://www.cs.vu.nl/~koole/research.

Stolletz R. and Helber S., 2004. *Performance Analysis of an Inbound Call Center with Skill-Based Routing. OR Spectrum*, 26, 331–352.

Thompson G.M., 1997. *Labor staffing and scheduling models for controlling service levels. Naval Research Logistics*, 8, 719–740.

Vogel S., 1994. *A Stochastic Approach to Stability in Stochastic Programming. Journal of Computational and Applied Mathematics*, 56, 65–96.

AUTHOR BIOGRAPHIES

ATHANASSIOS (THANOS) N. AVRAMIDIS is Researcher in the Département d' Informatique et de Recherche Opérationnelle at the Université de Montréal, Canada. He has been on the faculty at Cornell University and a consultant with SABRE Decision Technologies. His primary research interests are Monte Carlo simulation, particularly efficiency improvement, the interface to probability and statistics, and applications in computational finance, call center operations, and transportation. His recent research articles are available on-line from his web page: http://www.iro.umontreal.ca/~avramidi.

MICHEL GENDREAU is Professor of Operations Research at Université de Montréal, Canada. He is also the Director of the Center for Research on Transportation, a center devoted to multi-disciplinary research on transportation and telecommunications networks, since 1999. His main research interests deal with the development of exact and approximate optimization methods for transportation and telecommunications network planning problems, topics on which he has published more than 125 papers. Dr. Gendreau is the Area Editor of "Heuristic Search and Learning" for the INFORMS Journal on Computing and an Associate Editor of Operations Research, Transportation Science, the Journal of Heuristics, RAIRO-Recherche opérationnelle and the Journal of Scheduling. He is Vice-President of the International Federation of Operational Research Societies (IFORS) and Vice-President International of the Institute for Operations Research and the Management Sciences (INFORMS). He received in 2001 the Merit Award of the Canadian Operational Research Society.

PIERRE L'ECUYER is Professor in the Département d'Informatique et de Recherche Opérationnelle, at the Université de Montréal, Canada. He holds the Canada Research Chair in Stochastic Simulation and Optimization. His main research interests are random number generation, quasi-Monte Carlo methods, efficiency improvement via variance reduction, sensitivity analysis and optimization of discrete-event stochastic systems, and discrete-event simulation in general. He is currently Associate/Area Editor for ACM Transactions on Modeling and Computer Simulation, ACM Transactions on Mathematical Software, Statistical Computing, International Transactions in Operational Research, The Open Applied Mathematics Journal, and Cryptography and Communications. He obtained the E. W. R. Steacie fellowship in 1995-97, a Killam fellowship in 2001-03, and he was elected INFORMS Fellow in 2006. His recent research articles are available on-line from his web page: http://www.iro.umontreal.ca/~lecuyer.

ORNELLA PISACANE is currently a PhD Student in Operation Research at the Università della Calabria, Italy. She obtained her first degree in Software Engineering with the maximum vote at the same university in 2004. Her main research areas are concerned with simulation-optimization problems. Her web page is: http://dottorato-ro.deis.unical.it/area_dottorandi/pagedot.jsp?coddott=26. Her email address is: pisacane@hpcc.unical.it.

Modelization and analysis of trajectories in medical simulations

François-Xavier Magaud[1,2]
Claude Dussart[1,3]
Claude Petit[1,2]

[1] European Center for Health Humanitary University Claude Bernard - Lyon 1 - France
[2] Laboratory LIRIS "modelling health decision" CNRS UMR 5205 University Claude Bernard - Lyon - France
[3] Military Hospital Desgenettes - Lyon – France
e-mail : claude.petit@cesh.org

ABSTRACT

This paper proposes a description of a new tool for analysis the trajectories simulated and an application which illustrates the mechanisms in the field of health. The trajectories simulated are treated to obtain coherent bundles of trajectories. They are represented by meta-trajectories. These meta-trajectories are associated to explanatory factors. The interpretation of the results is easy. Several innovative concepts are presented: automatic classification of simulated curves, learning of meta-trajectories and association of explanatory factors. The application relates to the exploitation of the nosocomial infection in a hospital. It shows the capacity of decision-support of this tool.

KEYWORDS

Trajectories, clustering, meta-trajectories, simulations analysis, decision support

INTRODUCTON

This paper presents an original tool for analysis of the simulated trajectories. A trajectory is represented by a curve. It shows the evolution of an indicator for a phenomenon and a period. For influenza, the indicator is the temperature. The trajectory expresses the evolution of the temperature curve for a patient over a period. The application selected relates to the nosocomial infection rate. A group of expert simulates a great number of cases to understand the preventive decisions to take in a hospital. Each simulation provides a trajectory of the weekly rate of infection and the explanatory factors over the 52 weeks period. The model suggested analyzes simulations and gives the explanatory factors, of the principal tendencies. Before the description of this application, the concepts of the tool are described. The first problem relates to the choice of a clustering algorithm. The algorithm selected is presented and commented on. Top-down non hierarchical algorithm is selected. It implies interventions on behalf of the user. It was necessary to design a new algorithm which authorizes the completely automatic classification, without intervention of the user. Then the construction of coherent trajectories bundles is explained. The explanatory capacity calculated on each bundles gives useful information which is supplemented by the corresponding explanatory factors. Each bundle of trajectories is represented by only one trajectory (meta-trajectory) which simplifies the graphic reading.

SEEKING AN ALGORITHM FOR THE SIMULATIONS ANALYSIS

The built model answers clustering problems of individuals according to the evolution in the time of their pathologies. Thus it was necessary to choose the best and adapted clustering method. Three models stood as candidates: Bottom-up hierarchical clustering, non hierarchical method of the k-means and top-down non hierarchical clustering. Several criteria were discussed, search for the number of class automatically, preference of a non hierarchical method.

Bottom-up hierarchical clustering

The technique of bottom-up hierarchical clustering gathers the individuals according to their similarity formalized by a function of dissimilarity (between elements and group of elements). The first stage consists in building a partition of the individuals in singleton. Then closest individuals are gathered. A new partition is obtained. So the stage is reiterated until obtaining a partition of the space of the individuals. The method is hierarchical, that means that once two elements are gathered, they belong to the same class at the following stage. This element appeared to be less good than a non hierarchical method which allows unexpected regroupings. This method was to be modified to integrate an automatic criterion of stop to determine the number of classes.

K-means

K-means method aims at optimizing the inertia intraclass of the partition created. It initially will choose randomly a partition of the whole individuals in Q classes fixed by the user of the method. Then the method transfers individuals from a class in another to increase the explanatory factor. The algorithm stops when no individual allows increasing the explanatory capacity of the built partition. A theorem of convergence exists. This algorithm has three disadvantages. The algorithm always does not provide the same result. The result depends on the random partition built at the beginning. Many times thus should be repeated the algorithm to obtain the best result. The algorithm does not build Q classes systematically. It is possible to see disappearing classes. Finally the k-means algorithm does

not choose automatically the number of classes. It is thus necessary to use a method of groping to obtain the number of classes. It is for these reasons that we did not choose this method of clustering.

Top-down non hierarchical clustering

Top-down non hierarchical clustering appeared to be the most adapted in spite of the automatic absence of model to determine the number of classes. This lack was filled by the contribution of a technique of incremental non hierarchical clustering built by C Petit. This choice thus is based on two elements; addition of an automatic criterion to determine the number of classes and the characteristics related to the non hierarchical models.

ALGORITHM NHD FOR CLUSTERING

Hierarchical clustering produces series of encased classes which define a hierarchy. This method consists in providing a whole of more or less fine partitions, obtained by successive regroupings of parts or division successive of the parts. Non hierarchical clustering differs by not taking into account the concept of hierarchy in the construction of the parts. Clustering is known as top-down because it starts from a class containing all the individuals and disaggregates an element in two parts until obtaining the whole of the isolated individuals. The algorithm of top-down non hierarchical clustering named NHD is defined for quantitative data type. The algorithm is presented in a generic way.

NHD Algorithm

REQUIRES:
E : elements to be classified
D : function of dispersion
q_{max} : number of partitions defines by the user
ENSURES: partitions in q_{max} parts

$P_1 \leftarrow \{E\}$
$i \leftarrow 1$
do while $i \neq q_{max}$
Seek the element P^k_{cardP} of the partition of maximum dispersion
Seek the loose element $\{a_k\}$ of the unit
Create a new partition
$P_{(cardP)+1} \leftarrow [P^1_{cardP}, P^2_{cardP}, \ldots, (P^k_{cardP} - \{a_k\}), \ldots, P^{cardP}_{cardP}, \{a_k\}]$
Transfer the elements towards the new element from the partition in order to decrease dispersion
$i \leftarrow i+1$
end while
RETURN P_{qmax}

It should be noted that the algorithm of top-down non hierarchical clustering will seek a partition of the whole of the individuals E. We must then introduce two concepts: Dispersion is a mathematical measurement making it possible to calculate in a class of elements the difference (distance) between the elements and the center of gravity of their class. The more dispersion is raised and individuals (elements) are divided on a great surface. The loose element of a class is the element which contributes more to its dispersion. The algorithm builds in an iterative way a new partition at each stage, until obtaining the number of partition defines by the user. At each stage, the algorithm seeks the element of the partition which has maximum dispersion in order to split up it in two disjoined elements. The first fragment gathering all the individuals private of the individual bringing the most dispersion to the class defines by the element of the partition. Then a stage of transfer exchanges the elements towards the new class with an aim of decreasing total dispersion. We introduce first the concept of inertia which makes it possible to build the indicator explanatory capacity. We call inertia of A (implied compared to the centre of gravity) the quantity:

$$I(A) = \sum_{i=1}^{cardA} m_i . d^2(a_i, g(A))$$

where $g(A) = \dfrac{\sum_{i=1}^{cardA} m_i . a_i}{m_A}$ is the center of gravity of class A.

Particular case:

If for all $a_i \in E, m_i = \dfrac{1}{cardA}, I(A)$ is called variance of A and is noted V(A).

Inertia relative to a partition

We call intra-groups inertia or intraclass inertia the sum of inertias of elements of P.

$$I_{\text{intra}} = \sum_{j=1}^{cardP} I(P^j)$$

We call interclass inertia, the inertia of the unit $\{g(P^1), \ldots g(P^{cardP})\}$

$$I_{\text{inter}} = \sum_{j=1}^{cardP} m_{P_j} \cdot d^2\left(g(P^j), g(E)\right)$$

The total inertia of the system is constant. It is the sum of intra and inter classes inertias.

$$I(E) = \sum_{i=1}^{n} m_i . d^2\left(a_i, g(E)\right) = I_{Intra} + I_{Inter}$$

If clustering contains card E groups then $I_{\text{intra}} = 0$ and $I_{\text{inter}} = I(E)$. In the case of a partition in only one part of E, we have $I_{\text{intra}} = I(E)$ and $I_{\text{inter}} = 0$. The

concept of inertia makes it possible to introduce the concept explanatory capacity.

$$P_e = \frac{I_{\text{inter}}}{I(E)}$$

This value lies between 0 and 1. A value close to 0 indicates that groups are not very homogeneous and are badly separated. A value close to 1 corresponds to quite separate and compact groups.

Search for the maximum dispersion element of the partition

Definition: Dispersion is an application of the parts of E in the set of the real positive numbers:

(i) $D(A) = 0$ if and only if all the elements of A are confused (they have the same representation),

(ii) Superadditivity : ∀A, B ∈ P(E), A ∩ B=∅, $D(A \cup B) \geq D(A) + D(B)$.

Dispersion of a partition: Let be $P = \{A_1, \ldots, A_q\}$ a partition in q groups. Dispersion of P is called:

$$D(P) = \sum_{i=1}^{q} D(A_i)$$

We propose three dispersions usable by the top-down non hierarchical clustering algorithm.

Inertia :

$$I(A) = \sum_{i=1}^{cardA} m_i . d^2(a_i, g(A))$$

Spreading dispersion :

The spreading dispersion measures the distribution of the individuals in space. Measurement is based at the maximum distance separating all the individuals balanced by the number of element of the class.

$$D_e(A) = card(A).\max\{d(a,b), a, b \in A\}$$

Radiant dispersion :

Radiant dispersion seeks the individual nearest to the class, in term of dissimilarity. A compact and homogeneous class will thus gives a sum weaker than a very wide class, whatever the element on which one calculates the sum.

$$D_r(A) = \min\left\{ \sum_{i=1}^{cardA} \rho(a_i, x), x \in E \right\}$$

where ρ is a dissimilarity function.

Seek the loose element

Definition: We call loose element of a class any element x_A such as $D(A - \{x_A\}) = \min\{D(A - \{x\}), x \in A\}$. It is the element which contributes more to the dispersion. This element is obtained in card E operations.

Stage of transfer

The stage of transfer makes it possible to add new individuals to the new class. It seeks the elements having a strong attraction for the new class. This attraction is modelled by the function of following attraction.

Let be A, B∈P, such as A ∩ B=∅ and x∈A the attraction of A on X is defined by: $Att(A, x) = 0$, Attraction of B to x is defined by :

$$Att(B, x) = D(A) + D(B) - D(A - \{x\}) - D(B \cup \{x\})$$

If Att(B,x) > 0 the dispersion of the partition decreases when x go into B.

The method of transfer is simple. It consists in seeking the elements which decrease dispersion. We exchange the elements until obtaining a partition locally optimal in q classes, such as:

$$\forall \text{ i, j, } \forall x \in A_i$$

$$D(A_i) + D(A_j) \leq D(A_i - \{x\}) + D(A_i \cup \{x\})$$

Stage of transfer

Repeat until no element changes more class

For i = 1 to n **do**

Seek $A_f \in P_{cardP}$ such as

$Att(A_f, a_i) = \max\{Att(A_j, a_i), A_j \in P\}$

if $Att(A_f, a_i) > 0$

then $A_{ai} \leftarrow A_{ai} - \{a_i\}$;

A_{ai} represents the class of membership of a_i

$A_f \leftarrow A_f \cup \{a_i\}$

end if

End for

End repeat

Postulate: An element cannot turn over in a group which contained it during the stage of transfer. The postulate makes it possible to calculate complexity in worst case. We obtain C(n)= $q_{\max}.n^2$.

The k-means algorithm allows that one of the classes has disappeared at the end of the stage of transfer, i.e. they are

empty. The example below illustrates this case. Is it the same for NHD algorithm?

Let be a_1, a_2, a_3, a_4 four aligned points, with $d(a_1, a_2) = 1, d(a_2, a_3) = 10, d(a_3, a_4) = 1$ and it is supposed that the number of class required is three.

Figure 1. Initial stage

If at a given moment, the current partition is like illustrated by figure 1 then it is clear that at the end of a stage of transfer, we will obtain the partition in two classes such as figure 2.

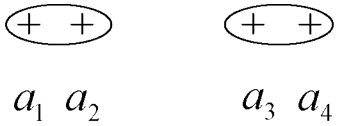


Figure 2. Final Stage

This situation is impossible because the stage represented on figure 1 cannot emerge at the time of the execution of the algorithm of incremental top-down non hierarchical clustering. Indeed this situation is not locally optimal. The fact that the partition is locally optimal makes it possible not to empty a class completely.

Discussion

The total complexity of the algorithm of the clustering in worst of the cases is about $q_{\max}^2 . n^2$. The explanatory capacity increases from one stage to another. That implies the reduction in intraclass inertia. This highlights the improvement of the result at the creation of a new class. A disadvantage of the system is the determination of the number of classes. The traditional approach obliges the user to test the algorithm on several tests, while varying the number of classes. By considering the reduction in intraclass inertia, we can stop the process of clustering: when this reduction is not significant any more, the number of class obtained is satisfactory. The incremental aspect of the algorithm is very practical, because it makes it possible automatically to find the partitions successive for different numbers classes (of the partition in two classes with that in $q_{\max}$ classes).

IMPROVEMENT OF NHD

Criterion of stop

We propose an automatic test of stop. We do not have to stipulate the number of classes of the final partition. The following criterion was proposed. The algorithm stops when

$$\frac{I_{\text{intra}}(P_n) - I_{\text{intra}}(P_{n+1})}{I_{\text{intra}}(P_n)} < d$$

where d is a threshold fixed by the user and I_{intra} the inertia intraclass of the partition. We propose using a batch of tests the value of 0.1 for the threshold d.
This criterion of stop can be written in the following form:

$$I_{\text{intra}}(P_n)(1-d) < I_{\text{intra}}(P_{n+1})$$

The series $I_{\text{intra}}(P_n)$ is decreasing. The criterion of stop thus consists in stopping the algorithm when the function of intraclass inertia decreases slightly. Indeed for d = 0.1 we obtain :

$$1 > \frac{I_{\text{intra}}(P_{n+1})}{I_{\text{intra}}(P_n)} > 0.9$$

Rejection step

A second modification is made to the algorithm. The idea is to introduce a phase of rejection of the loosest individuals, i.e. contributing a too strong share to inertia intraclass and unbalancing the group. Thus we eliminate the elements which have a percentage of the inertia of the class higher to a threshold fixed by the user. The algorithm becomes:

Incremental algorithm of top-down non hierarchical clustering

---

$P = \{E\}$

do while $\frac{I_{\text{intra}}(P_{n+1})}{I_{\text{intra}}(P_n)} < 0.9$

Seek the element P_{cardP}^k of the partition of maximum dispersion

Seek the loose element $\{a_k\}$ of the unit P_{cardP}^k

Create a new partition

$P_{(cardP)+1} \leftarrow [P_{cardP}^1, P_{cardP}^2, \ldots, (P_{cardP}^k - \{a_k\}), \ldots, P_{cardP}^{cardP}, \{a_k\}]$

Transfer the elements towards the new element from the partition in order to decrease dispersion

Rejection of the loosest elements of the last class created

end while

return the partition

Discussion

Modified algorithm NHD is called NHDI, or algorithm top-down non hierarchical incremental. The complexity of algorithm NHDI is limited by n^4. This algorithm is effective because it has several levels of recursivity:

• An incremental level: the number of classes of the partition increases and the groups take shape.
• A level of reallocation: a first coarse partition was created, and the algorithm seeks for each element the adequate class, by minimizing total dispersion.
• It reiterates, as in the k-means model, this second level until the dispersion of the classes is weakest possible (partition locally optimal).

Algorithm NHDI provides excellent clustering's. However we must note that the criterion of stop is based on the weak growth of inertia intra-groups, which can involve the creation of a number of partition higher than what is necessary. It is thus very important to quantify well, the criterion of stop to avoid any overflow.

Application to the trajectories

The mechanism proposed performs a top-down non hierarchical clustering applied to curves. It determines in an incremental way the number of partitions i.e. the number of bundles. It withdraws also the atypical trajectories which could not enter a bundle. Thus it makes it possible to eliminate a certain number of noises. The curves are memorized by using a discretization of time. The curves are then represented by a vector time of T length corresponding to the length of measurement. It should be noted that the step of the discretization is fixed. Thus each point of the vectors corresponds to the same X-coordinate. The whole states of e(i) corresponds to notation Ci of the definite curves. We call bundle a class of curves, or trajectories, gathered by the algorithm. We represent a class by his centre of gravity. Thus we have a graphic tendency for a group of individuals. The built model makes it possible to choose between various dispersions presented previously. These elements are skeletal by the choice of a distance in the space of the curves. It is pointed out that a distance is an application of E^2 in the whole of the positive numbers real such as:
$\forall x \in E, \forall y \in E, d(x,y) = 0 \Leftrightarrow x=y$
$\forall x \in E, \forall y \in E, d(x,y) = d(y,x)$
$\forall x \in E, \forall y \in E, \forall y \in E, d(x,z) \leq d(x,y) + d(y,z)$

Three distances were programmed:
- the euclidean distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{t} (x_i - y_i)^2}$$

- the Tchebychev distance:

: $d(x,y) = \max\{|x_i - y_i|, i = 1,\ldots,t\}$

- the rectangular distance:

$$d(x,y) = \sum_{i=1}^{t} |x_i - y_i|$$

Meta-trajectory representation

A trajectory is a curve. The method of clustering gathers a whole of curves in a new entity called bundle. This bundle corresponds to a meta-trajectory. Indeed, this element is knowledge on knowledge. The representation of the meta-trajectory is obtained by regression. It is calculated by the technique of the center of gravity. A meta-trajectory is thus represented by the trajectory "average" of the individuals composing the meta-trajectory. The curve representing the meta-trajectory is obtained by following calculation:

$$g(A) = \frac{\sum_{i=1}^{k} m_i . a_i}{m_A}$$

It would have been thus possible to carry out a polynomial regression with a polynomial of degree T by using an unspecified base of the polynomials (Legendre, Tchebychev...). The result obtained would have been a function representing the bundle.

NUMERICAL APPLICATION

The built numerical application makes it possible to analyze the simulations provided by the experts. These simulations relate to the prediction of the rate of nosocomial infection at the hospital. An infection is known as nosocomial (IN) if it missed with the admission at the hospital. This criterion is applicable to all the infections. When the situation specifies with the admission is not known, a time of at least 48 hours after the admission (or a time higher than the incubation period when this one is known) is commonly accepted to separate a community infection from an nosocomial infection.

Each expert provides two simulations of the rate of nosocomial infection. The IN rate is the rate between the number of patients infected by a disease nosocomial and numbers of patients admitted at the hospital. A pilot simulation is produced i.e. without investment and a simulation with investment. A simulation is 52 weeks long. It corresponds to one year. The trajectories thus obtained are represented on figure 3.

Before commenting the meta-trajectories, the tool indicates for this analysis a very satisfactory capacity explanatory; about 0.979 what means that the bundles are well formed (clustering of good quality). The algorithm of clustering gives the meta-trajectories which gather in the same bundle the "similar" trajectories. Three meta-trajectories are obtained (figure 4).

The meta-trajectory (blue group 3) shows a continuous growth of the IN rate without any reduction over all the year. The IN rate was maintained on the same level of growth. The yellow meta-trajectory (groupe2) increases gradually since the beginning of the year to reach a peak which appears about March-April. It decreases then gradually to arrive at the initial state about July. The red meta-trajectory (groupe1) increases gradually to reach a peak in the middle of the year. This peak is later than the yellow one. Then the curve decreases and stabilizes about

September-October but remains with the top of the value initial.

The study of the explanatory factors of the three groups shows that group 3 represents the pilot case, i.e. without investment for the hospital complex. If nothing is done, the IN rate increases on average by 9%. The explanatory factors of group 2 are an investment in small equipment and the staff training to the new material. At the time of the catch in hand of the new material the IN rate increases quickly (30% on average) then returns quickly to the normal. Finally group 1 corresponds to a strategy of investment in heavy equipment. A light retreat of the IN rate appears during a first time then an average increase of 20%. At the end of 52 weeks the situation still did not return to the normal. The three meta-trajectories obtained shows legitimacy to apply a correct action carrying to the IN rate. Indeed the pilot curve shows a regular increase of the rate. The corrective actions permit to obtain a IN rate lower than the rate predicted by the pilot group of simulation. This action results in investment.

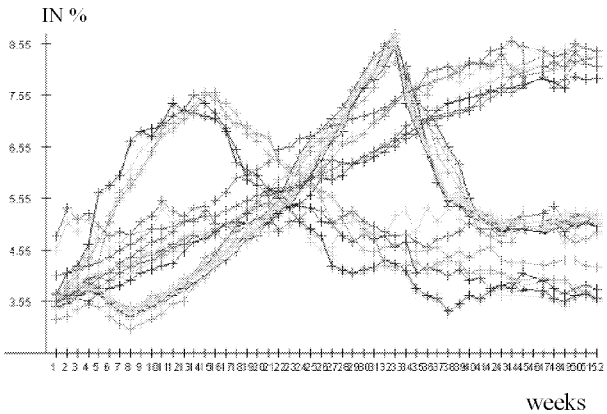


Figure 3. Expert's trajectories

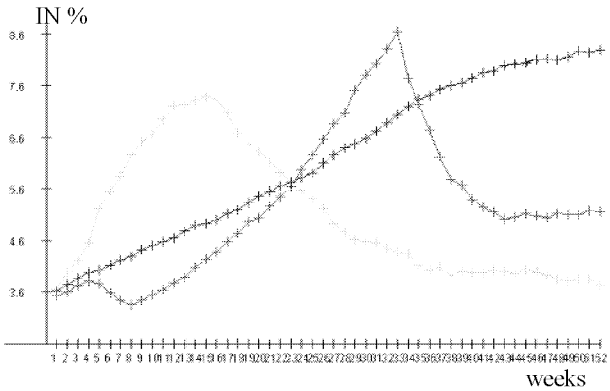


Figure 4. Meta-trajectories

CONCLUSION

The tool for simulations analysis gives to the analyst information for decision support which is very easy to interpret. It makes it possible to express tendencies associated with explanatory factors. In the panoply of the existing tools, it takes a special place. It is appropriate well for the problems which can be represented by curves over a period.

REFERENCES

Arthur D., Vassilvitskii S. "How Slow is the k-means Method?," *Proceedings of the 2006 Symposium on Computational Geometry (SoCG)* 2006

Berkhin P., Survey of Clustering Data Mining Techniques, *Accrue Software*, 2002.

Dussart C and Petit C, *Méta-apprentissage d'expériences, Nouvelles voies en data mining*, Vuibert. 2003.

Tounissoux D. Classification automatique, recherche de partitions. *Cours MASTER SITN*. Université Claude Bernard Lyon 1 - 2001.

Petit C. Méta-trajectoires *ICEIS* Angers, 2003.

Picard P. Classification sur des données hétérogènes. Rapport technique IREMA, Université de la Réunion. 2001

BIOGRAPHY

François-Xavier MAGAUD is a PhD student at University Claude Bernard – Lyon 1 in France. His research topic is about meta-learning. He works on meta-learning strategies.

Claude DUSSART holds a Ph. D. (2003) in medical decision making "meta-learning of distributed experiences" and a Pharm.D. (1998) by the pharmaceutical faculty of the Claude Bernard University (Lyon 1, France). He is researcher in the European Center for Health Humanitary (CESH- University Claude Bernard Lyon 1 France), free teacher at the faculty of pharmacy (Lyon) and works at the military hospital Desgenettes.

Claude PETIT : Habilitation for research director « approach for meta-model », Doctor of philosophy « Generativ modelisation », Doctor of philosophy « Global model for entreprise »
He is research director in the European Center for Health Humanitary (CESH Lyon) and researcher in the laboratory LIRIS "modelling health decision" (University Claude Bernard – Lyon 1 - France)

URBAN AND MOTORWAY TRAFFIC SIMULATION

URBAN TRAFFIC SYSTEMS – PROPOSITION AND APPLICATION OF GENERIC COMPONENTS

Michelle Chabrol
Michel Gourgand
Damien Hostin
Patrice Leclaire
LIMOS CNRS UMR 6158
Université Blaise Pascal, Campus des Cézeaux
63177 Aubière, France
E-mail: {chabrol,gourgand,hostin,leclaire}@isima.fr

KEYWORDS
Urban Traffic System (UTS), methodology, modelling, simulation, software components.

ABSTRACT

Urban Traffic Systems (UTS) are increasingly complex and problems dealing with maintenance or evolution of these systems are increasingly difficult. Simulation allows to evaluate the performances of new structural assumptions and, thus, to discuss the feasibility of these projects. We propose in this paper an implementation of the ASDI-mi modelling methodology dedicated to UTS described by D. Sarramia. New components are proposed and a case study is presented.

INTRODUCTION

Due to the complexity of Urban Traffic Systems, before any effective implementation, what if studies are needed to take into account structural and/or logistical modifications of urban networks. Simulation allows such studies by the evaluation of the performances of the future system. Our work is related to the study of the crossing of an average size town in full change, and more particularly the modification of habits of urban network users in order to unblock the town centre.
Our contribution deals with the scientific validation of assumptions on network installations by means of simulation models. The implementation of these models requires the use of a modelling methodology for analysis, specification, design and implementation.
The paper is structured as follows. In the first part, we present a state of the art on the modelling and simulation environments dedicated to the Urban Traffic Systems (UTS) and the ASDI-mi methodology. This methodology allows to study any kind of complex system, in a multiple and incremental way, in order to produce a decision-making aid environment. We note ASDI-mi UTS its instantiation for our field of study. The second part presents the generic components proposed to supplement the existing generic knowledge model of the methodology. The third part deals with the application of the enhanced generic knowledge model to our study, the unblocking of the town centre of Riom, France.

MODELLING AND SIMULATION ENVIRONMENTS

A modelling environment, represented by a target (figure 1), is constituted by a package of methods and tools divided into layers.

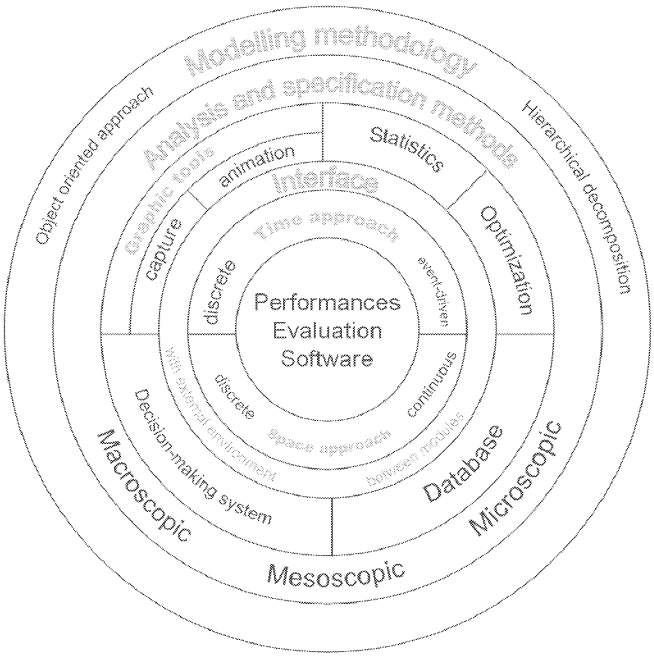


Figure 1: Software target

It includes:
• a **performance evaluation** software based on mathematical theories or on simulation,
• a **graphical** layer for the data capture of the conceptual model, the results exploitation and the animation,
• a **decision-making** layer providing dashboards for data analysis,
• a **data** layer,
• a **statistic and optimization** layer for the exploitation of results given by the models, providing optimisation methods,
• an **analysis and specification** layer containing tools for the description of the structure and the flows of the system,
• a **domain modelling methodology** layer for the construction of simulation models, without necessarily having competences with this respect.

State of the Art

We present some software offers specially developed for the domain of the Urban Traffic Systems, comparing them through the proposed modellings and abstractions.

VISSIM is currently one of the most complete modelling and simulation environment. Included in the PTV Vision (Scherr et al. 2003) pack of the German company PTV, this software, with a microscopic approach, aims at integrating any kind of infrastructure and most common means of transport (Cohen 1990). This model is based on the driver behaviour. Another software of the PTV Vision pack, called VISUM, is dedicated to macroscopic and strategic studies for modelling the traffic on large areas. The extensibility offered by the latter is twofold ; the first one, called vertical extensibility, to model a system from a simple intersection to a national network, the second one, the horizontal extensibility, is used to take into account every means of transport. The system load is modelled thanks to an Origin-Destination matrix or thanks to a temporal diagram of the transportation demand. Besides, the network is composed of traffic areas or of stations and nodes.
VISEM (Fellendorf et al. 1997), the software of the PTV Vision pack, working on mesoscopic level of details, models the load gathering individuals likely to have a homogeneous displacement behaviour and assigns to them an activity chain. These groups of individuals may be employed people, students with or without vehicle. The chains of activities are composed by displacements related to work, studies, consumption or leisure.
DYNASIM (Orcutt et al. 1976; Zedlewski 1990) is proposed by the DYNALOGIC firm, in this third version. This tool is composed of a graphic editor, a microscopic, stochastic and discrete event approached simulator, 2D/3D visualization tools and simulation results analyser. This software allows to gather several assumptions in a single study. An assumption, also called simulation scenario, gathers a flow scenario, a network one, and eventually fire and public transport ones.
URBANSIM (Waddell 2002) is an American project of academic origin, which can be used not only in a strategic way to plan installations to be realized on the system in order to make it more efficient, but also in a microscopical way to study the behaviour of a precise area. This project responds to new expectations and gives the UTS an economic dimension. The aim of the model is to represent the dynamic interaction between urban development and transport including the demographic growth and evolution of geographic areas.
SUMO (Krajzewicz et al. 2002) is a microscopic simulation environment for urban traffic, jointly developed in Germany by DLR/IVF (Institut de Recherche en Trafic du Centre de Recherche Aéronautique) and ZAIK (Koln University) laboratories from tools developed by the two organisms. The movement model, microscopic with continuous space, on time discretised, relies on Stefan Krau's studies (Krau et al. 1997; Krau 1998), the assignement model on Christian Gawron's one (Gawron 1998).

To conclude, we can notice that the modelling and simulation environments VISSUM / VISEM / VISIM allow to approach every level of abstraction. However, this offer is composed of three autonomous pieces of software, so that the coexistence of several abstraction levels in a same simulation is impossible. URBANSIM allows to specify a context for a microscopic simulation from data of microscopic abstraction level. SUMO is a promising environment which provides powerful tools for the definition of a network and the distribution of vehicles on routes. Nevertheless it is based on only one model per function in microscopic abstraction level. As far as the modelling methodology is concerned, the different pieces of software are always focused on the load of the systems. SUMO differs from the others because its integrating in its modelling methodology the definition of the network and the assignements of vehicle flows.

ASDI-mi

The ASDI methodology (Analysis - Specification - Design - Implementation) has been developed in order to offer a software environment for modelling and simulation (Gourgand and Kellert 1992). This methodology gives some elements for the design and/or management of complex systems and recommands the re-use of the knowledge of a domain for the study of the systems which compose it ; a domain can be defined as a class of systems with common technical and functionnal characteristics. The main contribution of this methodology consists in separating the knowledge gathering from the construction of the computer model, so as not to depend on a particular programming language and to promote the reusability of the components as much as possible.
The knowledge model describes the structure and the functioning of the system in a natural or graphic language (UML, Aris for instance), thanks to a decomposition in three subsystems:
• The logical subsystem (LSS): elementary services that the system have to treat,
• The physical subsystem (PSS): physical entities, their geographical repartition and their interconnections,
• The decision subsystem (DSS): management rules of the interactions between LSS and PSS.
This model is then translated in the action model thanks to a mathematical and/or programming language. Several action models (discrete event simulation model, ...) can be obtained from the same knowledge model.
An extension of this methodology, called ASDI-mi, have been proposed (Chabrol et al. 2006) so as to integrate the multiple and incremental aspects in modelling. This extension comes from the following observations:
• In practice, a system never has only one modelling, and, when several types of modelling coexist, each has its own advantages. It is then interesting to preserve various types of modelling, to be able to compare them, or combine them.
• According to the various operational, tactical and/or strategic objectives, several levels of detail are necessary to study a system as a whole. The combined use of different abstractions was often considered but not effectively implemented.

Thus, ASDI-mi allows to deal with the structural and functional complexity of a system step by step, to study it

according to various horizon lengths, to combine various action models, and even various models of different levels of abstraction.

ASDI-mi UTS

ASDI-mi UTS is the instantiation of ASDI-mi methodology for the UTS domain (Sarramia 2002).

Multiple and incremental model of the PSS
The structure of the physical subsystem of ASDI-mi UTS allows to model a part of the network at a macroscopic level and another part at a microscopic level, by combining them. The physical subsystem is the only one subsystem which is incremental and hierarchically arranged. The diagram of classes (figure 2) describes its structure. Multiple modelling is allowed by the delegation of the models of representation of the classes node and edge, macroscopic or microscopic, to the structures they contain. These structures allow the implementation of several models per specialization of the mother class structure while considering of the macroscopic and microscopic levels of detail.

Update of the models of simulation
The update of the simulation models is carried out by browsing the network (structure represented by figure 3). For each macroscopic node, then for each macroscopic edge associated to a given node, the treatment of update is invoked. At the update time of a macroscopic element (node or edge), the first stage consists in checking if a microscopic structure is defined. If it is, the update of the microscopic network is invoked ; otherwise, the macroscopic element is directly updated. ASDI-mi UTS, from its incremental structure, makes it possible to juggle between microscopic and macroscopic descriptions within the same representation of a system.

Except ASDI-mi UTS, there is no other software offer which approaches in a general way the modelling of network (PSS), or its management (DSS) or only allows to implement models different from those when the software is validated.

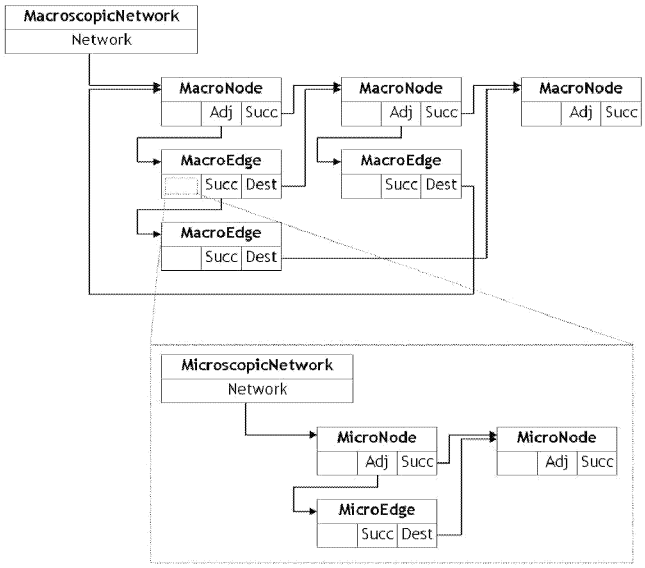


Fig. 3. Graph Structure

GENERIC COMPONENTS

This part deals with the design of the evolution of ASDI-mi UTS. The simulation core is not questioned, and the evolution of the software respects the ASDI-mi methodology which made it possible to model the domain. We will describe the contributions brought to the library of generic components used for the modelling of the urban traffic systems and which were necessary for our case study.

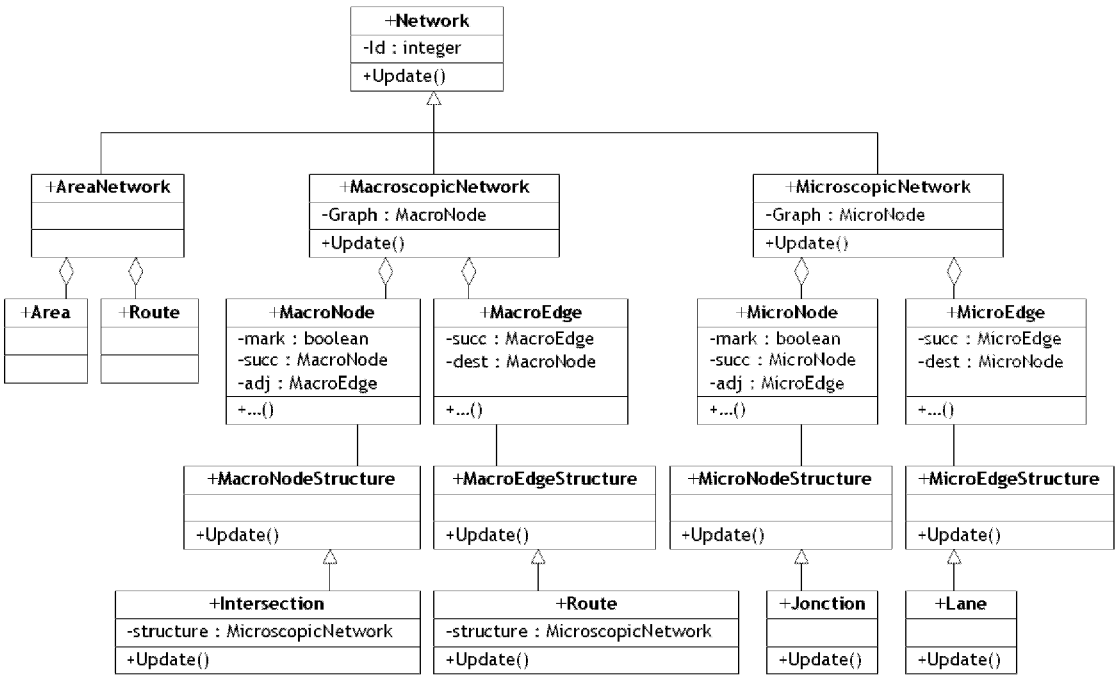


Figure 2: PSS Classes Diagram

Give Way Component

The work of D. Sarramia consists in studying the microscopic junction for the "give way". The current component cannot treat all types of crossroads, the number of lanes to be checked being limited.

Principle

In this study, we discuss a "give way" type junction and we consider all the possible configurations of lanes. The rinciple relies on a list of rules of checking of lanes. These rules take into account the destination of the vehicle which uses a "give way". For instance, if we consider the crossroads represented in figure 4, the vehicle located at the "give way" (end of lane 11) can choose two destinations: turn left (lane 8) or turn right (lane 9). The safety requirements to be applied in both cases are not the same. If the driver turns right, he will have to check if there is no vehicle coming from left (lane 4) and if the road to be crossed (5) is clear. If the driver turns left, he will also have to make sure that no vehicle comes from the right (lane 3) and that the lanes (2 and 7) to be crossed are clear.

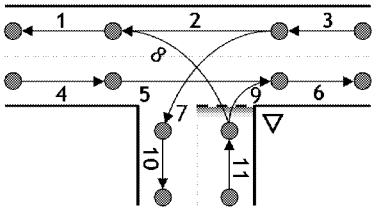


Figure 4: "Give way" Case

We can create a list of rules based on these safety requirements. A rule is a couple of consecutive lanes. On the first one we have to check the presence of a vehicle according to a safety distance and on the second one we have to check the presence of a vehicle over all the length of the lane. This allows to model the fact that a vehicle goes ahead if the crossroads area is clear and if nobody is about to enter it. In the case of figure 4, we have a list of three rules. The distribution of the rules according to the destination of the vehicle is presented in table 1.

Table 1: "Give way" Parameters Example

Rules indexes	Lane to be checked	Clear lane ?
0	4	5
1	3	2
2	3	7

Destination	Rules indexes applied
9	0
8	0,1,2

Some configurations can carry out to deadlock situations. The resolution of these sorts of problems is developed in the following part.

Resolution of some deadlock situations

Figure 5 presents crossroads where a deadlock can occur. When two vehicles (vehicle i on lane i, i ∈ {1;2}) arrive on the crossroads, if the vehicle i wants to turn right (lane i + 2), it will have priority and if the lanes to be checked are clear, it can go ahead. Deadlock situations can occur when each vehicle i wants to turn left onto lane 5−i. Each vehicle will then wait until the other leaves its lane, which cannot happen since they expect each other to clear their respective lane ; this is a deadlock situation.

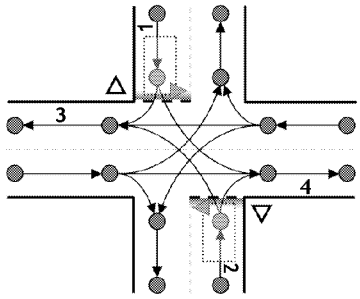


Figure 5: Crossroads with "Give Way" Signs

In practice, as the trajectories do not cross, the two vehicles would have gone ahead. The indicators would have informed each driver of the intention of the vehicle opposite. The considered solution is to take into account the destination of the vehicle opposite. The presence of the vehicle is ignored if it goes onto a lane that is not in the checking rule of the first one. Thus when vehicle 1 checks the presence of vehicle 2 and its destination, it can go on if the destination of vehicle 2 is not lane 4.

A borderline case appears in crossroads with multiple priorities including more than two vehicles having identical dates of arrival ; for example, crossroads in cross shape without signs where 4 vehicles arrive at the same time. The strict respect of the highway code prohibits the passage to all the vehicles ; in practice a person must decide to advance to solve the situation. In a simulation, a solution consists in recursively building a list of vehicles which can block the passage to another one. When we elaborate this list, the detection of a double indicates a deadlock situation and the double indicates which vehicle must go ahead and thus solve the situation.

Give way and light signals

Figure 6 illustrates crossroads presenting a case where it is important to know the colour of the light signal of the lane that a vehicle wants to cross. We denote by i the vehicle located on lane i, i ∈ {1, 2, .., 6}. When the light is green on lanes 1, 2, 3 and 4, the vehicles located on these lanes can circulate. Vehicles 1 and 3 can go straight on or turn right. Vehicles 2 and 4 must wait until lanes 1 and 3 are cleared. However, when the light is red, lanes 1 and 3 may not be cleared. At this stage, vehicles 2 and 4 will not move, and vehicles 5 and 6 will be blocked.

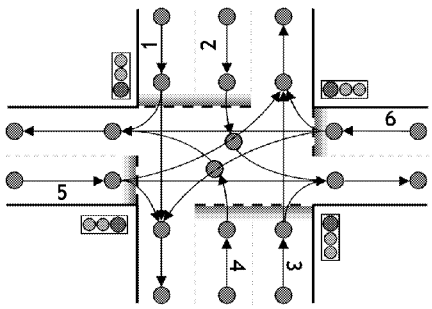


Figure 6: "Give Way" and Light Signals

The problem is solved if we take into account the colour of light. The component "give way" makes sure that there are light signals at the end of the lane to be crossed. If there are, and if the light is red, the vehicles of this lane are ignored.
The component "give way" has been improved, which revealed deadlock situations. Solutions have been suggested to avoid these problems. Another element of the urban system must be considered: the parking areas, as detailled in the following part.

Parking Area Component

Their design respects the framework introduced into ASDI-mi UTS (figure 7). Every area has a capacity. The management of the vehicles in this area depends on its specialization.

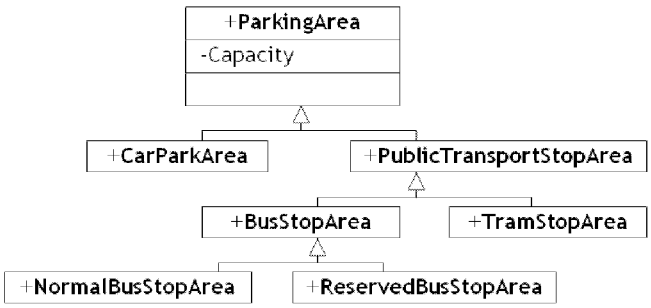


Figure 7: Parking Area Specialization

Carpark
The specialization in Carpark simulates a traditional carpark, with the following assumptions:
• an entry and an exit,
• the parking does not block circulation,
• the distance covered by all a vehicle to go to a place and to leave the carpark is an average of all the possible distances,
• the waiting time is an average of the waiting times usually observed.
These assumptions allow to assign to the Carpark class the only "waiting time" attribute. The attribute "capacity" is obtained by the heritage of the Parking Area class, and the attribute "length" by the heritage of a microscopic edge. If the occupation of the carpark allows it, the vehicle parks and waits. This action means that the vehicle object is put in a queue and that it becomes inactive until the end of its waiting time. This queue is of type FIFO because all the vehicles wait for the same time. The first vehicle ready to leave is the first one which is put in the file and whose waiting time is over. Consequently, the algorithm tests if the vehicle can advance as if it were at the end of the lane representing the carpark. If the test is validated, the vehicle is put out of the queue, it leaves the carpark passing onto the following lane. In case of a complete occupation of the carpark, the vehicles cannot be put in the queue any more. The vehicles then wait at the entry of the carpark for a free parking lot. This makes it possible to highlight cases of saturation.
An extension can consist in including users' behaviours and randomness on the parking times.

Bus Stop on non-dedicated lane
Specialization in Bus Stop models the stops of the buses on the roadway, which blocks the traffic during stop time. The selected approach is similar to the carparks' one, as far as the attributes are modelled. The bus stop class inherits in the same way the capacity of the parking area and length of the lane. The bus stop area is simulated by a null speed during a waiting time, to allow the users to get out of the bus at the end of the lane or behind the last bus on standby. If the way allows n buses to stop, a potential n + 1-th bus will have to wait until the last bus clears the area to be able to enter the lane, park and let its passengers get off. The certainty that only the buses will use this stop has been carried out beforehand thanks to a filter on the types of vehicle.

Router Component

To manage the flow distribution in a more advanced way, a component router was added to ASDI-mi UTS. This component is part of the logical subsystem because it makes it possible to control the flows of the vehicles. Its principle is simple: when a vehicle enters on a lane, a route is affected to it. During its evolution through the network, the vehicle will respect this route. This component allows to correctly model the passage of a complex crossroads or to apply the routes defined by Origin-Destination matrices.

Functionning
This component, made up of a list of lanes and their probability of application, is assigned to a lane. When a vehicle merges onto a lane with a router, one of the routes is affected to it. Only the lanes located at junctions which have several exit points are useful.

For the roundabout of figure 8, a router placed on lane 1 provides three routes with for example 10% of chance to take the first exit, 80% the second one and 10% the third one:
• route 1 (10%) : {7}
• route 2 (80%) : {9, 11}
• route 3 (10%) : {9, 13, 15}
The attribution of the route to a vehicule occurs during the updating process of the simulation.

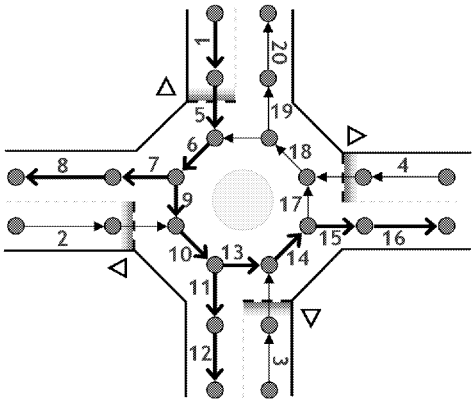


Figure 8: Roundabout with Router

Extensions
The distribution offered by the router has to be defined manually for each lane of each router. However it is possible

to create a tool making it possible to calculate its routes from various pieces of information like the departure and arrival points and possibly some intermediate stages. An Bellman-Ford-like algorithm is used to solve this problem. For a more intuitive capture, a graphic interface with a representation of the network can be used to initialize the lists of edges of the routes.

CASE STUDY

This part details the application of the modelling and simulation environment ASDI-mi UTS to the study of the town of Riom, France, realized in collaboration with the SORMEA (Société de Réalisation de Mesures d'Etudes et d'Analyses).

Analysis

In spite of the construction of a by-pass (figure 9), a few years ago, to divert the flow of vehicles crossing this city, most users continued to cross the town centre. To solve this problem, it is planned to create an obstruction, so that the time necessary to go through the town by crossing the town centre should be more important than presently, thanks to a reduction of the number of lanes and a speed limitation. To show that these installations increase significantly the time of the course by the town centre, we will proceed in two stages ; after modelling the various principal and secondary axes necessary to our study (approximately 200 macroscopic elements, detailed in microscopic graphs).

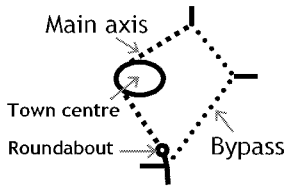


Figure 9: Study Case of Riom, France

The first stage consists in the simulation of the current situation. This simulation makes it possible on the one hand to make sure that the model of simulation is valid by checking that the flows and the global behaviour is conform to reality. On the other hand, it constitutes a comparison point.

The second stage consists in simulating the system with the same load, but with different installations. The microscopic level of details is selected in order to be able to calculate times of course of each vehicle on the two principal routes. The load of the network is modelled from countings raised on the various roads necessary to the simulation. The selected performance criteria are the times of the courses and the length of the queues. The two supervised courses are the axis which goes through the town centre and the by-pass, from the south to the north.

Evaluation

The results obtained by the simulation of the current state are interesting. The model of action reproduces the strong density of vehicles in the town centre and confirms that the time of the course through the by-pass is twice shorter than the time of the course through the town centre. The results of the simulation after the new installations confirm the existence of waiting times. The reduction of the number of lanes and of the speed contributes to increase of 7% on average the time of the course through the town centre. As far as the course through the by-pass, there is no major modification in the results which remain stable. The queues at the roundabout are more frequent and longer. With regard to the calculation performances, a 30-minute simulation of traffic is carried out in approximately 40 seconds on Pentium IV 2.8Ghz with 512Mo of read-write memory.

CONCLUSION

The purpose of the study was to make sure that, following modifications of installations, obstructions would be formed to the access to the town centre and that the time of the course by the town centre would be longer. The difference between the two simulations - before and after the installations - highlights the consequences on the traffic flow that installations will have if the users do not change their habits.

To realize this study, we improved the model of knowledge of methodology ASDI-mi UTS. New components were added: give way, various types of carpark and a new system of vehicles routing. They were validated on a real case but are reusable for the study of any type of Urban Traffic System.

REFERENCES

Chabrol M., D. Sarramia and N. Tchernev. 2006. "Urban traffic systems modelling methodology". *International Journal of Production Enconomics*, No. 99, 156–176.

Cohen S. 1990. "Ingénierie du trafic routier". *Presses Ponts et Chaussées*.

Fellendorf M., T. Haupt, U. Heidl and W. Scherr. 1997. "VISEM – An Activity Chain Based Trafic Demand Model". *Product Information*.

Gawron C. 1998. "Simulation-Based Traffic Assignment", *Inaugural Dissertation*.

Gourgand M. and P. Kellert. 1992. "An object-oriented methodology formanufacturing system modelling". *Proceedings of the 1992 Summer Computer Simulation Conference*, Reno, Nevada, 1123–1128.

Krajzewicz D., G. Hertkorn, C. Rössel and P. Wagner. 2002. "SUMO (Simulation of Urban MObility); An open source traffic simulation". Proceedings of the 4th Middle East Symposium on Simulation and Modelling (MESM2002), SCS European Publishing House, 183-187.

Krau S., P. Wagner and C. Gawron. 1997. "Metastable States in a Microscopic Model of Traffic Flow". *Physical Review E*, No. 55, 55–97.

Krau S. 1998. "Microscopic Modeling of Traffic Flow: Investigation of Collision Free Vehicle Dynamics". *Hauptabteilung Mobilitt und Systemtechnik des DLR Kln, ISSN 1434-8454*.

Sarramia D. 2002. "ASCI-mi : une méthodologie de modélisation multiple et incrémentielle. Application aux systèmes de trafic urbain". Ph-D Thesis.

Scherr W., A. Dick and T. Bauer. 2003. "An Integrated Model for Planning and Traffic Engineering". *9th TRB Planning Methods Applications Conference*.

Waddell P. 2002. "UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning". *Journal of the American Planning Association*, Vol. 68 No. 3 (Summer 2002), 297-314.

NEURAL NETWORK MICROSIMULATION OF MOTORWAY TRAFFIC

Abs Dumbuya, George Lunt, Jill Weekley, Anna Booth and Andrew Lewis
TRL, Crowthorne House, Nine Mile Ride, Wokingham, Berkshire, RG40 3GA, UK,
Email: adumbuya@trl.co.uk

KEYWORDS
AI-supported simulation, neural network, behavioural science, psychology, traffic microsimulation

ABSTRACT

This paper introduces new techniques for integrating re-identified Individual Vehicle Data (IVD) into an Artificial Neural Network (ANN) based traffic micro-simulation framework. Re-identified IVD is detailed microscopic traffic data describing thousands of individual vehicle trajectories passing a number of consecutive MIDAS (Motorway Incident Detection and Automatic Signalling) loops. Over 50,000 individual vehicle data collected from the UK motorway are used to train an ANN. Artificial neural network models generally use a mathematical model for information processing with a functional architecture that resembles the neuron structure of the human brain. These models are capable of learning from training examples and demonstrating learned behaviour in unseen situations, hence leading to more realistic behaviour. These concepts and techniques have wider implications for traffic microsimulation. Microsimulation plays an important role in traffic and transport research. In particular it can be used to appraise new transport schemes, help predict short/medium term traffic conditions (e.g. flow breakdown from traffic behaviour), improve the analyses (or assessment of the impact) of traffic flow and its characteristics. An animation visualisation tool is also described to integrate the IVD/ANN to aid interactive manipulation of the simulation and test the qualitative validity of the work. The results from the trained neural network model demonstrate learned behaviour by mimicking motorway traffic behaviour, e.g. car following.

INTRODUCTION

The UK Highways Agency (HA) has developed an automatic speed-control environment to protect slow-moving traffic. At its core is the Motorway Incident Detection and Automatic Signalling (MIDAS) system which monitors current traffic conditions, and sets signals according to pre-defined rules. The MIDAS detection system consists of inductance loop pairs typically every 500m down the length of the motorway. These loops are connected to outstations which contain signal processing and communications equipment. The inductance loop pairs respond to the magnetic field of passing vehicles, and the system is configured such that the following statistics of individual vehicles are measured:

- Time of day (timestamp)
- Speed
- Length
- Headway
- Lane.

Although the roadside control units (outstations) do not store the full details of individual vehicles, it is possible to intercept the individual vehicle data (IVD) before it is aggregated into one minute averages. This is currently achieved by physically connecting a laptop computer to a MIDAS outstation inside a roadside cabinet. Laptops are typically left for a few days recording the IVD.

Individual Vehicle Re-identification is the process of matching data from sensor systems at different locations so as to track an individual vehicle down the highway. Perhaps the most accurate method of re-identification is automatic number plate recognition (ANPR) (Frith and Pearce, 2002). However, another approach pioneered by Coifman and collaborators (Coifman, 1998; Coifman and Cassidy, 2002), is that re-identification may also be accomplished by the better analysis of data from existing inductance loop systems. In the past, the accuracy and proportion of vehicles that could be re-identified in this way has been quite limited. More recent work (Lunt and Wilson 2003; Lunt 2005 and Lunt 2006) discussed early designs and progress that has been made to produce hundreds of thousands of individual vehicle data.

The design and development of realistic neural network microsimulation models using such a unique data source offers the opportunity to produce credible models which can be robustly validated. Traditional development of traffic micro-simulation models has predominately focused on building the model from concept and then attempting to find data to validate the model (the top-down approach). This approach highlights the general criticism of microsimulation; that is, its lack of flexibility in validating the accuracy of the model due to the requirements for large datasets. New work (Dumbuya 2006; Dumbuya et al. 2005; Dumbuya et al. 2002) has attempted to address these criticisms by developing traffic simulation models, with driver models that are capable of, for example, perception (Dumbuya and Wood 2003) and learning (Dumbuya et al. 2006).

In Dumbuya et al. (2006), modelling concepts and techniques using ANN for improving behavioural intelligence and realism in driving simulation scenarios were introduced. The paper discussed the development of Neural Driver Agents capable of learning and successfully replicating human driving behaviour based on data collected from the TRL car simulator. The work reported here is an extension of this previous work. The primary aim of the work reported in this paper is to propose an approach that is significantly different from the traditional mechanistic or deterministic (e.g. rule-based) approaches. The approach proposed uses a novel traffic framework

developed from bottom-up and integrated with IVD and driver behaviour models that are based on Artificial Intelligence (AI) techniques (e.g. Artificial Neural Networks, ANN). The IVD is used to train an ANN and used to demonstrate the capability of the integrated IVD/ANN in learned behaviour by mimicking motorway traffic behaviour, e.g. car following.

NEURAL NETWORK MICROSIMULATION

The integrated traffic microsimulation framework

A summary of the framework is shown as packages in Figure 1. The traffic framework was designed in Unified Modelling Language (UML) notation, version 1.3 (The Object Management Group). The packages contain software elements or collection of classes of the framework that are explained below.

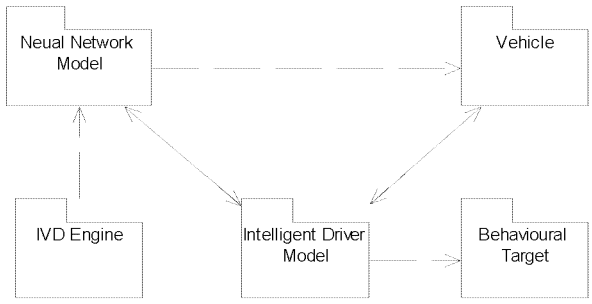


Figure 1: Package Diagram of Framework

Neural Network Model: This is an instance of the Intelligent Driver Model (IDM) class from which the NNM inherits some of its attributes and functionalities. The NNM class is responsible for estimating speed and headways using different configurations e.g. functional approximation for car following.

Vehicle: this package currently contains a simple class holding the attributes (e.g. speed, brake) and functionalities (update vehicle, update headway) of the vehicle model. It is associated with the IDM class. In principle the vehicle-driver classes are considered as a single unit. However, for the purpose of design, it is beneficial to separate the two so that each unit can be explicitly modelled in detail rather than the combined unit.

IVD Engine: This package holds the class for re-identified individual vehicle data which is used to train the neural network model. It is saved in the simulation model as a Dynamic Link Library (DLL).

Behavioural Target: The class within this package aims to identify two typical motorway traffic behavioural targets - car following and lane changing. This class is not yet fully implemented; however, the output from the simulation model is intended to produce car following behaviour.

Neural Network Model (NNM)

This section briefly describes the NNM. Figure 2 shows a typical multilayer architecture of the NNM. The concept of the model is based on Artificial Intelligence (AI) techniques, e.g. ANN, which, at a minimum level, aims to model the system such that it can exhibit human-like properties; for example, learning, knowledge, reasoning and decision making. The NNM is based on a supervised learning paradigm, with a *backpropagation* training algorithm. This means the inputs are propagated through the hidden layers and output layer. The error (or mismatch) between the output and the pre-specified desired output is minimised by using the *gradient descent* rule (which essentially attempts to avoid the local minimum of a function by taking steps proportional to the negative of the gradient of the function at the current point). This allows the errors to be signalled backwards from output to input nodes until the error approaches zero. In other words, the network learns by adapting interconnecting weights.

Input and output parameters of the network are defined below (with the subscripts indicating the time intervals). Note that these parameters have been carefully selected to allow straightforward interfacing with the IVD engine.

- Previous speed, $v_{t-4}, v_{t-3}, v_{t-2}, v_{t-1}, v_0$
- Headway from lead vehicle $hw_{t-4}, hw_{t-3}, hw_{t-3}, hw_{t-2}, hw_{t-1}, hw_0$
- New speed v_{t+1}
- New (time) headway hw_{t+1}

Using these principles, a generalised expression can be derived for the NNM architecture, using gradient descent to approximate the desired output values for new speed and headway. The mathematical derivation of the NNM is based on (Gurney, 1997). The rest of this section notes the typical algorithms implemented for training and testing the NNM.

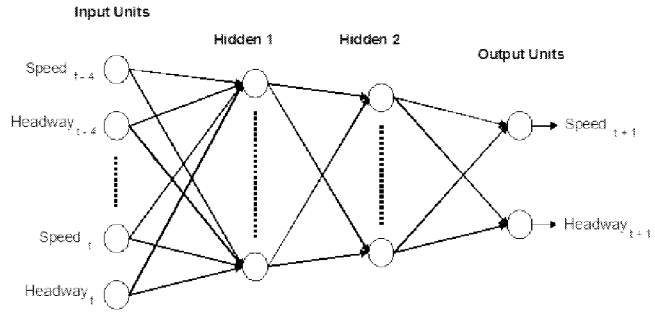


Figure 2: Neural Network Architecture

Training algorithm - to train the network, the following steps are performed:

1. *Feed the network with a set of inputs*
2. *Propagate these inputs through the network layers (forward pass)*
3. *Set desired outputs for these inputs*
4. *Let the network compute the error, propagate it backwards through the layers and adjust the link's weights to minimise the error (backward pass)*

Testing algorithm - after the network is trained, it can be tested with inputs that were not part of the training set and the

results observed. The algorithm for testing the network may be summarised as:

1. *Set the inputs*
2. *Propagate the inputs through the network (forward pass)*
3. *Get the output*

Four measures can be used to evaluate testing:

1. *Training time - overtraining may lead to poor generalisation*
2. *Error tolerance - deliberately incorporating noise within training data*
3. *Accuracy - measure by calculating error for the testing data points*
4. *Measures for validation (e.g. compare NNM with Gipps model, Gipps 1981) for headway and speed estimation*

The implementation section of this paper shows how these procedures are implemented in a commercial package.

Processing Individual Vehicle Data (IVD) to train the NNM

Several techniques were employed to extract and process the data from the IVD for training the NNM to analyse traffic behaviour e.g. car following.

Extracting raw data: The raw MIDAS data was extracted from six loops. For each of the six loop sites there is a record of each vehicle which passes it and its speed, length, lane, headway and the time at that point. For each vehicle record at each loop site, the length and the speed values were used to calculate the occupancy of the vehicle in tenths of a second. The first task was therefore to re-identify each vehicle which passes the first loop at each subsequent loop site and create a record of the change in speed, lane and headway of the vehicle over the whole stretch of motorway. The matching and re-identification of vehicle data was performed using a re-identification algorithm developed in an earlier project; for further details see (Lunt 2006). The algorithm was run on data spanning a time period of about two days in the centre of the period covered by the MIDAS data.

The algorithm was run to obtain vehicle matches between each pair of consecutive loops, and then the datasets were combined to produce a file with sets of six ID numbers, each representing a single vehicle crossing six loop sites. Each vehicle was also given a unique ID number. Each line of this matchings file had the format:

Unique veh id, loop 1 id, loop 2 id, loop 3 id, loop 4 id, loop 5 id, loop 6 id.

Vehicle re-identification and interpolation: The second task was to format these vehicle records for use in the neural network. This task was undertaken by writing a data processing program in MATLAB. The vehicle data at this stage in the process are evenly spaced by distance, i.e. information is given for each loop site. The neural network requires data spaced in equal time and therefore the next stage was to interpolate the data.

Figure 3 shows the data procedure to extract data to train the NNM for car following behaviour. Each vehicle record has a number of variables associated with it, the values of which are known at each loop site. For the car following model the variables are speed and headway and illustrated below using a general speed variable.

The interpolation program reads in the speed values at each loop site and interpolates to provide speed values at fixed temporal points. The input parameters are:

- Number of points (times) required by the neural network (not including the test "predicted" point);
- Timestep between these points;
- 'Prediction delay'- length of time between the last point provided to the neural network and the test predicted point values.

For the car following model the test predicted point values will be the time, speed and headway values of the vehicle at loop 6.

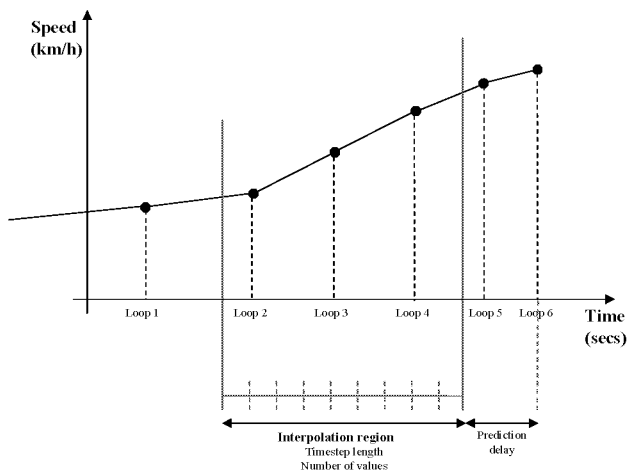


Figure 3: Interpolation Parameters

The interpolation region can be chosen to lie anywhere on the graph. If some or all of the required timepoints occur earlier than Loop 1, then interpolation can be made backwards to approximate these values. The interpolation region also does not need to stretch from Loop 1 to Loop 6; indeed this is not possible since different vehicles take very different lengths of time to cover this distance.

The interpolation program creates a text file for input into the neural network. Each line of this file represents a vehicle as before.

For the car following model each line has the format:

Timestep 1 speed, timestep 1 headway, timestep 2 speed, … … … … …, prediction speed, prediction headway.

Verification of data processing procedures: Figures 4-5 show the comparisons between original and interpolated data used to train the car following neural network model. Each graph shows the speed data against time for a typical

vehicle. Each graph shows the original data from the six loop sites ('original loop data') and the data obtained from the interpolation program ('interpolated data') ready for input into the neural network. Note that the data points for the original loop data are regular in space not time, and the data points for the interpolation data are spaced according to the input timestep value.

Only the interpolated data values for the required number of outputs are plotted on the graph; the neural network additionally requires a desired or predicted value as an input. For the car following model this is the speed value at loop site 6 and can be seen on the graph as the upper extreme of the loop data graph.

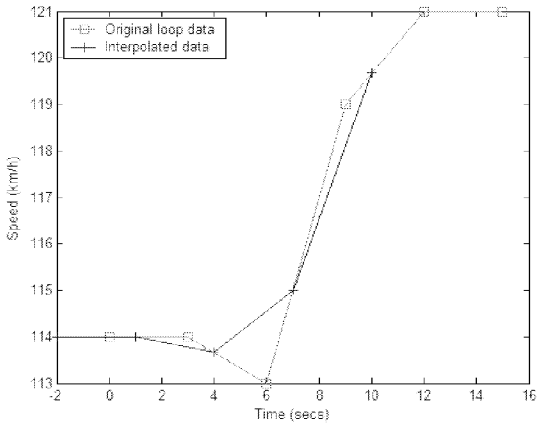


Figure 4: Car Following, Example Vehicle A, Input Parameters: Delay = 5 secs, Timestep = 3 secs, Number of Output Values = 5

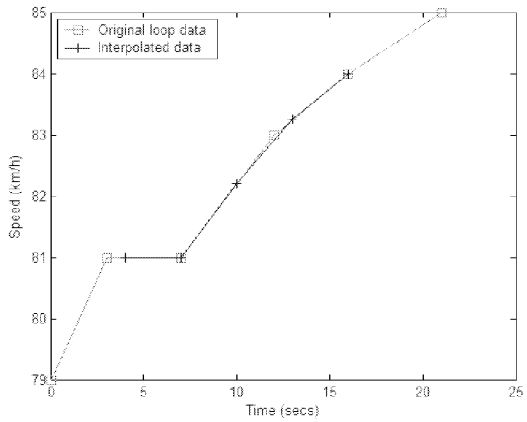


Figure 5: Car Following, Example Vehicle B, Input Parameters: Delay = 5 secs, Timestep = 3 secs, Number of Output values = 5

IMPLEMENTATION OF NNM

Car following

The processed data served as input on which to train the neural network. This section demonstrates a neural network model for car following using the re-identified IVD. For car following a functional approximation (estimate underlying relationship between input-output data) configuration with sigmoid transfer function, was used to predict new speed at 3 seconds in the future. The dataset was processed for training and testing following the procedures described in NNM section. This dataset represented 80% of the original IVD data. The remaining 20% of the data formed part of the validation data: this data was not used in the model development but simply to confirm and validate the initial analysis.

The model was implemented in NeuroSolutions software package from NeuroDimension (www.nd.com). A typical screenshot of the software is shown in Figure 6. In developing the model in NeuroSolutions typical parameters were configured for car following. The parameters were chosen after several 'trial and error' and the use of Genetic Algorithm (GA) to determine the optimal configurations that allowed quick convergence of the network.

The summary of NNM input parameters are:

- Number of layers - 4 layers: 1 input, 2 hidden and 1 output
- Number of processing elements per layer - Input layer: 10, Hidden layer 1: 36, Hidden layer 2: 11 and Output layer: 2.
- Type of transfer function - sigmoid
- Training algorithm - Supervised training using backpropagation with gradient decent performed on error and momentum to speed up learning
- A Genetic Algorithm (GA) method was used to optimise the number of units in the hidden layers, 'momentum value' and 'learning rates'
- Number of epochs (period of time) - typically 3000

The main clipboard of Figure 6(i) shows an example of the car following neural network implementation. The bottom left hand-window of Figure 6(iv) shows the learning curve of the network. The lines in this window correspond to the error of the cross validation set (top line) and the error of the training set (bottom line). The neural driver was trained using the supervised learning paradigm (sigmoid function Figure 6(iii)). It is noticeable that the learning curves approach zero with the cross validation error initially falling with the training error. This indicates successful training of the network.

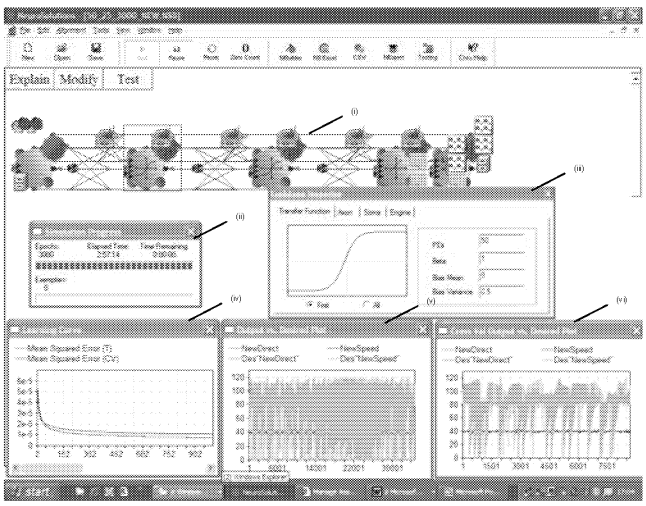


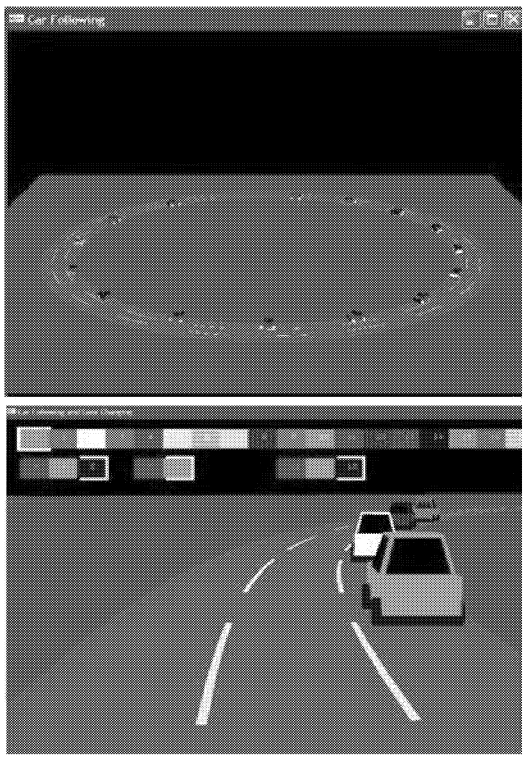
Figure 6: NeuroSolutions Software, Showing Learning Curve at 3000 Epochs

Simulation and animated visualisation

A simulation and visualisation module was developed to integrate both the Neural Network implemented in NeuroSolutions and the Gipps (1981) model (see results section). The integration of the trained network within NeuroSolution is achieved via C++ implementation. The NeuroSolutions package provides a Custom Solution Wizard which is an optional add-in product that takes a neural network, designed within NeuroSolutions, and transforms it into a dynamic link library (DLL). For example, this DLL can then be embedded into C++, Visual Basic, Excel, Access or Internet applications. In the work reported here the trained neural network DLL was attached to the simulation model.

The integrated framework models n vehicles on a circular road of length L, with 3 available lanes. At the start of the simulation loop, the headways for each of the surrounding vehicles are calculated and sent to the car following neural network DLL to update the weights. In addition, a Gipps model is also updated with the calculated headways. The simulation then runs either the car following neural network or the Gipps model version to predict the new speed for the vehicle. This speed is used to move the vehicle.

To aid visualisation of driver/vehicle behaviour, a graphical front-end developed using OpenGL has been added to the simulation model, Figure 7. The visualisation was initially developed as a simple bird's eye view of the vehicle movement. The graphical front-end was later enhanced to add capability to render views from each driver's perspective to demonstrate detailed driver interactions and data storage and analysis, to enable rapid post-processing of simulated behaviours.



(i)

(ii)

Figure 7: Animated Visualisation of Car Following: (i) Bird's Eye View (ii) Driver's View

SIMULATION RESULTS

Behavioural analysis

Following the integration of the trained NNM with the simulation model a number of simulation runs were implemented on a circular track to test different parts of the neural network model. Figure 8 shows the vehicle trajectories for 5 vehicles controlled by the neural network on a 1 km circular road, initially spaced 100 metres apart and with a starting velocity of 60 kph. Figure 9 shows the speed profiles for the first 70 seconds of the simulation.

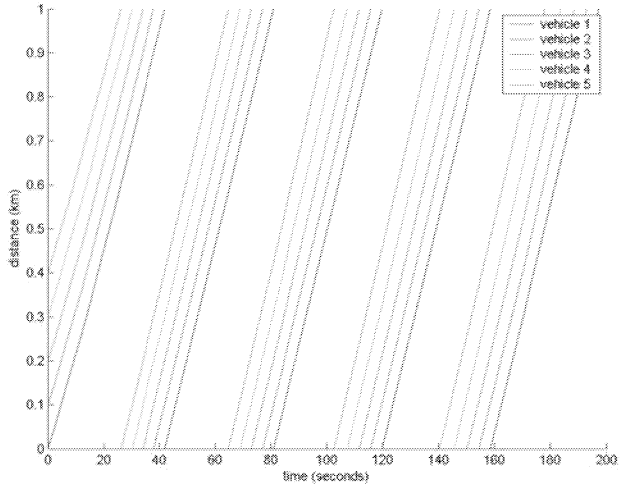


Figure 8: Vehicle Trajectories from Simulation Run

Figure 9 shows some oscillations which are most prevalent in vehicle 5 (the lead vehicle) compared with the other vehicles which have a smaller headway (as shown in Figure 8). The steady speeds for all the vehicles are about 95 kph.

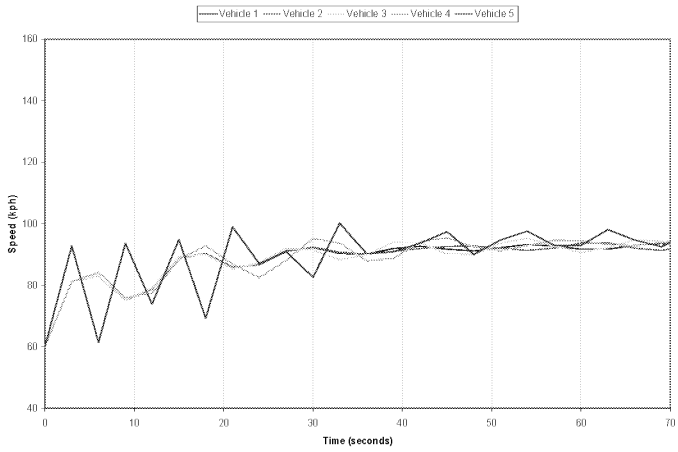


Figure 9: Speed Profiles for the first 70 Seconds of the Simulation

Simulating Gipps model in the framework

Gipps (1981) proposed a car-following model which assumes that (under car-following conditions) drivers select a speed such that they are able to bring the car to a safe stop in reaction to a vehicle ahead braking sharply. In light traffic, a vehicle will accelerate towards its maximum

desired speed. This principle and consequent driver behaviour can be translated to a car following equation given in Gipps (1981). This model was then implemented in the microsimulation framework.

Figure 10 shows vehicle trajectories from the Gipps model. 50 vehicles have been modelled on a single lane circular road of 1km. Each vehicle has been given a starting velocity of 100 kph, perturbed by $\pm$ 5 kph. Wilson (2001) proved that shockwaves (due to flow breakdown) should be propagated under the parameters used above, which is confirmed by the results. This result provides confidence that the microsimulation framework is operating correctly.

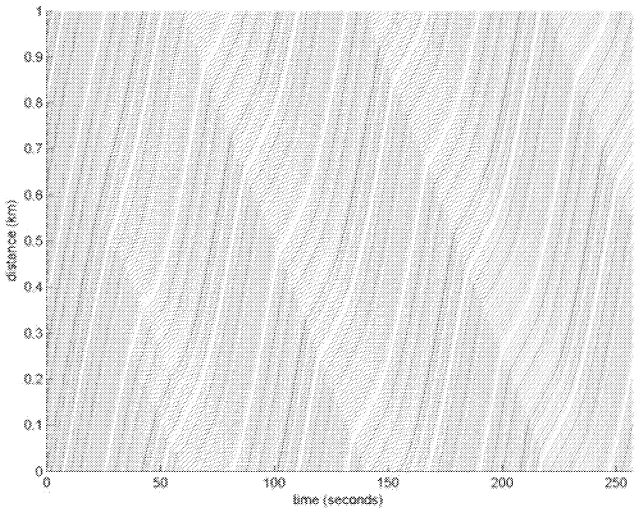


Figure 10: Implementation of Gipps's Car-Following Model in the Simulation Framework

DISCUSSION OF RESULTS AND CONCLUSIONS

In commenting on results, it is important to note that driving behaviour is a complex process involving many interacting parameters and processes. For example, a driver receives (e.g. visually, auditorially) information from the environment which is processed by the brain to lead to an action. Neural networks provide a potentially suitable modelling approach to represent this complexity by feeding the network with collected field data whilst avoiding the potential subjective interpretation associated with developing models in a top-down fashion.

One observation from the current work is that although the individual vehicle trajectories extracted from IVD provide a unique data source to train neural network, the data is collected between detectors separated by 100m. This poses problems for neural networks because interpolating between this long separation potentially produces too coarse data for the neural network, which may impact performance. It should also be emphasised that neural network configuration, training and diagnosis can prove extremely difficult because of its black box nature. It is therefore difficult, for example, to determine the right combination and optimisation of parameters and weights, although this project employed an optimisation search algorithm to facilitate this.

Fewer vehicles are modelled in the NN model (Figure 8) than the Gipps model (Figure 10) since the NN model demonstrates some anomalies at short headways. Consequently some drivers do not brake sufficiently to avoid collisions with the vehicle in front. It is thought that the NN model has not been supplied with sufficient data in this behavioural regime where drivers react to decreasing headways at close distances. This is an issue which will be resolved in subsequent research projects.

In concluding this paper, it is worth highlighting some key achievements. New modelling concepts and techniques for extracting re-identified IVD and configuring a neural network using over 50,000 individual vehicle data collected from the UK motorway to demonstrate learned behaviour e.g. car following have been presented. A number of conclusions can be drawn from this project.

- The project explored the use of Artificial Neural Network (ANN) to develop a Neural Network Model (NNM). The model was trained using captured data from the IVD.
- The project demonstrated learning capability by the NNM using thousands (over 50,000) of individual vehicle data points collected from the UK motorway network.
- A simulation and visualisation model was developed to integrate the IVD/NNM to aid interactive manipulation of the simulation and test the qualitative validity of the work.
- The analysis of the results showed a reasonable performance of the NNM for a car following model.

It is also worth briefly commenting on the potential application of the work reported in this paper:

- It might be possible to add an enhance version of this model into car simulators to allow realistic interactions between virtual drivers and simulator participants.
- This would further allow psychological factors that influence general driving and traffic to be evaluated more accurately. For instance, constraining neural network drivers to produce driving behaviour similar to that of impaired drivers, e.g. stressed and drunk.
- The work may assist greater behavioural realism in traffic micro-simulations models.
- Such models can be used to appraise new traffic management schemes, help predict short/medium term traffic conditions, improve the analyses (or assessment of the impact of) traffic flow and its characteristics.
- Prediction of incidents and flow breakdown from traffic behaviour.
- Prediction of congestion levels by quantifying and understanding how individual driver decisions lead to flow breakdown.

ACKNOWLEDGEMENT

The authors are particularly grateful to TRL Academy Director, Neil Paulley, for funding this work and to the Chief Scientist Research Director of Transportation, Alan Stevens, for his support and technical review of this work. The views expressed in this report are the authors' and not necessarily those of TRL.

REFERENCES

Coifman, B., 1998, "Vehicle Reidentification and Travel Time Measurement in Real-Time on Freeways Using the Existing Loop Detector Infrastructure". *Transportation Research Board*, pp 181-191.

Coifman B, Cassidy M., 2002, "Vehicle Re-identification and Travel Time Measurement on Congested Freeways". *Transportation Research: Part A*, vol36, no 10, pp 899-917.

Dumbuya A., 2006, "Modelling complex systems of traffic interactions using agent-based micro-simulation". In: *TRL Annual Research Review*, 2005. Wokingham: TRL, pp. 129-139, ISBN 1-84608-829-1.

Dumbuya, A., Booth, A., Reed, N., Kirkham, A., Philpott, T., Zhao, J. and Wood, R., 2006, "Complexity of Traffic Interactions: Improving Behavioural Intelligence in Driving Simulation Scenarios", *European Simulation and Modelling Conference Proceedings*.

Dumbuya, A., Reed, N., Rhys-Tyler, G., Zhao, Q.J., and Wood, R.L., 2005, "Interactive Scenario Modelling for Hazard Perception in Driver Training" In L.Dorn (ed.) *Driver Behaviour and Training Vol. II*, (Ashgate Publishing, Hampshire), Chapter 1, pp. 3- 17. ISBN-10: 7546 4430 8.

Dumbuya, A. D., Wood, R. L., Gordon, T. J., and Thomas, P. 2002, "An agent-based traffic simulation framework to model intelligent virtual driver behaviour", In *Proceedings, Driving Simulation Conference (DSC'02)*, Paris, France, pp. 363-373.

Dumbuya, A. D., and Wood, R. L., 2003, "Visual perception modelling for intelligent virtual driver agents in synthetic driving simulation", *Journal of Experimental and Theoretical Artificial Intelligence (JETAI)*, 15 (1): 73–102.

Frith B and Pearce D., 2002, "Driver Information on Journey Time Variability Generated using ANPR Data". *RTIC conference proceedings*.

Gurney, K., 1997, *An introduction to Neural Networks*, UCL Press.

Gipps, P. G., 1981, "A Behavioural Car-following model for computer simulation", *Transportation Research, Part.-B* 15, 105-111.

Lunt G.M. and Wilson R.E., 2003., "New Data Sets and Improved Models of Highway Traffic", *UTSG 35th Annual Conference*, Volume 2.

Lunt G.M., 2005, "Vehicle Re-Identification Using Induction Loop Data". *ECTRI-FEHRLFERSI Young Researchers Seminar 2005*.

Lunt, GM., Day, M.A., Wilson, R.E., 2006, "Enhancing Motorway Traffic Data with Novel Vehicle Re-Identification Techniques". *13th World Congress on Intelligent Transport Systems (ITS World Congress, 2006)*, London, UK, October 2006.

Wilson, R. E., 2001, "An analysis of Gipps's Car-following model of highways traffic", *IMA Journal of Applied Mathematics*, **66**, 509-537.

Object Management Group:http://www.omg.org/ (Accessed 17th December 2006).

OpenGL: http://www.sgi.com/products/software/opengl/ (Accessed 17th December 2006).

BIOGRAPHY

ABS DUMBUYA is a TRL Academy Fellow and Research Scientist in the Transportation Division at TRL. He has worked on a wide range of projects, including: leading research in agent based micro-simulation modelling of driver behaviour, origin-destination matrix estimation, routing assessment of navigation systems, systems engineering of road user charging. He is a Loughborough University graduate with BEng (Hons.) degree in Manufacturing Engineering and Management and PhD in traffic modelling and simulation both from the Wolfson School of Mechanical and Manufacturing Engineering. He has authored and co-authored several academic publications in the areas of agent based traffic micro-simulation, driver behaviour and driving simulation.

GEORGE LUNT is an Associate Academy Fellow within the Transportation Division at TRL. His research interests include analysis and interpretation of motorway traffic data, and the development of real-time diversion strategies. He also has experience in a wide of traffic modelling, having been a developer for English Highways Agency's microsimuation model, a user of CONTRAM a mesoscopic traffic assignment model, and a developer of the UK's government National Transport Model.

JILL WEEKLEY is a member of the Network Performance group in the Transportation division of TRL. She has worked on a wide range of projects in a variety of areas, in particular the development of models for real-time network management, and was elected as an Associate Academy Fellow of the TRL Academy in November 2006. Jill joined TRL in 2004 after completing a PhD in Fluid Dynamics at the University of Nottingham.

ANNA BOOTH is a TRL Academy Associate Fellow and Researcher in the Transportation Division at TRL. Anna has an MSc in Natural Computation from Birmingham University and a BSc in Mathematics and Computer Science from Nottingham University. She is currently studying towards an MSc Transport Planning and Practice at the University of Surrey. Anna has worked on a variety of projects at TRL including the development and training of Neural Network models to learn driving behaviours, the testing of a cell transmission model for traffic flow and the development of PCMOVA, an application to allow control of junctions in traffic microsimulations.

ANDREW LEWIS has recently gained a First-Class Honours Masters degree in Software Engineering at the University of Southampton. Working in the Software Development Group within the Transportation Division he has been involved in work on an Italian version of ARCADY as well as maintaining various TRL software products. He has been closely involved with the testing and proving of OSCADY Pro, a phase-based traffic signal optimisation program. Andrew worked as a software developer in the project and has interests in simulation, visualisation and real-time and embedded systems.

TRAFFIC AND LOGISTICS CASE STUDIES

SIMULATION MODELLING OF AIRPORT PASSENGER OPERATIONS: AN IMPROVEMENT CASE STUDY

Martin S. Barber
CORAS
University of Salford
Salford, M5 4WT, U.K.
E-mail: m.s.barber@salford.ac.uk

Peter Durie
Head of Business Projects
Manchester Airport Group
Olympic House
Manchester, M90 1QX, U.K.
E-mail: peter.durie@manairport.ac.uk

KEYWORDS
Discrete event simulation, passenger transport, decision support systems

ABSTRACT

This paper describes operational process improvements and the development of various simulation models for Manchester Airport. This project is a collaborative partnership between Manchester Airport and The University of Salford.

Several initiatives have been developed to improve airport operational processes and enhance the customer experience. These include, firstly the installation of Common Use Self Service (CUSS) terminals into the check in areas to increase the potential capacity of check in.

The second is a review of the Gating process rules and enhancement of passenger information to improve efficiency by reducing process time and increasing passenger dwell time.

INTRODUCTION

Manchester Airport is the largest U.K. airport outside London, the third largest in the U.K. and the eleventh largest in Europe. As frequently quoted by the media airports are currently growing by about 6% yearly in terms of passenger numbers. This is putting considerable pressure on all airports to efficiently handle the increased volume in passengers. In addition to this there are peaks and troughs during the day, week, and year to address. It is not like a factory that can simply produce level production over time and use a stores buffer to match supply to customer demand.

There is a need to breakdown the operations into their individual component parts and understand how each operates and then consider their interactions with operations to which they are interlinked; a systems analysis approach. In addition to this there are many different organisations involved in the process with reduces the flexibility in using free staff to assist in similar areas or to cover staff absence. Various processes are under major periods of change currently:

- Check in
- Security
- Gating

Self-service Check In for passengers to replace a manned check in desks especially as many low cost carriers have increasing the capacity and weight of walk on luggage for passengers.

Passenger security checks have increased due to 9/11 and the recent 10th August 2006 threat has eroded the process improvements made in the system at Manchester.

The Gating process has been revised to standardise the rules in place with the individual airlines. Airlines like to gate passengers early to ensure they don't wander off and delay the flight. This can lead to increased levels of stress for the passengers, excessive pressure on gating areas, lost retail sales and future loss of custom.

These initiatives must be paying off as Manchester Airport was voted Best UK Airport 2005 by readers of Travel Weekly and IATA Eagle Award winner 2006 in recognition of high customer service and reducing airline charges.

LITERATURE SEARCH

Airport modelling is usually performed as a sequential process with a capacity calculation of each processing stage and holding area. Detailed formulas are defined in The Airport Development Reference Manual (2004) produced by the International Air Transport Association. Peak capacity is simply matched at all stages of the process with a safety margin in some cases. Detailed contingency plans exist for bad weather and other emergencies which cause bottlenecks in the system.

Considering the airport passenger operations as a systems model (figure 6) the Arrivals and Departures processes are linked together by the turn round process. The turn round process is defined for an A380 large aircraft in figure 7 by Horstmeier (2001). The only difference between this and a normal aircraft such a 737 or 767 is the sub-process times are approximately half due to the increased number of passengers and fuel on the A380 compared to smaller aircraft.

The processes within Departures and Arrivals are listed in the figure 1 below. The arrival rate of passengers depends on the number of passengers booked onto each aircraft, the aircraft schedule and the typical arrival rate, such as short haul – no frills and long haul charter passengers.

DEPARTURES:	ARRIVALS:
Parking	Disembark
Check in and bag drop	Immigration
Landside retail	Baggage reclaim
Security	Customs
Airside retail	
Gating	

Figure 1: Overview of Departures and Arrivals Processes

The passengers are given a window of time to arrive in, e.g. check in opens three hours before flight and closes one hour before flight.

The authors have visited other airports to compare and study operating procedures and process times. These airports include Heathrow T5, Schipol, East Midlands, Berlin, Lyon and Liverpool.

A literature search has been made of airport passenger operations and simulation. Little airport simulation research has been published, as generally it is of a confidential nature and simulation provides a competitive edge for the airport concerned. Kiran and Cetinkaya (2000) modelled a New International terminal. Pendergraft et el (2002) used simulation to model Airport passenger security system. However, with recent events and changes these papers are out of date.

Torres et el (2005) has proved a link between passenger waiting time and retail expenditure. Yan et el (2002) has devised a simulation framework for airport gate assignment.

This section identifies some of the many different areas of Airport Simulation which are all relatively unexplored.

PROCESS CHANGES

The two processes that will be discussed in detail in this paper are CUSS and the Gating process.

Common Use Self Service

Self Service check in terminals have been in use for some time at many airports but instead of allowing airlines to continue to introduce their own dedicated terminals, that only that airlines customers could use, airports are now introducing flexible terminals that any authorise airlines customer can use. This philosophy is essential in any airport that has many different airlines sharing facilities to prevent unnecessary duplication of equipment and increase the equipments utilisation. It benefits the airlines in reducing the requirement for check in staff. It also benefits the airport in that it substantially increases the capacity of the existing check in halls which previously may have been a bottleneck, restricting the growth of passenger numbers. It also releases some of the existing check in desks to be designated fast bag drops for hold baggage. These bag drop desks are usually positioned alongside main access corridors to allow speedy access for passengers that use CUSS processes and to encourage usage due to increased visibility and customer convenience. The airlines benefit in that they have reduced staffing and check in rental costs with an improved customer service.

Terminal:	Dedicated Self Service	CUSS
1		12
2		28
3	14	2

Figure 2: Numbers of Self Service Check In Positions in Each Terminal at Manchester Airport

The advantage is that the CUSS terminals can be used by any passengers to check in; provided their airline is using the system. This can substantially reduce passenger queues at the departure check in area. This CUSS area however does require hosting to help introduce some customers to the technology. Even British Airways with a high percentage of frequent business flyers is still hosting its Self Service check in machines after some ten years of operation. This is in stark contrast to banks Automatic Telling Machines which the public use unassisted. So in airports some of the space which was previously used for queuing customers has become a flexible check in area.

This has a significant improvement for the passenger in that they have a significant reduction to their queuing time for check in if they use the CUSS machines. The Airline has reduced labour and check in costs. The Airport has increased capacity of what is usually a bottleneck operation, increased flexibility of the area concerned and increased the potential shopping time for passengers. Passengers are able drop their hold baggage and go through security checks. They can then relax and wait for their flight.

The Gating Process

Airlines generally prefer to gate passengers early to ensure that they do not get lost or delayed and then delay the aircraft take-off. They also have to reconcile the hold baggage to passenger list to ensure that there are no unescorted bags on the aircraft, if the passenger goes missing in the terminal building. This is simply a check of the hold baggage passenger list against passenger boarding cards.

Figure 3 below shows a typical passenger arrival rate at a boarding gate before any changes. All times are minutes before scheduled take off time. The green call to gate occurred at 60 minutes, 60% of the passengers are already there as they know the gate number already. The 223 passengers started boarding at 38 minutes and the boarding process was completed by 10 minutes. A total of 28 minutes boarding time. Also of note is that 80% of passengers were present at 44 minutes. Overall this demonstrates a long boarding procedure with a slow build up of passengers.

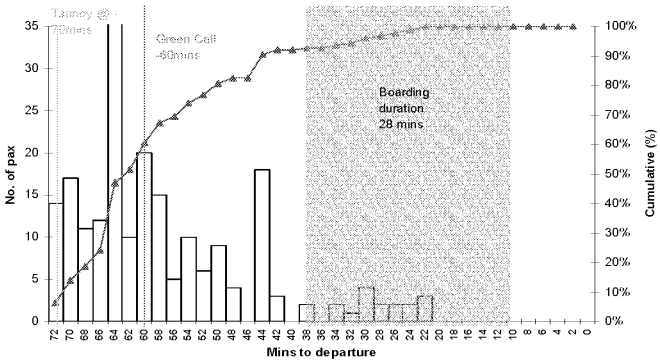


Figure 3: Passenger Arrival Rate at Gating Process

A study of 152 flights was made and the study showed that regularly 80% of passengers arrive at the gate 45 minutes prior to departure, boarding time typically takes 9 seconds per passenger and is complete 10-15 minutes prior to departure. There followed a review of the call to gate timing and an enhancement to the flight screen information. One major change for Manchester is that the boarding gate number is no longer printed on the passengers' boarding card. Instead the gating time is printed on the ticket and Passenger Information Display Screens (PIDS) at a predetermined interval of time. This holds the passengers in the Airside retail area until the scheduled time when the gate number is added to the screens and a traffic light system is employed. In addition if the aircraft is delayed then the call to gate time is updated, again saving unnecessary time confined in the gating area. Another change was to reduce this scheduled time by ten minutes with selected airlines. This gives the passengers an extra ten minutes in airside retail and not confined in a gating lounge. This ten minute reduction of the "green call" time had no adverse effect on passenger boarding and was accepted by the Airlines at the airport.

The following two figures (4 and 5) show the departure passenger flight information which alternate on each of the multiple display screens in the Airside retail area. These also use the industrial practice of traffic lighting information: green, amber and red. Green for a "call to gate" or "wait in lounge" message and red for "now boarding" message. Amber for a "flight delayed" message with a revised gating time.

The arrival of passengers to their gating areas has been surveyed before and after these changes. These changes have reduced passenger anxiety due to reduced waiting time in the restricted gating lounges, offer an improved customer experience, increase gating area capacity and increase the opportunity for airside retail sales. There is potential to reduce the gating time even further.

12.02	Flight	Destination	Gate
12.00	BA1662	Milan	
12.15	BA7605	Belfast City	148
12.25	BA1872	Glasgow	146
12.30	WW3901	Ireland West Airport Knock	56
12.50	BA1686	Dusseldorf	
13.00	BA1606	Paris-Cdg	

Figure 4: Passenger Information Display Screen: One

12.02	Flight		Gate
12.00	BA1662	Wait in lounge	
12.15	BA7605	Now boarding	148
12.25	BA1872	Go to gate	146
12.30	WW3901	Go to gate	56
12.50	BA1686	Gate opens in 35min	
13.00	BA1606	Gate opens in 25min	

Figure 5: Passenger Information Display Screen: Two

MODELLING APPROACH

Individual improvements have been tested at one of the three terminals at Manchester Airport and when successful rolled out to the other two. Only one of the three terminals has been modelled, Terminal 2.

Initial data was gathered from studying the current process and potential improvements identified. E.g. omit boarding gate number from card and upgrade PIDS screen display. A simple simulation model was built and tested to assist in choosing new operating rules. These changes were then presented to the willing airlines, tested and fully proven before rolling the changes out to all airlines across the airport.

DISCUSSION

There are detailed discrete event simulation modelling packages, such as ArcTerm which are used by many airports. These are very expensive to maintain and are often only used at the end of a new airport design or major expansion of an existing facility. Even the internal layout of the new Terminal 5 at Heathrow is being redesigned whilst the build is taking place as the processes and technologies continue to evolve over the life of the building project. There is a place for Discrete Event Simulation software packages such as Service Model to be used to evaluate the options for change without the exceptional graphics we have come to expect with more expensive packages.

Studies both mathematically and using simulation have been made into process times and various improvements made to individual process steps.

Way-finding signage is a critical part of any airport and the average passenger is overwhelmed by the commercial advertising when using any airport today. Correct positioning of signage is essential to prevent lost passengers and to avoid unnecessary staff costs to help passengers.

CONCLUSION

There is a place for the use of low-cost discrete event simulation software to model complex operations such as airports. Plus many small changes can be evaluated individually. It does however require a considerable amount of data gathering and testing of various individual changes. It is invaluable to gather the process data before and after process improvements are made to make objective decisions.

CUSS terminals installed into Check In areas enable passengers to check themselves onto their flight, thus avoiding queues. Passengers still have to use a fast bag drop for hold luggage.

The Gating process has been significantly improved for both customer and airport by revising call to gate times and updating Passenger Information Display Screens.

FURTHER RESEARCH

There are many changes taking place within Airports to cope with the increasing demand for air travel. Technology is being introduced to improve passenger handling efficiency. One common major bottleneck is the Outbound Security Check due to the increasing requirements placed on it and is the subject of ongoing process development and simulation modelling.

REFERENCES

Horstmeier T. and Hann F., 2001, Influence of Ground Handling on Turn Round Time of New Large Aircraft, Aircraft Engineering and Aerospace Technology, Vol.73, No.3, pp.266-270.

Kiran A.S. and Cetinkaya T., 2000, Simulation Modelling and Analysis of a New International Terminal, Winter Simulation Conference, pp.1168-1172.

Pendergraft D.R., Robertson C.V. and Shrader, 2004, Simulation of an Airport Passenger Security System, Proceedings of the 2004 Winter Simulation Conference, pp.874-878.

Pitt M., Wai F.K., and Teck P.C., 2001, Strategic Optimisation of Airport Passenger Terminal Buildings, Facilities, Vol.19 No.11/12, pp.413-418.

Rowley J. and Slack F., 1999, The Retail Experience in Airport Departure Lounges: Reaching for Timelessness and Placelessness, International Marketing Review, Vol.16, No.4/5, pp.363-375.

Torres E., J.S.Dominguez, L.Valdes and R.Aza R., 2005, Passenger Waiting Time in an Airport and Expenditure Carried Out in Commercial Area, Journal of Air Transport, Vol.11, pp.363-367.

Van Dijk N. and van der Sluis E., 2006, Check-in Computerisation and Optimisation by Simulation and IP in Combination, European Journal of Operational Research, Vol. 171, pp.1152-1168.

Yan S. and Shieh C. and Chen M., 2002, A Simulation Framework for Evaluating Airport Gate Assignments, Transport Research Part A: Policy and Practice, Vol.36, Issue 10, pp.885-898.

AUTHOR BIOGRAPHIES

MARTIN BARBER is a lecturer in Operations Management and Discrete Event Simulation at Salford Business School, The University of Salford. He previously worked as an Engineering Manager in the electronics industry and is a charter engineer before gained his MSc at Salford.

PETER DURIE (BSc. FRICS) is Head of Business Projects at Manchester Airport focusing on improving key operational processes in the passenger journey. He is a chartered surveyor with over 30 years experience in major public and private sector business projects. He delivered a multi-modal £60million station complex at Manchester Airport.

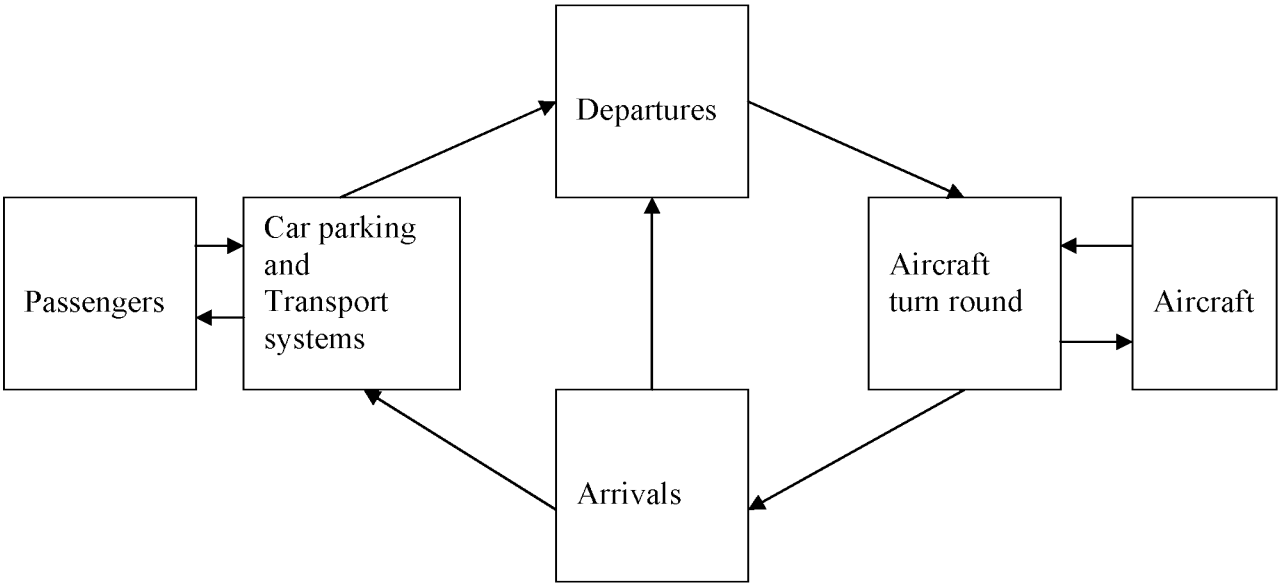


Figure 6: Passenger Operations System Overview Model

Sub-process:	Minutes:	0-15	30	45	60	75	90	105	120	135
De-boarding	15.9									
Boarding	35.0									
Loading/unloading	59.8									
Refuelling	70.3									
Cleaning	60.6									
Catering	87.1									

Figure 7: Calculate Timeline of A380 Turn Round Time
Source: Horstmeier, 2001

BULK TERMINAL MODELLING AND SIMULATION

Jaap A. Ottjes, Gabriel Lodewijks and Dingena L. Schott
Delft University of Technology
Faculty of Mechanical, Maritime and Materials Engineering
Mekelweg 2, 2628 CD Delft, The Netherlands
j.a.ottjes@tudelft.nl; g.lodewijks@tudelft.nl; d.l.schott@tudelft.nl

KEY WORDS

Transportation, Bulk Terminal, Process Interaction Modeling, Operational Control

ABSTRACT

A modeling approach for operational control and evaluation of large export bulk terminals is presented. The model works at individual equipment level including the equipment break down behavior. As key performance indicator the delays of the bulk carriers is taken. The results of an example case are shown determining the maximum capacity of a terminal for different stock pile configurations. A view is given on future research.

INTRODUCTION

Deep sea export bulk terminals are interfaces between mines and bulk carriers for deep sea transport. The heavy equipment and the substantial area needed for a bulk terminal require large investments and the operational costs are considerable. The demand for raw materials such as iron ore is expected to increase strongly in the next decades. As a consequence new terminals will have to be built and existing terminals are forced to expand.

The operational control of a terminal and the terminal design (layout and equipment configuration) are strongly related. It is anticipated that it is possible to improve operational control considerably by using modern techniques especially by taking into account dynamic aspects.

Minerals like iron ore or coal are extracted from inland mines at a large scale and transported, usually by train, to a deep sea export terminal for further transportation to import oriented bulk terminals. The bulk material is temporary stored on a stock yard and later reclaimed and loaded into bulk carriers. The key performance indicator of such a system is the waiting time of bulk carriers (Jagerman, 2003). The waiting times depend on the capacity and the reliability of the total chain of equipment components of the terminal. The reliability is a combination of reliability of all individual equipment: train unloaders, conveyors, stacker-reclaimers and ship loaders. The reliability of the individual components can be expressed as the availability and the distributions of down- and up-times of the component. An example of simulation of a specific terminal is described in Ottjes and Lodewijks (2004).

In order to determine the terminal performance as a function of many variables and to test alternative operational control methods, a simulation model has been developed. In this paper the modeling approach will be discussed and an example of the use of the model will be shown.

TERMINAL EQUIPMENT AND OPERATION

In this work we concentrate on export oriented dry bulk terminals. Because the design process of both import oriented and export oriented bulk terminals are the same and the layout design considerations are similar, the modeling approach discussed is applicable to both terminal types.

We consider a terminal that is fed by trains. Several material grades may be involved for example originating from several mines. Figure 1 shows a simple terminal lay out with equipment.

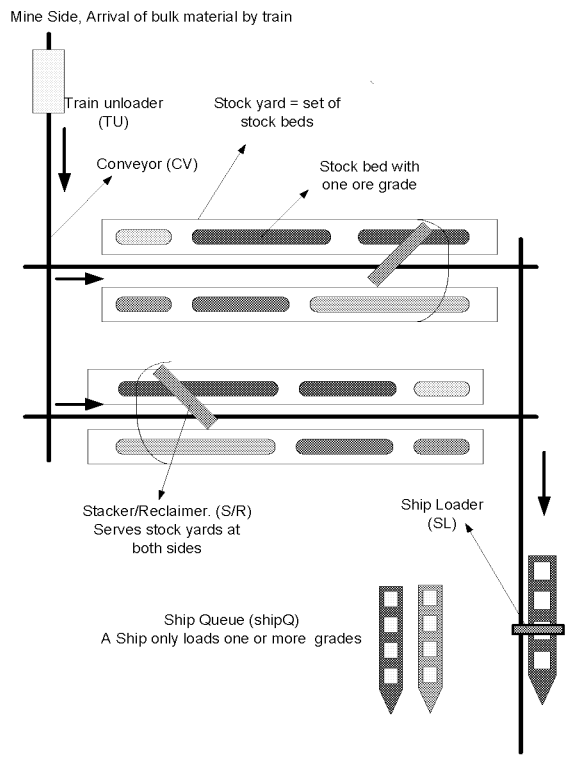


Figure 1: Terminal with Stock Yard and Stock Bed Lay Out

Train trucks are unloaded by a 'train-unloader', for example a so called a tippler. The unloaded material is fed as a continuous flow onto a belt conveyor line. If the material is to be stored, a proper 'stock bed' and a transport route to that stock bed have to be assigned. A stock bed is part of a stock yard and contains one material grade. The actual storage action is performed by a so called stacker.

Bulk carriers generate a transport demand from a stock bed to the ship loaders. A bulk carrier may demand batches of one or more material grades. For each batch one or more stock beds, including proper transport routes to the ship loader(s), have to be assigned. The material is retrieved from the stock bed by a machine called 'reclaimer'. In many cases stacker and reclaimer are combined in one machine, the so called stacker-reclaimer. Figure 2 shows the working of a stacker-reclaimer.

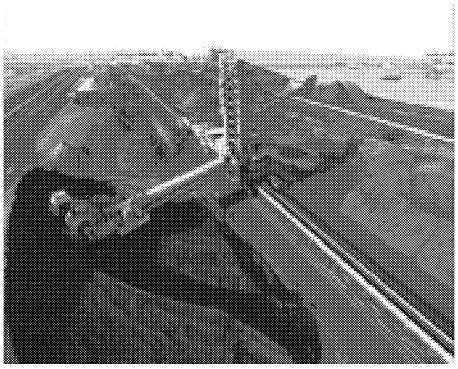


Figure 2 Stacker-Reclaimer (internet link 1)

The lay-out and operation of large scale terminals is much more complicated than shown in Figure 1. They may have several tipplers, stock yards, stacker-reclaimers, conveyors and ship loaders. Key issues in terminal operation are stock yard management, maximize utilization of equipment, dealing with equipment breakdowns, maintenance and keeping up the customer service level (Lodewijks 2003).
Train and ship (un)loading methods are discussed by Wöhlbier (1987) and Spanke (2000).

MODEL STRUCTURE AND ELEMENT CLASSES

The model is developed according the process interaction method (Zeigler 2000), providing the possibility to model very close to the real world system. The process interaction method can be summarized as follows: (1) decompose the system into relevant element classes, preferably patterned on its real world structure. An element class is characterized by its attributes. The state of each instance of a class is defined by the state or value of its attributes. (2) Determine the 'living' element classes and assign a process description to these classes making use of simulation-time consuming commands like 'hold' and 'standby' and process-interaction commands like 'interrupt', 'pause', and 'resume'. A process governs the behavior of each instance of the element class. The advantage of this modeling approach is that the model validity both structural and dynamically, can be assessed by specialists of the real system in an early stage of the project.

Next the model will be discussed and the crucial parts will be shown in pseudo code. The process-interaction approach best fits with the object-oriented paradigm.
Element classes in the model are supposed to descend from a simulation class called *SimClass*, having all the necessary properties and methods needed for process interaction modeling. Further we will use the concept of *queues* to create sets of elements. Apart from waiting line applications, the queue concept is used to create flexible object oriented data structures and to facilitate the modeling of control functions.

Job shop approach

The class of bulk terminals can be modeled very well as a modified 'job shop'. In a job shop, machine groups exist with machines having equal functions, for example a drilling group. In our model we distinguish 'Equipment line groups', containing 'equipment lines'. An equipment line consists of a sequence of pieces of equipment. It acts as a 'transport machine' that conveys a certain batch of bulk material between two points. The first machine of an equipment line always obtains the material in some way namely from a train or from a stock bed and the last machine of an equipment line delivers the material to a stock bed or a ship. The connection between the first and last machine is formed by a sequence of conveyors, a conveyor line. Each equipment line group has its own job queue (jobQ) with 'transport jobs' with corresponding 'From' and 'To' machines.

An equipment line group is defined as:

```
TEqLineGroup    = class(SimClass)
  JobQ      :Queue
  EqLineQ   :Queue
```

Transport job

Both the arrival of a train and a ship, trigger the generation of one or more 'transport jobs'. A transport job represents a transport demand of a batch of bulk material of one grade and can have three basic origin-destination combinations:

1: from train to stock bed
2: from stock bed to ship
3: from train to ship

The train-to ship connection may use stockyard conveyors as bypass. Though the related stacker-reclaimer is blocked during the transportation period, still the stacker-reclaimer system is relieved. If the batch would have been stacked and later retrieved, a stacker-reclaimer would be busy twice as long.
Another way is using special short cut conveyors as is indicated in Figure 3 but in that case cost will be higher.
The definition of a transport job is:

```
TJob     = class(SimClass)
    Grade
    TonsToMove
    FromMachine : TEquipment // has to be assigned
    ToMachine   : TEquipment // has to be assigned
    Origin      // train or stock bed, has to be assigned
    Destination // stock bed or ship, has to be assigned
```

Equipment and Equipment lines
In Figure 3 a terminal is shown schematically. There are 3 train unloaders (TU), 6 stacker-reclaimers (SR), 3 ship loaders (SL) and 18 conveyors (CV). We call a crossing of two conveyors, where material can be transferred from one conveyor to the other, a 'transfer point'. In practice, because of high investment costs, not every 'crossing' is a transfer point.
Three equipment line variants can be distinguished: Stacking lines, Reclaiming lines and Direct lines. Examples are:
Stacking line between TU1 and S/R1:
TU1, CV(1, 4, 13), S/R1
Reclaiming line between S/R1 and SL1:
S/R1, CV(13, 7, 10), SL1 and
Direct line between TU1 and SL1:
TU1, CV(1, 6, 18, 9, 10), SL1

Direct lines can also be formed using shortcut conveyors if available.

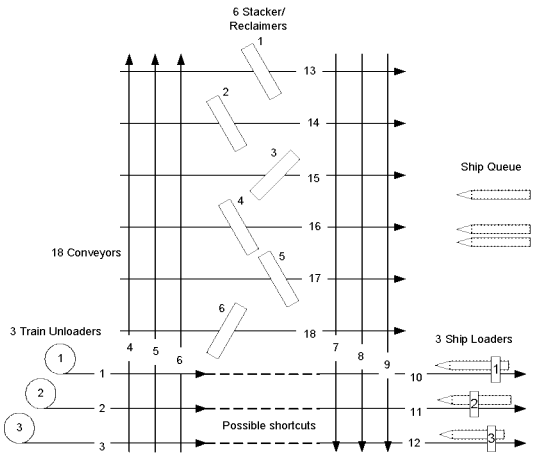


Figure 3: Schematic Terminal Lay Out with Equipment

Between the first and last machine of an equipment line in general multiple routes are possible composed of sequences of conveyors. Most equipment lines have one or more pieces of equipment in common with other equipment lines. The consequence of this is that if, for example, equipment line A is active, all other equipment lines containing one or more machines of A are not available.
The equipment line class is defined as

```
TEquipmentLine   = class(SimClass)
    EquipmentQ        :Queue        //set with all equipment needed
    EqAvailQ          :Queue        //sub set with equipment available
    MyGroup           :TEqLineGroup
    MyJobQ            :Queue        // reference to JobQ of MyGroup
    MyJob             :TJob         // job in process
    FromEq            :TEquipment   //first machine in line
    ToEq              :TEquipment   //last machine in line
    Flow                            //actual flow tons/hour
    Variant                         //Stacking, Reclaiming or Direct

  Methods
    SetEq_avail     //make own eq available for the other lines
    SetEq_Unavail   //withdraw own eq from other lines
    PauseEq         //interrupt all equipment in the eq. line
    ResumeEq        //resume process of all equipment in  the eq. line
    Process
```

The equipment class is defined as:

```
TEquipment   = class(SimClass)
    CurrentLine       :TEquipmentLine  //Eq line in which this eq is active now
    AllLineQ          :Queue           //all lines that need this equipment
    myUpDistr         :TDistribution   //distribution of Up times
    myDownDistr       :TDistribution   //distribution of down times
  Methods
    be_available      //enter EqAvailQ of all eq.lines  in AllLineQ
    be_unavailable    //leave EqAvailQ of all eq.lines exept the claiming one
    Process
```

PROCESS DESCRIPTIONS

In the processes, the next vocabulary is used:
Time consuming commands:
Hold(t) : suspend the process during t time units and proceed after that.
While (condition) standby: suspend as long as the condition is true, and then proceed.

Process-Interaction commands:
Elementx.Pause : Elementx is an element that is in a hold(t) status. Pause postpones the process of Elementx until Elementx.Resume is encountered from the process of another element. After that the remaining part of period t is continued.

'Repeat' in a process description means that the following block of indented lines is repeated continuously.
Repeat while (condition): repeat next indented block while condition is true.
Elements can be member of one or more queues at the same time.

Next the descriptions of the most relevant processes are given. Each representative of the class equipment follows a process that alternating samples a period of undisturbed work during, the 'uptime' and a down

period, the 'downtime'. Every time a down period starts, the equipment piece signals that to its current equipment line. (It always has a current equipment line, because else it would not be working and consequently could not break down). Physical dynamic aspects of long heavy loaded belt conveyor systems have not been taken into account (Lodewijks 2001).

```
TEquipment.Process
 Repeat
   UpTime    =myUpDistr.Sample     //sample up time
   Hold(UpTime)                    //work during up time
   DownTime=myDownDistr.Sample //sample down time
   currentLine.PauseEq             //have current eq.line pause all its equipment
   currentLine.Pause               //pause current eq.line itself
   hold(DownTime)                  //don't work during down time
   currentLine.ResumeEq            //have current eq.line resume all its equipment
   currentLine.Resume(now)         //resume current eq.line itself at time = Now
```

An equipment line waits in its equipment line group for a job. If a job is assigned, the equipment line preserves all the machines needed in its line and works during the time needed to transport the 'tonsToMove' of its job.
Every time however when one of its machines goes down, the equipment line process is interrupted and all other machines in the line are interrupted too. If the broken machine has been repaired, all machines are resumed and the equipment line resumes to work during the remaining transport time.

```
TEquipmentLine.Process
 Repeat
   PauseEq          //interrupt all equipment of the line
   SetEq_Avail      //make eq available for all eq. lines.
   //next wait for a job AND till all eq available
   while (MyJobQ.Length=0) AND (not all eq available) Standby
   //now this line is going to work, so its equipment has to be claimed
   SetEq_Unavail //call method be_unavailable of all eq. in equipmentQ
   ResumeEq      //put all equipment to work
   MyJob:=MyJobQ.FirstElement
   MyJob.LeaveQueue(MyJobQ)
   t:=MyJob.TonsToMove/flow      //batch transport time

   if destination is a ship
     If ship is in ShipQ ship.leave(ShipQ)// ships initially wait in the ShipQ
     Hold(t)
     If ship is ready ship.resume(TNow)   // ship will leave the terminal

   if destination is a stock bed
     Hold(t)
   // in all cases update system status (tons in stock, tons transported etc)
```

Other element classes and processes
Apart from the model elements mentioned the next element classes are represented in the model

- Train generator with train arrival and load information
- Trains, consisting of a set of loaded trucks
- Ship generator with ship arrival time data and load information
- Ships, with a set of batches with grade to be loaded
- Stockyard structure with a set of stock beds and one or more associated stacker-reclaimers (of class equipment) covering the yard

The trains and ships are created during the simulation run and provided with the proper attribute values. The generation of trains and ships can be based on statistical distributions or obtained directly from historical data measured in practice on the terminal under study or another similar terminal.

Operational control
During operation, decisions have to be made at several levels and at several moments.
There are many ways to assign equipment and storage locations to a job.
In case of a train job, for example a train unloader [1)] and a stock bed + associated stacker reclaimer [2)] have to be assigned.

1) Train unloader assignment options: Just cyclic or according earlier planning or 'just in time' according actual data such as availability of equipment or storage space.
2) Stock bed assignment options: According planning or 'just in time' on the basis of the content of all relevant stock beds and/or the (predicted) availability and/or location of the related stacker-reclaimer and/or stacker-reclaimer work loads.

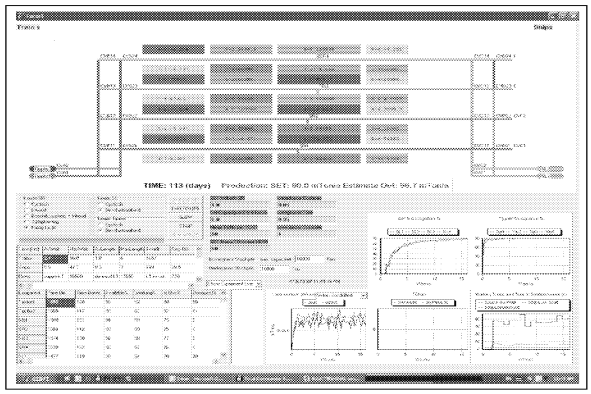


Figure 4: Screen shot of the test configuration with 4 Stock Beds per Material Grade

The same holds for the generation of transport jobs for loading a ship. A number of decision rules are built in the model and the model is made extendible for other rules.

A more strategic decision is how to arrange the stock yard lay out. Questions like: What is the influence of the bulk material throughput on the required total stock pile size and how many stock beds per grade have to be realized and of what size, have to be answered. The latter is tested in the following test case.

TEST CASE

The model is implemented in the process oriented simulation package Tomas (Veeke and Ottjes, 2000), (internet link 2). Figure 4 shows a screen shot of the model with one of the lay outs of the case. There are 8

grades involved and each grade has 4 stock bed positions available. Each stock bed is accessible by one of the 4 stacker-reclaimers. At first sight this is a logic lay-out because two trains and two ships can be served simultaneously. This is the case in Figure 4.
A disadvantage however is the fragmentation of the stacking area.

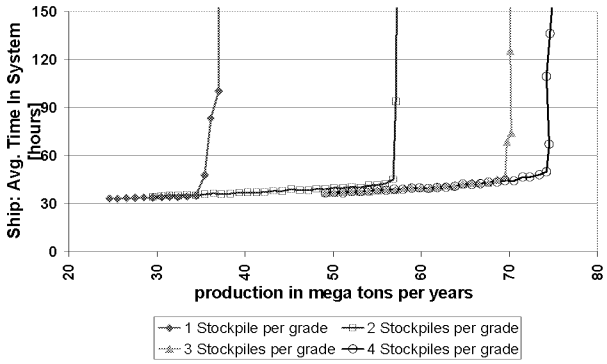


Figure 5: Results of Test Runs Varying the Number of Stock Beds per Material Grade

This lay out is compared with 3 other lay outs having 3, 2 and 1 stock beds per grade respectively. In Figure 5 the results are shown. They are as could be expected. The situation with 4 stock beds per grade performs best in terms of maximum throughput. Still it could be decided to choose the lay out with 3 stock beds per grade because of more efficient stacking area utilization.

CONCLUSIONS AND FURTHER WORK

A simulation model is discussed that can be used for design and improving operational control of large bulk terminals. The model is configurable with respect to material flow patterns, stock bed lay-out, equipment used, equipment reliability and a number of operational control methods. Break down behavior has been modeled at individual equipment level. The structure and the working of the model are explained and an example case is discussed.

Further work will focus on the development of intelligent procedures and algorithms for operational control of bulk terminals and to derive a new design approach. All results will be evaluated dynamically with simulation and validated using real world data of existing bulk terminals. The project finally has to result in a generic approach and software tools for design and control of deep sea bulk terminals.

REFERENCES

Jagerman, David; Altiok, Tayfur. (2003) "Vessel arrival process and queueing in marine ports handling bulk materials". *Queueing Syst.* 45 (2003), no. 3, 223--243.

Lodewijks, G., (2001), "Two decades dynamics of belt conveyors", *proceedings of the BeltCon 11 conference,* Johannesburg, Republic of South Africa.

Lodewijks, G., (2003). "Strategies for Automated Maintenance of Belt Conveyor Systems", *proceedings of the BeltCon 12 conference*, June 23-24, 2003, Johannesburg, South Africa.

Ottjes J. A. and Lodewijks G (2004). "Reliability of Large Scale Conveyor Systems", *Proceedings of the Industrial Simulation Conference 2004* (ISC2004) PP:324-329. June 2004. Malaga, Spain. ISBN: 90-77381-12-0

Spanke, M (2000). "Comparison of different types of ship unloaders with the focus on continuous technology". *Bulk Solids Handling,* vol. 20 Nr 4.

Veeke, Hans P.M., Jaap A. Ottjes, 2000. "Tomas: Tool for Object-oriented Modelling And Simulation". In *proceedings of Advanced Simulation Technology Conference (ASTC2000).* April 16-20, 2000, Washington, D.C. pp. 76-81. The Society for Computer Simulation International (SCS), ISBN: 1-56555-199-0

Wöhlbier, R.H. (1987), "The best of bulk solids handling, selection articles 1981-1985", *Bulk port Development & Operation.*

Zeigler B.P., Praehofer H and Kim T.G., 2000. "Theory of Modelling and Simulation 2nd Ed. Academic Press, San Diego.

Internet Link 1: **http://www.fam.cl/english/Products/Stockyard%2520systems/Stacker-reclaimers/detail.video.99999999997918.0.html**

Internet Link 2: **www.tomasweb.com**

SIMULATION OF A HOLDING UNIT BETWEEN FIRST AID CENTER AND SPECIALTY WARDS: A CASE STUDY

A.C. van Bruggen, H.P.M. Veeke, G. Lodewijks
Delft University of Technology, Faculty of 3mE
F.K. van Lambalgen
Leiden University Medical Center
e-mail: *H.P.M.Veeke@tudelft.nl*

KEYWORDS
Health care, Resource management, Discrete simulation

ABSTRACT

The Leiden University Medical Center wants to increase its productivity. One suggestion is a temporary separation of acute and planned patients by means of a holding unit between the First Aid Center and the specialty wards. Acute patients can wait in the holding unit for a free bed in the wards and some can already be released from it after treatment. A simulation model shows that a holding unit indeed increases the productivity due to a scale advantage in the holding unit and increased bed occupancies in the wards. Moreover, it also leads to an improved quality of care and an improved controllability of the system

INTRODUCTION

Health institutes in the Netherlands have been cut back retrospectively on their government budget in order to stimulate an efficiency increase. Furthermore, in the near future there will be increasing market forces in the health care sector and also an increase in the quest for health care while the available personnel is decreasing. These developments force a productivity increase at the Leiden University Medical Center (LUMC). One of the hypotheses is that a further separation between elective (planned) and acute care will lead to a higher utilization of the available bed capacity. More specifically, it is suggested to have a temporary separation between acute and elective admission patients. Instead of going directly to the ward after visiting the First Aid Center, acute patients will then stay in a holding unit for a while. In this holding unit the diagnosis for the patient should be determined and the treatment should also already be started.

Several other hospitals make use of similar units but the objectives are most often different from the objective in this case. Some objectives that are often mentioned are to relieve the overcrowded First Aid Center or to make sure acute patients are diagnosed within a certain amount of time [e.g. 1]. However, the transfer of acute patients to the specialty wards is found to be a bottleneck and a good transfer system is thus seen as one of the key prerequisites for a good functioning of a holding unit [2-6]. In this research project the focus is especially on the interaction between the holding unit and the wards because the main objective of the holding unit at the LUMC is to have a productivity increase in the clinical capacity of the hospital. Furthermore, there are also large differences between the hospitals regarding the maximum length of stay of patients in the unit. This question is also addressed in this research project.

APPROACH

In order to investigate the hypothesis first the current processes are analyzed with the help of a systems approach [7, 8]. From this analysis it is determined what the shortcomings are of the current situation. Then it is investigated if a temporary separation between acute and elective patients by means of a holding unit can help to overcome these shortcomings and especially if it will lead to a productivity increase. A computer simulation is made with the simulation package TOMAS [10] based on Delphi to give a quantitative judgment of the advantages of a holding unit compared to the current situation.

ANALYSIS CURRENT SITUATION

In the current situation the acute patients first visit the First Aid Center of the LUMC. There it is determined whether an admission in the hospital is needed. If this is the case the patient will be transferred to the ward of the

concerning specialty or the patient will be transferred to another nearby hospital. On the specialty wards the acute and elective patients come together.
For every acute admission patient it is tried to find an admission spot in the LUMC but this is not always possible. Therefore, part of the acute admission patients has to be transferred to another hospital. Because the LUMC is an academic hospital it should at least be able to admit the acute patients who have a complex need of care. However, it also occurs that acute patients who actually need an admission in an academic hospital have to be transferred. These are called forced transfers. Therefore, more choices should be made regarding which patients should be admitted to the LUMC and which patients could just as well be transferred to a nearby non-academic hospital.

To find out how at the wards the inflow of acute patients is dealt with, the processes of several specialties have been further analyzed. Four specialties were chosen for this analysis, which all have a high number of acute patients. It happens to be that there are large differences between these specialties in the way they handle the inflow of acute patients. For example, one specialty uses all kinds of flexibility to be able to create an admission spot for an acute patient, among other things by placing a patient at the ward of another specialty or by calling off an elective patient. Another example is a specialty that does not have many possibilities for flexibility and therefore uses some overcapacity to be able to cope with the peaks in the inflow of acute patients.
However, overcapacity leads to a decrease in productivity while a high amount of flexibility in the current situation leads to a decrease in the overall quality of care as well as in the controllability of the system. So both an overcapacity and a high flexibility lead to unwanted side effects.
The question is now, can a holding unit lead to a productivity increase while at the same time contributing to overcome the unwanted side effects in the current situation? The proposed function structure in the LUMC with a holding unit is presented in Figure 1.

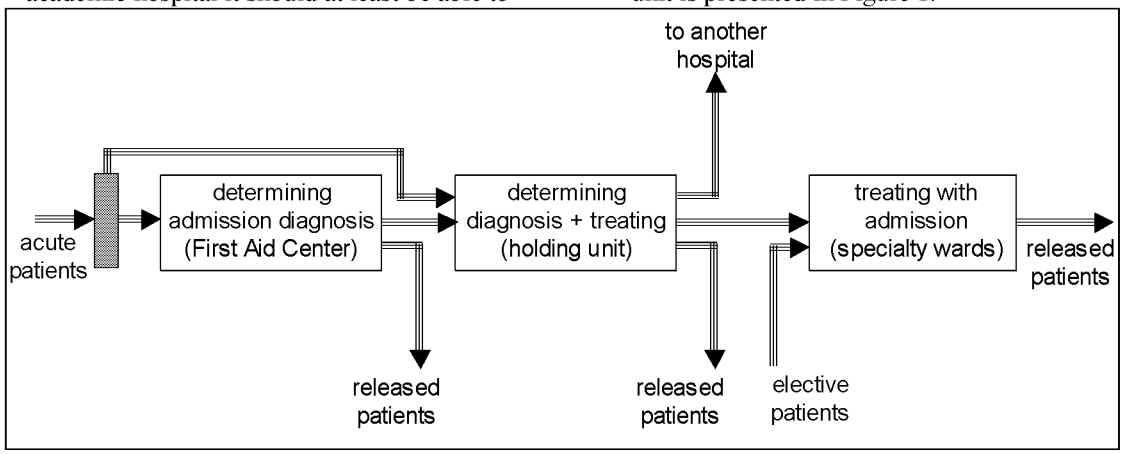


Figure 1: proposed function structure

As explained before most acute patients in the current situation first visit the First Aid Center after which they are transferred to a specialty ward if necessary. With the holding unit there will be one step in between. Upon arrival, part of the acute patients can be sent immediately to the holding unit if it is clear that the patient needs an admission. This will not be clear for each patient at the start so other patients first go to the First Aid Center and from there possibly to the holding unit and to the ward.

The holding unit would have two main functions as seen from the perspective of the wards, firstly filtering and secondly buffering. Filtering means that acute patients with a short length of stay can be diagnosed and treated in the holding unit and be released after treatment. These patients are thus kept away from the wards. For example, approximately one quarter of the acute patients is released on the same day as they were admitted. Buffering means that acute patients with a higher length of stay can stay in the holding unit until a bed is available at the ward.

A productivity increase with a holding unit is expected because of two reasons. Firstly, with a holding unit the number of acute patients at the wards will decrease and the length of stay of the acute patients at the wards will also decrease. Therefore, the bed capacity at the

wards can be decreased and this capacity can be used in the holding unit. Because of a scale advantage relatively less capacity is then needed in the holding unit. Secondly, the bed occupancy at the wards is expected to increase with a holding unit because acute patients can wait in the holding unit for a bed at the ward.

The non-occupied beds at the wards can then easily be filled with these acute patients.
With a holding unit the acute admission patients actually become semi-elective patients because there is some time available to plan a bed for the acute patient at the ward. A holding unit in combination with a variable planning horizon for elective patients is therefore expected to lead to an even higher productivity. Instead of planning all the elective patients one week in advance some beds are kept free in the bed planning for possible acute patients. If it turns out that acute patients will not be occupying these beds it is then possible to call up some elective patients one day in advance. This system would then lead to less forced transfers of acute patients in combination with a high bed occupancy at the wards. One of the specialties already has positive experiences with such a variable planning horizon.

OBJECTIVE OF SIMULATION

To find out if the above-mentioned presumptions about the holding unit are true a computer simulation is made with the simulation package TOMAS based on Delphi. Two main questions will have to be answered with the simulation. The first question is whether the holding unit will indeed lead to a productivity increase, with and without a variable planning horizon. If this turns out to be true the second question is which holding time will be most advantageous. The holding time is the maximum length of stay of acute patients in the holding unit. Similar units in other hospitals mostly have a maximum length of stay varying from a few hours to two days. Therefore it will be useful to see if the simulation results can lead to a justified choice for the holding time.

OUTCOMES AND KEY PERFORMANCE INDICATORS

For several holding times it is determined with the simulation program if a productivity increase can be accomplished. That is, having the same amount of admissions as in the current situation only with less means. Therefore, the following outcomes are needed for this:

- The number of admissions.
- The number of means. The number of beds is used as a standard for the number of means. Both the beds at the wards and the beds in the holding unit get a weighting factor that indicates the number of nursing capacity needed per bed. For a holding bed this is two times as much as for a bed at the ward because of the higher level of care that is needed in the holding unit.

Besides the productivity increase, the holding unit is also expected to lead to an increase in the quality of care. The following outcomes are relevant for measuring the quality of care:

- The number of forced transfers. These are the patients for whom no bed is available in the LUMC and who have to be transferred to another hospital in the region. In the current situation the number of forced transfers as a percentage of the total number of acute admission patients is on average 15%. In the simulation this percentage will be set at a maximum of 3%.
- The number of elective patients who have to be rescheduled because there is no bed available for them at their admission date as a result of disturbances in the bed planning. This number as a percentage of the total number of elective patients is also set at a maximum of 3%.

The above two Key Performance Indicators are used to determine the number of beds needed at the wards. To determine the number of beds that is needed in the holding unit the following Key Performance Indicator is needed:

- The average waiting time of acute patients for a bed in the holding unit. This average waiting time is set at a maximum of half an hour.

The variables in the simulation are the number of beds at the wards, the number of beds in the holding unit and the holding time. The way each simulation run is executed will be explained later.

PROCESS DESCRIPTION

Figure 2 displays the patient flows in the simulation. In this figure only one ward is displayed but in the simulation for each of the four considered specialties one ward is used.

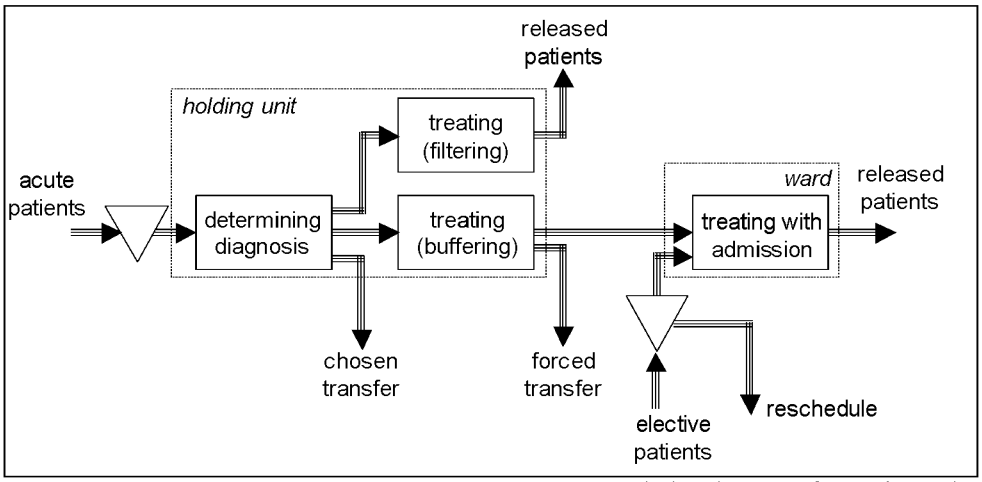


Figure 2: simulation description

The flowing elements in the simulation are the acute patients and the elective patients. All patients get an estimated length of stay on the basis of a distribution. They also get a real length of stay that is determined by adding or subtracting a correction to the estimated length of stay. This correction is also based on a distribution. The elements that are executing a process are as follows:

- Holding unit:
  Acute patients enter the holding unit according to an arrival pattern and they possibly have to wait until a bed is available in the holding unit. After determining the diagnosis, which takes a certain amount of time, for some of the patients it is chosen to transfer them to another hospital. For the other patients it is determined whether the estimated length of stay of the patient is smaller than the holding time. In that case the patient can be treated in the holding unit and be released from it (filtering). The other patients have to wait in the holding unit for a free bed at the ward while they are already treated (buffering). The Bed planning of the ward is informed that a bed has to be found for the acute patient. If no bed becomes available before the patient exceeds its maximum length of stay in the holding unit the patient has to be transferred to another hospital. This patient then has a forced transfer.
- Intake planning for elective patients (one for each ward):
  Every day a number of elective patients is planned and their admission date is set seven days later.
- Bed planning (one for each ward):
  For elective patients the bed at the ward is searched for that will become available as late as possible before the admission time of the patient. For acute patients it is first determined if there is already a bed available at the ward and if not the bed is searched for that will become available as soon as possible. If a bed is found it is reserved for the patient.
- Ward coordination (one for each ward)
  This process will release patients from the wards and admit acute or elective patients at the ward. Moreover, there may also be disturbances in the planning when the real length of stay of a patient is longer than its estimated length of stay. In that case it may be that other patients have to be rescheduled because their reserved bed will become available too late.

VALIDATION

Validation of the simulation model was done by a conceptual model validation, data validation and an operational validation [9]. The most important assumptions and simplifications in the simulation model are as follows:

- One ward per specialty. In reality most specialties have more wards.
- Constant number of beds at the wards. In reality this number may vary for example in holiday periods or because of ill personnel.
- The absence of flexibility. In reality patients can for example be released earlier or acute patients are placed temporary on another ward. In the simulation model this is not possible.

- A five-day working week. The influence of weekends is not accounted for.
- For the operational validation the holding time is set at 3 hours, which simulates the current situation. The percentage of forced transfers is taken as the indicator for the correctness of the model. With the current number of beds and the current number of patients this percentage would have to be equal to the real percentage of forced transfers. The operational validation is based on historical figures of both 2003 and 2004. The outcomes of this validation are shown in Table 1.

1. Make a simulation run with a high number of beds for both the holding unit and the wards.
2. Decrease the number of beds in the holding unit step by step until the average waiting time for acute patients reaches the maximum value of half an hour.
3. Decrease the number of beds at the wards step by step until one of the Key Performance Indicators reaches its maximum value.
4. If needed iterate step 2 and 3 until both the holding unit and the wards reach one of its maximum values for the Key Performance Indicators.

	Specialty 1 (23 beds)		Specialty 2 (21 beds)		Specialty 3 (30 beds)		Specialty 4 (45 beds)	
	2003	2004	2003	2004	2003	2004	2003	2004
Real percentage	8%	7%	26%	23%	0%	0%	15%	9%
Simulated percentage	8%	6%	27%	23%	2%	7%	10%	9%

Table 1: operational validation

The validation results show that the simulation model is performing well for specialty 1 and 2. Specialty 3 has a slight deviation from reality. In reality specialty 3 uses multiple wards, a variable planning horizon and there are special beds for acute patients. A similar validation for specialty 4 was not possible because specialty 4 often uses beds from other specialties to admit acute patients. Therefore, the current number of beds cannot be determined. It is only known that specialty 4 normally has 38 beds. The validation for this specialty is therefore done the other way around. It is thus determined what number of beds is needed in order to have the real percentage of forced transfers. Considering that specialty 4 often uses beds from other specialties the outcome of 45 beds seems acceptable as compared to the 38 beds that specialty 4 normally has of its own.

SIMULATION PLAN

For several holding times it is determined what number of means is needed and the number of beds is used as a standard for this. For each holding time the following steps should be taken:

First a reference run was made with the simulation model to determine the number of means in the current situation, which was found to be equal to 91. Although there is no holding unit in the current situation, the current situation can still be simulated with the model by setting the holding time at a maximum of only 3 hours as was just explained with the operational validation.
The main results of the simulation are shown in Figure 3. In each simulation run the number of elective and acute patients is the same as in the current situation.

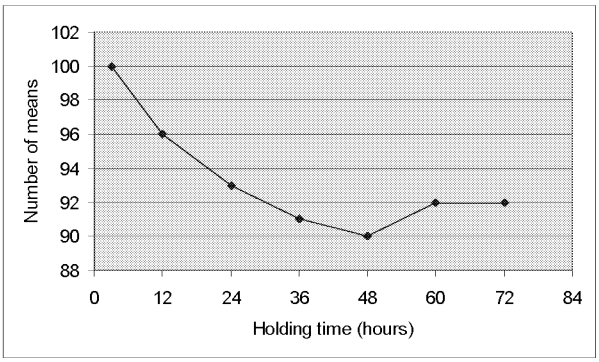


Figure 3: simulation results

The higher the holding time the more acute patients can be treated and be released from the

holding unit. The scale advantage in the holding unit will therefore increase with increasing holding times. When the holding time increases it will also be easier to find a bed at the ward for an acute patient because the patient can stay longer in the holding unit. Therefore the bed occupancy at the wards can increase. Until a holding time of 48 hours those two effects lead to a decrease in the total number of means that is needed. For higher holding times there is a slight increase in the number of means. One of the reasons for this is that the bed occupancy at the wards is optimal for a holding time of 48 hours and decreases again with higher holding times. This decrease in bed occupancy is caused by the fact that the number of acute patients that has to go to the ward decreases so therefore, there are fewer possibilities to fill up the free beds at the wards.

As can be seen from Figure 3 the minimum number of means is with a holding time of maximum 48 hours, where the number is equal to 90. Compared with the reference run, where the number of means was 91, it seems like there is no productivity increase with the holding unit. However, the reference run concerns the current situation with approximately 15% forced transfers of acute patients to other hospitals while in the simulation model this percentage is reduced to 3%. This reduction was introduced in the simulation model because less forced transfers will lead to an increase in the overall quality of care. More acute patients are thus admitted with almost the same amount of means. This is a productivity increase but not as is wanted by the LUMC because they want to have the same number of admissions as in the current situation only with less means. In the analysis of the current situation it was mentioned that more choices should be made regarding which patients to admit in the LUMC and which patients to transfer to another hospital. Therefore a scenario is considered with the simulation model with a higher number of these chosen transfers. The results are shown in Figure 4 together with the first results.

This figure shows that with less means it is indeed possible to have the same number of admissions as in the current situation. Several sensitivity analyses with the simulation model have further shown that these results are reliable.

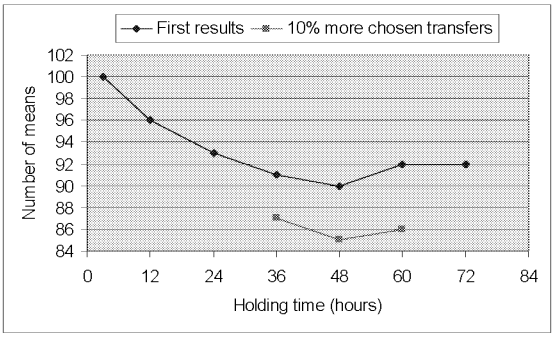


Figure 4: simulation results with more chosen transfers

The results in Figure 3 and Figure 4 are based on the historical figures for the year 2004. It follows from these results that the optimum holding time is 48 hours. To see if this is a reliable outcome the simulation results for 2003 are also determined. These are presented in Figure 5.

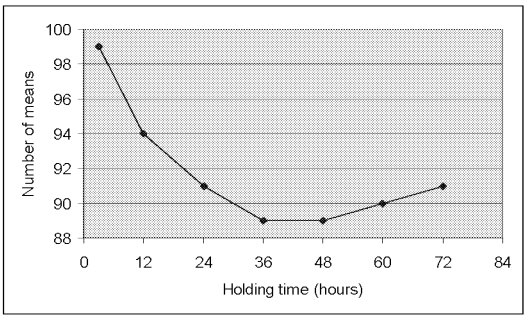


Figure 5: simulation results for the year 2003

The simulation results for the year 2003 are similar with the results for 2004. The optimum holding time is also around 48 hours.
It was suggested before to set up the holding unit in combination with a variable planning horizon for elective patients. Depending on the number of available beds extra elective patients can be called up for the next day. The expected productivity increase with this variable planning horizon is also investigated with the simulation model.
The results are shown in Figure 6 for the optimum holding time of 48 hours. The number of patients shown in the legend of the figure indicates how many patients at the most are called up for the next day per specialty.

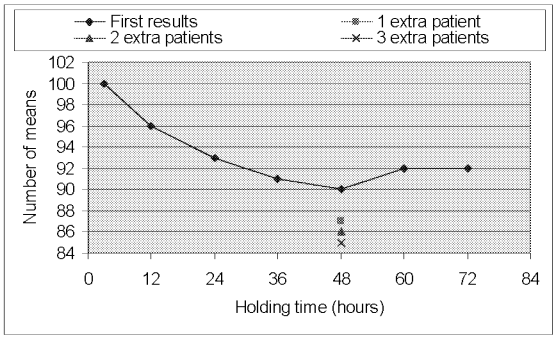


Figure 6: simulation results with a variable planning horizon

Figure 6 shows that a variable planning horizon can indeed further increase the productivity. This productivity increase is caused by higher bed occupancies at the wards.

CONCLUSIONS

It can be concluded that a further separation between elective and acute care will lead to a higher utilization of the available clinical capacity. The simulation has shown that with a separation of maximum 48 hours of acute and elective patients by means of a holding unit less clinical capacity is needed to admit the current number of patients in the LUMC. This is firstly because of a scale advantage in the holding unit and secondly because the bed occupancy at the wards can be increased. A variable planning horizon might further contribute to a productivity increase. Moreover, a holding unit will lead to an increase in the quality of care because of less forced transfers to other hospitals, less elective patients to be called off and less transfers of acute patients between the wards. Finally, the overall controllability of the system will increase compared to the current situation.
On the basis of this research it was therefore recommended to the LUMC to set up a holding unit with a maximum length of stay of two days. Furthermore, it was recommended to have directives for each specialty indicating which acute patients to admit in the LUMC and which patients to transfer to a regional non-academic hospital. Ideally, these directives should also be part of an agreement system between the hospitals in the region. Regarding the variable planning horizon it was recommended to do some further research into the possibilities for an introduction of this measure. This will depend mainly on the willingness of the patients to be called up one day in advance and on the possibility to plan other processes in the same time.

References

1. Davies, M., 2003, Proposal to Establish a new Medical Assessment Unit at the Royal United Hospital Bath
2. Alberti, G., 2003, Emergency Assessment Units – a Checklist, Department of Health UK
3. Brennan, R., 2004, Medical Assessment & Medical Admission Units, The Irish College of General Practitioners
4. Clinical Epidemiology & Health Service Evaluation Unit, 2004, Models of Care to Optimise Acute Length of Stay, Melbourne Health
5. Huang, X.-M., 1998, Decision making support in reshaping hospital medical services, Health Care Management Science 1, 165-173
6. Levy, J., 2001, Medical Assessment & Planning Unit (MAPU) – Outcomes, Royal Brisbane Hospital, Proceedings of the National Demonstration Hospitals Program Phase 3 Conference
7. Bikker, H., 2001, Analysis and Design of Production Organizations, Lecture Notes, Delft University of Technology
8. Veld, J. in 't, 2002, Analyse van Organisatieproblemen, een Toepassing van denken in Systemen en Processen, 8th edition, Stenfert Kroese, ISBN 90-207-3065-7
9. Sargent, R.G. 2000. Verification, Validation, and Accreditation of Simulation Models, Proceedings of the 2000 Winter Simulation Conference, 50-59.
10. Veeke, H.P.M., Ottjes, J.A., 2000, "TOMAS: Tool for Object oriented Modeling And Simulation",Proceedings of Advanced Simulation Technology ASTC 2000, pp. 76-81, WashingtonDC, SCS International

VR IN DESIGN AND MANUFACTURING

PRODUCT DESIGN WITH VR

VIRTUAL ERGONOMICS ANALYSIS TO VALIDATE PRODUCT DESIGN

Giorgio Colombo
Dipartimento di Ingegneria Meccanica
Politecnico di Milano
Via La Masa, 34
I 20156 Milano
Italy
E-mail: giorgio.colombo@polimi.it

Demis Bronco, Caterina Rizzi and Daniele Regazzoni
Dipartimento di Ingegneria Industriale
Università di Bergamo
Viale G. Marconi, n. 5
I 24044 Dal mine (BG)
Italy
E-mail: caterina.rizzi@unibg.it

KEYWORDS
Ergonomics, virtual manikin, internal and external visibility.

ABSTRACT

This paper presents an approach to evaluate product ergonomics using virtual humans. This work has been carried out to demonstrate that virtual humans are an important tool to improve virtual prototyping functionalities and, above all, to increase ergonomics and safety of products. A test methodology has been defined and several simulations have been made, varying anthropometry and workplace conditions. Two case studies that can demonstrate the validity of the approach are described. The first concerns instruments of a car for which the geometric model of the interior and a commercial virtual human have been used. The second case study concerns the external visibility of a farm tractor that can be equipped with different tools to accomplish to various tasks (e.g. harvesting and fruit picking). The results obtained are images that represent what the virtual human sees and permit to compare and validate different design solutions.

INTRODUCTION

During last years, virtual prototyping (Lee K., 1999) approach has become more and more important within the product development process; in fact, it is a well known technology that allows designers to realise the digital mock-up of a product. In this context, simulation techniques permit to reproduce machines and devices motion and analyse static and dynamic stresses and strains. Models and tools to simulate motion and tasks done by human body are also available.
In particular, the development of realistic models for human body is attracting more and more interest in different domains. For example, in the industrial context, applications are various and important: ergonomics in product design and maintenance, tasks simulation, personnel training, etc.
The first applications of human models started in the '80s in the movie domain; application in ergonomics were immediately subsequent, as demonstrated by research activities described in (Karwowski et al. 1990) (Case et al. 1990) (Feyen et al. 2000) (Berger et al. 2004) (Sundin and Örtengren 2004).
A virtual human is a kinematic chain composed by a number of rigid links connected by joints; each virtual joint reproduces d.o.f. of the physical one, with limits corresponding to the allowable motion of a human being. Algorithms for direct and inverse kinematic have been developed so that virtual model can replicate human body movements; they are derived from those applied in multibody analysis (Shabana 1994).
A virtual human may have virtual senses, such as sight, and can interact with the surrounding environment.
Nowadays, several frameworks with human models of different complexity are available, such as Ramsis (www.human-solution.com) ManniquinPRO™ (www.nexgenergo.com/ergonomics/mqpro.html), *Jack* (http://www.ugs.com/products/tecnomatix/human_performance/jack/); they can be used to define complex scenes with virtual manikins and objects and to simulate many tasks.
In this context, we focused the attention on virtual human modelling as a tool to integrate virtual prototyping and to carry out ergonomics analysis. To model virtual humans and realise simulations we adopted the commercial software package *Jack* (Badler et al. 1993).
Two case studies have been considered, both requiring the simulation of sight. The first is centred on the analysis of internal visibility of a car, the second on the external visibility of a farm tractor equipped with a loading shovel.

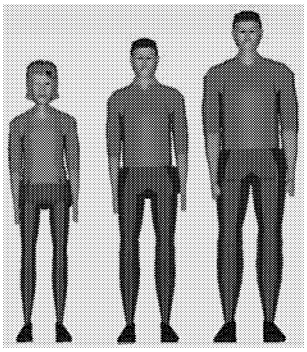
OBJECTIVES AND METHODOLOGY

The objectives of the work have been, first to define a methodology for mentioned ergonomics analyses in a virtual environment, and then to demonstrate the usefulness of human body models in virtual prototyping activities for products ergonomics evaluation.
According to the purpose of the ergonomics analysis and the Jack software package, following five main steps have been identified:

1. **Definition of the virtual scene**. During this phase the designer identifies the product parts involved in the ergonomics analysis and eliminates the others so that the simplified geometry of the virtual product allows reducing computational times. Then he selects the mobile parts (e.g., driver's seat) and those ones that can be considered fixed (e.g., window). Additionally, reference elements, like a grid, can be defined to better evaluate simulation results.
2. **Choice of virtual manikin model.** The user selects virtual models of different sizes from the anthropometrics databases integrated within the system. For our purpose we considered three virtual humans corresponding to 5%ile (female), 50%ile (male) and 95%ile (male) percentile according to SAE (Society of Automotive Engineers) standards (Fig. 1) since both

test-cases consider it as reference. Besides, the user can freely define specific manikins.

3. **Insert manikin in the virtual testing environment**. First the manikin posture is defined according to the ergonomics standard considered, the SAE standard for this work, which provides for each joint of the virtual human the range of angular values to assure a comfortable position (Fig.2). Then it is located in the right position within the virtual scene.
4. **Plan and prepare ergonomics analysis**. According to the goal of ergonomics analysis to carry out (e.g. internal or external visibility), the user plans and assigns the tasks to the virtual human.
5. **Run simulation and evaluate results.** During this step, the designer executes the analysis planned in the previous step and results obtained for each virtual human are evaluated and compared among them and with current national and international standards.



%ile	Height (cm)	Weight (kg)
5	160,27	55,27
50	175,49	77,69
95	186,65	98,07

Figure 1: 5%, 50% and 95% percentiles

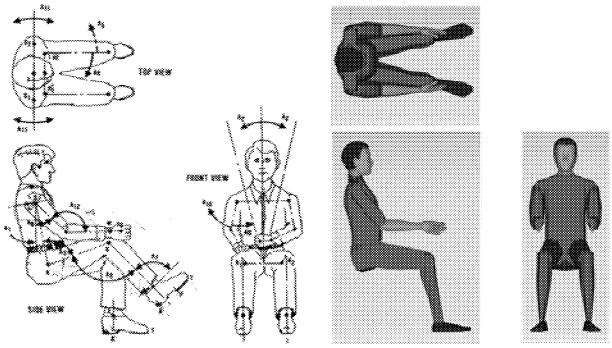


Figure 2: SAE Postures and related manikin

In the following we describe the results obtained for the two case studies.

INSTRUMENTS VISIBILITY FOR CAR INTERIOR DESIGN

The first study case is related to automotive domain and concerns car interior design. A specific ergonomic aspect in car interior design is the visibility that can be further subdivided into more sub-problems: external visibility through the windscreen, windows and mirrors, and the visibility of instruments and controls.

As far as concerns external visibility, many studies define the requirements that must be satisfied (Woodson et al. 1992). National and international standards establish the field of visibility and the limits for the impediments due to door pillars and other car parts.

For internal visibility, the ergonomic analyses concern those instruments requiring sight sense of the virtual human.

Figure 3 shows the model of car interior used to perform the simulation. The interior virtual model has been recovered from a design database and can be imported in *Jack* using a proper data exchange format (e.g., IGES, STL or VRML).

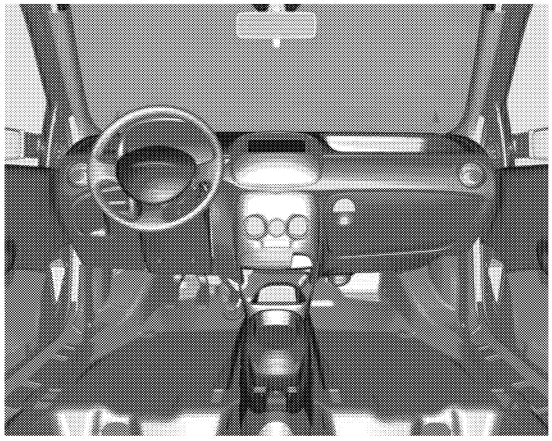


Figure 3: Virtual prototype of the car interior

The virtual driver is in a normal driving posture, with the right foot leaned upon the accelerator pedal and with hands on the steering wheel (Fig. 4).

Figure 4: Virtual human at the driver's seat

The analyses to verify instruments visibility, in particular the panel behind the steering wheel, have been conducted following a procedure that considers, for a given size of the manikin, possible adjustments for the seat and the steering wheel. In detail, for each wheel position, the seat is moved between furthermost positions, and for each defined positions we can control how the manikin sees the equipment and the panel.

The sight direction is fixed towards the instruments, by imposing the head position. In this way, when the seat is moved, the virtual human moves only the eyes to stare at the same point. *Jack* has not got a binocular view; therefore we can verify only what the right and left eyes see separately.

A lot of tests have been carried out; images showing what the manikin sees have been acquired for different virtual humans (from 50%ile to 95%ile) and adjustments of driver workstation (three positions for steering wheel and five for seat). Figure 5 portrays a sequence of images related to tests with the 95%ile manikin and for specified adjustments of the driver's seat and steering wheel at low position.

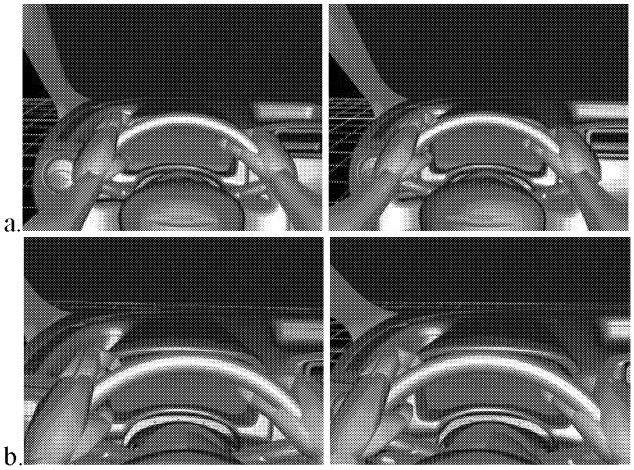


Figure 5: Instruments visibility for a 95%ile: a) seat completely outstanding, b) seat completely put forward. Left and right images show the shot of corresponding eye

EXTERNAL VISIBILITY FOR A FARM TRACTOR

This case study is related to a farm tractor specifically designed for vineyard and orchard. This type of tractor usually operates in narrow environment and is equipped with specific appliances. A specific ergonomic aspect in tractor interior consists in the external visibility of mounted equipment, in our case a loading shovel, through the front window and high-visibility roof. Figure 6 shows the simplified model of the tractor (parts not related to the simulation, such as engine, bolts and nuts have been not considered), the manikin and three different positions of the loading shovel (identified by different colours). In addition, a horizontal grid has been defined to evaluate quantitatively the visible area of the equipment.

The posture of the manikin has been defined according to SAE standards (Fig. 3). The wheel has been considered fixed while the seat can be moved horizontally along the X axis (-70mm, +80mm) and vertically along the Z axis (-50mm, +50mm). For each given manikin size, the seat is moved between furthermost positions in both directions for a total of nine different positions (Fig. 7).

According to the goal, different simulations have been carried out in order to analyse:

(1) The visibility of the loading shovel through the front window with the virtual human sight fixed to the horizontal direction and imposing the head stationary;
(2) The visibility of the loading shovel through the high visibility roof;
(3) The visibility through the front window of the loading shovel located at the lowest position (ground) with the virtual human looking down and imposing the head stationary.

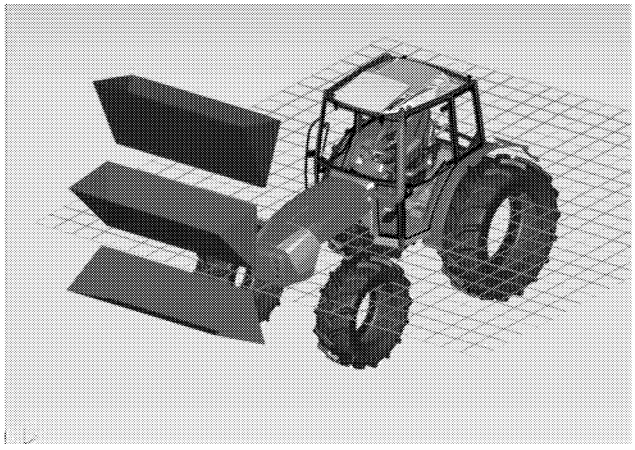


Figure 6: Virtual prototype of tractor

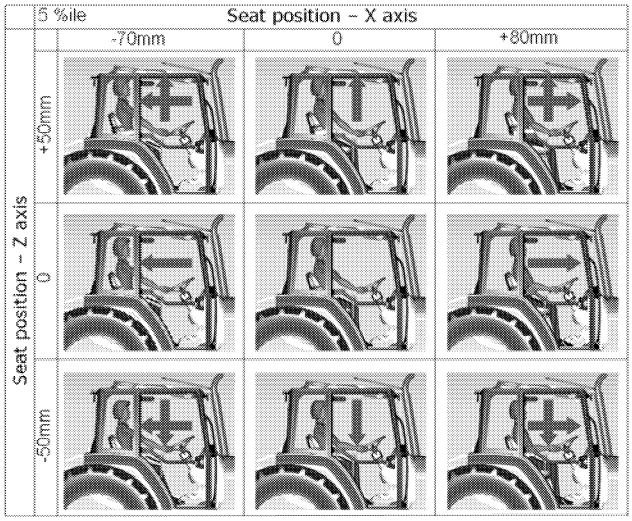


Figure 7: Seat Position for each manikin size

The case (1) has been performed in two steps. First we verified the optimal position for each manikin size, then, using these positions, we identified the meaningful positions of the loading shovel corresponding respectively to: the angle at which it begins to be visible, to be hidden by the shade curtain or it totally disappears, and so on. For instance, Fig. 8 shows the results obtained for a 50%ile for which the central position seems to be the best one, while Fig. 9 shows a comparison between the different manikin sizes for an X position of the seat equal to 0 and moving it along Z-axis direction.

For the case (2) the goal has been to verify the influence of under-roof geometry while looking through the high-visibility roof. Fig. 10 portrays for the optimal position of each manikin size, the results of the analysis carried out with and without the under-roof part.

For the last case (3) for each optimal position we considered the driver looking down along the central line and on the left corner. Fig. 11 portrays the results for the second situation.

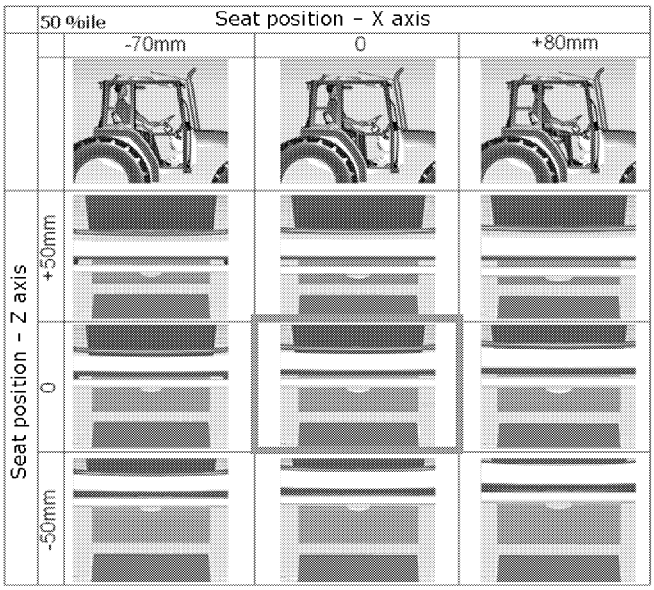


Figure 8: Frontal visibility for 50%ile

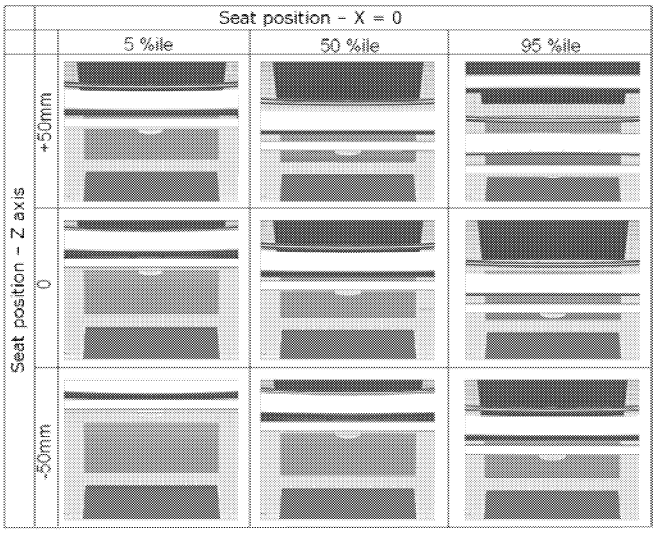


Figure 9: Comparison among manikin sizes with X=0 and Z variable

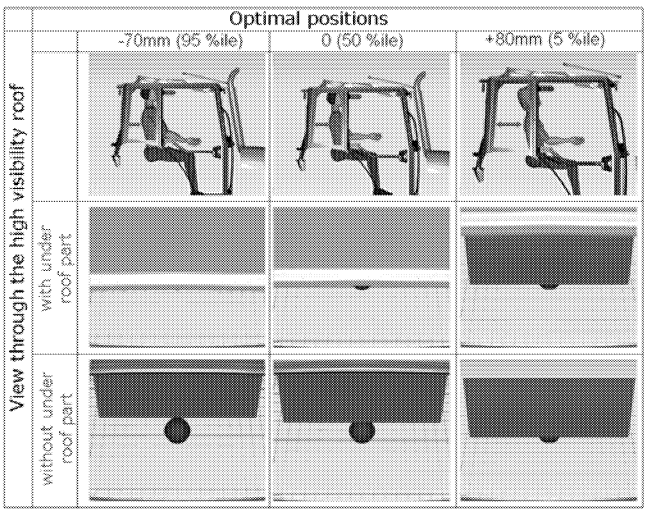


Figure 10: Influence of the under-roof part

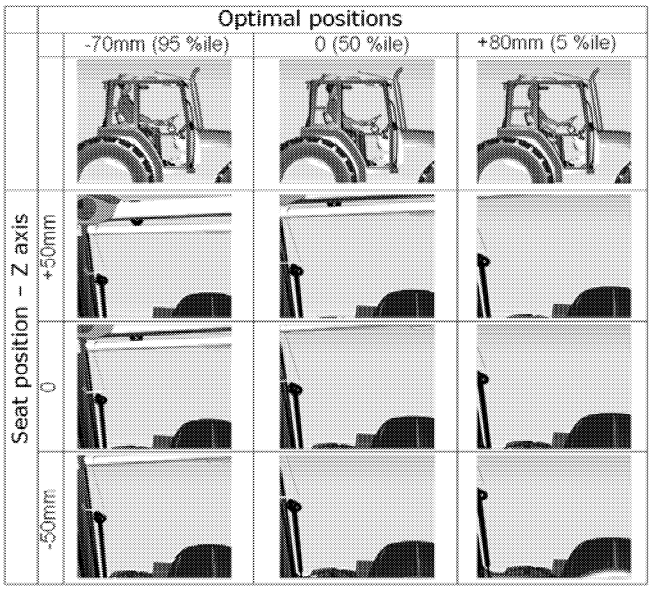


Figure 11: Views looking down toward the left corner.

CONCLUSIONS

This paper has focused the attention on a methodological approach based on the use of virtual humans to improve product ergonomics and safety and their applicability in an industrial context. Two case studies have been considered: one regards the car interior design, precisely the control panel visibility, while the other one the external visibility of a farm tractor. As far as concerns results evaluation, we have not defined a methodology to analyze images since, as already said, this study was mainly devoted to verify the feasibility of virtual analysis. Our objective has been primarily methodological and relies on the accommodation of human considerations in supporting design process. The definition of the virtual environment and the planning of the simulations are not particularly complex. On the other hand, it can be a little more difficult defining the correct position and posture of the virtual human and this can lead to wrong results. Anyway, we think that, in general, problems concerning visibility analysis can be faced using human models and performing simulations analogous to those described. Some examples can be typical problems related to vehicles and public buildings, such as theatres, cinema and stadium.

Future activities have been already planned. First a detailed experimental analysis of tractor driver's behaviour and movements will be performed using also body and eye tracking systems. Then, testing protocol will be defined and tests carried out considering different types of tractor.

ACKNOWLEDGEMENTS

The authors would like to thank STOLA and SAME FAHR Duetz companies, in particular Ing. Bruno Paggi, for the study cases proposed, and M. Culubret, G. Tovar for their collaboration.

REFERENCES

Badler, N.; C. Phillips and B. Webber. 1993. *Simulating Humans: Computer Graphics Animation and Control*. New York, Oxford University Press.

Berger, U.; R. Lepratti and H. Otte. 2004. "Application of digital human modelling concepts for automotive production". In *Proceedings of the TMCE 2004* (Lausanne-CH), 365-373.

Case, K.; J.M. Porter and M.C. Bonney. 1990. "SAMMIE: a man and workplace modelling system". *Computer Aided Ergonomics*, 31-56.

Feyen, R.; Y. Liu; D. Chaffin; G. Jemmerson and B. Joseph, B. 2000. "Computer-aided ergonomics: a case study of incorporating ergonomics analyses into workplace design". *Applied Ergonomics*, No. 31, 291-300.

Karwowski, W.; A.M. Genaidy and S.S. Asfour. 1990. *Computer Aided Ergonomics*. Taylor & Francis.

Lee, K. 1999. *Principles of CAD/CAM/CAE Systems*. Addison-Wesley.

Shabana, A.A. 1994. *Computational Dynamics*. Wiley-Interscience.

Sundin, A. and R. Örtengren. 2004. *Digital human modeling for CAE applications*. Chap. 39, Handbook of human factors and ergonomics. Edited by G. Salvendy.

Woodson, W. E.; B. Tillman and P. Tillman. 1992. *Human Factors Design Handbook*. 2nd Ed., McGraw-Hill Professional.

BIOGRAPHY

DEMIS BRONCO is a junior mechanical engineer who has worked in the field of Virtual Prototyping and Ergonomics applied to automotive industry. Currently is employed as mechanical designer of NC machines.

GIORGIO COLOMBO is an associate professor at Politecnico di Milano, Italy. His main research areas concern: parametric modelling, Knowledge Based Engineering, Simulation techniques, Computer graphics and Virtual Reality, Methodologies for Systematic Innovation and Robotics. He participated to international projects and coordinated national projects on the Virtual Reality and haptic system for ergonomics validation. He is author and co-author of more than 70 publications.

CATERINA RIZZI is full professor at the Faculty of Engineering, Università degli Studi di Bergamo, Italy. Her main research area concern: Virtual prototyping, Physics based modelling and simulation of non rigid products, Methodologies and tools for systematic innovation, BPR methodologies and ICT tools for product development processes, PDM/PLM systems to manage the product life-cycle. She participated and coordinated national and international projects in the area of product design and systematic innovation applications. She is author and co-author of more than 120 publications.

DANIELE REGAZZONI is researcher at the Faculty of Engineering, Università degli Studi di Bergamo. His main research areas concern Systematic Innovation (mainly TRIZ and derived methods and tools), Intellectual Property Management, Modelling of highly deformable products and PDM/PLM solutions. He has participated to national projects of research and is author or co-author of more than 20 papers.

A Simulator of Agent-Based Household Appliances

Konrad Steblovnik
Gorenje d. d.
Partizanska 12, 3320 Velenje, Slovenia
konrad.steblovnik.point@gorenje.si

Damjan Zazula
University of Maribor, Faculty of EE and CS
Smetanova 17, 2000 Maribor, Slovenia
zazula@uni-mb.si

INTRODUCTION

Ambient intelligence [1,7], ubiquitous computing, intelligent multiagent systems, all combined with the Internet capabilities they bode significant changes of human dwelling environments in a very near future. At the same time, the idea of intelligent home has matured, embedded computer technologies have reached the price/performance ratio acceptable for wide consumer community, while fast communication infrastructure spread out to all levels of today's society. This is bringing forth a new type of consumers that expect the way of usage and operation of technological devices will draw as close to the natural as possible.

Consequently, new families of household appliances will have to follow these trends. In the first place, this applies to home entertainment facilities and the products of white-ware. Some of such high-end appliances can already be found on the market: washing machines, refrigerators, or ovens are controlled by 16-bit microprocessor technologies, touch-sensitive screens, and built-in speech-recognition interfaces [15]. Higher levels of integration and more computing performance are becoming imperatives. Trend-setting companies are developing embedded systems with 32-bit Arm RISC technology, along with more powerful image and speech processing algorithms.

These developments are going to mitigate the most frequent household activities, such as laundry, cooking, baking, refrigeration, etc. New appliances will show a certain degree of intelligence, they will be able to act autonomously, to plan and induce actions in order to achieve the requested goals, and even to learn. Communication with them is going to be user-friendly, mainly based on visual and speech acts in both directions. A variety of new sensors are assumed to enable this kind of devices to obtain a wide spectrum of crucial pieces of information, such as the quality of washing powder, water turbidity, the level of liquid in the refrigerated bottles, etc. The machine will be able to complement the user's knowledge and wishes with its technical aspects, so that any action will be optimized through an initial user-appliance dialogue and hand-shaking procedure.

At the same time, agent-based and multiagent theory [8,14,18,19] and computer-software technologies [6] have also achieved the stage, where different technical and technological solutions are successfully implemented [8,9] and convenient design methods [10] and execution engines [12] became available. The fundamental features of agents are aimed at autonomous, collaborative, and intelligent operation [20]. Thus, combined with powerful computer systems they can serve as a firm basis for the implementation of intelligent household appliances.

In the sequel, we are going to reveal how an agent-based simulator of intelligent household appliances can be constructed and used in the verification and validation procedures for specifications, design and implementation of such devices as industrial products. We are going to briefly survey the most important features of rational agents and their architecture, introduce selected design and implementation tools for building the multiagent systems, and exemplify our approach by a washing machine simulator.

DESIGN AND SIMULATION ENVIRONMENTS FOR AGENT-BASED SYSTEMS

From the generic point of view, an agent is supposed to be able to perform adapted and autonomous actions within a dynamic, unpredictable and open environment [14]. Considering today's computer systems, they certainly meet this definition. By looking in the reverse order, this means that a computer installation may qualify as an agent. Agents sense the stimuli from their environment and they react accordingly. They also are proactive, persistently pursuing their goals, they adapt to unpredictable situations, and they are robust in the sense they detect and try to correct their own faults and faults of their collaborative agents. Another important property of agents is their ability to interact and build up multiagent communities [20].

Yet, the multiagent systems need more human-like characteristics in order to interact in a rational and deliberative manner. They have to mimic mental states, such as to proceed from believes, have their own wishes and goals, and be intentionally oriented to achieving them. By attributing these kinds of states to the machine, we can describe it as a rational agent with *believes*, *desires*, and *intentions*. This defines it as the so called BDI agent or BDI agent model [13, 18, 20].

Multiagent systems

Multiagent systems most often control complex environments and perform a variety of rational acts. It is therefore of great importance to have a powerful design methodology at one's disposal. On the other hand, the specification and design phases are inseparably linked to the implementation phase. Therefore, choosing a methodology

support and design tools influences heavily the implementation decisions. As will be explained in the next section, we were faced with the problem of a Multiagent Washing Assistant (MWA) design and simulation. We wanted to construct a rational system which will be apt of simple induction and learning. The BDI agent architecture turned out to be the most suitable one. This decision was alleviated by finding several parallels and good matching between the Prometheus design methodology and the Jadex simulation environment.

Agent-oriented design methodologies

Many standardized approaches exist in the field of agent-based technologies. A special association called AOSE (Agent Oriented Software Engineering) joins activities of agent design methodologies, development tools and programming languages. Regular conferences on AAMAS (Autonomous Agents and Multiagent Systems) focus on a variety of agent-related problems [17]. Inter-agent communication protocols are defined by the FIPA standard [2] and JADE platform [2].

The Prometheus methodology covers the main steps in a multiagent system design procedure, and is based on the following development phases [11]:

- *System specifications*

Define system goals, conceive system scenarios, and identify basic system functionalities, i.e. system roles.

- *System architecture design*

Define types of agent in the system, connect believes (data in the databases) and roles with agents, determine agent interlinks and system architecture, and set communication protocols between co-operating agents.

- *Detailed design of the agents*

Define a detailed internal structure of each agent, incorporate roles they are supposed to act, implement agent plans, connect each agent to its environment, and determine the overall hierarchy of goals to induce the sequences of planed executions.

This general Prometheus framework is always adapted to the concrete problem under consideration. The abovementioned steps are applied iteratively so that the obtained hierarchy of goals, interconnections of goals and roles, scenario structures, agents, and eventually the overall system architecture suit our target system most effective.

Development and simulation environments for multiagent systems

Several such development environments have been reported [6]: these are 3APL, AgentSpeak, and JASON programming languages, and JACK, JADE, and Jadex development environments, just to mention a few.

JADE and Jadex

Jadex is called a BDI-based reasoning machine [4]. Jadex resides on the JADE platform. It integrates BDI agent features by using object-oriented approach and XML-based descriptions. JADE respects the FIPA-introduced standards and takes care of all the system communications. Jadex is actually a virtual machine which interprets a user-defined hierarchy and structure of multiple interconnected BDI agents. It thoroughly supports the internal structures of believes, plans and goals.

Jadex is not only a multiagent system development tool, for it is also a sophisticated testing tool, model simulation environment, and agent-oriented execution engine.

DEVELOPMENT OF MULTIAGENT WASHING SIMULATOR

When a new technology is studied as a potential basis for new products and manufacturing, thorough verification and validation is of paramount importance for all the phases from requirements and specifications, to the design, and to prototypes. Most cost-effective solution is achieved by simulations prior to final development of real components and systems.

We were faced with the same problem. When a new, agent-based generation of household appliances appeared challenging, it was clear many dilemmas exist that should be answered before the crucial decisions on new approaches of the design, development, production, marketing, and maintenance can be reached. This is why we constructed a multiagent system which, in general, can be upgraded into either a simulator or an implementation of a particular household device. We tested several design, development and execution environments by developing a simulator of multiagent washing assistant. The approaches we adopted are exemplified by a few details in the sequel.

Prometheus methodology and Multiagent Washing Assistant design

This section exemplifies the conceptual design of a MWA–a household appliance whose intelligent behaviour and environmental perception is meant to make laundry easy and effective. As we are designing an agent-oriented system, we resort to the agent-oriented software methodology entitled Prometheus [11]. Eventually, it is important to consider which parts of the system should be treated as agents. With Prometheus, this outline is generally prepared in the *architectural design phase* that follows *system specification phase*, while in the last *detailed design phase* all agents are already evident from the system overview. In our example of MWA design we made a preliminary decision in favour of a multiagent system structure. We specified one supervisory agent that is helped by the agent assistants, because of two reasons: firstly, suitable physical/hardware structure of the target system, and secondly, the agents in such organization correspond to the basic tasks of an MWA. Our desired structure of MWA is shown in Figure 2. Other Multiagent Home Assistants (MHA) can be exemplified in the same manner: for instance, the Multiagent Cooling Assistant or Multiagent Cooking and Backing Assistant.

We are using Prometheus for the system specification, design and consistency test phases, whereas we plan the implemention phase of an agent-oriented system using the Jadex development tool [4].

Jadex supports four types of goals: '*achieve*', '*perform*', '*maitain*', and '*query*' [4]. This is why we decided to adopt the same type of goals in our Prometheus-based design of MWA. Our design is therefore strongly goal-oriented. Each goal can have a number of plans that can be used to achieve it. Each plan can have a number of sub-goals that themselves can have multiple applicable plans.

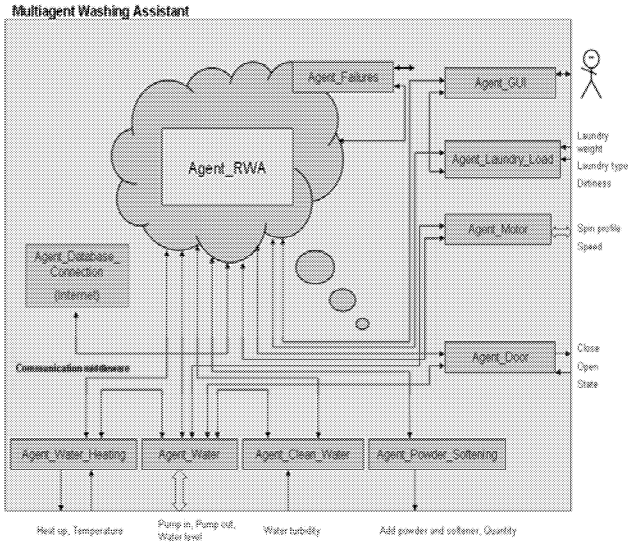


Figure 1: The MWA organization as induced by today's physical/hardware appliance structure.

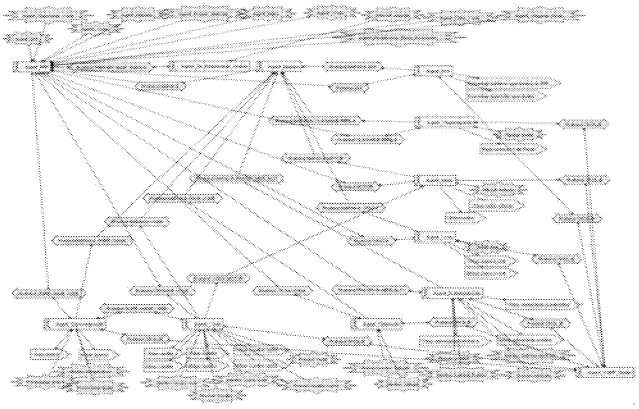


Figure 2: The MWA system overview diagram as designed by Prometheus

Prometheus instructs the user to the *architectural design*, going through *agent types*, *agent descriptors*, *interaction diagrams* and *interaction protocols* to the *system overview diagram* as depicted in Figure 2. There are also some other steps in this basic design, such as *agent acquaintance diagram, inter-agent messages definition, shared data repositories,* but because of limited space we are not going to describe them in detail.

The Prometheus *detailed design* leads to a progressive refinement of each agent with its *capabilities, internal events, detailed data structure,* and *process specification.* Fig. 3 depicts the aforementioned simple water heating agent, *Agent_Waterheating*, whose structure exemplifies a simple goal-plan tree. It is also obvious the agent's placement in its environment introduces perceptions, actions and messages to other agents. Internal agent messages take care of its goals that basically maintain the water in the drum heated and also perform the system check and report failure conditions.

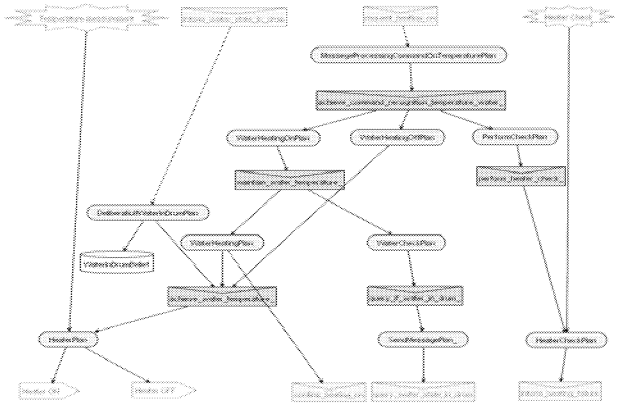


Figure 3: Detailed design of the *Agent_Waterheater* agent

Simulator development using Jadex in JBuilder's environment

The MWA design explained in the previous subsection can now be directly used in the construction of the simulator of multiagent washing assistant. We found a close correspondence between the Prometheus design methodology and the Jadex development and execution environment. For this reason, our simulator development originates on goal-oriented agent structure as offered by BDI architecture in Jadex. Brief examples of subsequent agents give a more concrete picture of the approaches applied.

Agent_Waterheating definition

Designing by Prometheus produced a clean and systematic overview of the agent-based washing assistant (Fig. 2). Details in Fig. 3 refer to the water-heater agent structure. Perceptions, believes, goals, and plans are interlinked in such a way that the heater is properly operated.

It is rather straightforward to build a Jadex-based agent out of the Prometheus definitions. Two separate files must be created. One is in the XML format and defines the agent's structure based on its beliefs, goals, and plans. The basic structure of this file is of ADF type (Agent Definition File), as introduced in [12]. It includes a basic definition of *Agent_Waterheating* from Fig. 1 referring to the Prometheus design from Fig. 3 and according to the Jadex basic rules. Its code is depicted in Fig. 4.

```
<!--  Agent Waterheating -->
<agent xmlns="http://jadex.sourceforge.net/jadex"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://jadex.sourceforge.net/jadex
file:///c:/jadex/jadex-0.94.xsd"
        name="Agent Waterheating"
        package="multiagentwashingassistant agent rational w
ashing_assistant">
        <imports>
                <import>java.util.*</import>
        <import>jadex.adapter.fipa.*</import>   <import>
multiagentwashingassistant_agent_rational_washing_assistant
*</import>
        </imports>
        <capabilities>
<!—There are no capabilities included -->
        </capabilities>
        <beliefs>
 <!—Belief definition axample -->
```

```
            <belief name="real_temperature" class="int">
                    <fact>20</fact>
            </belief>
          </beliefs>
          <goals>
<!— Goal definition example -->
            <achievegoal name="achieve_water_temperature">
              <contextcondition>
               $beliefbase.heating_state &amp;amp;&amp;amp;
               $beliefbase.water_level &amp;amp;&amp;amp;
               !$beliefbase.working_error
              </contextcondition>
               <deliberation>
                <inhibits ref="watercooling"
                 inhibit="when_in_process"/>
               </deliberation>
              <targetcondition>
               $beliefbase.dejanska_temperatura)>
               ($beliefbase.nastavljena_temperatura)
              </targetcondition>
            </achievegoal>
          </goals>
          <plans>
<!-- Java plan definition example-->
            <plan name="heater1">
              <body>new Heater1Plan()</body>
              <trigger>
                    <goal ref="achieve_water_temperature"/>
              </trigger>
            </plan>
          </plans>
          <events>
<!—External message definition example -->
            <messageevent name="heateron" direction="send"
                          type="fipa">
              <parameter name="performative" class="String"
                         direction="fixed">
                    <value>SFipa.REQUEST</value>
              </parameter>
            </messageevent>
          </events>
   <properties>
<!—Properties and instructions -->
   </properties>
   <initialstates>
<!—Initial goals, plans, and events states -->
      <plans>
          <initialplan ref="messageprocessing"/>
      </plans>
   </initialstates>
</agent>
```

Figure 4: The structure of water heating agent as constructed in Jadex

Together with ADF for every agent's plan, a separate Java class must also be created. All messages exchanged between the Jadex BDI-agents must correspond to the FIPA protocol. A message received by the agent (*Agent_Waterheating* in this case) triggers appropriate plan which, for example, switches on the heater to heat the water. Messages carry the parameters which can be used to drive corresponding actions, for example a temperature to be achieved. These parameters actually update agent's beliefs and are further used and processed by internal plans. A simple plan is exemplified by Fig. 5.

```
package jadex.RHP;
/*
* Plan heater on
 */
public class WaterHeatingOnPlan extends Plan
   {
    //
    public WaterHeatingOnPlan ()
     {
     }
     //-------- methods --------
     // Plan execution part.
     public void body()
     {
      String t = (String)getParameter("temperature").getValue();
        int It = Integer.parseInt(t);
         getBeliefbase().getBelief("temperature_setting").setFact(
         new Integer(It));
        IGoal goal = createGoal("achieve_water_temperature")
              try
               {
               dispatchSubgoalAndWait(goal);
               }
               catch(GoalFailureException e)
               {
               getLogger().info("Heater Error");
               };
      }
   }
```

Figure 5: A short example of the water heater plan

ADF files include definitions for agent beliefs, goals and plans and correspond to Jadex schema defined in *jadex-o.94.xsd.* Fig. 5 gives only a short example and does not include a complete agent definition.

Jadex is a strict goal-oriented tool, whereas the goal life cycle is defined by the following conditions:

- *creationcondition* which defines conditions for a goal creations,
- *contextcondition* which defines the goal activity context,
- *dropcondition* which defines conditions to abandon the goal, and
- *targetcondition* which defines the target goal conditions.

SIMULATOR TESTING AND RESULTS INTERPRETATION

Once having a multiagent simulator built, it can be used also to examine temporal behaviour of single agents, cooperation among different agents, built-in capabilities of reasoning, planning and learning. The inter-agent message density and system resource consumption is measured, too. Using the washing assistant simulator, a series of statistically independent washing runs was recorded. The obtained results, based on different laundry samples, were processed by an external machine-learning agent as training set. This learning agent computed regression models. These are implemented in further simulation, or real, runs to predict particular washing parameters. The predicted parameters extend the reasoning engine database, so that the rational washing assistant always infer from the latest learned knowledge about the washing process. Results of our simulations will be shown in the sequel.

Methodologies used in the frame of agent-oriented program engineering (AOSE) [10] mainly propose tools for multiagent systems analysis, design, and implementation. Little attention is dedicated to multiagent testing. R. Coelho [5] overviews some development methodologies (MaSE, MASCommonKAD, AGILE) which define clear verification

processes. It presents also a special agent developed for the JADE environment which is responsible to test the roles of individual agents. We verified our simulator capabilities by these tools, while extending them by some special and to our application adopted testing methods.
Our main attention was paid to the Jbuilder's and Jadex's testing and debugging tools. The simulator itself is designed in such a way that acting and behaving of individual agents can be observed with the help of GUI interfaces of every agent. A special GUI for testing and development was implemented to enable the observation of internal agent events. To simulate the washing agent environment a special *Agent_Environment* was developed. This agent simulates the washing machine environment and records the MWA internal states and beliefs at the same time. A graphical snap-shot of the selected belief throughout the entire washing process can be recorded and used for later analysis. Such snap-shots bring a lot of important information about the agent's behaviour in real time.

Jbuilder working environment

Jadex, which we employ as our multiagent development, simulation and execution tool, has no integrated working environment. Therefore, we installed Jadex and its libraries into the Borland's Jbuilder [3]. Jbuilder enables hierarchical and ordered project organisation, which we benefited from when developing our MWA. A systematic deployment of all the agents as defined in Fig. 1 was easily made. In this way, the agents are inserted in tree structures of the agents' Java classes. A list of all Java classes of every project in Jbuilder appears below the agent list. There are more Java classes then agents, since some of the Agent_RWA capabilities are organized as separate Java classes. All data structures of the Agent_RWA are organized separately and handled as the agent's belief: reasoning rules, basic washing programs, new program templates, and washing history.

Jbuilder can be used for hierarchal project organisation and, at the same time, as an effective debugging tool for realtime system testing.

Jadex testing environment

Jadex comprises different tools for developing and debugging the agents in real time. Jadex is started in the Jbuilder's environment. The Jadex control centre menu appears as in Fig. 6. From this menu, the user interface, *Agent_GUI*, is started. Also any other agent compatible with Jadex environment can be started from this menu.

Four different tools can be started in this environment:
- the agent's starting and stopping tool;
- the internal agent's state supervision tool;
- a conversation window to send messages between agents.

All these tools are very effective and useful, but with the intensive use on a large number of agents slow down the execution speed of the system and become almost useless. The complete system becomes unstable and unreliable. Because these capabilities have to be used in a limited way, an additional special user interface to record and display the agent's internal states was developed.

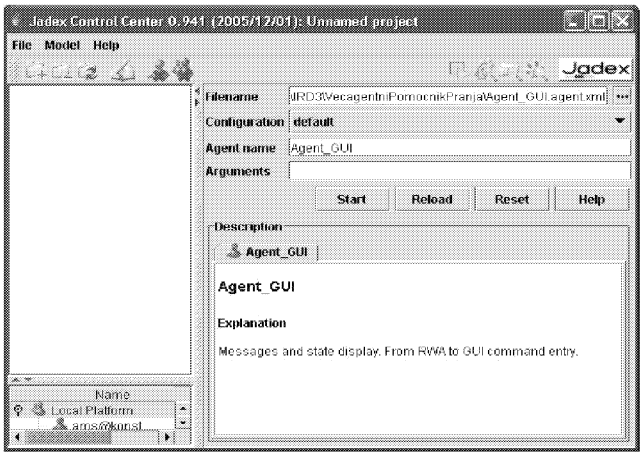


Figure 6: Jadex control centre.

Multiagent Washing Assistant simulation

The MWA simulator was developed on the portable PC Fujitsu Siemens with built-in processor Intel Pentium 1,70 GHz and 1 GB memory. This processor is capable of running 656 MIPS, benchmarked with the CPU test from [16].

Our simulator represents the PG5 washing machine of a Slovenian white goods producer [21]. Our goal was to simulate the machine thoroughly, which includes the washing process simulation, the dialog with the user, and also the washing machine environment simulation. The architecture of this simulator is depicted in Fig. 1. Indeed, it simulates real household appliance capabilities, but it also extends them by the capabilities of reasoning, planning and learning. The simulator was used to study the following approaches:
- verification of suitability of agent technologies and multiagent system structures in the filed of white goods,
- recognition procedures of simple laundry samples,
- reasoning procedures about new washing parameters,
- simple planning procedures,
- washing processes,
- and validation of new reasoning rules based on the washing history and created by an external learning agent.

Fig. 77 shows the MWA graphical interface. It consists of a user interface represented by *Agent_GUI*, which communicates the information between the user and *Agent_RWA*. This interface is used to run *Agent_RWA*, the laundry insertion procedures, and washing (upper part of RWA, left and right in Fig. 7). *Agent_RWA* reports about the laundry type, weight, and dirtiness, suggests the washing program type, washing parameters and takes care about messaging the current internal agent states. The user can change the suggested washing parameters, but only in the permitted frames. The described Agent_RWA interface is encompassed by several interfaces of the agent which simulates the appliance's environment, *Agent_Environment* (lower part of RWA, bottom and right in Fig. 7). This agent simulates water filling in and out, water heating, clean water measurement, powder and softening dosage, safe locking the door, drum rotation, and laundry insertion. To display different parameters from the environment, more user interfaces are built in the simulator, such as *envHeating*, *envWater*, *envClearwater*, *envPowderSoftener*, *envDoor*, *envMotor*, and *envInsertion*. There is a special agent called *Agent_Errors*, which is used to collect possible exception

conditions from all the agents assistants and reports the errors to *Agent_RWA*.

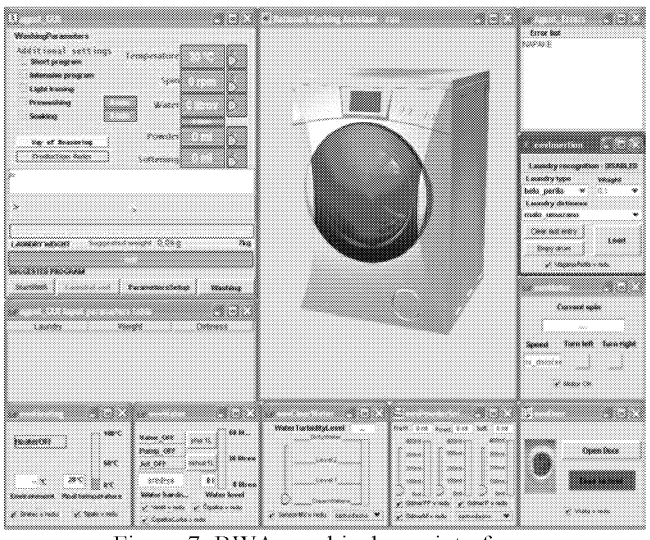


Figure 7: RWA graphical user interface.

This is the way to prevent irregularities in time. Additional user interface to test the system is built in. It helps to start or stop different agents, to send messages and commands to different agents and, thus, to attend its responses. Internal agent's states can be attended in real time, such as clear water level, water temperature, water hardness, etc. This control window can be adapted to the current needs. Assistant agents' behaviour can be monitored, how they react to the environment, and also how they react to the supervisory agent instructions. This information helps to infer about the behaviour of:

- the complete multiagent structure,
- individual agents,
- and the RWA environment.

Agent_GUI shows internal states of *Agent_RWA* in the form that reports the washing parameters and procedures to the user. The user can use these menus to arrange current and future washing processes.

Dynamic multiagent structure of RWA ensures easy future improvements of system execution performances by replacing agents and their capabilities, for example, an upgrade of user interface by image and speech recognition. The current environment simulating agent can be replaced by a real, physical washing machine environment. A real multiagent washing assistant can be constructed this way. It inherently adopts multiagent structure and intelligent capabilities.

Evaluation of communication and processing performance

The MWA simulator is executed on a PC. The real washing machine price does not allow a direct inclusion of such a powerful computer. Thus, it is important to assess the minimum hardware and software requirements to run a simulator like MWA.

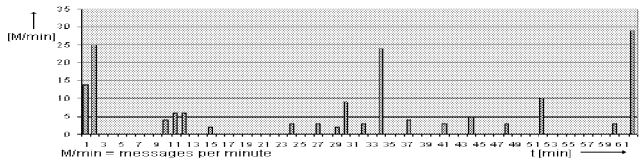
From Fig. 2, it is evident that there are many different protocols of internal communication built in MWA. To calculate the necessary communication-channel bandwidths, we measured the average message density in the MWA first.

The Jadex built-in protocols are based on FIPA. The message content in the text format has the following form:

(**request**: **sender** (:**name** Agent_Ratuonal_Washing_Assistant@konst)
:**receiver** (:**name** Agent_Water_Heating)
:**ontology** fipa-pta
:**language** SL
:**protocol** fipa-request
:**content**
(heater on 65)

The average message length is around 200 characters or 1600 bits. Jadex control centre enables the measurement of the average message density in the period of one simulation run. Several different runs have been measured (Fig. 8 shows two of them). The abscissa corresponds to the complete washing time in minutes, whereas the ordinate shows the number of messages accumulated in minute bins. Different laundry insertions have been applied, which initiated different washing programs. This explains the different washing durations. The distributions from Fig. 8 show that average message density is low. Maximum density is not larger then 30 messages per minute. The average number of messages is about 5 per minute. This leads to a maximum bandwidth of 60,000 bits/minute, or 1,000 bits/s. These results clearly show that modern communication capabilities of the LAN or WLAN type surpass this bandwidth by a few hundred times. So, the communication burden of one multiagent household appliance is negligible.

a) Coloured laundry, normally soiled, low quantity.



b) Coloured laundry, very soiled, low quantity.

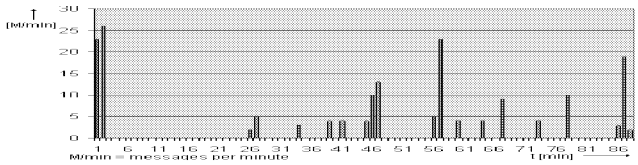


Figure 8: Messages density charts.

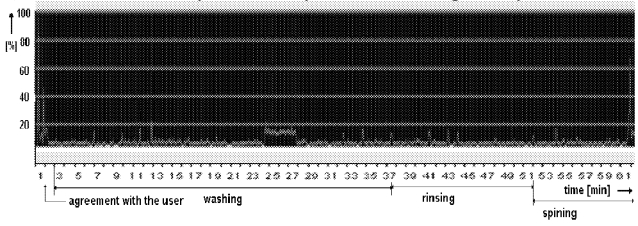
The processor and memory utilization are the next important factors that influence a system throughput. All agents in the MWA simulator are executed on the same processor, while in the real system they may be executed on different systems. In our evaluation, only a single processor and memory were monitored.

The processor and memory utilization (processor Intel Pentium and 1GB of memory) was measured by Windows Task Manager, which is built in the operating system Windows XP. Fig. 9 shows a few screen-shots of the processor utilization as measured during simulation of different washing procedures. The processor is heavily utilized only in the period of the *Agent_RWA* most demanding tasks, which is reasoning about new washing parameters–it uses all available processing time. Machine's reasoning accompanies the laundry insertion and machine-user dialog. This is a non-critical period of the machine operation with practically no time constraints. The only side effect of this situation is that in the case of a less powerful

processor the user waits longer for the machine response. Fig. 9 further shows that in the washing simulation period the processor is busy less then 20% on average, which means very low workload. Even more favourable situation can be achieved in the real situation, when the agent to simulate the machine's environment, i.e. *Agent_environment*, does not exist because it is replaced by the machine itself.

The established processor utilization figures demonstrate our simulator can run in real time even on a 350 MHz processor.

a) Coloured laundry, normally soiled, low quantity.



b) Sports laundry, normally soiled, little quantity.

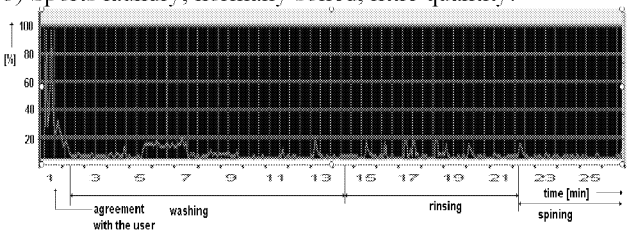


Figure 9: Processor utilization.

Using the same monitor, we observed the memory utilization, too. Windows "PF Usage" and "Physical Memory" were observed during Jadex simulation. These measurements are not very accurate and mean only a coarse assessment to help future design of a real MWA. The obtained results are as follows:

- the Jadex platform occupies about 35 MB of memory,
- *Agent_RWA* occupies about 8 MB of memory,
- while assistant agents occupy 1MB of memory each, on average.

These results show that memory requirements are high if we wanted to port the Jadex simulation and run-time environment to the real MWA. They exceed the capabilities of an averaged microcomputer system, used in today's home appliance. To keep Jadex and Java as development tools, enough powerful hardware with a processor and memory of an average PC must be used. However, this does not mean that the revealed approach cannot be ported to a practical solution directly. On the contrary, since the system is based on a multiagent structure, the core of it, including the rational supervisory agent, machine reasoning, and machine learning, can be executed on a powerful home computer system, whereas in the real appliance only a simple microcomputer runs simple agents that communicate with the main supervisory system and control simple local tasks.

CONCLUSION

We briefly described the key aspects of the design and implementation of multiagent simulator of household appliances. The main tools used in this process were the Prometheus agent-design methodology and Jadex simulation and agent-execution environment. We found the combination of Prometheus and Jadex entirely compatible and indispensable in the process of constructing the agent-based systems. Because the both tools are strongly goal-oriented, the results of using them in BDI agent-based systems proved to be synergetic and gave promising results. One of the advantages of this approach is also the number of cases where the automated tools can be used for the consistency checking against various artefacts of the design process.

Our simulator of multiagent washing assistant was built in the form of a supervisory agent and a set of subordinate assistant agents. For testing and evaluation purposes, the simulator was run with different laundry loads in order to evaluate its hardware requirements. Average processor utilization of an Intel Pentium 1.70 GHz combined with 1 GB of memory was less then 20%. The memory utilization was rather high with a few ten MB, and more, of memory. At the same time, data communication flows were found very low, at maximum 10,000 bits/s. Taking these figures into account, we can claim communication bandwidths and processor power is are already available in today's connected household appliances. White goods producers already use processors like ARM7 [15] with 130 MIPS of processing speed, which is quite enough to process the workload as measured in our simulations. This is especially true if the new versions of the ARM9 processor with 300 MIPS are considered. The only trouble when direct porting of Jadex environment to the real household appliances should be applied is high memory requirements. But, as we have explained, the multiagent architecture itself warranties a distributed solution with heavier workloads on more powerful computers and only rather stripped assistant agents in the real appliance to do the basic tasks and communication.

REFERENCES

1. J. Ahola, *Ambient Intelligence*. Internet publication, 2001. Accessible: http://www.ercim.org/publication/Ercim_News/enw47/intro.html
2. F. ,Bellifemine, G. Rimassa, A. Poggi, "JADE – A FIPA-compliant agent framework," 4th International Conference on the Practical Applications of Agents and Multi-Agent System, London, pages 97-108, 1999
3. Borland Jbuilder 2007. Accessible: http://www.borland.com/
4. L. Braubach, A. Pokahr, D. Moldt, W. Lamersdorf, "Goal Representation for BDI Agent Systems," Second International Workshop on Programming Multiagent Systems: Languages and Tools, New York, pages 9-20, 2004. Accessible: http://vsis-www.informatik.uni-hamburg.de/publications/view.php/208
5. R. Coelho, U. Kulesza, A. von Staa, C. Lucena, *Unit Testing in Multi-agent Systems using Mock Agents and Aspects*. Computer Science Department, Pontifical Catholic University of Rio de Janeiro. Accessible: http://portal.acm.org/citation.cfm?id=1138079&dl=acm&coll=&CFID=15151515&CFTOKEN=6184618
6. M. Dastani, R. H. Bordini, B. van Riemsdijk, "Programming Languages for Multi-Agent System," 7th Agent Systems Summer School Utrecht, 2005
7. G. Riva, *Ambient Intelligence, The evolution of technology, communication and cognition towards the future of human-computer interaction*. Accessible:

http://www.emergingcommunication.com
8. N. R. Jennings, M. Wooldridge, "Application of Intelligent Agents," In N. R. Jennings and M. Wooldridge, editors, *Agent Technology: Foundations, Applications, and Markets*. Springer-Verlag, New York, 1998.
Accessible:
http://agents.umbc.edu/introduction/jennings98.pdf
9. D. Kazakov, D. Kudenko, *Multi-Agent Systems and Application, chapter Machine Learning and Inductive Logic Programming for Multi-Agent Systems*. Springer-Verlag, New York, pages 246-270, 2001.
Accessible: http://www-users.cs.york.ac.uk/%7Ekazakov/papers/acai01.htm
10. P. Massonet, Y. Deville, C. Nève, "From AOSE Methodology to Agent Implementation," Proceedings of the first international joint conference on Autonomous agents and multiagent systems, Bologna, pages 27 – 34, 2002.
Accessible:
http://portal.acm.org/citation.cfm?id=544747
11. L. Padgham, M. Winikoff, *Developing Intelligent Agent Systems*. John Wiley & Sons, Ltd., 2004
12. A. Pokahr, L. Braubach, W. Lamersdorf, "Jadex: A BDI Reasoning Engine," In R. Bordini, M. Dastani, J. Dix and A. Seghrouchni, editors, *Multi-Agent Programming*. Springer Science+Business Media Inc., USA, 2005
13. A. S. Rao, M. P. Georgeff, *BDI Agents: From Theory to Practice, Technical Note 56*. Australian Artificial Intelligence Institute, 1995.
Accessible:
http://www.agent.ai/doc/upload/200302/rao95.pdf
14. S. Russel, P. Norvig, *Artificial intelligence: A Modern Approach, Second Edition*. Prentice Hall, 2003
15. Konrad Steblovnik, "*New Technologies and Ergonomics in White Goods*" Proceedings of the 2005 International appliance technical conference: 56th International Appliance Technical Conference & Exhibition: March 28-30, 2005, Ramada Plaza Hotel O'Hare Rosemont, Illinois U.S.A. Lexington, pages. 84-95, 2004
16. *TestCPU procesor detection and benchmark*.
Accessible: http://www.fi.muni.cz/
17. J. Thangarajah, L. Padgham, Michael Winikoff, "Prometheus Design Tool," Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, pages 127-128, 2005.
Accessible:
http://goanna.cs.rmit.edu.au/~winikoff/Papers/aamas05pdt.pdf
18. *Thorne: The BDI Model of Agency and BDI Logics, Technical report*. Laboratory for Applied Ontology, Trento, May 2005.
Accessible: http://www.loa-cnr.it/Files/bdi.pdf
19. M. J. Wooldridge, *An Introduction to Multiagent Systems*. John Wiley & Sons, Ltd., 2002
20. M. J. Wooldridge, *Reasoning about Rational Agents*. The MIT Press Cambridge, Massachusetts, London, England, 1998
21. PS PG AKO/12 *User manual*. Label 660427/103/052. Gorenje Velenje.

BIOGRAPHIES

KONRAD STEBLOVNIK studied electronics and computer science at the University of Ljubljana. He has 30 years of experience in different fields of industrial research and development: colour TV, computer peripherals, electronics for white goods, and intelligent home field. He manages Development team for Intelligent home in the company Gorenje in Velenje of Slovenia.

DAMJAN ZAZULA received Dipl. Ing., Masters and Doctor of Science degrees in electrical engineering from the University of Ljubljana, Slovenia, in 1974, 1978 and 1990, respectively. After being involved in industrial R&D for 12 years, he joined the Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia, in 1987. He holds a Full Professor position in computer science there, while from 1998 to 2003 he was also appointed Associate Dean of Research. Dr. Zazula has spent several months as visiting professor at the ETH in Zurich, Switzerland, and Ecole Centrale de Nantes, France. His main research interests are compound signal decomposition, biomedical imaging, and virtual training tools.

VISUALIZATION FOR AN AUGMENTED REALITY CONSTRUCTION MACHINERY SIMULATOR

Alvaro Segura
VICOMTech
Mikeletegi 57
E-20009 San Sebastin, Spain
email: asegura@vicomtech.es

Aitor Moreno
VICOMTech
Mikeletegi 57
E-20009 San Sebastin, Spain
email: amoreno@vicomtech.es

Gino Brunetti
Fraunhofer IGD
Fraunhoferstrae 5
D-64283 Darmstadt, Germany
email: gino.brunetti@igd.fraunhofer.de

Thomas Henn
Fraunhofer IGD
Fraunhoferstrae 5
D-64283 Darmstadt, Germany
email: thomas.henn@igd.fraunhofer.de

KEYWORDS

Interactive Simulation, Augmented Reality, Virtual Reality

ABSTRACT

This work presents a versatile real-time simulator for training workers in the operation of construction machinery. With the aim of reducing the accident rate in the construction sector, a realistic AR scene improves the user experience.

Our work addresses two stages of the training process. First, the instructor designs the working environment and the exercises with a specifically developed Scene Editor tool. Then, these exercises are run by the simulator. The simulation setup consists of a real cabin placed on a motion platform in order to provide a realistic interaction with the system and a stereoscopic augmented reality system for visualization. Visualization has been implemented as a chroma-key-based mixed reality system.

This paper describes the development of the Visualization subsystem of this simulator.

INTRODUCTION

The work described in this paper was carried out as part of project VAR-Trainer (Barrera et al. 2006). The general objective of VAR-Trainer is to develop a versatile real-time simulator, based on the integration of mechanics, electronics, automated technologies, and mixed reality (virtual augmented reality) in order to train workers in the operation of construction machinery. Special attention is paid to safety aspects with the aim of reducing the accident rate in the construction sector.

Although the use of machine simulators for training purposes is expanding nowadays, no commercial simulator exists for training workers in the use of construction machinery. Likewise, there is no European certified training in matters of safety when using this machinery. Most of the available simulators are specifically oriented towards a single machine or process type and are not open platforms that could be reconfigurable to cover different machine types, different levels of realism or different pedagogical aspects. Besides, machine training simulators usually aim to cover training in all the machine's functionalities and are not focused on specific training points, such as safety issues, for instance.

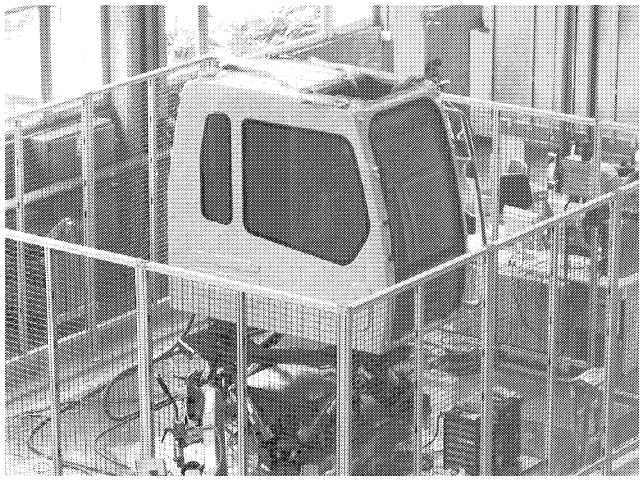


Figure 1: The excavator cabin on the motion platform

The system is *versatile* in two aspects: the same simulator core has to adapt to different machine types (from wheeled excavators or dumpers to mast climb platforms), and the same simulator core has to run on an average desktop PC, and on a complete integrated platform that includes a real vehicle cabin with its real control devices, a motion platform and an immersive AR visualization helmet. Four machine types have been included in the simulator: a wheeled excavator, a dumper, a mast climb platform, and a work goods

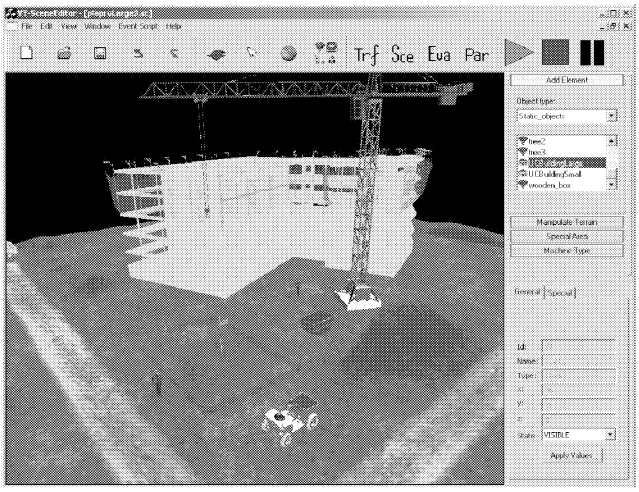


Figure 2: The Scene Editor main screen with the control panel on the right, tool bar on top, and 3D virtual view on the left

lift.

Most other comparable simulators are focused on one specific machine type and model, and often use large projection screens to display the environment. Our design is innovative in several aspects including its versatility and the special use of augmented reality, as will be explained, to provide immersion into the virtual world while interacting with true cabin controls.

As part of the versatility of the developed system, the instructor can design the virtual environments where the training exercises take place. A specific *Scene Editor* tool was created for this purpose, which is described in its own section.

The simulator has a modular architecture whose main parts are the *Exercise Manager*, the *Dynamics Module* and the *Visualization Module*. The Exercise Manager controls and evaluates the tasks that the user has to perform and the behavior of the external elements such as other vehicles. The Dynamics Module (Pujana et al. 2006) is responsible for simulating the dynamics of the machine by mathematically modelling its motors, hydraulics, joints and actuators. This paper discusses the design and development of the Visualization Module.

THE SCENE EDITOR

The Scene Editor is a graphical interactive tool that allows instructors to design the environment for the simulated exercises. It is not a generic 3D modelling application, but a system that knows the specific elements that are relevant to the simulation.

The first step in designing a scene consists of defining the shape of the terrain. When starting a new scene the editor presents the user a flat square area that he can interactively deform, adding slopes, hills, etc.

Then the designer would add objects, buildings, vehicles and workers to the scene. Mobile elements such as vehicles and people are assigned a path that they will move along during the simulation. A screenshot of the Scene Editor is shown in figure 2.

Some object types are configurable: when the user adds a building, the system will ask for the kind of building and its number of floors. The editor generates the appropriate geometry.

The user can also define special areas in the environment with specific meanings. Examples of such areas are start or parking zones, waypoints and excavation areas. Only specially designated areas can be excavated by the driven excavator.

Finally, the instructor uses separate module, the *Exercise Editor*, to create an exercise defining the behaviour of the elements in the scene and their reactions to driver actions. A scripting system specifies what will happen when the driven machine approaches another vehicle, or what task the trainee will have to perform after driving into an excavation zone.

VISUALIZATION MODULE

The visualization module is responsible for different tasks:

- Realize a mixed reality visualization environment by rendering the virtual environment and the visible parts of the driven machine.
- Report to the dynamics simulation module detected collisions of the driven machine with the virtual environment, e.g. with static objects like buildings, but also with moving objects like other vehicles or virtual workers.
- Report the penetration of wheels in the ground to drive the ground wheel interaction.
- Manage all animated and simulated objects (workers, vehicles, etc.) in the virtual construction site.
- Display a realistic impression of materials, e.g. soil, and different weather conditions.
- Visualize augmented system and instructor messages.

Rendering of the virtual scene is implemented with the OpenSG graphics library. In order to simulate the physics of some objects and detect collisions and interactions with the driven machine the *Open Dynamics Engine* (ODE 2007) was integrated into the visualization module. This interaction events are sent as input to the Dynamics Module for it

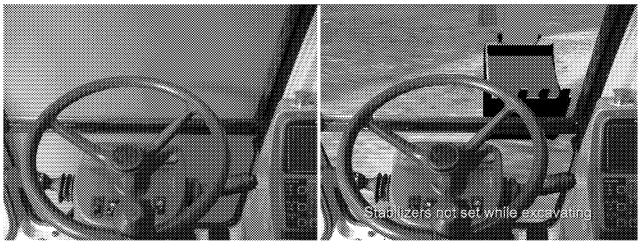


Figure 3: The view captured by the HMD cameras (left) and the view as seen by the user (right): the virtual background, the real interior and the information messages layer

to react accordingly, or are reported as risks to the Exercise Manager. For every visible scene object the module creates an associated scene graph node containing its visual representation and an associated collision geometry.

Visualization Hardware Setup

As we said there are two types of hardware setups where our versatile simulator must work: a common desktop PC, and our especially equipped integrated platform.

Visualization for the PC target consists of a virtual reality 3D view based on the *OpenSG* (OpenSG 2007) rendering library. All elements in the view, including the interior of the machine's cabin, are virtual representations.

The integrated platform, on the other hand, implements a novel Augmented Reality approach based on the chroma-key technique. The platform integrates a high-end PC with a professional Nvidia Quadro graphics card, and a specifically designed AR-oriented head-mounted display.

The adopted mixed reality setup for visualization is motivated by the different layers of information that should be visible to the trainee, maintaining low hardware costs. The layers are (in back to front order):

1. the virtual construction site scenario and the virtual part of the vehicle being the virtual background of the scene
2. the real parts of the cabin in front of the virtual scene, i.e. the real foreground
3. the augmented messages on top (in front) of the real foreground to display messages or to provide hints.

This AR approach is quite novel because classical AR has a real background with superimposed virtual elements (Seibert and Dhne 2006). Our system reverses the real and virtual layers and adds an additional layer for information messages fir the user. Figure 3 shows what the user of the simulator sees whle driving the excavator.

Because of the costs, driving simulators typically are equipped either with monoscopic non-immersive visualization systems, or with a head mounted display (HMD), the latter reducing the visual quality because of the typically low resolution and limited field of view. For VAR-Trainer, however, the decision had to be in favor of a low-cost fully stereoscopic immersive setup, as it is mandatory to provide a full panoramic view outside the cabin and because of the importance of depth when accomplishing a given training task.

The used HMD partially overcomes or improves some of the limitations stated above. It is the *eMagin Z800 3D Visor* that contains two high-contrast SVGA 3D OLED micro-displays delivering a bright, crisp image, and weighs less than 230 g. The HMD has a diagonal field of view of approximately 40 degrees.

To cope with the situation that the virtual background has to be overlaid with the real foreground, the *chroma-key* technique has been adopted. Chroma- key is a filmmaking technique of shooting foreground action against an evenly- lit monochromatic (often blue) background for the purpose of removing the background from the scene and replacing it with a different image or scene. The term also refers to the visual effect resulting from this technique.

Applied to VAR-Trainer, the chroma-key video technique is combined with the *AR Video See-Through* technique, where real images are captured into a computer through two cameras mounted on the HMD, one for each eye. The captured real images are mixed with computer-generated virtual images, and the mixed images are shown to the user in the HMD. In order to generate the correct view, the computer needs to know the current position and orientation of the user's head. We finally used an optical pose tracking system based on a ceiling mounted camera (Santos et al. 2006a) (Santos et al. 2006b) for this purpose. Initial trials with a commercial magnetic tracker were very problematic inside the metallic cabin (Nixon et al. 1998). Figure 4 shows our prototype Augmented Reality HMD visor.

The chroma key effect in our setup is realized by coating the windows of the cabin with a translucent blue foil (as seen in figure 1). The real interior of the cabin (and the trainee him/herself) are then extracted from the video images captured by the two cameras mounted on the HMD, whereas the blue portions are substituted by the virtual background environment.

For non-driven machines such as mast climb platforms we had a completely blue cubic box. Here the trainee stands inside and feels surrounded by the virtual environment by means of the same AR-HMD.

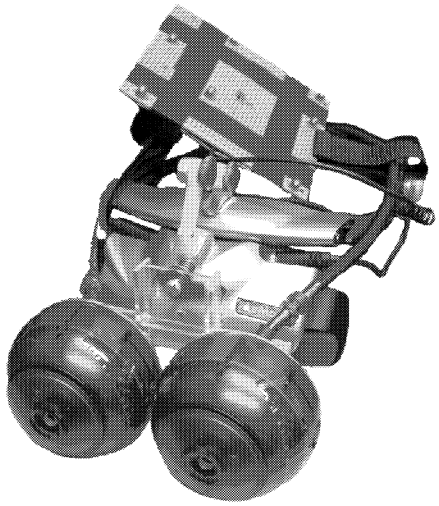


Figure 4: The prototype AR-HMD developed with its front-mounted cameras and a square fiducial marker needed for tracking

Figure 5: Scattered sky light simulation rendered with GLSL shaders

Weather Conditions

It was an important requirement to represent environmental conditions as accurately as possible, in order to train workers in reduced visibility situations, for example. We thus developed special algorithms to render the virtual world in different configurable weather conditions. The simulator can work in varying rain intensities, with different fog densities, clouds, wind and time of day.

Since the system must work on average PCs and on our high-end integrated platform, we developed separate techniques for each. For the integrated platform, where we can control the installed hardware, we developed programmable GPU shaders to render rain (including drops hitting the windshield), clouds and sky light. Shaders were implemented in the GLSL language as described in (Kalantary et al. 2006) and (Engel 2005). Figure 5 shows a rendering of the sky in special conditions. But since many average PCs may not support these advanced graphics features simpler methods were also implemented as a fallback solution based on more traditional OpenGL rendering methods.

Excavation

In our project, the most often demanded construction machine for training simulation is the retroexcavator, and its main task is obviously to excavate. We developed a graphical simulation algorithm to allow the machine dig on our virtual terrain.

The algorithm is inspired on the 2.5D material removal methods used in machining simulators. The diggable zone is subdivided into a fine regular grid of height samples. When the excavator's bucket intersects the ground surface, the corresponding height values in the matrix are reduced down to the intersection points. Additionally the system estimates the amount of removed material and visually fills the bucket accordingly. The volume of material in the bucket is passed to the dynamics module that takes its weight into account when simulating the whole machine behaviour.

Dropping the soil excavated is also simulated. This is accomplished through a particle system emitted by the bucket and controlled by gravity, providing a convincing effect. In the area where particles reach the ground the soil is accumulated forming a visible heap. Particles falling on the excavation area affect the heights matrix by adding a small value specially calculated to give the appearance of a growing heap.

In the integrated platform rendering the soil on the excavation areas also uses specially written GLSL shaders to improve realism. The shader implements a bump and displacement mapping technique. Moreover, it takes depth into account when rendering each pixel producing a color gradient looking like strata in the digged holes.

Avatars

Another important element for training in safety aspects is of course human characters, i.e. workers moving around. Their behaviour is defined in the paths and scripts created with the Scene Editor. We developed a simple animation algorithm that allows these avatars to walk. The movement characteristic itself was recorded with a motion capture system and applied to the 3D characters. Figure 6 shows a couple of these avatars in the work environment.

Avatars can walk on the free-form terrain and in the floors of

Figure 6: Animated workers and loadable objects as part of the environment.

under- construction buildings.

Loads

Some of the simulated machines can transport goods and loads, such as the dumper or the mast climb platform. The whole system is aware of this and there are special objects representing loads. The loads can be placed in the environment with the Scene Editor and participate in the exercise tasks. For example, an exercise may require the dumper to pick and drop a certain load in a specific place. The editor can define loading and unloading areas in the environment where this process is allowed.

We implemented methods to load and unload the driven machine at the request of the user. An animation shows the user the loading or unloading process. Figure 6 shows a few loads on the ground ready to be picked up.

CONCLUSIONS

We have presented here the work carried out to provide a suitable visualization subsystem to a versatile construction machinery simulator. Our developments in this subsystem include graphical simulations such as simple excavation and realistic representation of weather conditions.

Professionals from the construction sector have examined and tested the simulator. The system was presented to associations participating in the project specification (some focused on risk prevention), and recently in the German BAUMA 2007 Construction Fair. The impressions were positive and a good potential for training was seen by professional visitors.

This work proves the feasibility of implementing versatile simulation cores which can be adapted to different machines and run different hardware platforms. Four machines have been integrated into this platform but more could be simulated on the same simulation and visualization core.

Acknowledgements The results presented in this work have been partially funded by the European Comission under the Collective Research Project "VAR-TRAINER" (EC-Contract no. COLL-CT-2003-500452).

REFERENCES

Barrera C.; Sebastin R.; Brunetti G.; Henn T.; Esnaola A.; Pujana A.; Moreno A.; and Segura A., 2006. *VAR-Trainer: Versatile Construction Machinery Simulator for Security Training*. In *Technische Universitt Dresden. Institut for Frdertechnik, Baumaschinen und Logistik*. 167–181.

Engel W., 2005. *Shaderx3: Advanced Rendering with DirectX and OpenGL*. Charles River Media. ISBN 1-58450-357-2.

Kalantary B.; Brunetti G.; Segura A.; Moreno A.; and Hofmann U., 2006. *Realistic Rendering of Environmental Conditions in a Construction Machinery Simulator Applying GLSL Shaders*. Tech. rep., Fraunhofer Technical Report 06i004-FIGD.

Nixon M.; McCallum B.; Fright W.; and Price N., 1998. *The effects of Metals and Interfering Fields on Electromagnetic Trackers*. In *Presence: Teleoperators and Virtual Environments Vol. 7, No. 2*. 204–218.

ODE, 2007. *http://www.ode.org/*.

OpenSG, 2007. *http://www.opensg.org/*.

Pujana A.; Martinez A.; Martinez F.; Nicols C.; and Landaluze J., 2006. *Construction Machinery Dynamic Models for Real-Time Training Simulators*. In *ISC 2006*. 0–0.

Santos P.; Stork A.; Buaes A.; and Jorge J., 2006a. *Innovative Geometric Pose Reconstruction for Marker-based Single Camera Tracking*. In *ACM SIGGRAPH International Conference on Virtual Reality Continuum and its Applications VRCIA 2006*. 237–244.

Santos P.; Stork A.; Buaes A.; and Jorge J., 2006b. *PTrack: Introducing a Novel Iterative Geometric Pose Estimation for a Marker-based Single Camera Tracking System*. In *IEEE Virtual Reality 2006*. 143–150.

Seibert H. and Dhne P., 2006. *System Architecture of a Mixed Reality Framework*. In *International Conference on Computer Graphics Theory and Applications - GRAPP 2006*. 25–28.

VIRTUAL MANUFACTURING SIMULATION

A VIRTUAL MANUFACTURING ENVIRONMENT FOR INTEROPERABILITY TESTING

Charles R. McLean, Sanjay Jain, Andreas Craens, and Deogratias Kibira
National Institute of Standards and Technology (NIST)
Gaithersburg, MD 20899-8260
Email: charles.mclean@nist.gov

KEYWORDS
Virtual manufacturing, simulation, information models, interface standards, interoperability, testing

ABSTRACT

Studies have documented the high cost of the lack of software interoperability on manufacturing. New interface standards are being developed by various organizations to address interoperability issues. Unfortunately these standards often overlap and conflict with each other. Adequate testing facilities are not available for evaluating the suitability and effectiveness of existing and candidate standards for application to specific manufacturing domain areas. The Virtual Manufacturing Environment (VME) Project of the NIST Manufacturing Interoperability Program is focused on developing integrated manufacturing simulations for supporting future industry interoperability testing needs. This paper presents a summary of interoperability testing opportunities and needs, an overview of a project to develop a virtual manufacturing environment to support interoperability testing, and the current status of simulation development efforts.

INTRODUCTION

Manufacturing systems often tend to be large, complex, and expensive to construct, operate, and maintain. Studies have documented the high cost and other impacts of the lack of software interoperability on manufacturing; see (NIST 1999). Software will continue to evolve and interoperability is expected to remain a problem area for manufacturers. New interface standards are being developed by various organizations to address interoperability issues. Unfortunately standards often overlap and conflict with each other.

The U.S. National Institute of Standards and Technology (NIST) has invested considerable effort in the development of test methods and tools that address manufacturing interoperability issues, see (Kulvatunyou 2004). To date, these methods and tools have largely focused on static testing, e.g., correctness of schemas and message formats. Adequate testing facilities are not available for evaluating the suitability and effectiveness of existing and candidate standards for application to specific manufacturing domain areas. New, dynamic, manufacturing domain-specific testing capabilities will also be needed to evaluate the suitability of suites of standards for selected applications, identify and resolve conflicts between standards, and evaluate compliance of vendor implementations with standards. Dynamic testing capabilities would enable the live testing of multiple manufacturing subsystems working together using various interface standards and protocols that have been developed by different standards organizations.

Software developers, research institutions, universities, and testing facilities cannot afford to duplicate real manufacturing systems in their laboratories due to the high costs of manufacturing hardware and software, system maintenance, and required space. Researcher hands-on experiences with manufacturing systems are often limited to individual or small groups of machine tools in laboratory shops, prototype work cells, or tabletop manufacturing systems. Manufacturing software development activities, training programs, research experiments, and testing activities could be significantly enhanced if manufacturing systems could somehow be brought into the laboratories of research institutions, universities, and software developers. Computer simulation technology now allows us to construct large, realistic virtual worlds in software. The military has made extensive use of this technology for a number of years. Virtual manufacturing environments could be used by a variety of organizations involved in manufacturing for training, experimentation, and testing purposes.

The Virtual Manufacturing Environment (VME) Project of the NIST Manufacturing Interoperability Program is developing integrated manufacturing simulations that can be used to support future industry interoperability testing needs. The rationale for this project is that it is impractical to use real industrial systems to support manufacturing interoperability research and testing due to:

- *Access issues* – Manufacturing facilities are normally not open to outsiders because businesses are concerned that proprietary data and processes may be compromised.
- *Technical issues* – Existing operational manufacturing and support systems are not likely to be properly instrumented to support testing needs.
- *Cost issues* – Manufacturers lose productivity when systems are taken offline to support testing.

No publicly available facility with open interfaces currently exists to support interoperability testing for a broad range of manufacturing interface standards and software applications. Prohibitive development costs and other priorities prevent most software vendors, research, and standards organizations from developing systems to support

such testing. Non-proprietary systems and neutral test case data sets are needed to support fair and open competition and accelerate the systems integration process.

TESTING OPPORTUNITIES AND NEEDS

Simulations can be used to model the behavior of and take the place of real manufacturing systems. What types of simulation applications are needed to support testing? A number of different clusters of manufacturing simulators are envisioned. Each possible cluster and its associated simulation applications are briefly introduced below:

- *Supply chain simulators* can be used to model the organization and management of supply chains. Organizations that may be simulated include supply chain headquarters, manufacturing primes, suppliers, transportation networks, warehouses, distribution centers, retailers, and customers. Some of the issues that may be addressed include lead times, inventory levels, production capacity, operations under surge conditions, and information flows.
- *Enterprise organizational simulators* can be used to model the internal business processes of various departments within the manufacturing organization, such as customer order servicing, design, engineering, production, and inventory management. Business process modeling techniques may be used to analyze order flow and processing times in order to streamline operations and minimize non value-added functions.
- *Manufacturing system and equipment simulators* can be used to model the normal operations, failure modes, and maintenance of various manufacturing equipment, such as fabrication, assembly, material handling, quality, and packaging systems. Examples of some of the equipment making up these systems includes machine tools, coordinate measuring machines, robots, storage and retrieval systems, and conveyors. Discrete event simulation techniques may be used to analyze operation times, capacity, queue lengths, bottlenecks, buffer storage requirements, inventory levels, etc.
- *Physical process simulators* can be used to create accurate models of the physical transformations that products and tooling undergo in various manufacturing industries. Industries that will have unique process simulations include metalworking, electronics, food, textiles, plastics, and chemicals/refining. For example, a physical process simulator for metalworking may model processes associated with a machine tool's operation. Information obtained from the simulation may include changes to work piece geometry, chip formation, tool wear, chatter, thermal and mechanical variations to the machine.

How might manufacturing simulations be used to support interoperability testing? Some examples of possible simulation-based testing applications include:

- Evaluate effectiveness of new interface standards and protocols to meet manufacturing industry needs.
- Evaluate conflicts and inconsistencies between standards developed by different organizations.
- Perform interoperability testing with models of systems being integrated. For example a model of a robot controller may be integrated with a model of the robot for testing purposes to ensure interoperability.
- Perform interoperability testing with emulated physical equipment. For example, a physical programmable logic controller may be tested with an emulated conveyor system before the physical conveyor system is installed or even delivered.
- Evaluate the capability of the delivered process, system or design to meet interface specifications.
- Perform conformance and acceptance tests using simulations to create the specified range of inputs for a delivered system or process.
- Evaluate whether new systems, processes or designs meet performance specifications. For example, test programs for robots and other machinery using simulations.
- Develop metrics to allow the comparison of predicted performance against "best in class" benchmarks to support continuous improvement of manufacturing operations.

Testing applications will also require that simulations be technically correct, i.e., accurately reflect the interfaces and behavior of real manufacturing systems. Models will need to be carefully validated; however, the procedures used may be more focused on functional and deterministic validation rather than statistical validation used for system-level research applications that use stochastic factors. The validation procedures should be defined to ensure common practices. Supporting applications that exercise the models through the range of parameters defined in the specifications should be provided to facilitate the process. The development of test cases would help manufacturers by providing a baseline that could be used for the initial screening of vendors. If a vendor's software passed initial tests, customers could proceed with testing candidate applications using company-specific data.

A number of different testing and support tools will be needed that may not normally be required in an actual production environment. What types of testing and support tools are needed? Some examples of interoperability testing instrumentation is briefly described below:

- *Module integration infrastructure* – This software is needed to allow simulation modules and other applications to be interconnected in such a way as to enable time synchronization and data sharing.
- *Communications channel monitors* – Mechanisms are needed which will allow test personnel to observe communications traffic between modules.
- *System and module status displays* – Displays are needed that will allow testers to track the health and status of applications under test.
- *Logging and reporting tools* – Various diagnostic tools are needed for recording traces of the execution sequence of modules being tested.
- *Message and file syntax checkers* – Syntax checkers are needed to validate that messages and data structures

written to files are properly formatted according to standards.

- *System initialization, control, and rollback utilities* – Functions are needed to initialize tests, place systems in proper modes, set checkpoints, and rollback tests and data sets to these checkpoints to avoid restarting entire systems unnecessarily during testing.
- *Configuration management and build software* – Support tools are needed to provide configuration management and system build functions for testing tools and test case data sets.
- *Testing policies, procedures, and checklists* – Testing procedures must be consistent and repeatable, standard policies and procedures must be established and documented to ensure that systems are evaluated fairly.

PROJECT OVERVIEW

The goal of the VME Project is to: 1) establish a virtual manufacturing environment based upon simulation technology that enables dynamic interoperability testing for manufacturing software applications, candidate interface specifications, protocols, and standards, and 2) provide interoperability testing support to software developers, manufacturers, research institutions, consortia, and standards organizations for selected manufacturing product domains, facilities, systems, operations, and processes.

The objectives that have been identified as steps to achieving these goals include:

- Identify and select high priority manufacturing product domains for implementation within the virtual manufacturing environment.
- Identify system functions and test case data for selected supply chain, facilities, systems and processes.
- Develop simulation models of manufacturing supply chains, facilities, and process lines.
- Identify and implement relevant neutral interfaces and standards to enable integration and data transactions between simulations and with external systems.
- Instrument simulations with monitoring, testing, and diagnostic tools to support interoperability testing.
- Establish interoperability testing services, procedures, and reports for internal and external customers.

What benefits may be expected from this project? The creation of a virtual manufacturing environment for interoperability testing:

- Establishes a baseline reference model for targeting interface standardization needs and helps minimize redundant specification efforts.
- Provides open, neutral test-based evaluations of interface standards and conforming software applications.
- Enables collaboration and cost sharing between competing organizations (e.g., manufacturers and software vendors) that have difficulty cooperating or sharing information in other venues.
- Provides neutral models, test case data, and support tools that can be used by software developers for self-testing and academic institutions for research purposes.
- Shortens development time for new standards and improves the interoperability of commercially developed manufacturing software applications.

The major components of the VME test bed are:

- Simulations of manufacturing systems at various hierarchical levels, e.g., supply chain, manufacturing plant, and shop floor.
- Design, engineering, and analysis systems for creating manufacturing data and populating life cycle databases,
- Instrumentation and testing tools for conducting interoperability tests.
- Test case data and databases in neutral format for conducting interoperability tests on real manufacturing systems.
- Interface standards, protocols, and communications mechanisms for connecting simulation modules and modules under test.
- Policies and procedures for providing interoperability testing services.

See Figure 1 for an illustration of the conceptual architecture of the VME test bed.

A phased approach is underway for the development of the VME test bed. The focus of Phase I was to develop test case data and stand-alone simulation models of an automotive supply chain, vehicle final assembly facility, and paint process line. This phase has been completed. Phase II, currently underway, is focusing on extending the simulations using selected interface specifications and standards to integrate the simulations with each other and other manufacturing software applications. In Phase III, monitoring, testing, and diagnostic tools to enable interoperability research and testing will be incorporated into the test bed and integrated with test bed systems. These will be designed to support the testing opportunities and described in the first section of this paper. Phase IV will open interoperability testing facilities and services for use by software developers, manufacturers, research institutions, consortia, and standards organizations.
Phases V and later will expand the VME test bed to include simulations of other manufacturing supply chains, facilities, systems, operations, and processes as well as additional interface specifications, testing tools, and data sets.

CURRENT STATUS

In this section, a summary of the current status of the simulation development is provided. Work on a standard manufacturing simulation data model is also highlighted.

Supply Chain Simulation

The purpose of the supply chain simulation is to provide a representation that generates dynamic information exchanges that would be created in a real life supply chain in order to test standards and supply chain applications'

compliance to standards. The simulation executes a model of interactions and material and information flows through a defined supply chain network extending from suppliers to customers.

The scope of the supply chain model includes manufacturing facilities with multiple stages of suppliers on the input side and multiple stages of distribution on the network side. Each of the supplier facilities is modeled at an abstract level based on the capacity of bottleneck and the lead-time through the facility. The flow of material is tracked at the supplier at three major stages: raw materials, work in process, and finished goods. Suppliers can send their outputs to multiple consuming facilities. For example, a tier II supplier can send its products to a tier I supplier and to the manufacturing facility directly.

The manufacturing facility itself is modeled in a bit more detail, with major sections (lines or departments) of the facility modeled with their individual bottlenecks. The flow of product is tracked through the stages of raw material (components), work in process within the major sections and in-between the sections, and finished goods. The production activity is modeled at shift level.

The distribution network can be modeled to include flow of product to distribution centers, retailers and customers either linearly through these stages or directly to any of them. Customer purchase activity through the retailers can be modeled by specifying appropriate distributions. The logistics is modeled at an abstract level with travel times defined in integer days based on a from-to matrix.

The current implementation of the model is based on a generic automotive supply chain data set with the final assembly plant at the center as the manufacturing facility, tier I and tier II automotive suppliers on the supply side, and distribution centers, car dealers and customers on the consumption end.

The model mimics the dynamics of the supply chain and associated interactions between the supply chain nodes. These interactions can be executed through transactional messages between the nodes consistent with a standard that may be under evaluation. With the current implementation of an automotive supply chain, the interaction messages use data fields consistent with those defined in OAGIS/AIAG Business Object Documents (BODs) for Inventory Visibility and Interoperability (IV&I) (OAGi, 2007). For example, the orders for vehicles from dealers to assembly plants are defined using XML fields defined in the "order" BOD of the OAGIS IV&I standard. Similarly, the shipment notifications that are sent from the assembly plant to dealers use XML messages that are formed using the corresponding shipment notice BOD specification.

The supply chain simulation has been developed using Rockwell Automation's ARENA discrete event simulator. The simulation is data driven with all the supply chain parameters defined in variable structures. Figure 2 shows a screenshot of the model with the generic logic modules identified. The data for the automotive supply chain has been defined in an Excel file and brought into ARENA to create the model. A graphical user interface (GUI) module has been developed using C## for flexibility in accessing and presenting information beyond that offered by ARENA. The GUI module communicates with the simulation model through an MS-Access database using .NET framework.

The input data for the automotive supply chain has been developed as a neutral data set. It is representative of industrial data based on discussions with automotive manufacturers and examples in the open literature, see (Schmenner 1981). It does not include any proprietary information and can be shared freely among researchers. The data includes a high level description of automotive bill of material, the suppliers for the major components, the assembly plant, the dealers and the logistics network connecting all the nodes. A more detailed, external simulation model of a manufacturing plant, described in the next section, can replace the assembly plant node in the supply chain model.

Manufacturing Plant Simulation

The purpose of the manufacturing plant simulation model is to generate dynamic information representative of a real life assembly plant to enable testing of standards and applications for plant level information and decision-making systems. The simulation mimics the flow of product and associated information in a manufacturing plant. The interfaces to the model can be built using selected standards. Similar to the supply chain simulation, the current plant simulation implementation uses interactions messages defined in OAGIS BODs for IV&I.

The scope of the manufacturing plant simulation model includes all the major sections of the plant with the key workstations represented. In the current implementation of an automotive assembly plant, its three major sections namely, body shop, paint shop and general assembly are modeled with a number of workstations for each connected by the appropriate material handling system. The body shop representation includes body assembly stations, conveyors and various kinds of buffers allowing stacking of auto body panels. As the car bodies move through the stations, the processing is modeled through passage of cycle time. The movement of equipment at each workstation such as robots is not modeled. The paint shop and general assembly area are modeled similarly except that the conveyance mechanism is different. A power and free conveyor is modeled with car carriers appropriate for moving the assembled car body through the stations.

The plant model includes tracking of component inventories in storage areas and on the assembly line. The inventories are used as the components are assembled on to the cars. Completely assembled cars move on their own power through test stations and any required rework to a

shipping area. The shipment from the plant to distribution centers and/or car dealers is modeled in the supply chain simulation described above.

The automotive assembly plant simulation model was developed using the Delmia QUEST discrete event simulation system, based on an internally generated neutral data set. The model includes high-level representations of body shop, paint shop and the general assembly area of an automotive plant.

Figure 3 shows a screenshot of the 3-D view of the modeled plant. The neutral data set for the assembly plant includes a more detailed bill of material than used for the supply chain simulation, definitions of stations on the line, and policies controlling the scheduling and flow of vehicles through the facility. In the future, the simulation will include a capability to replace line segments with an external detailed model briefly described next.

Line and Work Cell Simulations

The purpose of detailed simulations of parts of the manufacturing plant is to generate dynamic control messages representative of real life manufacturing lines and work cells to enable development and testing of standards and applications at that level. The current line level implementation uses CMSD based data files for reading in a description of a more detailed system. The paint process line of an automotive plant has been incorporated that uses Enterprise Dynamics software. The model is based on proprietary data and hence its details and screen images are not included. The experience will be useful for developing a publicly accessible model in the future.

Standards Development

If software applications including simulations are going to share data, they should have a common understanding of the meaning and structure of that data. The primary objective of the Core Manufacturing Simulation Data model (CMSD) is to develop a structure for exchanging shop data between various manufacturing software applications, including simulation. The idea is to use the same data structures for managing actual production operations and simulating the machine shop. The rationale is that if one structure can serve both purposes, the need for translation and abstraction of the real data would be minimized when simulations are constructed. The machine shop data model contains twenty major elements. The data elements are called: *Organizations, Calendars, Resources, Skill-definitions, Setup-definitions, Operation-definitions, Maintenance-definitions, Layout, Parts, Bills-of-materials, Inventory, Procurements, Process-Plans, Work, Schedules, Time-Sheets, References, Revisions, Units-of-measurement*, and *Probability-distributions*

Maintaining data integrity and minimizing the duplication of data is an important requirement. For this reason, each unique piece of information appears in only one place in the model. Cross-reference links are used to avoid the creation of redundant copies of data. The mapping of real world data into simulation abstractions is not, for the most part, addressed in the current data model. Due to space limitations, the entire model is not discussed in detail.

This information model and associated data formats are undergoing standardization under the Simulation Interoperability Standards Organization (SISO 2005). For a more detailed discussion of the CMSD model, see (Lee 2003) or (McLean 2005a).

CONCLUSIONS

The implementation of the Virtual Manufacturing Environment provides a test bed industry, government, and academic researchers to evaluate manufacturing interface standards. The VME can be used to test the interoperability of manufacturing applications including enterprise resource planning, scheduling, manufacturing execution systems, machine and material handling equipment control programs, and robot programs. It can also be used to test proposed standard interfaces for such applications.

The test bed will be highly effective if supported with additional test case data that is based upon real industrial scenarios and problems. The repository of test case data can also serve as a benchmark for comparison of alternate approaches for similar applications and thus further spur development. Test cases may also be used by industry to establish baselines for evaluating vendor offerings. It will also eventually provide a new set of capabilities not previously available to standards development organizations.

NIST researchers have prepared draft specifications for shop floor data and are working with the Simulation Interoperability Standards Organization for their formal acceptance. Work on the CMSD and implementation of the VME is expected to help identify the need for other manufacturing data standards including data models, interfaces, distribution and synchronization mechanisms.

REFERENCES

IEEE Standards Association. 2000. 1516-2000 IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) - Framework and Rules. (Accessed at http://standards.ieee.org/)

Kulvatunyou, B., K.C. Morris, J. Buhwan, and P. Goyal. 2004. "Development Life Cycle and Tools for XML Content Models," *XML Conference 2004*, International Digital Enterprise Alliance, Alexandria, VA.

Lee, Y. Tina, C. McLean, and G. Shao. 2003. "A Neutral Information Model for Simulating Machine Shop Operation." *Proceedings of the 2003 Winter Simulation Conference*, eds: S. Chick, P.J. Sanchez, D. Ferrin and D.J. Morrice, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2003, pp.1296-1304.

Leong, S. , Y.T. Lee, and F. Riddick. 2006. "A Core Manufacturing Simulation Data Information Model for Manufacturing Applications," *Proceedings of the 2006 Fall*

Simulation Interoperability Workshop, Simulation Interoperability Standards Organization.

McLean, C., Y.T. Lee, G. Shao, and F. Riddick. 2005. "Shop Data Model Interface Specification," *NISTIR 7198*, National Institute of Standards and Technology, Gaithersburg, MD, January 2005.

McLean, C., F. Riddick, and Y.T. Lee. 2005. "An Architecture and Interfaces for Distributed Manufacturing Simulation," *Simulation: Transactions of the Society for Modeling and Simulation International*, Volume 81, No. 1, Sage Publications, San Diego, CA, pp. 15-32.

National Institute of Standards and Technology (NIST). 1999. "Planning Report 99-1: Interoperability Cost Analysis of the U.S. Automotive Supply Chain." Gaithersburg, MD.

OAGi, 2007. Open Applications Group Integration Specification (OAGIS), Release 8.0 with Automotive Industry Action Group (AIAG) Overlay 1.0a, Document Number 20040923, Open Applications Group, Incorporated. Available online from the website: http://www.oagi.org/downloads/oagisaiag/loadform.htm [Last accessed online on Feb. 15, 2007]

Schmenner, R. 1981. "A Machine-Paced Line Flow Process; General Motors Corporation, GM Assembly Division, Tarrytown, NY," Production/Operations Management: Concepts and Situations, Science Research Associates, Chicago, pp. 75-95

Simulation Interoperability Standards Organization (SISO), Core Manufacturing Simulation Data Product Development Group, (accessed at http://www.sisostds.org/ on 8-18-05)

ACKNOWLEDGEMENT

Work described in this report was sponsored by the National Institute of Standards and Technology (NIST) Manufacturing Interoperability Program, Gaithersburg, Maryland. The Manufacturing Interoperability Program supports NIST projects applying information technologies and standards-based approaches to manufacturing software integration problems. The work described was funded by the United States Government and is not subject to copyright.

DISCLAIMER

Software architecture, models and languages are identified in context in this paper. This does not imply a recommendation or endorsement of the associated commercial software products by the authors or NIST, nor does it imply that such software products are necessarily the best available for the purpose.

BIOGRAPHY

CHARLES R. MCLEAN is a computer scientist and Group Leader of the Manufacturing Simulation and Modeling Group at the National Institute of Standards and Technology. He has managed research programs in manufacturing simulation, engineering tool integration, product data standards, and manufacturing automation at NIST since 1982. He has authored more than 50 technical papers on topics in these areas. He is on the Executive Board of the Winter Simulation Conference and the Editorial Board of the International Journal of Production, Planning, and Control. .He holds an M.S. in Information Engineering from University of Illinois at Chicago and a B.A. from Cornell University. His e-mail address is charles. mclean@nist.gov.

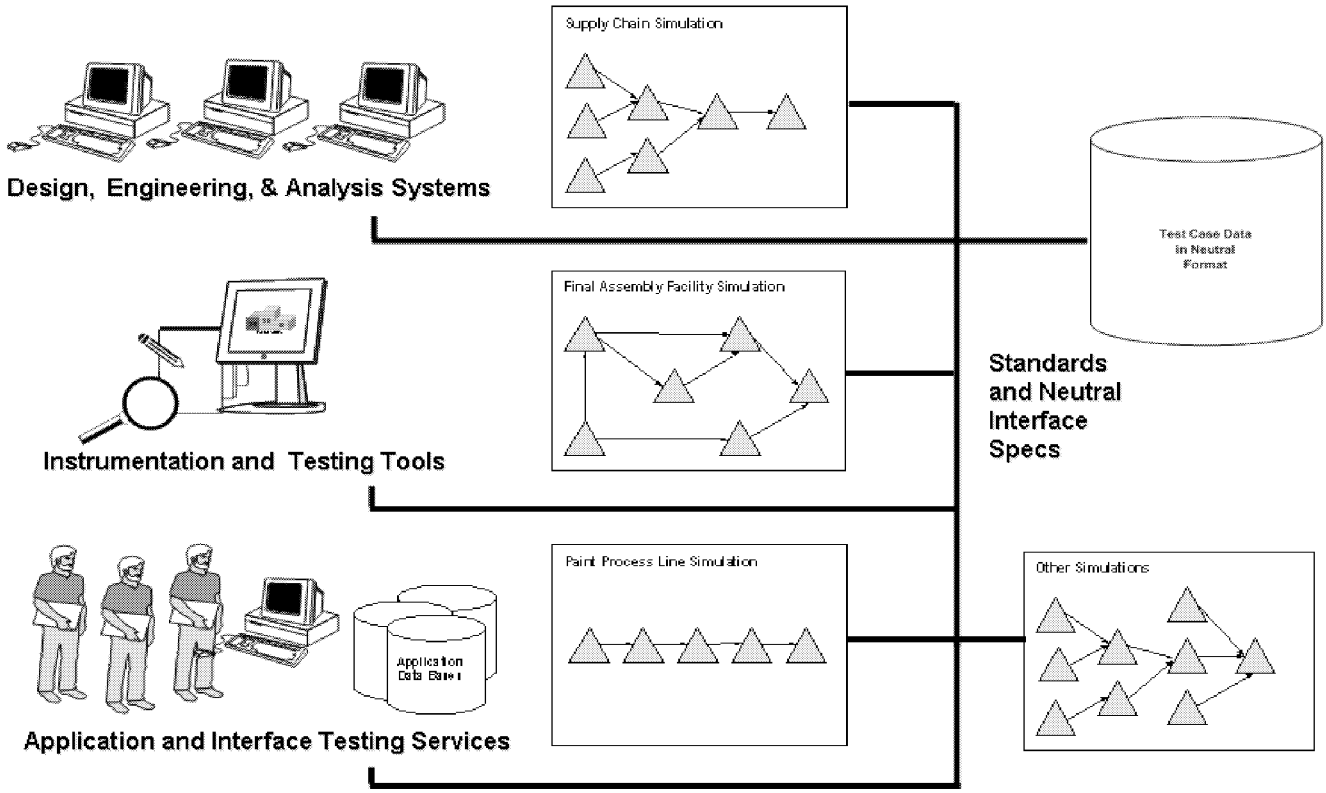


Figure 1: Conceptual architecture of Virtual Manufacturing Environment for interoperability testing

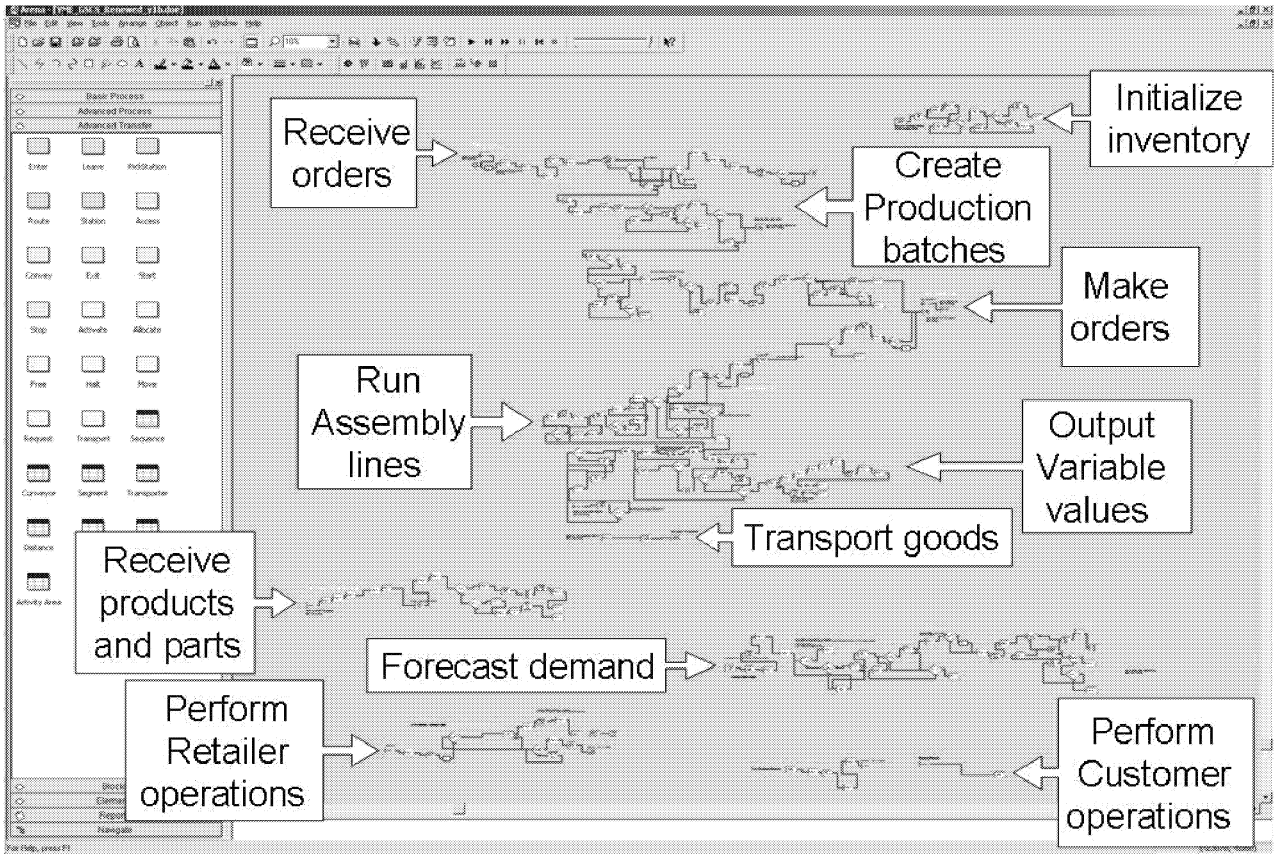


Figure 2: Screen image of supply chain simulation model logic in ARENA with annotations

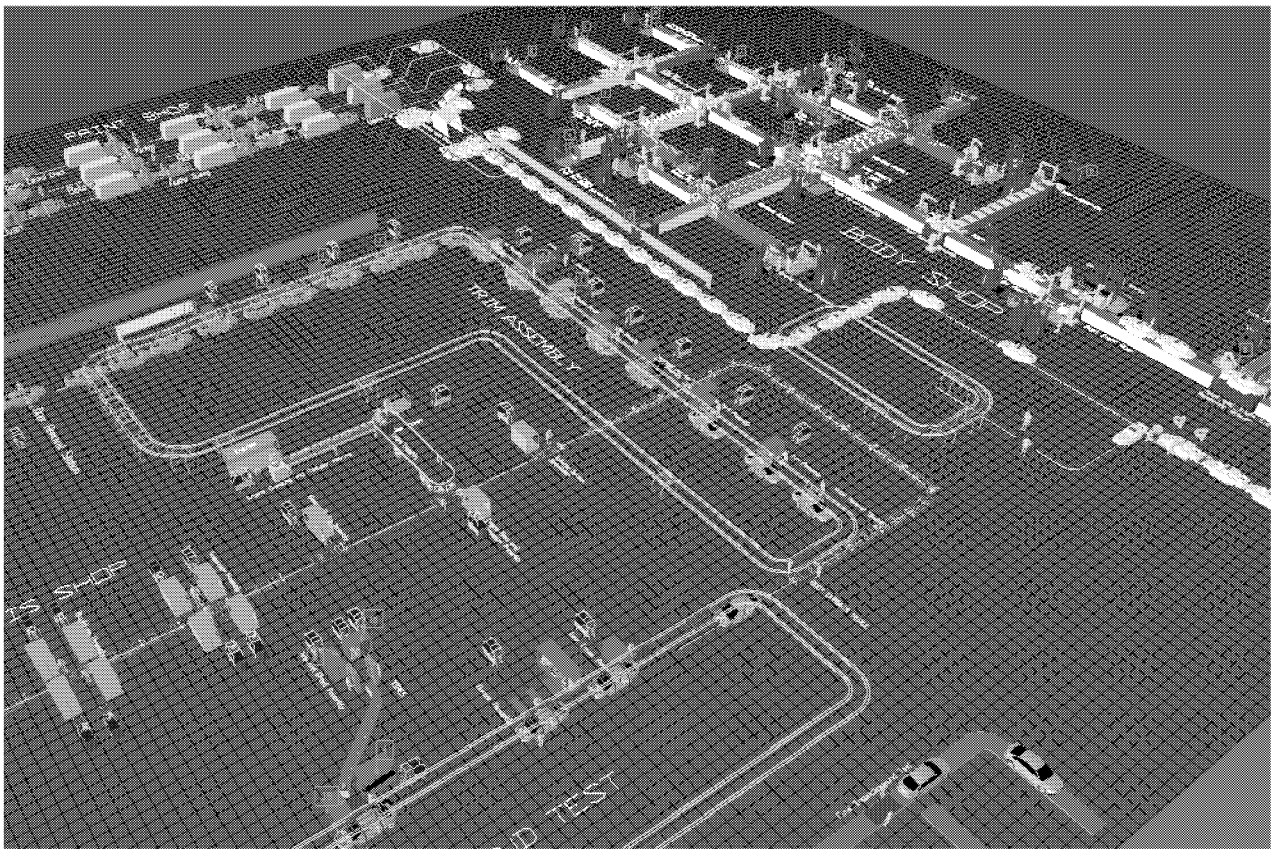


Figure 3: Screen image of the automotive manufacturing assembly plant simulation in QUEST

A software simulator for manufacturing plant

Osvaldo Barbarisi and Carmen Del Vecchio, Giovanni Palmieri and Alessandra Parisio*

Abstract

This paper presents a manufacturing system simulator MASYS suitable for solving machine conflicts in manufacturing production scheduling in the computer integrated manufacturing system (CIMS) environment.

The model is developed with a UML language and takes into account both production status and structural constraints of the production process.

1. Introduction

Modern industries are moving towards continuous, high-speed and automated production process with large devices. The production focus is placed on high quality, low cost and small lot of different varieties. In order to enhance competitive power, many international corporations are devoted at developing computer integrated manufacturing systems (CIMS) which can improve productivity of large devices, reduce waiting-time between operations, reduce material and energy consumption, and reduce production costs [1].

Production scheduling is a key component of CIMS. Its task is to determine the starting times and the ending times of production on the machines so that a chosen measure of performance is optimized [2].

In this paper production scheduling problems are investigated and a computerized scheduling system is presented. The manufacturing plant is considered as a discrete system and its logic is described by event-based model[3].

The industry we consider in this paper is very general: it produces several products, belonging to a macro-typology; every time a machine changes typology, it might be necessary a setup time, thus the optimal production sequence which minimize the setup time is crucial for the operation management.

The simulation can be defined both as a methodology to develop a model that represents the real system and as a test tool: using scheduling strategies it is possible to understand the performance of process. Using the simulator described in this paper it is possible to detect plant bottlenecks, moreover it is possible to evaluate the effects due to changes in layout, input and production. This allows to save resources and to avoid experiments on real system, dangerous for operators and plant.

The simulation helps the production management in solving problems regarding : machinery and employer assignment [4] [5], transport typology, position and buffers dimension, management of time for one task, evaluation of throughput, lead time, total queue time of product, measures of single queue at machine, production scheduling, control strategies.

We developed a general purpose MAnufacturing SYstem Simulator (MASYS) that can be customized to represent many production processes. The simulation model has been developed using UML language. The aim of this paper is to describe the simulator and to show the application to a production process.

2. Plant Description

The manufacturing plant we here consider has a structure commonly encountered in production industries: discrete entities, called parts, move through a network of machines (or group of them) where they are processed according to their production sequence.

Without loss of generality we here consider the production process consisting of four sequential phases. The working time of each machine, in each phase, is variable and it depends on the product type. Moreover each phase could be carried out by different machines, in this case there are alternative sequence production for the same product type.

Consider each production sequence as a job. In the plant can be done L different jobs $\mathscr{J} = \{J_1, J_2, ...J_j, ...J_L\}$ which differ for colors, material, dimensions, production sequence process duration [6]. Moreover parts belonging to the same job can have the same production sequence code; parts with equal production sequence code are processed by the same set of machines. It is also possible that for some step, a job has two different machines that can process it. The plant consists of M + 3 machines $M = \{M_1, M_2, ..., M_m, ..., M_M\} \bigcup \{M_D, M_H, M_F\}$, M operative machines which process parts, and three "dummy" machines, M_D, M_H and M_F , useful to the process descrip-

*The authors are with Dipartimento di Ingegneria, Università del Sannio, Italy – E-mail: {barbarisi, c.delvecchio, palmieri, aparisio}@unisannio.it.

tion: the dispatcher machine M_D directs each part to the first working machine in its production sequence; the machine M_H is necessary to guarantee the movement of product between two different transport equipment and the machine M_F holds finished parts. Associated to each operative machine M_m there is a capacitated buffer. All machines do the same sequence of actions: they take a part from their own buffer and work it.

Parts are chosen according to a predefined priority order; parts with higher priority are worked first. At the end of the process the part is sent into the buffer of the next operative machine. Unfitting parts are sent to a correction loop (i.e., some machines rework the parts to make them fitting quality requirements); discarded parts are sent to the discarded machine. The goal of simulation is to find the production scheduling that guarantees that different products will be produced in different typology optimizing the time and the resources.

We introduce four temporal variables for the production plant:

- working time: the total time spent by the part into the machine
- waiting time: the total time spent by the part waiting to be processed by a machine
- setdown time: time needed to at the end of a production type sequence to reset the machine
- setup: time needed to set the machine for next operation

The waiting time could be a critical parameter; sometimes it is constrained by a maximum value, that can not be exceeded otherwise parts deteriorates. The working time can be variable. The waiting time is a critical parameter and it cant exceed a maximum value otherwise part can be deteriorated. The working time can be variable.

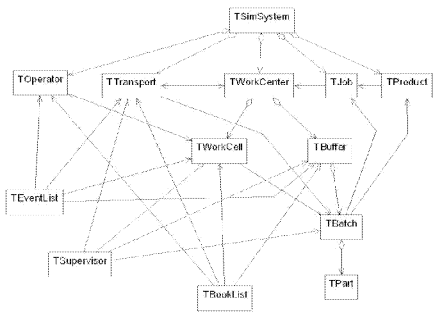


Figure 1. Principal diagram class

3. Simulation model

In this section we will present the structure and the behavior of the UML simulation model [7]. The class diagram describes the model objects and the static relationships among them; the statechart diagram represents the dynamics of each object. The principal classes of model (showed in figure 1) are described below:

TPart. The TPart class is used to model the single unit to produce. A batch is a union of several parts. When the simulation starts, each part will have its instance of TPart class, and it will insert in the batch. The number of parts and batches are decided by production plain. Two codes identifies each part: a code referees to the product which the part belongs; the second code identifies the part itself.

TProduct. The TProduct class is used to model the type of product that the simulator has to manage. It is an extension of TPart with several properties that identify the part that have to be worked. Each instance has the sequence of operation that have to be done on batches of products that the instance represents.

TBatch. The TBatch class is used to model the group of part that move together in the plant. One batch has to follow the product process specified in the corresponding TProduct.The figure 2 shows the finite-state machine of this object and the conditions that have to be validated to active one transition.

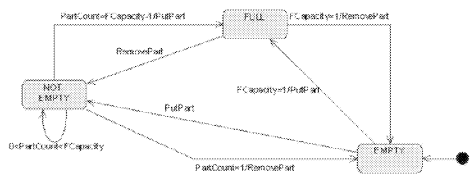


Figure 2. State diagram of TBatch.

An object of TBatch can be in three states. Empty: there aren't parts. Not Empty: there are some parts, but it is possible put other. Full: there are the maximum number of parts

TWorkCenter. The class TWorkCenter represents a set of WorkCells. A Workcell is the elementary unit processing a TBatch. Each WorkCenter is represented by a class-instance which it is possible to assign one or more bridgecranes to. Workcenter of different department can be connected by one or more vehicles, while in each department inner connection are realized with bridgecranes or conveyer.

All WorkCells of one WorkCenter share common

resources such as inner buffer. Figure 3 shows the generalization of WorkCenter:

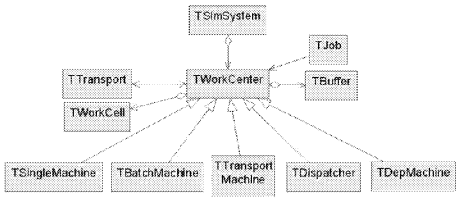


Figure 3. Class diagram of TWorkCenter class.

- TDispatcher models the Dispatcher. It simulates the part of process plant before the simulated one. It also can represent the input buffer of whole plant. During the simulation, the scheduler scans all batch into the own buffer and sends each one to the buffer of the first WorkCenter that will process it. TSingleMachine is used to model a particular WorkCenter that works a single part at a time.
- TBatchMachine is used to model a particular WorkCenter that works more parts together at glance.
- TTransportMachine is used to model a particular WorkCenter that needs of transport equipment to send finished part. Each object has a list of assigned transport equipment.
- TDepMachine is used to model a storage [8], a place where the batch wait for being worked. There is one storage for each department of system.
- THideMachine is used to model a virtual WorkCenter, that is necessary to model the transfer of parts between different departments of system.

The WorkCell is modeled as a state-finite machine, each state represents a working state of object:

- Idle: WorkCell is ready to start with a new part;
- WOpSetup: WorkCell is waiting for operators necessary to set-up the machine;
- WBatchSetup: part loading is done;
- Busy: WorkCell is working a part;
- Blocked: WorkCell has finished its process, but it is not able to send the part to next machine;
- Fault: WorkCell cannot process;
- Maintenance: WorkCell is in maintenance.

TOperator. The class TOperator models the resource human operator. Each operator is enable to work on different machines and he can be skilled to several operations.

It is possible assigned different shifts for each typology of operator.

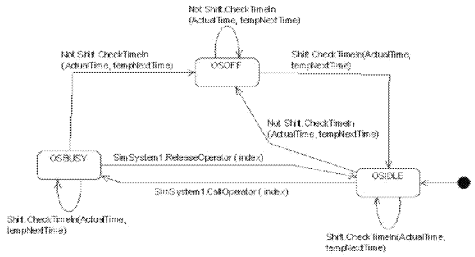


Figure 4. State diagram of TOperator class.

TTransport. The TTransport class is used to model the transport equipment; it is used to move whole batches or single part between WorkCenter; these machines are implemented with instance to inheritance classes from TTransport: TBridgeCrane, TConveyer and TRailCar. The transfer between two machines belonging the same department is done by bridgecrane or by the conveyer. At this stage conveyer can transport only between two machines.

The transport between two different department is done by railcar: the bridgecrane of departing department put the batch on truck, the truck arrives to the destination department where there is another bridgecrane that brings it. When a machine has to send a part to a machine in a different department, it calls a bridgecrane and railcar that can do the path in the minum time. When it find them, the bridgecrane brings the part to the THidemachine of destination department and drop the part into the railcar. This one transport the part to the THidemachine of the destination department where another bridgecrane will bring the part to the destination machine. The logical scheme is shown in figure 5.

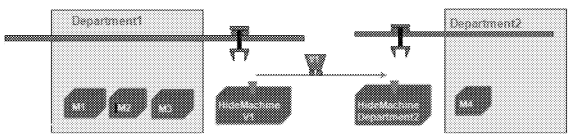


Figure 5. Transfer between two different departments.

TJob. The TJob class models the sequence of operations that have to be executed on batch belonging to the same product's type. For one sequence it is possible

to specify an alternative machine that is used when primary workcenter is busy.

TBuffer. The TBuffer class is used to model the places where the batches are positioned waiting for enter in a workcell (inner-buffer) or to move to transport equipment (output-buffer).

A TBuffer object can be in three different state. Empty: there is no part to be processed. Not Empty: there are some parts waiting to be processed but the number is less than the maximum. Full: there are the max admissible number of parts.

TSimsystem. The TSimSystem class represents the whole system, so there is only one instance of this class. It contains all the references to all the components of the system.

TSupervisor. This class is the core of the simulator. It manages all the rules that are not local to each object but interferes with two or more of them. As example it manages the shifts and skills of operators; it decides which product has to be worked and which production cycle has be done; it manages all storages and transport equipments.

4. Interface User-Simulator

The interface of simulator [9] is implemented in two levels: the designer of plant and the manager of production. Each interface has different level of access. It is possible to create or modify a XML file where are stored all information about plant and production plan. The configuration file of the system is done by simulator that creates an TSystem instance.

After the simulation several statistics allows to analyze the system performance. The interface of the simulator is divided in two part: on the left side there is the browser of the plant: here, with hierarchical list, the structure of plant is represented. It contains the name of all object that the simulator manages (Workcenter, Job, Part, Transport, Shift (and Skill); on the right side there is the graphical visualizer of the system state. There are seven parts: Input, WorkCenter, Output, Discard, Operator, Bridgecrane, Railcar and Conveyer. The change of state for each object is visible with change of colors of rectangles related to each object.

4.1. Plant layout

An hypothetical plant of ten machines is depicted by its layout in the figure 6. The plant is divided in four departments (blue rectangles); each department has own machines and bridgecranes. The departments are connected by railcars.

The result shown in the next section refers to a

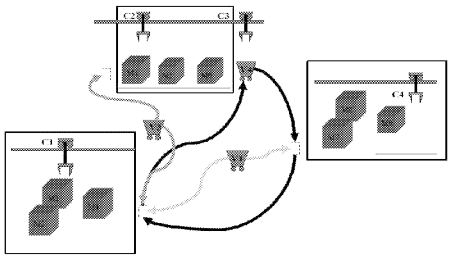


Figure 6. Layout of simulation.

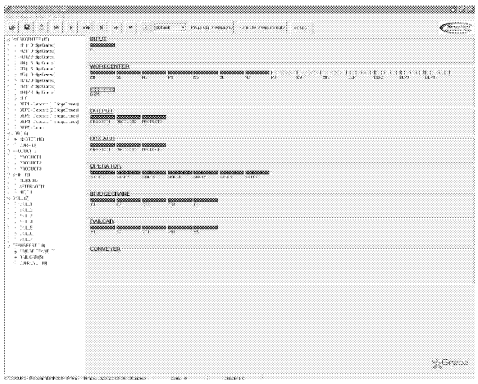


Figure 7. User interface of Masys simulator.

simulation of 655 parts belong to 2 different macro-typology of product. PRODUCT1 is produced in 11 different jobs and PRODUCT2 in 9, each jobs identifies a production sequence. Each sequence is composed by three o four phase and the working time for each phase depends on the machine and on the typology of product.

5. Statistic after simulation

In this section some statistics obtained after a simulation with MASYS simulator are shown. In each simulation step the information of systems are memorized in several variables such as the permanence in each state, the occupation of buffers in the time, the typology of part carried from the transport equipment, and so on.

The first result of stabilities is the Gantt diagram of whole production, as shown in figure 8. In the abscissa axis there is the time while in the ordinates axis there are all machines in the plant. Each part being processed is represented by a rectangle (the color depends by its product type) where is written the relative code. At the end of one product sequence there is a setdown time that is drawn with black-line.

With the statistic it is possible to analyze the behavior of element of the system to estimate the performance of plant.

In figure 9 it is shown the permanence of all Work-

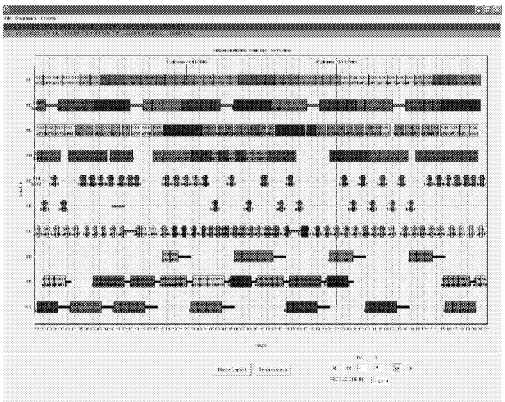


Figure 8. Gantt diagram of production.

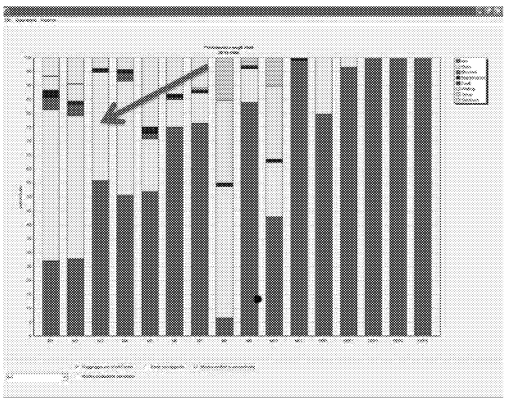


Figure 9. Permanence of stated for workcell M3.

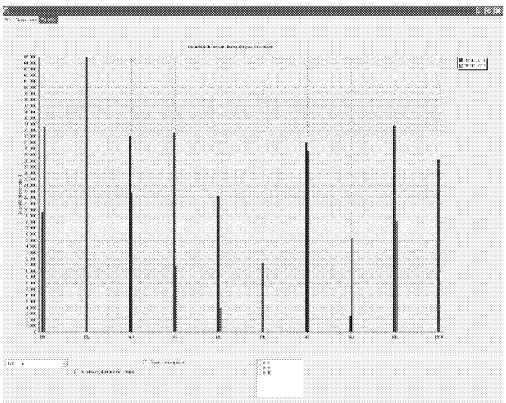


Figure 10. Working part of each machines in the plant.

Cells in its possible state. On ordinate axes there is the time duration as percentage along the object has been in the same state. For example in the illustrated diagram there are multicolor bars: it's means that during this interval the machine has been for 39% in Busy (yellow), for 4% in Blocked (red), 2% in Maintenance (blue), 3% (grey) in waiting operators and for 52% Idle (green).

Figure 10 depicts the number of working parts by each Workcenter. Each workcenter has a bar for each worked part. Each part has a owner color and the height is proportional to amount of worked parts by that machine. So in this case the Machine M6 has worked 19650 part of product PRODUCT1, 33600 of PRODUCT2.

6. Conclusion

In this paper we illustrated a simulator for complex real system such as manufacturing plants. After the definition of fundamental components of the system, the simulator MASYS can simulate the whole production of the system and the evolution of produced process. The output of the simulation are several statistics that allow to analyze the behavior of system, underline the critical point, verify the respect of process restrictions, such as priority, waiting times, and evaluate the effects of layout changes. The next step will be to resolve the problem of scheduling in the production process by introducing optimization strategies. For this aim the simulator will be tested with data of a real plant systems.

References

[1] L. Tang, J. Liu, A. Rong, and Z. Yang, "A mathematical programming model for scheduling steelmaking continuous casting production," *European Journal of Operational Research*, pp. 423–435, 2000.

[2] D. Pacciarelli and M. Pranzo, "Production scheduling in a steelmaking-continuous casting plant," *Computers and Chemical Engineering*, pp. 2823–2835, 2004.

[3] I. Harjunkoski and I. E. Grossmann, "A decomposition approach for the scheduling of a steel plant production," *Computers and Chemical Engineering*, 2001.

[4] A. M. Law, Michael, and G. McComas, "Simulation of manufacturing systems," *Proceedings of the Winter Simulation Conference.*, 1999.

[5] J. W. Fowler and O. Rose, *Grand challenges in Modeling and Simulation of Complex Manufacturing Systems*. The MITRE Corporation, McLean, VA.

[6] O. Barbarisi and C. Del Vecchio, "Uml simulation model for hybrid manufacturing systems." 13th IEEE Mediterranean Conference on Control and Automation, June 2005.

[7] O. Barbarisi, C. Del Vecchio, L. Glielmo, and F. Vasca, "Why adopting uml to model hybrid manufacturing systems?" IEEE International Conference on Systems, Man and Cybernetics, October 2004.

[8] S. Zanoni and L. Zavanella, "Model and analysis of integrated productioninventory system: The case ofsteel production," *Science Direct*, 2005.

[9] O. Barbarisi, C. Del Vecchio, and M. D. Placido, "Strategie di scheduling applicate ad un simulatore ad oggetti per impianti di produzione." Anipla, November 2005.

COMMUNICATIONS MODELLING AND SIMULATION

MODULAR APPROACH BASED DECISION SUPPORT FOR COMMUNICATION SYSTEMS

Calin Munteanu
Simona Caramihai
Janetta Culita
„Politehnica" University of Bucharest
Automatic Control and Computer Science Faculty
Splaiul Independentei 313, Sector 6, Bucharest, Romania
e-mail: mc_aurel@yahoo.com

KEYWORDS
Hybrid Petri nets, Decision Support Systems, performance evaluation, simulation.

ABSTRACT
It is well known that the complex system operation requires the use of new scientific tools and computer simulation. This paper presents a modular approach for modeling and analysis of the complex systems (in communication or transport systems area) using Hybrid Petri nets. The performance evaluation of the hybrid model is made by a simulation methodology that allows building up various functioning scenario. A Decision Support System based on the above mentioned methodologies of modeling and analysis will be designed for performance evaluation and time optimization of large scale communication and/or transport systems.

INTRODUCTION

Complex systems represent a fast growing area of interest for control and optimization. Domains as communication and transport are important fields of interest from this point of view, due to the fact that the present tendency is to build aggregated networks either informatics or of transport (as the European unified railway system project).

Performance evaluation, analysis and optimization of these networks necessitate particular methodologies and formal instruments. Since pure mathematical models are not easy to manage for these categories of systems, decision is often taken on the basis of the simulation of various functioning scenarios. Decision Support Systems (DSS) ensuring the appropriate framework for modeling, analysis and comparison of those scenarios are valuable tools for network managers.

The paper aims to propose a DSS architecture for complex network analysis, based on hybrid Petri Nets modeling and analysis.

Petri Nets (PN) represent a formal tool that was defined in 1962 by Karl Adam Petri, for modeling and analysis of information systems. Their modeling power, capacity of representing concepts as parallelism, synchronization, resource sharing a.s.o. in a clear, intuitive graphical format, have contributed to their wide development and utilization, for very different domains..

The paper will use the Hybrid Petri Nets (HPN) formalism, defined in (Le Bail, *et al.*, 1991; David and Alla, 2005), as the basic approach for designing a Decision Support System for communication and/or transport networks.

The next section will present the specifications of the network analysis problem, the modeling tool and the consequent structure and operation of the DSS.

Section 3 will present a short case study, illustrating the methodology of analysis implemented by the DSS. Finally, the conclusion section will present some future research directions.

THE MODEL-BASED APPROACH

Communication systems analysis – problem context

The paper is considering large communication networks with high dynamics of server load, where large amounts of information have to be transferred with given time constraints. It is assumed that there are several paths for information transfer between two given servers, usually none of them being sufficient for the transfer in due time.

The analysis has as an objective to find an acceptable combination of paths, that used together will allow to accomplish the transfer problem.

Communication, as well as transport networks, have in common the fact that they have to transfer a certain flow of entities (informational or physical) from one starting point to a destination, via a net of possible paths, which will be called connections, usually with some time constraints. The availability of connections can vary in time, as well as the amount of entities to be transferred and the time limits, so a solution will satisfy only a given set of constraints and will be unique. On the other hand, usually there is searched only a valid solution and not the best one, even if

some optimization is desirable.

Since usually the global network is constituted from different sub-networks, there was taken into account the possibility of composing sub-models into a global one, to be analyzed.

The information to be transferred, consisting of a given number of packages, has to reach the destination into a given amount of time. The packages can be sent by different routes, as the destination node has the possibility to order them as necessary for reconstructing the information. Therefore, an efficient distribution of packages on different paths, starting from the source node could improve the overall transmission time. The problem to be analyzed is how to distribute packages on different routes, according to their availability and transfer speed, in order to meet the overall time constraints. Obviously, the routes will not consider twice the same node.

Modeling framework

Taking into account the above presented problem context, it is necessary for the modeling tool to reflect the amount of information to be transferred, the availability of every path and the transfer speed dynamics. Consequently, Hybrid Petri Nets (HPN) were considered as the most appropriate choice.

Continuous PNs were defined recently (David and Alla, 1987). Autonomous continuous PNs have been shown to be a limit case of discrete PNs. In a continuous PN, the markings, arc weights and firing quantities are non-negative, as in a discrete PN, but are *not necessarily integers*. In a timed continuous PN, maximal speeds are associated with transitions. Other authors have proposed various results concerning these PNs (Demongodin, *et al.*, 2000).

The initial motivation leading to the concept of continuous PN was an approximate modeling of discrete systems with a large number of states, as a consequence of the management of a large number of entities by the modeled process. By this approach it was possibly to analyze the flow of entities instead of following the evolution of every one. Another domain of application of this type of nets is for continuous system modeling.

Hybrid PNs contain a discrete part and a continuous part, usually interacting (Le Bail, *et al.*, 1991; David and Alla, 2005). Given a timed hybrid PN (timings associated with discrete transitions and maximal speeds with continuous transitions), instantaneous behavior is analyzed in the following way: a stabile marking of the discrete part is sought, then the instantaneous firing speeds of the continuous transitions are computed.

The semantics related to instantaneous firing speeds is relatively easy to define. However, automatic calculation for the whole continuous PN is difficult. Iterative algorithms presented in (David and Alla, 1987) and (Alla and David, 1998) do not work in all cases. Calculation by resolution of a linear programming problem (LPP) was used for some specific cases in (Balduzzi, *et al.*, 2000).

The bases of a speed calculation method for continuous PN are presented in (Munteanu, *et al*, 2004). More details are given in (David and Alla, 2005) and software for analysis of Hybrid and Continuous Petri nets is available at { HYPERLINK "http://www.lag.ensieg.inpg.fr/sirphyco/telecharger/sirphyco.zip" }.

DSS Structure

The block architecture of the Decision Support System is presented in figure 1. Its structure was designed taking into account the specifications of the problems it has to resolve.

Consequently, the DSS will include a net-editor that will allow the user to design models of networks represented in HPN formalism.

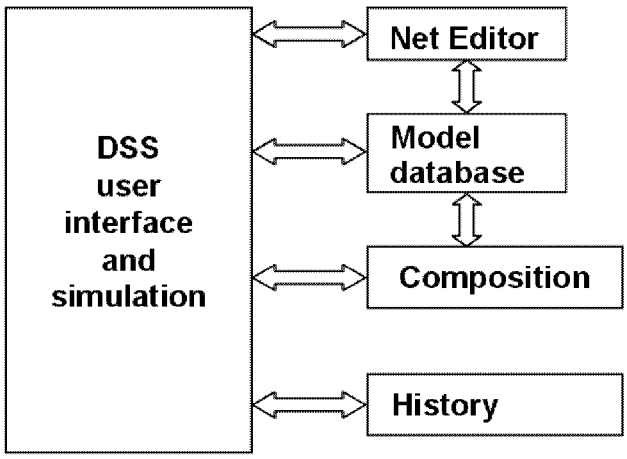


Figure 1: DSS Structure

The nets can be either directly analyzed or stored in a model database. With the models from the database, the composition block permits the building of larger global nets. Finally, the history module let the user to either store analyzed scenarios or to verify and compare previous analyses.

The user interface module allows the user to initialize the structural models with actual parameters and especially to simulate the net functioning.

CASE STUDY

Figure 2 presents a communication sub-net that aims to transmit a number of information packages from the node 1 to node 4, eventually using internal nodes 2 and 3. Thus, the possible transfer routes are: 1 -> 4, 1 -> 2 -> 4;

1 -> 3 -> 4; 1 -> 2 -> 3 -> 4 and 1 -> 3 -> 2 -> 4 as specified in the figure.

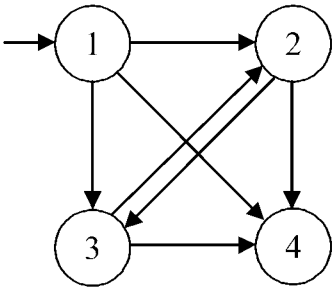


Figure 2: Communication system structure

As a part of a larger communication system, it is assumed that nodes 1, 2 and 3 perform also other jobs more or less important than the transfer activity. Moreover, for a particular system configuration some physical connections between nodes could be not available. The priorities of jobs and the availability of connections will be modeled by the continuous transitions speeds and priorities.

Figure 3 illustrates the hybrid Petri net structural model of the communication sub- system. For simulation purposes the marking will be initialized in order to reflect the associated connection state. It should be noted that even if a connection is available at a certain moment, it could physically breaks-down. The presence/absence of a connection is modeled by the discrete transition - place group.

In figure 3, transitions $T_1 \div T_3$ model the working speed of the nodes 1 ÷ 3. Continuous place P_1 is associated to node 1. The continuous transitions $T_4 \div T_6$ model the transmission speed from node 1 to nodes 2, 4 and 3 respectively. T_{15} is used for modeling the other jobs the node 1 can execute, other than the transmitting activity, with higher priority. Similarly T_{16} models the supplementary jobs of node 1, other than the transmitting activity, having lower priority. During the simulation of different activity scenarios, speeds and priorities can be modified in order to compare functioning regimes.

Place P_5 models the number of packages to be transmitted to the destination. The modeling approaches for both nodes 2 and 3 are similar: P_2 models node 2 and P_3 models node 3. From these nodes the information could be transmitted towards node 3 or 4 (from node 2), respectively nodes 2 or 4 (from node 3) by transitions T_7, T_8 and T_9 respectively T_{10}, T_{11} and T_{12}. The transition T_7 models the information transmission from node 2 to node 3 and T_{12} from node 3 to node 2. For the route 2 -> 4 transitions T_8 and T_9 are used that distinguish between the ways 1-> 2- > 4 and 3 -> 2 -> 4. Identically, in order to model the transmission on the route 3 -> 4 T_{10} and T_{11} are used (corresponding to connections 1-> 3 and 2 -> 3).

Places $P_6 \div P_9$ are intermediate and model the information flow transferring from node 1 to node 2 (P_6), 1 to 3 (P_7), 3 to 2 (P_8) and 2 to 3 (P_9). The transitions T_{14} and T_{17} model the tasks more important than the transmission of the nodes 2 respectively 3; the transitions T_{13} si T_{18} model the tasks less significant than the transmission of the nodes 2 respectively 3. The place P_4 corresponds to node 4. The transfer towards node 4 is already encoded in the transitions T_5, T_8, T_9, T_{10} and T_{11}.

All low priority transitions (T_{13}, T_{16} and T_{18}) will have infinite maximal execution speed so that the places P_1, P_2 and P_3 will not accumulate tokens.

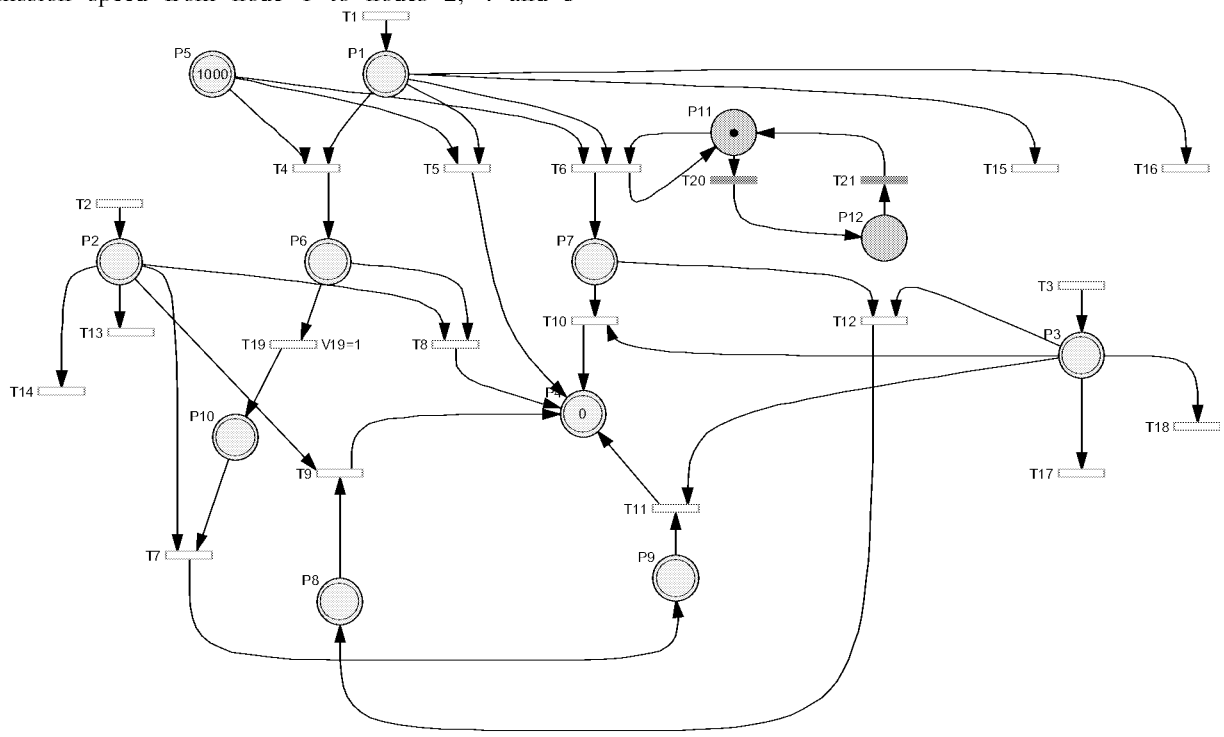


Figure 3: Hybrid Petri net model

System analysis (scenario analysis)

Each transition $T_4 - T_{11}$ is enabled also by a discrete place. By unmarking a discrete place, the absence of the associated connection between two nodes is evidenced. There are two configuration possibilities: one of them is the setting of the discrete marking; the other consists in time association to discrete transitions (temporizations), so as enabling /disabling (of a connection) is realized during the system analysis

Different analysis scenarios could be obtained by setting the maximal speeds associated to either the transfer activities or other processing jobs. Its values could be constant a priori established, piecewise constant or stochastic (generated by the computer) on time intervals. Also by setting certain priorities/sharing for different possible transfer route various functioning scenario will be constructed.
If a certain route is preferred, the associated modeling transition will be associated with higher priority than the other conflicted transitions modeling the other routes. Should no route is favorite (main), all the transitions will have the same weight in a common sharing group. The sharing situation must be carefully chosen, as in the RPC analysis algorithms there are only three priority/sharing combinations (David and Alla, 2005).

The analysis of the communication /transport system assumes the following steps:

- setting the configuration of a communication network (the user sets the initial discrete marking that corresponds to available connections);
- setting the package transmission speeds on each direct link. This could be done either by the user or by the system. In the last case stochastic values could be associated to the speeds.
- choosing the number of packages composing a message;
- setting the maximum time of message transmission.

Besides these initializations, before the simulation begins, a priority level has to be assigned to each transition.

Example

A numerical example will be analyzed in order to illustrate the effective speed computing. It is assumed that all connections are available. The continuous marking is set for position P_5 (1000). The maximal speed vector is :

V_j=[4, 3, 5, 1, 1.5, 2, 1, 0.5, 1, 1, 1, 2, -1, 0.5,0.5,-1,1,-1,1],

The transition priorities are:

- for P_1 place: $T_{15}<T_5<T_4<T_6<T_{16}$;
- for P_2 place: $T_{14} < T_8 < T_9 < T_7 < T_{13}$;
- for P_3 place: $T_{17} < T_{12} < T_{10} < T_{11} < T_{18}$;
- for P_5 place: $T_5 < T_4 < T_6$;
- for P_6 place: $T_8 < T_{19}$;
- for P_7 place: $T_{12} < T_{10}$;

Consequently, the instantaneous speed vector is:
V = [4, 3, 5, 1, 1.5, 1, 0.5, 0.5, 1, 0, 0.5, 1, 0.5, 0.5, 0.5, 0, 1, 2.5, 0.5].
Then, the marking evolution (for places with non-zero marking evolution) is:

m4(t) = 3.5*t
m5(t) = 1000 - 3.5*t

From the instantaneous speed vector values, it could be observed that, even if a connection is available, it may not be used due to time consuming server utilization. The other parallel routes realize a transmission time of 286 t.u.

Taking into account two modes of solving the transition conflict, the decision support system will analyze all the possible configurations. The first configuration meeting the time criteria will be selected.

If the time criteria is not accomplished then:

- if there is a place that accumulates markings its input transition will be set with lower priority;
- a larger number of parallel routes are considered;

Before simulation starts the Decision Suport System will chose first (for transitions that model information transmission) the priority level according to the transitions maximal speeds: the transition with biggest maximal speed will have the highest priority level. But this doesn't always lead to best results, as it is shown in the following example:

Suppose that the Petri Net's maximal speeds are:

V = [4, 3, 5, 3, 2, 2, 1, 0.5, 1, 1, 1, 1, ∞, 0.5, 0.5, ∞, 1, ∞]

Because T_4 has the biggest maximal speed (from T_4, T_5 and T_6), it will receive initially the highest priority level (from the three transitions).

Case A

Conflict resolutions are: $T_{15} < T_4 < T_5 < T_6 < T_{16}$ (for conflict resolution associated to P_1 place) and $T_4 < T_5 < T_6$ (for conflict resolution associated to P_5 place). With these maximal speeds and priority levels, the continuous Petri net will have the following evolution:

The first instantaneous transition speed vector will be:

V = [4, 3, 5, 3, 0.5, 0, 0.5, 0.5, 0, 0, 0.5, 0, 1.5, 0.5, 0.5, 0, 1, 3.5, 0.5]

Because the input transition of place P_6 (i.e. T_4 - 3) has an instantaneous speed greater than the sum of maximal speeds for output transitions of P_6 (i.e. $T_{19} - 0.5$ and $T_8 - 0.5$), the markings are accumulated in P_6. This means that:

- node 2 must have a buffer to store the parts that it cannot deliver (due to speed limitations);

- a second evolution phase is needed in order to deliver the parts from P_6. This evolution is characterized by the instantaneous speed vector: V = [4, 3, 5, 0, 0, 0, 0.5, 0.5, 0, 0, 0.5, 0, 1.5, 0.5, 0.5, 3.5, 1, 3.5, 0.5]

First phase will end in 286 time units and the system will send all the packages in 857 time units.

Case B

A lower priority level will be associated to transition T_4 .? Giving to T_4 a priority level lower then T_5, the conflict resolution rules would be: $T_{15}<T_5<T_4<T_6<T_{16}$ (for P_1) and $T_5<T_4<T_6$ (for P_5).

With these new priority levels, the first phase of the continuous Petri net will be characterized by the following instantaneous transition speed vector:
V = [4, 3, 5, 1.5, 2, 0, 0.5, 0.5, 0, 0, 0.5, 0, 1.5, 0.5, 0.5, 0, 1, 3.5, 0.5]

The input transition of P_6 will still have an instantaneous speed (i.e. 1.5) bigger than the sum of maximal speeds for output transitions of P_6 (i.e. 0.5+0.5=1), so a second evolution phase is needed. This is characterized by:

V = [4, 3, 5, 0, 0, 0, 0.5, 0.5, 0, 0, 0.5, 0, 1.5, 0.5, 0.5, 3.5, 1, 3.5, 0.5]

First phase will still end in 286 time units and the system will send all the parts in 429 time units.

Case C

Transition T_4 will be set with the lowest priority level. Transition T_5 keeps its higher priority and T_4 will have a priority level lower then T_6; the conflict resolution rules would be: $T_{15}<T_5<T_6<T_4<T_{16}$ (for P_1) and $T_5<T_6<T_4$ (for P_5).

With this new priority levels, the first phase of the continuous Petri net will be characterized by the following instantaneous transition speed vector:
V = [4, 3, 5, 0, 2, 1.5, 0, 0, 1, 0.5, 0, 1, 1.5, 0.5, 0.5, 0, 1, 2.5, 0]

This is the only phase needed to transfer all the packages from node 1 to node 4. The phase time is 286 time units.

From all three cases it is noticeable that (because of node 1 limitation) only one intermediate node is selected for message transmission and the total number of routes is three.

As it was mentioned before, there is the possibility of increasing the number of routes. This situation appears if neither node 2 nor node 3 could send the packages as fast as they receive them.

CONCLUSIONS

The paper presents a decision support system constructed on a modular approach for modeling and analysis of the complex systems in communication/transport area. The partial models of intermediate nodes could be composed in order to obtain the whole system model. RPH was chosen as modeling and analysis tool due to a significant modeling power appropriated for complex systems thought as hybrid systems (David si Alla, 2005).

The decision support system inspects the scenarios provided by the system analysis and proposes a time suited solution.

The future research trend is to search the optimal solution corresponding to the minimum transmission time. In this purpose, all the possible situations for priorities/sharing allocation will be analyzed.

REFERENCES

Alla, H. and R. David. 1998. "A Modelling and Analysis Tool for Discrete Event Systems: Continuous Petri Net", *Performance Evaluation*, vol. **33**, pp. 175–199.

Balduzzi, F., A. Giua, and G. Menga. 2000. "First-Order Petri Nets: A Model for Optimization and Control", *IEEE Trans. on Robotics and Automation*, vol. **6,** N° 4.

David, R. and H. Alla. 1987. "Continuous Petri Nets", *8th European Workshop on Appli. and Theory of Petri Nets*, Saragosse (E), pp. 275-294.

David, R. and H. Alla. 2005. *Discrete, Continuous, and Hybrid Petri Nets*, Springer, Heidelberg, ISBN: 978-3-540-22480-8.

Demongodin, I., M. Mostefaoui, and N. Sauer. 2000. "Steady State of Continuous Neutral Weighted Marked Graphs", in *Proc. Int. Conf. on Systems, Man and Cybernetics*, IEEE / SMC, USA, pp. 3186-3194.

Le Bail, J., H. Alla, and R. David. 1991. "Hybrid Petri Nets", *Proc. European Control Conference*, Grenoble (F).

Munteanu, C, David, R and H. Alla – "Algorithmic speed calculation for a timed continuous Petri net" - *7th Workshop on Discrete Event Systems (WODES)*, Reims, France, 2004.

Petri, C. A. 1962. *Communication with Automata*, Supplement 1 to Technical Report RADC-TR-65-337, N.Y., 1965. Translation by C.F. Greene of Kommunikation mit Automaten, PhD Dissertation, University of Bonn.

Recalde, L. and M. Silva. 2000. "PN Fluidification Revisited: Semantics and Steady State", *Proc. Int. Conf. on Automation of Mixed Processes: Hybrid Dynamic Systems* (ADPM 2000). Dortmund (D), pp.279-286.

Special Issue on Hybrid Petri Nets (2001). *Journal of Discrete Event Dynamic Systems: Theory and Applications*, vol. **11**, A. Di Febbraro, A. Giua, and G. Menga Eds.

COMBINATION AND INTERWORKING OF FOUR MODELLING METHODS FOR INFOCOMMUNICATIONS AND BUSINESS PROCESS SYSTEMS

Gábor Lencse
Department of Telecommunications
Széchenyi István University
Egyetem tér 1.
H-9026 Győr, Hungary
e-mail: lencse@sze.hu

László Muka
Elassys Consulting Ltd.
Bég utca 3-5.
H-1026 Budapest, Hungary
e-mail: muka.laszlo@elassys.hu

KEYWORDS

discrete-event simulation, traffic-flow analysis, entity flow-phase analysis, information and communication systems, business process systems

ABSTRACT

This paper describes how we can use four modelling methods together. We model Information and Communication Technology (ICT) systems and Business Process (BP) systems. The aim of the modelling is to experiment with the models, that is, simulation. Two of the methods are the detailed modelling of the ICT and of the BP systems, as we usually do that for discrete event simulation. The other two methods are those that enable us to make rapid performance analysis of these systems, namely Traffic-Flow Analysis (TFA) for ICT systems and Entity Flow-phase Analysis (EFA) for BP systems. The solutions for the combination and interworking of these methods are described.

INTRODUCTION

Mixed Simulation Projects

Simulation projects aimed at supporting the design of Information and Communication Technology (ICT) systems and Business Process (BP) systems in an organisation are traditionally independent, separate projects, in spite of the fact that these systems may have significant influence on each other. Common analysis of these systems may have advantages, but in this case we need to have methods appropriate for both types of systems. That is, we need models of the two types of systems that can interact with each other, because the two types of systems interact with each other in the real world, too.

Detailed Modelling of ICT Systems: DES-IT

As for the detailed modelling of ICT systems, we usually model some servers (database, web, etc.) and the communication networks. We model the networks as a graph of nodes and lines. The load of the systems we describe by applications that use some services. They cause load for the servers by their requests, and for the network by communication: sending and receiving packets to/from the servers (or sometimes to/from each other). Discrete event simulation can be used to make experiments with this type of models. We use the abbreviation DES-IT, to distinguish it from the simulation model of the BP systems.

There are well known simulators for ICT systems, especially for communication networks, such as: OPNET Modeler, BONeS or ImiNet (Elassys 2007).

Detailed Modelling of BP Systems: DES-P

As for BP systems, we use the definitions we have given in (Lencse and Muka 2006):

"Business processes are related to enterprises and they define the way in which the goals of the enterprise are achieved.

A Business Process is a set of Enterprise Activities linked together to form a process with one or more kinds of input to produce outputs."

We model a BP system by a graph built up from activities (as nodes) and links (as edges) between them. The load is modelled by entity generators, which generate entities of different types according to given probability distributions. The entities travel through the network according to the routing decisions of the activities. The activities use resources of given types with limited capacities. The activities forward the entities after some processing delay that depends on the type of the entity, the type of the activity and the availability of the limited resources.

The simulation model we use for the detailed discrete-event simulation of a BP system is named DES-P.

Simulators for BP systems include SIMPROCESS, ARIS and ImiFlow (Elassys 2007).

Similarities & Differences Between DES-IT and DES-P

Hopefully, the reader has already noticed the formal similarity between DES-IT and DES-P. In both cases we have a graph, and some objects are travelling through the graph. However, the interpretation of the model elements is very different, that is why the simulators of the two types of systems are quite different.

Rapid Modelling of ICT Systems: TFA

The Traffic-Flow Analysis (TFA) (Lencse 2001) is a simulation-like method for the fast performance analysis of

communication systems. TFA uses statistics to model the networking load of applications.

In the first part, the method distributes the traffic (the statistics) in the network, using routing rules and routing units.

In the second part, the influences of the finite line and switching-node capacities are calculated.

The important features of TFA:

- The results are approximate but the absence or the place of bottlenecks is shown by the method.
- The execution time of TFA is expected to be significantly less than the execution time of the detailed simulation of the system.
- TFA describes the steady state behaviour of the network.

As TFA is a less well-known method, it has only one partial implementation, which is a part of the aforementioned ImiNet network expert system.

Rapid Modelling of BP Systems: EFA

The Entity Flow-phase Analysis (Lencse and Muka 2006) has been derived from TFA. This derivation is based on the formal similarity of the DES-IT and DES-P models. EFA uses the same two phase method as TFA, only the interpretation of the model elements is different. The statistics represent entities (not messages) and the interpretation of the routing is also different. While the packets of a network usually do not multiply, the entities may fork (and the descendants must meet somewhere) or split (and the descendants live their own life separately); see more details in the aforementioned paper.

An implementation of EFA is planned as an extension for the ImiFlow system.

Our Goal is to Use the Four Methods Together

Why do we believe, that the combination of the four methods may be beneficial?

Combining ICT and BP Models
As we have already mentioned, the ICT and BP systems are not independent, but interact with each other. If we model and simulate them separately, we usually consider the effects of the other one to the modelled one only as parameters or random variables. Modelling the two systems together (and their interactions) may give a richer picture of the whole system containing ICT and BP parts.

Combining the Detailed and the Rapid Models
As a motivating example, let us consider an ICT system: a high-availability server (realised by a cluster of multiple servers), its immediate network neighbourhood and the whole network of the company that uses the server. To study the behaviour of the high-availability server, we need the detailed simulation of the high-availability server and its immediate network neighbourhood, but the detailed simulation of the whole network of the company (consisting of hundreds of switches and thousands of computers) would require so much computing power (and perhaps even memory) that we can not afford. However, the omission of the network and direct connection of all the client applications to the high-availability server would oversimplify the situation. The application of DES-IT for the high-availability server and its immediate network neighbourhood and the application of TFA for the rest part of the network may be beneficial.

In the rest of this paper we examine how the four modelling methods can be used together for more precise or for more computation efficient performance analysis of a system containing both ICT and BP parts.

COMBINATION OF THE FOUR MODELLING METHODS

To be able to use two (or more) methods together, we must be able to teach them 'how to speak each others language'. Figure 1. shows four possible interactions between the neighbouring methods. These are:

1. DES-IT – TFA
2. DES-P – EFA
3. DES-IT – DES-P
4. TFA – EFA

The first one has already been discussed in details in (Lencse 2004), now we just give a short summary.

Combination of DES-IT and TFA

We have an ICT system. One segment of the system is modelled for DES and the other segment is modelled for TFA. Of course, traffic exists between the two segments. To enable the two models to speak each others language, we need bidirectional conversion between their different traffic representations – messages and statistics – on the boundary of the two segments. When the traffic travels from the DES segment to the TFA segment, the representation mode is changed from messages to statistics. It means that we need to collect the appropriate statistical characteristics of the message flow, and perhaps it is necessary to transform the results to the kind of statistics that are used in the TFA segment. In the opposite direction, where the traffic travels from the TFA segment to the DES segment, messages should be generated on the basis of the statistics (called *traffic model* in the original TFA terminology) of the TFA segment. The characteristics of the message flow such as the distribution of the message length, inter-arrival time, the source and destination of the packets are representing the same information that is coded in the statistics travelling from the TFA segment to the DES segment.

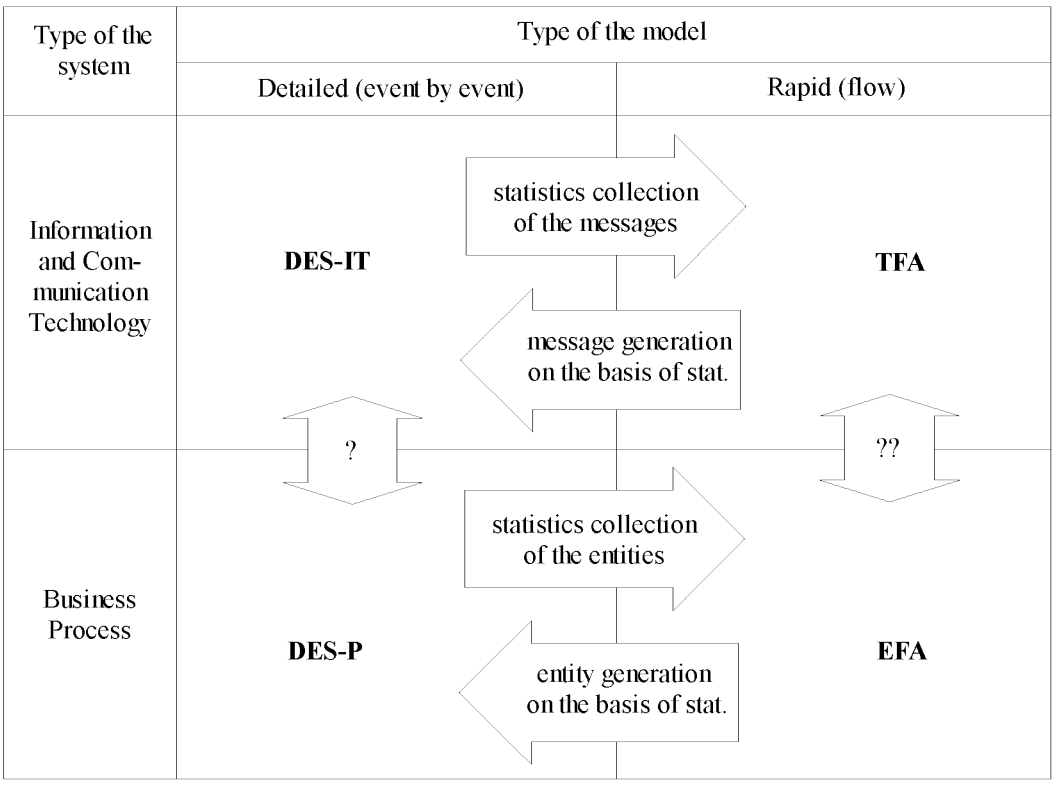


Figure 1. - Relationships between modelling approaches – we have two open questions

Combination of DES-P and EFA

Using our previous observation of the formal similarity of DES-IT and DES-P, plus the fact that EFA is nothing else than the application of the TFA principles for a BP system, we can reuse our previous results for this case. We can say the same about the combination of DES-P and EFA that we have just said about the combination of DES-IT and TFA, we must only replace the words "messages" and "traffic model" by the words "entities" and "entity-load model", respectively.

Combination of DES-IT and DES-P

Despite the formal similarity of DES-IT and DES-P, the combination of the two models is not trivial at all. The formal replacement of messages by entities and vice versa would result in a nonsense system! Their semantics are very much different!

Now let us identify, what kind of interactions are possible between the ICT and the BP systems. We have found the following three possible cases:

1. An activity of the BP system uses the ICT system.
2. An activity of the BP system sends information to another activity of the BP system using the ICT system.
3. The ICT system acts as an initiator towards the BP system.

Let us examine the points above one by one in detail.

1. An Activity of the BP System Uses the ICT System

As an example, let us consider a help desk system, where a consultant (part of the BP system) would like to know a client's e-mail address and makes a database query. (See Figure 2.) This query causes a load for the ICT system (for both the database server and the communication network between the consultant's computer and the server). The response time depends on the type of the given query and the other load of the ICT system, that is the other traffic that uses the same network and the other queries sent to the same database server.

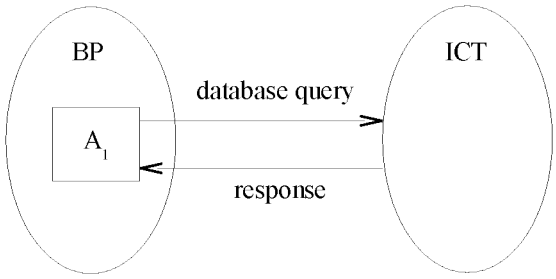


Figure 2. Database request and response

By modelling the BP and the ICT systems separately, both models would be less accurate. As for the BP system, we could model the response time with a random variable. As for the ICT system, we could model the different database queries by a packet source, that sends requests to the database server through the network according to a given time distribution (e.g. according to a Poissonian process).

The combined modelling of the BP and the ICT systems can be much more accurate.

2. Two Activities in the BP System Use the ICT System for Communication

To continue the previous example, let the consultant send an e-mail to the client. Similarly to the previous case, the BP systems causes load to the ICT system. Similar things can be said about the situation as we said before, but there is a difference: there is communication between the two activities of the BP system. This communication is also modelled within the BP system even when the two systems are simulated separately.

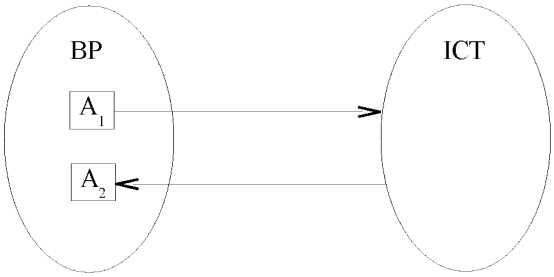


Figure 3. Communication through the ICT system

3. The ICT System Acts as an Initiator Towards the BP System

Let us begin with two examples. When a car driver uses GPS navigation, he regularly gets information from the navigation software (e.g. by voice messages or by checking the screen of his PDA). When something crucial happens, e.g. the cooling water of the car boils, the driver gets a warning message.

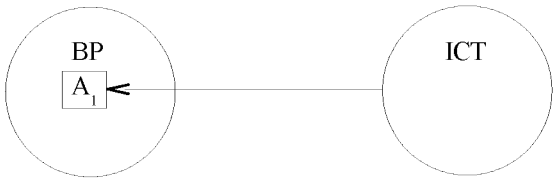


Figure 4. The ICT system initiates towards the BP system

In both examples, the ICT system produces entities (carrying information) that are processed by an activity (performed by the driver). As we have seen it before, the quality of our model of the BP system was improved by connecting it to the ICT model.

We have not found any other cases that are significantly different from the three cases above. All other interactions can be described by their combination. Let us see examples for the composition. If the above mentioned consultant mistypes the e-mail address of the client, he will probably get an error message from the outgoing SMTP server of his company. We found that the generation of this error message is similar to case number one: if the e-mail address is invalid, the generation of the error message will surely happen, and its time depends on the load of the mail handling subsystem, like in the case of the database query mentioned before. However, there can be a different way of receiving an error message. The client may further process, e.g. redirect his mails to a different address. If the error message is produced because of the client's improper mail settings, we might model the situation by case three, where the ICT system acted as an initiator towards the BP system. As a third example, if the consultant asks for a read confirmation when sending the e-mail to the client, the sending of the read confirmation we see as another action of the client (as an activity), and as it is an e-mail sending, it can be categorized as case two.

Combination of TFA and EFA

The reader may ask: Why do we deal with this question at all? At an intersection of two roads three pedestrian crossings are enough. Really? See figure 5. Probably all of us would like to choose the shortest path. That's why we must cope with this problem of conversion between TFA and EFA. However we can use the detour for the solution. First, we interpret the entity statistics of EFA with the help of the DES-P model. Second, we examine the effect of the entities to the DES-IT model. Third, we collect statistics of the messages for the TFA model. To explore the effect of the TFA statistics to the EFA part, we follow the same detour in the opposite direction. However, we need to do this only once per message/entity types in the phase of model building, that is, before the execution of the simulation. We make rules (or tables) for the conversion, and we can use them during the simulation.

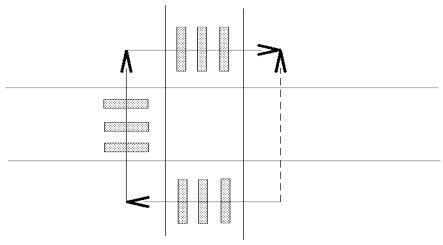


Figure 5. Are three cross-walks enough?

THE PROBLEM OF INTERWORKING

So far, we were concerned only with the conversion between the different traffic representations (e.g. between DES-IT and TFA) or interpretations (e.g. between DES-IT and DES-P) of the models. It was crucial to bridge the gap between them, otherwise we would have had no chance to make them work together. Is there any other problem to solve? Yes! There is the problem of model time (also called virtual time). This problem requires three different handling methods depending on which two models we want to cooperate.

Interworking of DES-IT and DES-P

If we are lucky enough and use the same modelling framework for the ICT and BP parts of the system and we also execute the model as a single process, then the event-driven discrete-event simulation kernel handles the events

correctly: the causality is ensured between the two parts of the model. (This is the situation in the case of ImiNet and ImiFlow.) If we use two different frameworks for the ICT and BP parts of the system, we must equip them with proper input and output interfaces that provide the data for and receive the data from the other model. Plus we have to deal with time synchronisation. We can use one of the well known synchronisation methods for parallel discrete-event simulation (Fujimoto 1990), namely conservative or optimistic. If the applicability criteria (Lencse 1999) are met, we may consider SSM-T, the time-driven version of the Statistical Synchronisation Method (Lencse 1998).

Interworking of DES-IT and TFA

Here the problem arises from the different usage of virtual time of the two methods. Virtual time elapses in the DES-IT, but TFA describes the steady state behaviour of the system. This problem was explored in full depth and two different solutions were proposed in (Lencse 2004). These solutions can be used for the interworking of DES-P and EFA, too.

Interworking of TFA and EFA

The virtual time handling of the two methods is the same, we only have to take care of the good interweaving of the traffic of the models, that is their own traffic and the traffic coming from (or caused by) the other part. For more details see (Lencse 2001), (Lencse 2004) and (Lencse and Muka 2006).

Interworking of more than two types of models

The interworking of more than two types of models seems to be possible and can be the direction of future research. The parallel execution of multiple models by multiple computers may result in a good speed-up, if we consider the results in (Lencse 2005), where only two of the methods (DES-IT and TFA) were used together.

CONCLUSIONS

We have briefly introduced four methods (DES-IT, DES-P, TFA and EFA) for the performance analysis of ICT or BP systems.

We have shown the way for their combination: how they can be combined in spite of their different traffic representation or interpretation.

We have also given the solutions to the problem of their interworking, ie. the synchronisation of their virtual time or the reconsiliation of their different usage of virtual time.

We conclude that the combination and interworking of the four methods may result in more accurate modelling, better simulation results and may also give a good chance to efficient parallel execution of the simulation of a system containing ICT and BP parts.

REFERENCES

Elassys Consulting Ltd. 2007. ImiNet and ImiFlow Systems http://www.elassys.hu

Fujimoto, R. M. 1990. *Parallel Discrete Event Simulation.* Communications of the ACM 33, no 10, 31-53

Lencse, G. 1998. *Efficient Parallel Simulation with the Statistical Synchronization Method*" Proceeding of the Communication Networks and Distributed Systems Conference (CNDS'98), (San Diego, CA, USA, January 11-14) SCS, 3-8

Lencse, G. 1999. *Applicability Criteria of the Statistical Synchronization Method* Proceedings of Communication Networks and Distributed Systems Conference (CNDS'99), (San Francisco, CA, USA, January 17-20.) SCS, 159-164

Lencse, G. 2001. *Traffic-Flow Analysis for Fast Performance Estimation of Communication Systems* Journal of Computing and Information Technology 9, No. 1, 15-27.

Lencse, G. 2004. *Combination and Interworking of Traffic-Flow Analysis and Event-Driven Discrete Event Simulation* Proceedings of the 2004 European Simulation and Modelling Conference (ESM®'2004) (Paris, France, Oct. 25-27. 2004.) EUROSIS-ETI, 89-93.

Lencse, G. 2005. *Speeding up the Performance Analysis of Communication Systems* Proceedings of the 2005 European Simulation and Modelling Conference (ESM®'2005) (Porto, Portugal, Oct. 24-26.) EUROSIS-ETI, 329-333.

Lencse, G., Muka, L. 2006. *Expanded Scope of Traffic-Flow Analysis: Entity Flow-Phase Analysis for Rapid Performance Evaluation of Enterprise Process Systems* Proceedings of the 2006 European Simulation and Modelling Conference (ESM'2006) (Toulouse, France, 2006. Oct. 23-25.) EUROSIS-ETI, 94-98.

BIOGRAPHIES

GÁBOR LENCSE received his M.Sc. in electrical engineering and computer systems at the Technical University of Budapest in 1994 and his Ph.D. in 2000. The area of his research is (parallel) discrete-event simulation methodology. He is interested in the acceleration of the simulation of info-communication systems. Since 1997, he works for the Széchenyi István University in Győr. He teaches computer networks and networking protocols. Now, he is an Associate Professor. He is a founding member of the Multidisciplinary Doctoral School of Engineering, Modelling and Development of Infrastructural Systems at the Széchenyi István University. He does R&D in the field of the simulation of communication systems for the Elassys Consulting Ltd. since 1998. Dr Lencse works part time at the Budapest University of Technology and Economics (the former Technical University of Budapest) since 2005. There he teaches computer architectures.

LÁSZLÓ MUKA graduated in electrical engineering at the Technical University of Lvov in 1976. He got his special engineering degree in digital electronics at the Technical University of Budapest in 1981, and became a university level doctor in architectures of CAD systems in 1987. Dr Muka finished an MBA at Brunel University of London in 1996. Since 1996 he has been working in the area of simulation modelling of telecommunication systems, including human subsystems. He is a regular invited lecturer in the topics of application of computer simulation for performance analysis of telecommunication systems at the Multidisciplinary Doctoral School of Engineering, Modelling and Development of Infrastructural Systems at the Széchenyi István University of Győr.

MODELING AND SIMULATION OF SELF-SIMILAR TELETRAFFIC

Dimitar Radev
University of Rousse
8 Studentska Street
7017 Rousse, Bulgaria
E-mail: dradev@abv.bg

Izabella Lokshina
SUNY Oneonta
324B Netzer Administration Bldg.
Oneonta, NY 13820, USA
E-mail: lokshiiv@oneonta.edu

KEYWORDS
Modeling and simulation in telecommunication networks, Self-similar teletraffic, Random number generation.

ABSTRACT

The paper examines self-similar (or fractal) properties of real teletraffic data over a wide range of time scales. These self-similar properties are very different from the properties of traditional models based on Poisson and Markov-modulated Poisson processes. The algorithms for modeling the sequentional generators and the fixed-length sequence generators that are used to simulate self-similar behavior of real teletraffic data are developed and applied. Numerical examples and simulation results are provided.

INTRODUCTION

The growth of broadband networks and the Internet has been exponential in terms of users and user-end systems as well as in traffic in recent years. High-speed communications networks are able to support a wide range of multimedia applications, such as audio, video and computer data that differ significantly in their traffic characteristics and performance requirements (Giambene, 2005). One possible goal for telecommunication developers is to build a unified high-speed communications network platform capable of carrying diverse traffic and supporting diverse levels of quality of service (QoS). Recent studies of real teletraffic data have shown that teletraffic (technical term, identifying all phenomena of control and transport of information within the telecommunications networks) exhibits self-similar (or fractal) properties over a wide range of time scales (Boxma & Cohen, 2005). The properties of self-similar teletraffic are very different from properties of traditional models based on Poisson, Markov-modulated Poisson, and related processes (Trivedi, 2001). The use of traditional models in networks characterized by self-similar processes can lead to incorrect conclusions about the performance of analyzed networks (Faraj, 2000). Traditional models can lead to over-estimation of the network performance/quality, insufficient allocation of communication and data processing resources, and hence difficulties in ensuring the QoS (Rubinstein & Melamed, 1998).

Self-similarity can be classified into two types: deterministic and stochastic (Jeong, 2002). In the first type, deterministic self-similarity, a mathematical object is assumed to be self-similar (or fractal) if it can be decomposed into smaller copies of itself (Park & Willinger, 2000). That is, self-similarity is a property, in which the structure of the whole is contained in its parts. This work is focused on stochastic self-similarity. In that case, probabilistic properties of self-similar processes remain unchanged or invariant when the process is viewed at different time scales. This is in contrast to Poisson processes that lose their burstiness and flatten out when time scales are changed. However, the time series of self-similar processes exhibit burstiness over a wide range of time scales. Self-similarity can statistically describe teletraffic that is bursty on many time scales.

One can distinguish two types of stochastic self-similarity. A continuous-time stochastic process $\mathbf{Y}_t$ is strictly self-similar with a self-similarily parameter H ($1/2<H<1$), if $\mathbf{Y}_{ct}$ and $c^H\mathbf{Y}_t$ (the rescaled process with time scale ct) have identical finite-dimensional probability for any positive time stretching factor c (Kushner, 2001). This definition, in a sense of probability distribution, is quite different from that of the second-order self-similar process, observed at the mean, variance and autocorrelation levels. The process $\mathbf{X}$ is asymptotically second-order self-similar with $0.5<H<1$, if for each k large enough $\rho_k^{(m)} \rightarrow \rho_k$, as $m\rightarrow\infty$, where $\rho_k=E[(X_i-\mu)(X_{i+k}-\mu)]/\sigma^2$. In this work exact or asymptotic self-similar processes are used in an interchangeable manner, which refers to the tail behavior of the autocorrelations (Heyes & Ganesh Babu, 2004). Modeling and simulation of self-similar traffic is performed with the generators of synthetic self-similar sequences, divided into two practical classes: the sequential generators and the fixed-length sequence generators. Both classes of generators are considered in this study.

SEQUENTIAL GENERATORS OF SELF-SIMILAR TELETRAFFIC

It is possible to construct a sequential Markovian model that imitates a self-similar sequence. A disadvantage of this approach is in the connection between the model parameters and its self-similar properties that is difficult to understand. Markovian models for self-similar traffic require including several control parameters with a wide range of input values. As a result, controlling these values in sequential generators is much more complex than in generators of fixed-length sequences of self-similar processes with a given Hurst parameter (Stathis & Maglaris, 2000).

We considered and investigated the following efficient candidate sequential generators, based on:

- Fractal-binomial-noise-driven Poisson processes;
- Superposition of fractal renewal processes (SFRP);
- M/G/∞ processes (MGIP);
- Pareto-modulated Poisson processes (PMPP);
- Spatial renewal processes and fractional Gaussian noise;
- Superposition of autoregressive processes (SAP).

For the standard fractal renewal process (FRP), inter-event times are independent random variables (Bobbio et al, 2003). The marginal probability density function (PDF) of such a fractal renewal process can be defined as (1),

$$f(t)=\begin{cases}0, & t\le A,\\ \delta A^{\delta}t^{-(\delta+1)}, & t>A\end{cases} \tag{1}$$

where $0<\delta<2$. Selecting δ in this interval proves far superior to $0<\delta<1$ for the same required values that the inter-event time PDF can be further improved (Borst & Mitra, 1998). The improved PDF of the FRP decays as a power law, as shown in (2),

$$f(t)=\begin{cases}\delta A^{-1}e^{-\delta t/A}, & 0<t\le A,\\ \delta e^{-\delta}At^{-(\delta+1)}, & t>A,\end{cases} \tag{2}$$

which is continuous for all t, producing smoother spectral density function (Radev, 2005).

A method based on the Fractal-binomial-noise-driven Poisson processes (FBNDP) adds M independent and identically distributed (iid) alternating FRPs to generate a fractal binomial noise process that serves as the rate function for a Poisson process. The FBNDP requires five input parameters to generate self-similar sequences: A, δ, R, Δt, and M. The resulting Hurst parameter H assumes the value $(\alpha+1)/2$. The suggested algorithm is advancing with the intervals Δt. If S is a simulation clock, which advances in time, and $S^{(j)}$ is the elapsed time of the j-th FRP sequence, then $S^{(j)}=\tau_0^{(j)}+\tau_1^{(j)}+\ldots+\tau_k^{(j)}$ for some k and $j=1,2,\ldots,M$, where $\tau_k^{(j)}$ is the inter-arrival time. The sequence of self-similar pseudo-random numbers $X_0, X_1,\ldots$ is generated through the following steps of Algorithm 1.

Algorithm 1:

Step 1: *For each $j=1,2,\ldots,M$, generate $\tau_0^{(j)}$ from (3),*

$$\tau_0^{(j)}=\begin{cases}-\delta^{-1}A\log[U(\delta^{(V-1)})(\delta^{(V-U)})^{-1}, & V\ge 1,\\ AV^{1/(1-\delta)}, & V<1,\end{cases} \tag{3}$$

where

$$V\equiv\frac{1+(\delta-1)e^{\delta}}{\delta}U \tag{4}$$

and U is an iid uniformly distributed random variable over the unit interval [0,1]; *set $S^{(j)}=\tau_0^{(j)}$.*

Step 2: *Find j^* and $S^{(j)}$ such that $j^*=\mathrm{argmin}_j\{S^{(j)}\}$.*

Step 3: *Calculate (5).*

$$x=\begin{cases}0, & if\quad S^{(j^*)}<A,\\ 1, & if\quad S^{(j^*)}\ge A.\end{cases} \tag{5}$$

Step 4: *If $x=1$, then X_0 should be drawn from a Poisson probability distribution with $\lambda=1$. If $x=0$, then $X_0=0$.*

Step 5: *Set $i=1$, and $y=0$. Advance the simulation clock, i.e., $S\leftarrow S^{(j^*)}$.*

Step 6: *Construct a new inter-event time $\tau_i^{(j)}$ from (6),*

$$\tau_i^{(j)}=\begin{cases}-\delta^{-1}A\log[U], & U\ge e^{-\delta},\\ e^{-1}AU^{-1/\delta}, & U<e^{-\delta},\end{cases} \tag{6}$$

and set $S^{(j^*)}\leftarrow S^{(j^*)}+\tau_i^{(j)}$.

Step 7: *Find a new j^* such that $j^*=\mathrm{argmin}_j\{S^{(j)}\}$, and compute $S^{(j)}-S$.*

Step 8: *Repeat Step 6 through Step 8 to obtain x as in (5).*

Step 9: *Advance the simulation clock, i.e., $S\leftarrow S^{(j^*)}$, and set $y=y+x$.*

Step 10: *Repeat Step 6 through step 10 within time slot of length Δt.*

Step 11: *Compute $X_i=POISS(y)$, set $y=0$, and $i=i+1$.*

Step 12: *Repeat Step 6 through step 11 until $i=n$, where n is the number of sample points.*

An approximate self-similar sequence $\{X_0, X_1, X_2,\ldots\}$ was obtained. It took 1 min. 38 sec. on the Centrino-based HP PC (1.76 GHz) to generate a traffic sample sequence of 1,048,576 numbers (e.g., about $524*10^6$ inter-event times). Table 1 shows the mean values of estimated Hurst parameters H and 95% confidence intervals for the mean in parentheses, obtained using the wavelet-based H estimator for the FBNDP method with four input values: A=9.92, R=200, M=4 to 14, and H=0.6 to 0.9. The results confirm that the most appropriate aggregate level is M=10.

Table 1

M	Mean Values of Estimated H			
	0.6	0.7	0.8	0.9
4	0.6023 (0.575, 0.630)	0.6864 (0.659, 0.714)	0.7817 (0.754, 0.809)	0.8469 (0.819, 0.874)
6	0.6085 (0.581, 0.636)	0.6871 (0.660, 0.715)	0.7804 (0.753, 0.808)	0.8496 (0.822, 0.877)
8	0.6080 (0.581, 0.636)	0.6870 (0.660, 0.715)	0.7803 (0.753, 0.808)	0.8489 (0.821, 0.876)
10	0.6089 (0.581, 0.636)	0.6875 (0.660, 0.715)	0.7827 (0.755, 0.810)	0.8502 (0.823, 0.878)
12	0.6053 (0.578, 0.633)	0.6903 (0.663, 0.718)	0.7832 (0.756, 0.811)	0.8501 (0.823, 0.878)
14	0.6049 (0.577, 0.632)	0.6895 (0.662, 0.717)	0.7842 (0.757, 0.812)	0.8497 (0.822, 0.877)

A method based on a superposition of the fractal renewal processes (SFRP) uses a group of independent and identical fractal renewal processes (FRP). This method requires three parameters, i.e., α and A from the individual FRPs, and M, which is the number of FRPs superposed. The resulting Hurst parameter H, and the mean μ and the variance σ^2 of the marginal output distribution in related counting process during the unit time interval, can be defined according to (7):

$$\begin{aligned}&H=(\alpha+1)/2\\ &\mu=\mathrm{E}[X_n]=\lambda\\ &\sigma^2=Var[X_n]=(1+(1/t_0))^{\alpha}\lambda\end{aligned} \tag{7}$$

where

$$\lambda=M\delta[1+(\delta-1)^{-1}e^{-\delta}]^{-1}A^{-1} \tag{8}$$

is an aggregate arrival rate of events in the unit time interval. The sequence of self-similar pseudo-random numbers $X_0, X_1,\ldots$ is generated by the following steps of

Algorithm 2:

Step 1: *For each $j=1,2,\ldots,M$, and $i=0$, generate $\tau_0^{(j)}$ from equations (3) and (4); set $S^{(j)}=\tau_0^{(j)}$.*

Step 2: *Find j^* such that $j^*=\mathrm{argmin}_j\{S^{(j)}\}$ and set $X_0=S^{(j^*)}$.*

Step 3: *Advance the simulation clock, i.e. $S\leftarrow S^{(j^*)}$.*

Step 4: *Set i=i+1. Construct a new inter-event time* $\tau_i^{(j)}$ *from equation (6) and set* $S^{(j^*)} \leftarrow S^{(j^*)} + \tau_i^{(j)}$.

Step 5: *Find a new j* such that j*=* $\mathrm{argmin}_j\{S^{(j)}\}$, *and compute* $S^{(j)}$-S.

Step 6: *Advance the simulation clock, i.e.* $S \leftarrow S^{(j^*)}$.

Step 7: *Repeat Step 4 through Step 6 until i=n is reached, where n is the number of sample points.*

The obtained results from the simulation with M=10, A=3.8 and other conditions that similar to the FBNDP method, and with the mean values of estimated H and 95% confidence intervals for the mean in parentheses, are presented in Table 2.

Table 2

M	Mean Values of Estimated H			
	0.6	0.7	0.8	0.9
4	0.6076 (0.580, 0.635)	0.6986 (0.671, 0.726)	0.7929 (0.765, 0.820)	0.8603 (0.833, 0.888)
6	0.6076 (0.580, 0.635)	0.6986 (0.671, 0.726)	0.7929 (0.765, 0.820)	0.8655 (0.833, 0.888)
8	0.6192 (0.592, 0.647)	0. 7084 (0.681, 0.736)	0.7997 (0.772, 0.827)	0.8655 (0.838, 0.893)
10	0.6166 (0.589, 0.644)	0. 7091 (0.682, 0.737)	0.7986 (0.771, 0.826)	0.8686 (0.841, 0.896)
12	0.6134 (0.586, 0.641)	0. 7085 (0.681, 0.736)	0.7978 (0.770, 0.825)	0.8718 (0.844, 0.899)
14	0.6143 (0.587, 0.642)	0. 7063 (0.679, 0.734)	0.7983 (0.771, 0.826)	0.8708 (0.843, 0.898)

The mean values for 0.6 and 0.7 of estimated H with 95% confidence intervals for the means in parentheses for FBNDP algorithm are 0.6086 (0.581, 0.636) and 0.6875 (0.66, 0.715), respectively. Their ΔH(%) is +1.44 and -1.789, respectively. For H=0.6, 0.7 and 0.8 the overall relative error is less then 5%, but for H=0.9 it is greater then 5%. The same characteristics, obtained with SFRP algorithm are 0.6166 (0.589, 0.644) and 0.7091 (0.682, 0.737), and their ΔH(%) are +2.759 and -1.296, respectively.

The SFRP approach shows more accurate results and the relative error for H=0.6, 0.7, 0.8 and 0.9 is +2.76%, +1.29%, -0.17% and -3.49%, respectively. Similar to FBNDP algorithm, H values, estimated with SFRP algorithm, ranged from positively biased to negatively biased, as the H value increased.

A method based on the superposition of autoregressive processes (SAP) generates asymptotically self-similar sequences when aggregating several independent autoregressive processes. In the simplest case this can be presented as the sum of two autoregressive processes of the first order, as shown in (9):

$$\begin{aligned} z_{1i} &= A_{1i} z_{1,i-1} + y_{1i} \\ z_{2i} &= A_{2i} z_{2,i-1} + y_{2i}, \quad i = 1,2,... \end{aligned} \tag{9}$$

where A_{1i} and A_{2i} are randomly chosen from a beta-distribution $B(\alpha_1,\alpha_2)$ on [0,1] with the shape parameters α_1 and α_2, where α_1>0, α_2>0; y_{1i} and y_{2i} are a pair of independent and identically distributed sequences of random variables with mean of zero and variance σ^2=1.

Using the least-square fitting we can find that α_2=7.7929∗log(H)+4.9513. Then, the Hurst parameter H is linearly dependent on the shape parameter α_2 of a beta-distribution, while α_1 can be selected arbitrary; for example, α_1=1 in all cases that are investigated.

The PDF $f(x)$ of the beta-distribution is defined as (10),

$$f(x) = \begin{cases} \dfrac{x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}}{\beta(\alpha_1,\alpha_2)} & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $\beta(\alpha_1,\alpha_2)$ is given as (11).

$$\beta(\alpha_1,\alpha_2) = \int_0^1 x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1} dx = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} \tag{11}$$

This method, also based on a superposition of the autoregressive processes, consists of the following steps.

Algorithm 3:

Step 1: *Set i=i+1. Determine* z_{1i} *and* z_{2i} *using (9).*

Step 2: *Calculate the sum* $X_i = z_{1i} + z_{2i}$, i=1,2,…..

Step 3: *Repeat Step 1 and Step 2 until i=n, where n is the number of sample points.*

The asymptotically self-similar traffic sequence $\{X_1, X_2, \ldots\}$ of 1,048,576 numbers was generated in 3.5 seconds on the Centrino-based HP PC (1.76 GHz).

Unlike other sequential generators, the SAP generator does not require an aggregation level to be assumed as an input parameter. But it requires the shape parameter α_1 as input parameter. Mean values of estimated H obtained with algorithms for sequential generators are shown in Table 3 and Table 4, respectively, for Hurst parameter, equal to 0.6, 0.7 and 0.8, 0.9.

Table 3

Methods	Mean Values of Estimated H and ΔH			
	0.6		0.7	
	$\hat{H}$	ΔH	$\hat{H}$	ΔH
FBNDP	0.6086 (0.581, 0.636)	+1.440	0.6875 (0.660, 0.715)	-1.789
SFRP	0.6166 (0.589, 0.644)	+2.759	0.7091 (0.682, 0.737)	+1.296
SAP	0.5989 (0.595, 0.603)	-0.182	0.6852 (0.680, 0.690)	-2.112

Table 4

Methods	Mean Values of Estimated H and ΔH			
	0.8		0.9	
	$\hat{H}$	ΔH	$\hat{H}$	ΔH
FBNDP	0.7827 (0.755, 0.810)	-2.157	0.8502 (0.823, 0.878)	-5.538
SFRP	0.7986 (0.771, 0.826)	-0.174	0.8686 (0.841, 0.896)	-3.491
SAP	0.7845 (0.781, 0.788)	-1.937	0.8971 (0.894, 0.900)	-0.325

The cumulative relative error for the FBNDP method was less then 5%, except for H=0.9. For the SEFR method the cumulative relative error was +6.21%, +1.78%, -1.43% and -5.37%, for H=0.6, 0.7, 0.8 and 0.9, respectively. Similar to the FBNDP method, estimated H values ranged from positively biased to negatively biased, as the H value increased. For α_2=2.9 and 8.1 all values of the Hurst parameter from the traffic sample sequence of the SAP method were higher than the required values. The relative error was +24.14% and +10.45%, respectively; in that

case, the results were overestimated. For α_2=21.3 and 71.5, the relative error was -0.01% and -8.45%.

FIXED-LENGTH SEQUENCE GENERATORS OF SELF-SIMILAR TELETRAFFIC

The most frequently studied discrete-time models of self-similar traffic belong to fractional autoregressive integrated moving-average (F-ARIMA) and fractional Gaussian noise (FGN) processes, and possible fixed-length sequence generators are based on:

- Fractional autoregressive integrated moving-average processes (F-ARIMA);
- Fast Fourier transform (FFT);
- Fractional Gaussian noise and Daubechies wavelets;
- Random midpoint displacement (RMD);
- Successive random additions.

Let consider F-ARIMA(0,d,0) method for generating self-similar sequences, where d is the fractional differencing parameter, $0<d<1/2$, and let generate the process $\mathbf{X}=\{X_i: i=0,1,2,\ldots,n\}$ with a normal marginal distribution, the mean of zero and the variance σ_0^2, and the autocorrelation function (ACF), $\{\rho_k\}$ ($k=0,\pm1,\ldots$) that defined as (12),

$$\rho_k = \gamma_k/\gamma_0 = \frac{\Gamma(1-d)\Gamma(k+d)}{\Gamma(d)\Gamma(k+1-d)}, \qquad (12)$$

where

$$\gamma_k = \sigma_0^2 \frac{(-1)^k\Gamma(1-2d)}{\Gamma(k-d+1)\Gamma(1-k-d)}$$

The fixed-length sequence of self-similar pseudo-random numbers is generated through the following steps of Algorithm 4.

Algorithm 4:

Step 0: *Set N_0=0 and D_0=1. X_0, the firs pseudo random element in the output self-similar sequence, is generated from the normal distribution $N(0,\sigma_0^2)$, where σ_0^2 is the required variance of the X_i.*

Step i: ($i=1,\ldots,n$-1). *Compute $mean_i$ and var_i of X_i recursively, using the following equations:*

$$N_i = \rho_i - \sum_{j=1}^{i-1}\phi_{i-1,j}\rho_{i-j},$$

$$D_i = D_{i-1} - N_{i-1}^2/D_{i-1},$$

$$\phi_{ii} = N_i/D_i,$$

$$\phi_{ij} = \phi_{i-1,j} - \phi_{ii}\phi_{i-1,i-j}, \qquad j=1,\ldots,i-1,$$

where ϕ_{ij}, $i=0$, $j=0,\ldots,n-1$, is given by

$$\phi_{ii} = -\binom{i}{j}\frac{(j-d-1)!(i-d-j)!}{(-d-1)!(i-d)!},$$

$$mean_i = \sum_{j=1}^{i}\phi_{ij}X_{i-j},$$

$$\mathrm{var}_i = (1-\phi_{ii}^2)\mathrm{var}_{i-1},$$

Generate X_i from $N(mean_i, var_i)$. Increase i by 1. If $i=n$, then stop.

A self-similar traffic sequence $\{X_1,X_2,\ldots,X_n\}$ was obtained. It took 4 hours, 1 min. and 24 sec. on the Centrino-based HP PC (1.76 GHz) to generate the F-ARIMA traffic sample sequence with 1,048,576 numbers. The algorithm, based on the F-ARIMA method, is too rigorous to be used to generate long sample sequences.

In this work we propose a new generator of pseudo-random self-similar sequence based on fractional Gaussian noise (FGN) and Daubechies wavelet (DW), called the FGN-DW method. The use of Daubechies wavelet makes possible to produce more accurate self-similar sequences, where wavelets are closer to the true values. Wavelets can provide compact representation for a class of FGN processes, because the structure of wavelets naturally matches the self-similar structure of long-range dependent processes (Daubechies, 1992).

The wavelet analysis transforms a traffic sequence onto a time-scale grid, where the term scale is used instead of frequency, because the mapping is not directly related to frequency as in the Fourier transformation. The wavelet transformation delivers good resolution in both time and scale, as compared to the Fourier transformation, which provides only good frequency resolution. The developed algorithm consists of the following steps.

Algorithm 5:

Step 1: *Given: H, Start for $i=1$ and continue until $i=n$. calculate a sequence of values $\{f_1,f_2,\ldots,f_n\}$ using* (13),

$$f(\lambda,H) = c_f|\lambda|^{1-2H} + O(|\lambda|^{\min(3-2H,2)}), \qquad (13)$$

where $c_f = \sigma^2(2\pi)^{-1}\sin(\pi H)\Gamma(2H+1)$, $O(\cdot)$ represents the residual error and $f_i = \hat{f}(\pi i/n;H)$,

the value of the frequencies f_i corresponds to the spectral density of an FGN process for f_i ranging between $\pi/n \div \pi$.

Step 2: *Multiply $\{f_i\}$ by realizations of the independent exponential random variable with the mean of one to obtain $\{\hat{f}_i\}$, because the spectral density estimated for a given frequency is distributed asymptotically as the independent exponential random variable with the mean $f(\lambda,H)$.*

Step 3: *Generate a sequence $\{Y_1,Y_2,\ldots,Y_n\}$ of complex numbers such that $|Y_i|=\sqrt{\hat{f}_i}$ and the phase of Y_i is uniformly distributed between 0 and 2π. This random phase technique preserves the spectral density corresponding to $\{\hat{f}_i\}$. It also makes the marginal distribution of the final sequence normal and produces the requirements for FGN.*

Step 4: *Calculate two synthetic coefficients of orthonormal Daubechies wavelets. The output sequence $\{X_1,X_2,\ldots,X_n\}$ representing approximately self-similar FGN process is obtained by applying the inverse Daubechies wavelets transformation of the sequence $\{Y_1,Y_2,\ldots,Y_n\}$.*

It took only 2 seconds on Centrino-based HP PC (1.76 GHz), to generate a sequence of 1,048,576 numbers using this algorithm.

ANALYSIS OF NUMERICAL RESULTS

The Hurst parameter estimates for the wavelet-based H estimator are shown in Table 5 (H=0.6 and 0.7) and Table 6 (H=0.8 and 0.9) with different wavelet coefficients.

Table 5

Wave-let coeff.	Mean Values of Estimated H and ΔH			
	0.6		0.7	
	$\hat{H}$	ΔH	$\hat{H}$	ΔH
2	0.6019 (0.574, .630)	+0.323	0.6984 (0.671, 0.726)	-0.229
4	0.6026 (0.575, 0.630)	+0.433	0.7039 (0.676, 0.731)	+0.554
8	0.6026 (0.575, 0.630)	+0.430	0.7031 (0.676, 0.731)	+0.445
16	0.6013 (0.574, 0.629)	+0.214	0.6987 (0.671, 0.726)	-0.185

Table 6

Wave-let coeff.	Mean Values of Estimated H and ΔH			
	0.8		0.9	
	$\hat{H}$	ΔH	$\hat{H}$	ΔH
2	0.7943 (0.767, 0.822)	-0.709	0.8898 (0.862, 0.917)	-1.137
4	0.8055 (0.778, 0.833)	+0.684	0.9074 (0.880, 0.935)	+0.821
8	0.8039 (0.776, 0.831)	+0.486	0.9049 (0.877, 0.932)	+0.545
16	0.7962 (0.769, 0.824)	-0.474	0.8938 (0.866, 0.921)	-0.694

The obtained results are averaged over 30 sequences, and they show that for all input H values the F-ARIMA and the FGN-DW methods produce sequences with less biased H values than other methods. The means of estimated H values, obtained with the suggested wavelet-based algorithm, are shown in Table 7.

Table 7

wavelet coefficient	Mean Values of Estimated H and ΔH			
	0.6	0.7	0.8	0.9
2	1.949e-04	2.154e-04	2.393e-04	2.670e-04
4	2.508e-04	2.540e-04	2.586e-04	2.645e-04
8	2.440e-04	2.432e-04	2.438e-04	2.456e-04
16	2.116e-04	2.055e-04	2.007e-04	1.973e-04

Analysis of the mean times required to generate sequence of a given length demonstrates that the sequential generators are more attractive for the practical simulation studies of computer networks than the F-ARIMA-based generator, since they are much faster. However, these generators require more input parameters, and selecting appropriate values is a problem that remains. In addition, the problem of how to define the relationship between the Hurst parameter and two shape parameters of a beta-distribution in the case of SAP also remains.

CONCLUSIONS

Teletraffic exhibits self-similar properties over a wide range of time scales that are very different from the properties of traditional models. Self-similar models are appropriate for teletraffic as they provide capacity to estimate the network performance/quality, allocate the resources, and ensure the QoS. One of the problems that computer network researchers face during simulation is how to generate long synthetic sequential self-similar sequences. To solve the problem three aspects must be considered: (a) how accurately self-similar process can be generated; (b) how quickly the methods generate long self-similar sequences, and (c) how appropriately self-similar processes can be used in sequential simulation. The algorithms developed on the base of different methods for the sequentional generators and the fixed-length sequence generators perform the modeling and simulation of self-similar behavior of real teletraffic data. Numerical examples and simulation results are provided.

REFERENCES

Bobbio, A., Horváth, A., Scarpa, M., Telek, M. (2003). Acyclic discrete phase type distributions: Properties and a parameter estimation algorithm. - Performance Evaluation, **54**(1), 1-32.

Borst, S.C., Mitra, D. (1998). Virtual Partitioning for Robust Resource Sharing: Computational Techniques for Heterogeneous Traffic. - *IEEE Journal on Selected Areas in Communications*, **16**, (5), 668–678.

Boxma, O., Cohen, J. (2000). Self-Similar Network Traffic and Performance Evaluation. John Wiley & Sons, New York.

Daubechies, I. (1992) *Ten Lectures on Wavelets, vol.61 of CBMS-NFS Regional Conference Series in Applied Mathematics*. SIAM Press, Philadelphia, Pennsylvania.

Faraj. R. (2000). Modeling and analysis of self-similar traffic in ATM networks. PhD thesis.

Giambene,G.(2005). Queueing Theory and Telecommunications: Networks and Applications. Springer, NY. 585 p.

Hayes, J., Ganesh Babu, T. (2004), *Modeling and Analysis of Telecommunications Networks*, NY, John Wiley and Sons.

Jeong, H.-D. (2002). Modelling of Self-Similar Teletraffic for Simulation. PhD Thesis. University of Canterbury. USA.

Kushner. H., (2001). Heavy Traffic Analysis of Controlled Queueing and Communications Networks. Springer, NY.

Mehdi, J. (2003), *Stochastic Models in Queueing Theory*, Academic Press.

Park, K., Willinger, W. (2000). Self-Similar Network Traffic and Performance Evaluation. John Wiley & Sons. New York.

Radev, D. (2005), Fluid Flow Analysis for Long-Range Dependent Traffic, *Proceedings of the Computer Science'2005*, Chalkidiki, Greece, vol. 2, 19-24.

Rubinstein, R., Melamed, B., (1998), *Modern Simulation and Modeling*, New York, John Wiley & Sons.

Stathis, C., Maglaris, B. (2000). Modeling the Self-Similar Behaviour of Network Traffic. *Journal of Computer Networks*, **34** (1): 37 – 47.

Trivedi, K. (2001), *Probability and Statistics with Reliability, Queueing, and Computer Science Applications*, New York, John Willey & Sons.

BIOGRAPHY

DIMITAR RADEV, D.Sc. is Associate Professor at Department of Communication Technique and Technologies at University of Rousse, Bulgaria. His main research interests are Teletraffic Theory, Simulation and Modeling of Communication Networks, and Performance Analysis of Queuing Systems; dradev@abv.bg

IZABELLA LOKSHINA, PhD is Associate Professor of MIS and chair of Management, Marketing and Information Systems Department at SUNY Oneonta, USA. Her main research interests are Artificial Intelligence and Complex System Modeling and Simulation; lokshiiv@oneonta.edu

NUCLEAR ENERGY AND PETROCHEMICAL ENGINEERING

Using Web Services in Nuclear Power Plant Simulators

José A. Cortés, Manuel Díaz, Daniel Garrido, Isabel Parrado and Enrique Soler
University of Málaga
Department of Languages and Computing Science
CP.29071 Málaga, Spain
E-mail: {jacortes, mdr, dgarrido, isapc, esc}@lcc.uma.es

ABSTRACT

Web Services is a new technology that uses standard transport and encoding protocols to exchange information. This technology is hugely extended and widely renowned in the current market which makes it more trustworthy. Its main feature is the interoperability, given that it allows communication among systems from any platform across intranets, extranets or the Internet and it supports at the same time a safe and reliable service. This paper presents the use of Web Services in the context of Distributed Nuclear Power Plant Simulators and more specifically in a subsystem called the Instructor Console, where communication plays a really important role since many components have to interoperate with each other. The application developed allows the instructor full control and supervision of a simulator session and it can also be adapted to any simulator simply by adjusting an intermediate interface. Despite of the fact that Web Services are so extended nowadays, they have not been used in simulator environments so this is an innovative contribution to this market.

INTRODUCTION

Previously developed simulators for Nuclear Power Plants were based on old techniques and languages hindering scalability and reusability because of the dependence and coupling to the platform, language and operative system, among other factors. Nevertheless, by applying new technologies such as Web Services (Web Services Architecture, 2007), more advantages and future scalability can be achieved.

Web Services are basically a set of programming libraries, templates and projects that allow exposure of the component functionality via Internet or intranet connections, using a combination of http and SOAP. A Web Service is a collection of protocols and standards used to exchange data between applications. In (Mark and Brunton, 2005), Web Services have been used in the context of the High Level Architecture (HLA) (Dahmann, J.S.; Fujimoto, 1997), but centered on Grid computing (Buyya, 2007).

Since Internet Interoperability is reached by adopting open standards, different software applications developed in different programming languages and executed over any platform can use Web Services to exchange data through computer networks. The OASIS and W3C committees are responsible for the WS architecture and regulations.

The software described in this paper has been developed as a result of joint projects between the company Tecnatom S.A. and the department of Languages and Computer Science at the University of Málaga. The work has been related to the development of simulation software with special emphasis on communications based on Web Services and the building of components that could be reused in future projects.

The main intention of the simulators is the training of future operators for the operations and maintenance, allowing for the practice of different situations, from the most usual situations such as temperature monitoring, valve operation, etc. to most unusual ones like emergency situations. The Instructor Console can be used in any training centre wishing to acquire it. Previous versions of this application were developed at the request of a certain nuclear power plant to assist in the training of their future workers. This version tries to migrate to PC and at the same time it aims to be as general and reusable as possible so it can be adapted to any specific environment and easily scalable in the future. It can be integrated with either FSS (Full Scope Simulator) or IGS (Interactive Graphic Simulator), which are the two kinds of simulators used in the Tecnatom S.A projects.

This paper analyses the integration of Web Services in a simulator environment, which is presented in the following sections. This new technology is also described together with the impact on the system. A comparative study with previous technologies is carried out and the results are graphically shown and commented on. Finally, the conclusions extracted from this whole study and the benefits of this new integration are presented.

SYSTEM ARCHITECTURE

As stated before, the simulation projects of Tecnatom usually include two simulators which influence the hardware and software architectures. The first simulator is called *Interactive Graphic Simulator* (*IGS*), which through graphic applications (see Figure 1) allows the training of future operators.

The second simulator is called the *Full Scope Simulator* (*FSS*), which is an exact replica of the control room of the power plant (see Figure 2). These simulators share parts of

the hardware architecture and a large part of the software components, especially the components related to communications.

For further reading on the hardware architecture, see (Diaz and Garrido, 2004).

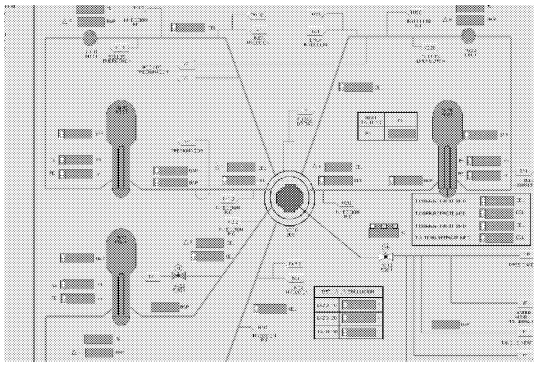


Figure 1. Example of *IGS* application

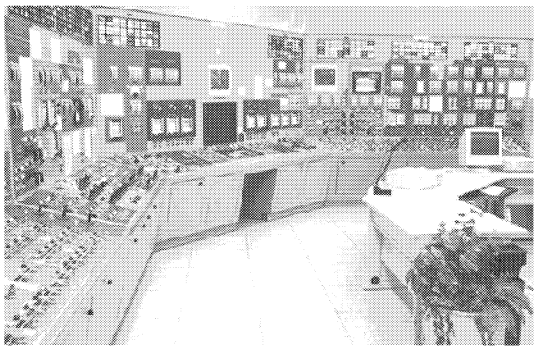


Figure 2. *FSS* simulator

The main hardware elements of *FSS* and *IGS* are:

***Simulation Computers*:** These computers are responsible for the simulation process executing the simulation models and providing data to the rest of the software and hardware components. They are the main elements of the simulators.

Physical panels: The *Physical panels* are exact replicas of those existing in the control room.

Subsystems: Depending on the nuclear power plant, there will be several subsystems that have to be simulated.

***Student workstations*:** The *IGS* simulators additionally include the hardware needed for the *student posts* of the simulator.

Instructor Console: The *Instructor Console* only exists in the context of the simulators. This element is used by the instructor of the simulation session and it allows the creation of scenarios which have to be solved by the students.

In this paper, we will focus on the instructor console: architecture, functionality and how, by applying Web Services, previous versions can be improved upon.

INSTRUCTOR CONSOLE

The Instructor Console (IC) is the subsystem that allows the instructor to control and supervise a simulation session. The IC will be capable of attending to the connected peripherals, communicating with the simulation programs and fulfilling functional requirements.

The main goal of this project is to develop a generic Instructor Console that can be easily integrated into any training centre simulator. To achieve this goal, an interface dependant on the specific Simulator to which it will be applied is used. This will also be in charge of communicating the orders from the console to the specific simulator in the required format.

The global architecture where this project is integrated can be seen in Figure 3. The system has a client-server architecture. In the server part is the simulator with the attached simulation variables. The IC is located in the client side together with an intermediate console which will offer the IC an interface with the functions offered by the simulator. There is also a local copy of the most used variables to speed up access to them.

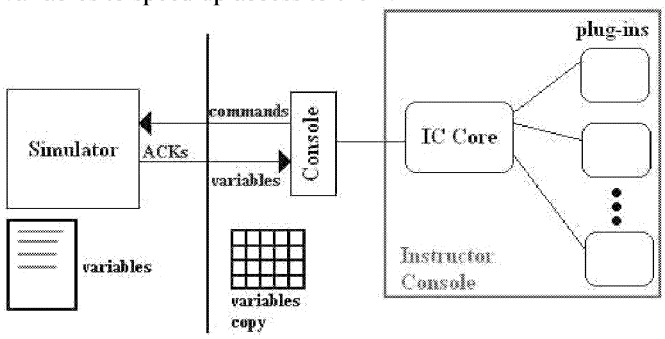


Figure 3. Simulator Architecture

The IC is constituted by a main core to which several plug-ins, representing the different functionalities to be offered to the user, will be added. This structure makes the software extensibility easier so it isolates the integration of the functionalities from the rest of the code. This software is developed using Microsoft Visual Studio .NET C#.

The IC can be used according to the following using modes:

- Fixed local or network
- Mobile local (PDA)
- Remote (Internet)
- Simultaneousness of use (two different consoles in the same workstation)
- On-line use (with the simulation workload) and Off-line use (without simulation workload by reading historical variables, actions and other data files, recorded in previously executed scenarios).

The main reason for using WS here is to achieve all these use modes, avoiding too much extra work and reusing as many components as possible. The WS technology is capable of supporting all the versatility requirements here as the development of just one version of it is possible, which can be consumed by different kinds of clients such as another web service, a windows form or a web page. The application can be locally executed if it is first installed in the machine or accessed through a web browser from any device connected to a network.

COMMUNICATION ISSUES

Simulators are composed of several soft real-time applications, which simulate the Power Plant Control Room, where Web Services are the basis for the communication. In the following sections, the main contributions using Web Services in the field of distributed simulators will be described.

Introduction to Web Services

According to the W3C committee (Web Services Architecture, 2007), WS can be defined as a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL).

Other systems interact with the Web service using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.

As mentioned, WS are based on HTTP over TCP in port 80 which is a significant issue since organizations protect their networks with firewalls by closing all the TCP ports in order to block a large part of Internet traffic except port 80. Therefore WS cannot be blocked.

From CORBA to Web Services

Nowadays, it is becoming more and more common to develop distributed applications running on different network-connected machines. Typical technologies that make this communication possible are DCOM (Microsoft Distributed Component Object Model) or CORBA (Common Object Request Broker Architecture) which are reliable and scalable to some extent in an intranet environment but not over the Internet. The main reasons are that they do not present interoperability and their communication architectures (explicit invocation methods) lead to strongly coupled systems. In (Díaz and Garrido, 2004b), an approach for simulator development based on RT-CORBA (Schmidt and Kuhns, 2000) is presented.

Nonetheless, Web Service applications expose their functionality on the Internet which makes these kinds of applications loosely coupled (the client does not need to know anything about the service implementation which (s)he is using, only the WSDL definition) and it is (also) outstandingly interoperable.
The application of WS to the simulator allows these advantages to be applied to the software development, thereby allowing all the interoperability and security mentioned.

Simulator – Web Service communication

Once it was decided to introduce WS to this whole system, there appeared a problem trying to obtain data information from the simulator since it is written in C while WS is in C#. After investigating a way to intercommunicate applications executed at different times across different systems and networks which may be temporarily offline, the best solution found was Microsoft Message Queue Server (MSMQ). As mentioned in its web site (Message Queue Server, 2007), MSMQ is a message infrastructure and a development tool for building distributed message applications for the Microsoft Windows Operative Systems.

Applications send messages to MSMQ, and MSMQ use message queues to ensure that messages reach their destinations. MSMQ guarantee message delivery, reliability, safety and priority use in the messages. It also allows communication among applications written both in C and C# using .NET, as both languages have message send and receive primitives.

Implementation Details

The implemented Web Service has all the required functionality, allowing any application (in this case, the instructor console), that needs information from the simulator, to obtain it simply by requesting it from the service. Among the implemented functions we can highlight the following:

- *send_run*: Start the simulator
- *send_freeze*: Pause the simulator
- *load_ic*: Load an initial condition with the saved simulator variables
- *save_ic*: Save the current simulator state
- *getVars*: Retrieve the simulator variable values

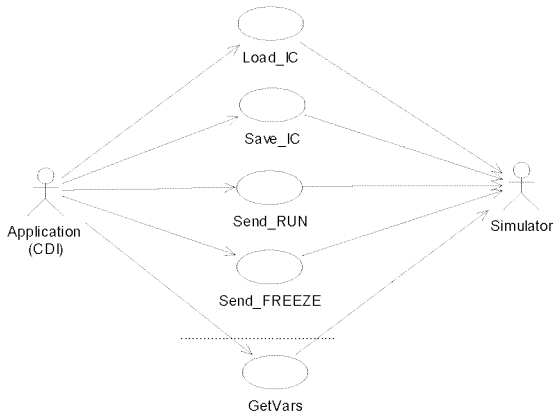


Figure 4. CDI use cases

Another important issue to point out in the web services implementation is the use of sessions. It allows the server to store the session details for each client accessing the service. This is necessary since some applications exist which require the storage of variables states.

As mentioned in (Powell, 2002), ASP.NET session compatibility for each web method is disabled by default, so it has to be expressly enabled for each web method using the session state. To make this possible, the property EnableSession is added to the attribute WebMethod of the function; subsequently, .NET runtime will be in charge of all the information administration management to enable session use.

To enable the use of sessions in the web services, it is necessary to process this way:

```
public bool send_RUN() { ............... }
public bool send_FREEZE() { ............... }

[WebMethod(EnableSession = true)]
public double[] getVars(){...............}
```

Using Microsoft Message Queuing

Due to communication efficiency and the fact that the queuing system allows asynchronous communication on the transmitter side (the simulator will not be blocked while waiting for the receptor, the web service, to read the message); it has been decided to use the Windows Queuing Messaging System (also known as MSMQ) instead of other possible technologies such as sockets.

Both C and C# languages contain *send* and *receive* primitives for MSMQ. To send information to a MSMQ system the ***MQSendMessage*** primitive is used, which sends a message to the queue previously specified. To receive data, the ***Receive()*** primitive is used**,** which writes the first available message in the queue referenced by ***MessageQueue***. This call is synchronous and blocks the current execution subprocess until a message is available.

System Operation

Each second (refresh frequency) the Simulator sends the simulation variables values to the message queue using the ***MQSendMessage*** primitive. Before sending those values, it checks whether there are any previous unread messages in the queue system (variable values previously sent) and, in that case, it deletes these values from the system. This behaviour is essential since each time variable values are updated, the previous ones are no longer needed so they can be substituted with the new ones. Concurrently, web services, by the instructor console petition, request from the queue system the current variables values, updated by the simulator using **Receive().** Later, the instructor console will use this information to show the state of the simulator. It will also send data to the simulator such as a new value for the backtrack interval variable, as an instance.

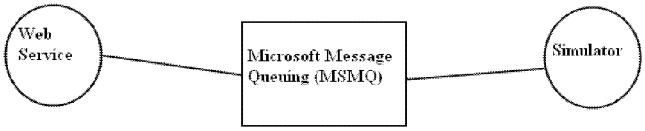


Figure 5. Messsage queues and Web Service

CASE STUDY: SOCKETS VS WEB SERVICES

To test the performance of the proposed architecture, we have developed a TCP client and a server, running on different machines and using sockets to communicate with each other. To follow the same model as WS do, the server waits for clients to request a connection. The client, once accepted, requests a service provided by the server, in this case, 4000 double numbers that represent simulation variables. The server reads these variables from memory and sends them to the client.

To contrast the results achieved here, we have also developed the WS version; a service that will provide the same data (an array of 4000 doubles) is published so that any WS user can discover and use it. In this case, the client accesses the WS and gathers the required information.

Each test has been repeated 100 times in order to obtain more reliable statistical measurements. The maximum response time of each scenario can be observed in Figure 6.

The next table shows a brief summary of the machine configuration where the tests have been performed:

CPU	PentiumR D CPU 3.00Ghz
Ram	1 GB
Network Card	Realtek RTL8169/8110 Family Gigabit Ethernet NIC
Operative System	Windows XP Service Pack 2

As Figure 6 shows, socket response times are much lower than WS ones. The reason for this considerable difference is that socket communication over the TCP transport protocol is really more efficient than over HTTP and since the transmitted data is binary, there is no need to either encode or decode it so the processing overhead is low here.

In this scenario, the amount of information to be transmitted is not high so the use of WS is affordable despite the lower response time. If the main goal were performance or a more complex object-oriented programming were needed, sockets would be the best suited solution but since here one of the important requirements is the interoperability, Web Services will be the best solution.

Taking into account that the simulator refreshing time is 1 second and that it will be a human interacting with the Instructor Console, WS response times are enough for this project.

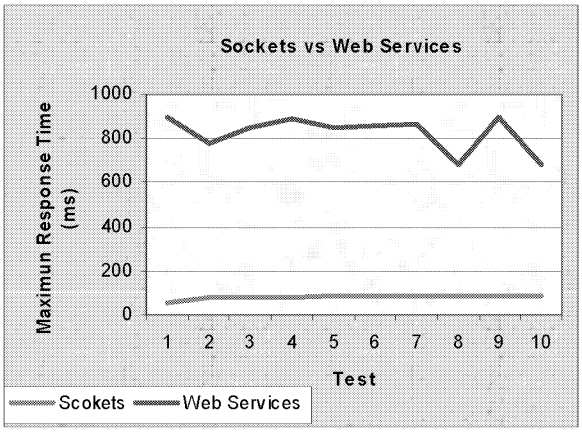


Figure 6 Sockets vs Web Services

CONCLUSIONS

Many companies, while trying to adapt to growth of technology, study how to take advantage of the new research and technologies in terms of the time cost benefits.

In this paper, the adoption of new technologies such as Web Services has been presented in a project for the creation/development of the Instructor Console of Nuclear Power Plant Simulators. The main contribution is the use of Web Services in Distributed Simulators. In this way, the development of the communication infrastructure is easier than with traditional techniques like RPC or sockets; it would substitute traditional communications using sockets with a newer technology, remarkably extended, reliable and efficient. The structure has also significant advantages, overall in terms of scalability and the application maintenance.

Besides the previously mentioned advantages, in this Project WS hugely simplifies the adaptability of the console both for the Internet and PDAs. Therefore, as stated, the web service would be hosted in a server and the console could be used from any platform independently of its specific details. This does not involve any extra work on the development of the Instructor Console since it can be developed in the same way as other applications in Visual Studio .NET. This workspace also makes the development of web Services easier since it includes specific projects for this; in essence, the WS is like a class library.

At local execution level, the Web Service is exactly the same as other applications running in the same machine; no differences will be appreciated here. When access is achieved remotely through Internet, the components required for the application (Instructor Console) would be downloaded to the specific device, which slows down the performance a little but it enables access from any platform without either previous installation of software or the development of several versions of the same software.

The loss of performance in comparison with other applications or a normal Web page is not very significant in this application where the data transfer is constant but not really high, so it can be considered a viable option. A detailed performance study can be observed in the previous section.

REFERENCES

Buyya, R.: Grid Computing Info Centre. Online documents at http://www.gridcomputing.com, 2007.

Dahmann, J.S.; Fujimoto, R. and Weatherly, R.M.. "The Department of Defense High Level Architecture", Winter Simulation Conf., pp. 142-149, 1997.

Díaz, M. and Garrido, D. "A Simulation Environment for Nuclear Power Plants". 8th IEEE Intern. Symp. on Distributed Simulation and Real Time Applications, pp. 98-105, 2004.

Díaz, M. and Garrido, D. "An Experience Report: RT-CORBA in Nuclear Power Plant Simulators". 2nd Industrial Simulation Conference, pp. 243-247, 2004.

Mark Pullen, J.; Brunton R.; Brutzman, D.P.; Drake, D.; Hieb, M.; Morse, K.L. and Tolk, A. "Using Web services to integrate heterogeneous simulations in a grid environment". Journal of Future Generation Comp. Syst., vol. 21, num. 1, pp. 97-106, 2005.

Message Queue Server, Microsoft Technet. Online document at http://msdn2.microsoft.com/en-us/library/ms711472.aspx, 2007.

Powell, M. "Using ASP.NET Session State in a Web Service". Microsoft Corporation. Online document at http://msdn2.microsoft.com/en-us/library/aa480509.aspx, 2002.

Schmidt D.C. and F. Kuhns. "An overview of the Real-time CORBA Specification", IEEE Computer special issue on Object-Oriented Real-time Distributed Computing, June 2000.

Shodjai, p. "Web Services and the Microsoft Platform". Microsoft Corporation. Online document at http://msdn2.microsoft.com/en-us/library/aa480728.aspx, 2006.

Web Services Architecture, W3C Working Group. Online documents at http://www.w3.org/TR/ws-arch/, 2007.

INSTRUMENTATION AND CONTROL SYSTEM RENEWAL USING SIMULATION AT PAKS NUCLEAR POWER PLANT

János Sebestyén Jánosy
MTA KFKI Atomic Energy Research Institute
Head of Simulator Development Department
P.O.Box 49, H-1525 Budapest Hungary
E-mail: jjs@sunserv.kfki.hu

KEYWORDS
Instrumentation and Control refurbishment, Nuclear Power Plant, Full-scope Simulators

ABSTRACT
Replacing the aging instrumentation and control system (I&C) in a nuclear power plant (NPP) is always a difficult task, even without the present energy-distribution disturbances in Europe and the more and more frequent plant-life-extension and power-up-rating projects.

The NPP plant life-cycle tends to reach 50-60 years, while that of the I&C system is only 8-10 years. Not only the systems become obsolete: spare parts for the old system are practically non-existent.

1. INTRODUCTION

The Paks Nuclear Power Plant produces more than 40% of electrical power of the Hungarian energy network. The country has no significant fossil energy resources and practically all the rest is made from imported oil and gas.

The first unit of the NPP was put into operation in 1982 followed by three more units. Some years ago the old, relay-based reactor protection system became less and less reliable, that means unnecessary reactor trips happened without any reasonable cause. It has been decided to replace the old system with a new computer-based and advanced one with automatic self-diagnostic and full 100% backup reserve.

This was the first occasion that the testing phase of the project was based mostly on the full-scope, replica simulator, and it was accomplished in due time, without significant increase of the outage time.

In the mean time, the power of the PWR units has been increased to 505 MWe from the original 440 MWe in several consecutive steps.

The original planned life-time of the units is 30 years, which for the first unit expires in 2012. It is rather common nowadays to extend the life-time of the NPPs all over the world; it started in the USA, with the oldest NPPs and continues all over in USA and Europe.

Based on the experience of successful NPP life-time extensions, similar project started in Hungary for the Paks NPP, too. No reason can be seen so far preventing the successful accomplishment of this project.

Based on the successful Reactor Protection System Refurbishment Project, we prepare a proposal to set up several simulators in Paks for the planned I&C replacements.

2. THE STRUCTURE OF THE SIMULATOR

The basic structure of the original full-scope simulator is shown on Fig. 1. There are several processors executing the model programs in parallel. All state variables are stored in the common memory, accessible to all processors. The time step of the integration for the models is 0.2 sec; that means, in every 0.2 sec we reach a consistent new state of the simulated processes. The term 'state variable' means: variable, the previous value of which is necessary to calculate the new actual state.

The 0.2 sec time step is too large for some models, especially for the models of the neutron kinetics and the two-phase-flow thermo-hydraulics. These model are executed sometimes with significantly smaller time steps but the data exchange with other models takes place once every 0.2 sec.

The simulated measurements are calculated from the state variables with interpolation and added noise every 0.1 sec. All the Control Room actuators and switches are scanned with the same 0.1 sec time step in order to make the response in the Control Room acceptable for the operators.

There are several other parts of the simulator not shown on Fig. 1 because they are not important for the simulator-based Instrumentation and Control System Refurbishment: Plant Computer system, Reactor Core Surveillance system, the Instructor's system, etc. The Control Room of the simulator is shown on Fig. 2.

The system used for the Reactor Protection Refurbishment Project is shown on Fig. 1. Blocks representing parts of the original simulator are shown in black colour. The blue boxes are software modules, belonging to the model of the new I&C units. Red colour indicates the hardware modules (the real instrumentation - and the I/O system connecting it to the simulator).

The whole structure is based on the requirement that the functionality of the new system should remain the same as that of the old system. It means that both systems should have the same reaction to the same events and their output should be identical; therefore they could be compared.

This requirement greatly enlightens the commissioning procedure: as far as the old functions were already accepted by the Authorities, only the similar behaviour must be demonstrated and verified.

3. THE TESTING PROCEDURE

During the verification and validation there were four stages of testing by the simulator:

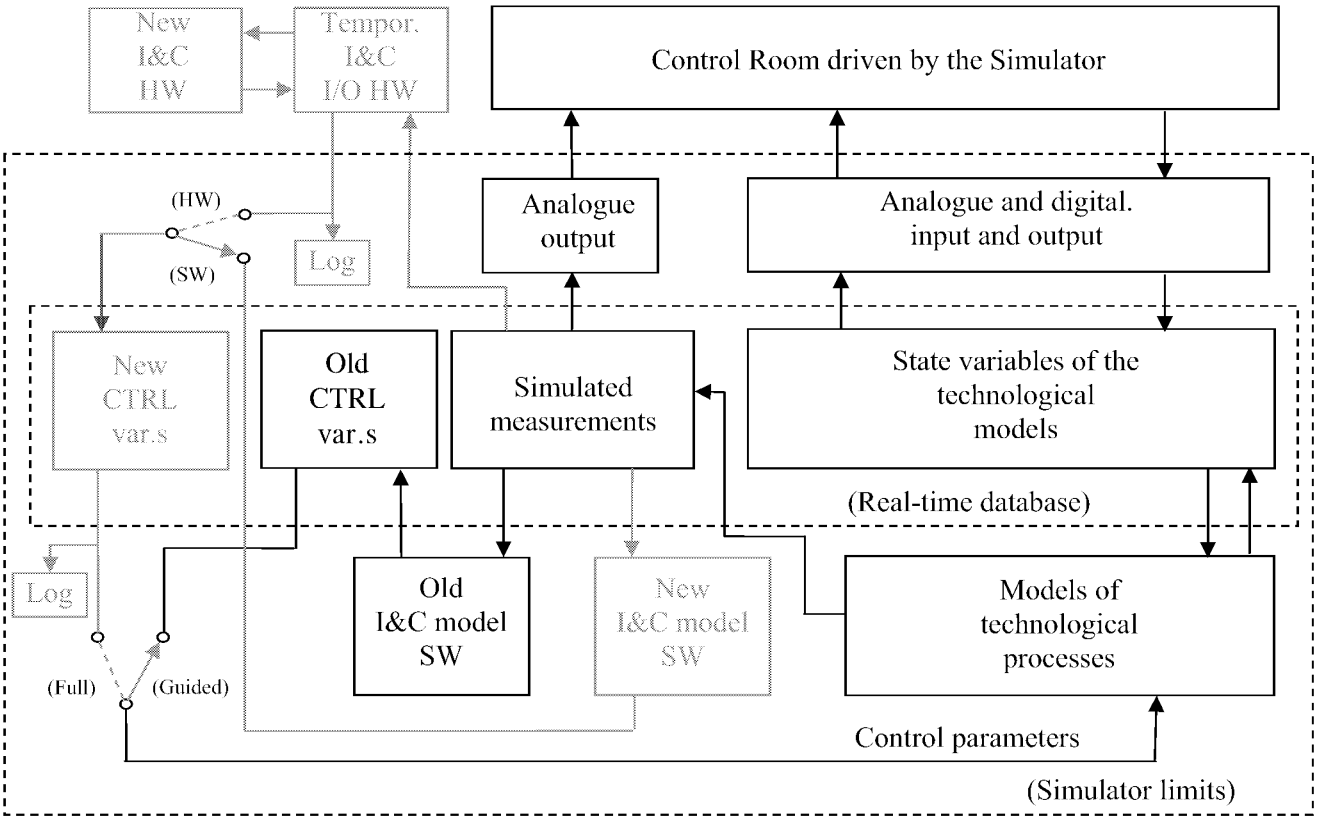


Fig. 1. Testing of the of the new I&C by means of the simulator

<u>First stage:</u> the SW Switches are in (SW) and (Guided) mode

The functions of the new I&C are modelled by the 'New I&C model SW' box. We have tools with convenient Graphic User Interface (GUI) for this purpose [1]. This module is driven by the simulated measurements, the same way as the modules of the old system are. The real control is still evaluated by the old system. It is not shown at the picture, but the outputs of the new system are not only carefully recorded ('Log)' but compared with those of the new system.

This 'guided' mode is very convenient, because every transient is performed in the old, very well known way. If the new - and probably imperfect - actions of the new I&C would be fed back into the technological models, the system could move to unknown, unrealistic and therefore meaningless states - due the strong feed-back.

The investigation of these unrealistic states would mean just loss of effort and time.

If the behaviour and the control actions of the new system are very similar to those of the old system, the second stage can be entered.

<u>Second stage:</u> the SW Switches are in (SW) and (Full) mode

The control of the technology is handed over to the model of the new I&C model software. This is the so called "software in the loop" configuration. The most important transients are to be repeated again, in order to make sure: the differences (caused by the slightly different behaviour of the new I&C algorithms) are not leading the system to unacceptable state trajectories and final steady states.

This is the first occasion that the Authorities are requested to issue a preliminary permit to continue. Completing successfully the first two stages one can decide whether the functions of the projected new system are acceptable and it can be purchased/manufactured.

Being manufactured, the ready-made new I&C units are put into operation at the simulator first. New, temporary I/O system has to be constructed to connect the I&C HW with the simulator (red modules on Fig. 1.) Sometimes this can be omitted, if the reserve channels found in the stationary I/O system are satisfactory for the job. It is very important that the time resolution of the temporary I/O system should not influence the performance, therefore it has to be fine enough compared with the time response of the control system. Then the third stage commences.

<u>Third stage:</u> the SW Switches are in (SW) and (Full) mode

Generally, it is similar to the second stage: still the model of the new I&C governs the technological models. However, now both the model SW and the real HW of the new I&C are driven by the same measurements, and the outputs are compared.

The differences in the HW and SW control signals can be originated partly by implementation errors, partly by time delays caused by the temporary I/O system - in each case it has to be carefully analyzed and clarified.

If the real HW seems to operate in the same way, the fourth and final testing-by-the-simulator stage can be entered.

Fig. 2. Control Room of the Full-scope Replica Simulator

Fourth stage: the SW Switches are in (HW) and (Full) mode

Now the full control is made by the external hardware of the new I&C located next to the simulator. This mode is called as the "hardware in the loop" configuration. Every transient, every mode/state change can be carefully studied and evaluated - while the real power generating unit of the plant is still running. If some parameters are to be tuned, it has to be done on the software module first: the necessary transients of the previous stages should be repeated, too. Even accidental situations can be triggered and analyzed - this is not possible on the real unit. Completing the fourth stage means finalizing the tests on the simulator.

Fifth stage: Installation and training

Having the new I&C modules thoroughly tested, tuned and verified, the plant outage time can be minimized. During the installation of the new I&C on the real unit, the simulator is ready for the training of the personnel for the usage of the new system (SW modes, of course). During the start-up usually it is enough to test the new I&C only in the most critical modes of its operation: in general it is enough to show, that it is performing in the same way as on the simulator.

Summarizing: the replacement of the whole - very sophisticated - instrumentation, control and protection system of the four reactors in Paks NPP could be done in four consecutive years, using the normal refuelling outage time. During this four years the simulator was able to provide training to the both (old and new) systems.

4. BASIC ASSUMPTIONS

A special study has been prepared to determine: what to do, how to make the I&C refurbishment if the license will be granted by the Authorities for another 20 or 30 years of operation. The experience of other NPPs - especially the Loviisa NPP in Finland - has been taken into account during this study [2]. The Loviisa NPP is very similar to the NPP in Paks - they have been constructed approximately in the same time and they are of the same type, using Soviet VVER-440 Pressurized Water Reactors (PWRs) fitted to Western standards.

According to the results of the feasibility study:

- We have to keep the functionality of the existing old I&C systems in full detail. The existing control and protection functions are approved by the Authorities, are well known for the staff and are proven as fully satisfactory during the last 20 years of operation. The last significant change has been made in the beginning of the nineties when our energy system became temporarily independent and we had to regulate the frequency in the network even with our NPP units, operating on full power whenever possible.

- together with the structure and functions, the existing cabling has to be preserved, too. These cables are laid in air-conditioned corridors and tunnels, separated by fire gates and equipped with automatic fire extinguishing features. The cables are covered by fire-protecting coating and are in good condition.

The greatest challenge to them is the laying of new cables, or the changing of something: in this case the fire gates are to be demolished temporarily and that goes with a lot of carving-out; this is extremely dangerous for the existing cables, therefore should be avoided or at least minimized

- The refurbishment has to be divided into several stages, as independent from each other as possible. It is a lot easier to accomplish the replacement of only smaller parts of the existing I&C systems at a time.

- The testing, verification and validation has to be done by means of the simulator and the replacement of the old I&C with the tested and approved new one has to be done during outages for refuelling, once in a year.

- The refurbishment in four or five steps will last therefore four or five years. The sequence starts on one of the units and it has to be repeated three times, for the three other units, with a time delay of one year for each.

- The whole process - from starting with the first, up to the finishing on the last - may take nine years, but a given system will be replaced only on one unit at a time, and the experience gained with the first can be used on the others.

- The training simulator should be able to train the staff for all phases, for any constellation of the I&C refurbishment.

- The usage of the full-scope replica simulator should be the same as it has been used for the successfully accomplished Reactor Protection System Refurbishment project.

5. SIMULATOR REQIREMENTS

The simulator is used already in two shifts even without any reconstruction project. There are over 20 staffs for the four units to be re-trained - twice a year for a week - that takes over 40 weeks each year just for the re-training. That together with some repair and development takes one shift of the simulator completely.

Teaching newcomers and providing re-training due upgrading the qualification - e.g. turbine operator becomes shift supervisor - takes the second shift.

The Rector Protection System Refurbishment Project could be done only with significant cutting-back of the training on the simulator.

6. STRUCTURE OF ADDITIONAL SIMULATORS

The approach described above is not suitable for a new nine-year project. Additional simulators are to be set up for the I&C refurbishment projects. In order to unload the 'main' training simulator, several others are to be constructed to facilitate the I&C refurbishment projects at their different stages.

These simulators differ basically from the existing one by the lack of the replica Control Room. Construction of these simulators is far less expensive than it seems. First, the license provided by the simulator manufacturer allows the usage of unlimited copies of the simulator software inside the fence of the NPP. That means, duplicating the training simulator only a rather small software development is necessary, initiated by the lack of the replica Control Room. Practically two types of the simulators can be developed: the ***Engineering Simulator*** and the ***Development Simulator***. The structure of both are shown on Fig. 3.

The black SW blocks are coming from the full-scope replica simulator; the blue ones are developed to model the new I&C system. The red HW blocks ensure the "hardware in the loop" testing, the same way as it was done during the Reactor Protection Refurbishment Project.

As it can be seen, the basic difference between these and the full-scope replica simulator is that the very costly Control Room and the real-time I/O system is missing. Both simulators are identical with the full-scope one concerning the model software and the real-time data base. This means, that even snapshots and steady states are easily transferable from one simulator to the others. The testing of different I&C systems does not mean a very great flux of information coming and going to and from the Control Room.

Usually just a small disturbance or malfunction is generated and after that the performance of the system is observed.

Some "control quality" type variables (integrals of error signals, etc.) are calculated, logged and plotted together with some other characteristic state variables.

There is a small amount of variables changing and few actions are to be taken manually: the Control Room can be represented by the standard Plant Computer screens, the usual Reactor Core Surveillance screens, the Instructor's screens (Fig. 4.) providing the possibility to control the simulator and a special "actuator screen" replacing the usual control knobs and switches.

7. THE ENGINEERING SIMULATOR

The so called ***Engineering Simulator*** is just limited to the configuration described above. The basic tasks are to perform the model development, the Stage 1 and Stage 2 testing procedures (see Point 3.). These simulators are the cheapest and can be distributed in several copies.

8. THE DEVELOPMENT SIMULATOR

The ***Development Simulator*** differs only by the software switch and the I/O software reserved for the "hardware in the loop" testing (Stage 3 and Stage 4 in Point 3.) After the hardware of the new I&C system is manufactured, it has to be delivered to the simulator together with the corresponding I/O system (at Paks we are using the VME industrial standard which is a modular, very flexible system). Reserve cards and units of the main replica simulators can be used.

We discussed the possibility to deliver similar simulators to the manufacturers of the I&C systems to their site in order to facilitate the in-house testing. It would be advantageous but in this case the usage of the software is not allowed by the original SW license: this solution could be expensive and therefore it has been rejected.

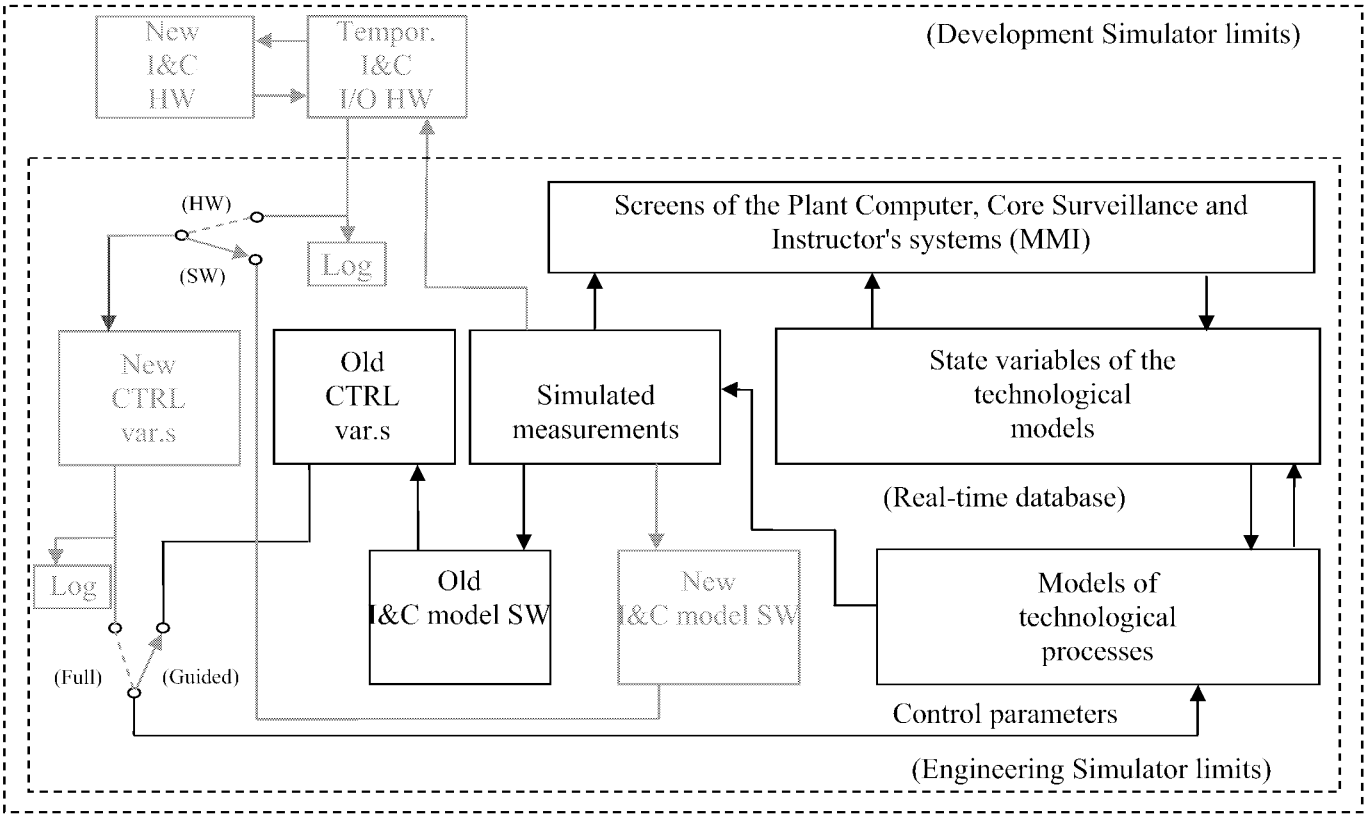


Fig. 3. Engineering and Development Simulators

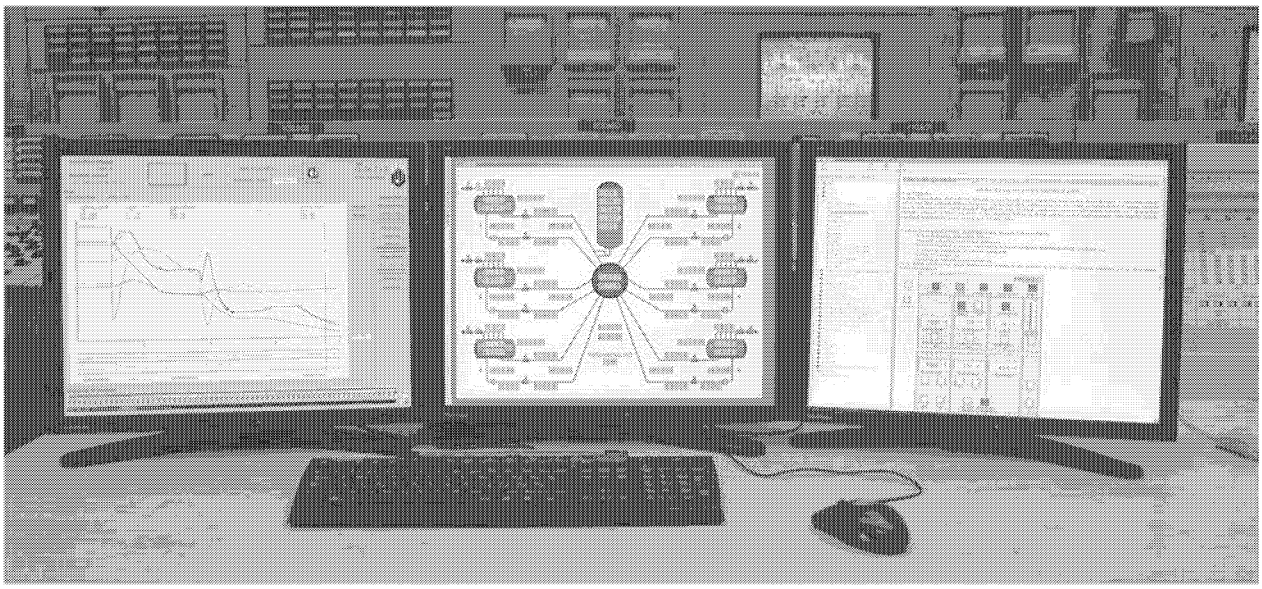


Fig 4. Instructor's System of the Full-scope Replica Simulator

9. CONCLUSIONS

The upcoming I&C system refurbishment project is now in the pre-planning stage. The experience gained during the Reactor Protection Refurbishment Project shows that a lot of time and money can be saved with extensive use of on-line simulation provided by different additional simulators. Putting the models together allows the training of the operators to their future tasks already during the implementation phase. Moreover, the whole process converts the simulation and the simulator facilities into "every-day-tools", accepted and hopefully admired by large number of people in the plant, increasing the efficiency and reliability of future developments and plant operating practices.

10. REFERENCES

[1] J.S.Janosy, G. Hazi, L. Seregi, B.K. Szabo: *"GRASS - the Graphic Simulation System"*. The 2nd CSNI Secialist Meeting on Simulators and Plant Analyzers. 29.09-02.10.1997, Hanisaari Cultural Centre, Espoo, Finland.

[2] Kari Porkholm, Aino Ahonen, Olli Tiihonen: *"Utilization of the Simulators in I&C renewal project of Loviisa NPP"*. IAEA Technical Meeting to Develop a Technical Report on Upgrade and Modernization of NPP Training Simulators, September 19-22, 2005, KSG, Essen, Germany

REAL TIME SYSTEM FOR A HEAVY WATER DETRITIATION PLANT

Florin Hartescu
National Institute for R&D in Informatics
8-10 Averescu Avenue, 71316 Bucharest 1, Romania
E-mail : flory@ici.ro

KEYWORD
Real time systems, process control, optimization, control of distributed parameter systems, simulation, intelligent cell

ABSTRACT

The paper is focused on an integrated system designed for the heavy water detritiation process control and optimization. The system is based on a RT-ARCH (Real Time ARCHitecture)[1], an architecture of software tools used in process control. The system is composed by classical algorithms running on a network of PLC-s and controlling algorithms implemented in a process computer. Some of them are typical numerical algorithms, and the others are adaptive control algorithms.

INTRODUCTION

Today all the systems that manage the security are characterized by the complexity of their major functions like identification, authentication, access control and data protection. The implementation of these functions is usually and objectively based on a trusted model that uses a trusted architecture, which is the platform of the security architectures.
Automation systems use more and more communication systems to be able to carry out right their task. Therefore, today it's already available a wide catalogue of products that allows the embedded systems used for the automation to interface to one or more field bus networks and to one or more wide range communication networks (telephone network, Internet, etc.)
But currently the systems both for the industrial automation and for other kinds of control need several devices to be able to work within the network.
In particular, the main device containing the CPU foresees some interfaces for being able to be programmed, some interfaces for being able to talk with the I/O devices, some interfaces for being able to talk with the communication lines.
Heavy water detritiation is one of the most important industries in the world nowadays.
The Open Source movement has developed new concepts of making process control algorithms based on transparent and co-operative ownership of software.

DESCRIPTION OF TECHNOLOGICAL PROCESSES

Tritium (symbol T or ^3H) is a radioactive isotope of hydrogen. The nucleus of tritium (sometimes called triton) contains 1 proton and two neutrons, whereas the nucleus of protium (the most abundant hydrogen isotope) contains no neutrons. Its atomic weight is 3.0160492. It is a gas (T_2 or $^3H_2$) at standard temperature and pressure. Tritium combines with oxygen to form a liquid called tritiated water T_2O or partially tritiated THO.
Tritium is radioactive with a half-life of 12.32 years. It decays into helium-3 by the reaction

$$^3_1\mathrm{T} \longrightarrow {}^3_2\mathrm{He}^+ + \mathrm{e}^- + \overline{\nu}_\mathrm{e}$$

releasing 18.6 keV of energy. The electron has an average kinetic energy of 5.7 keV, while the remaining energy is carried off undetectably by the electron antineutrino. The low-energy beta radiation from tritium cannot penetrate human skin, so tritium is only dangerous if inhaled or ingested. Its low energy also creates difficulty detecting tritium labelled compounds except by using liquid scintillation counting.
The paper presents a system with large capacity about calculation and communication in real time but at low cost. The low cost, compared to the devices currently available on the market, is obtained thanks to the fact that the "Intelligent Cell" has an architecture very simple constituted by a microprocessor very powerful with only three single connectors: one for the power supply and two for the communications. The

"Intelligent Cell" is produced by the factory only with communication abilities and it can immediately be installed where it will have to reside during its operation activity.

HEAVY WATER DETRITIATION

The core of the nuclear steam supply system of a nuclear power plant is a large cylindrical vessel called the calandria. This vessel is filled with cool, low-pressure heavy water. The vessel houses 380 horizontal tubes, loaded with natural uranium fuel bundles. With the uranium fuel surrounded by heavy water, chain reaction fission takes place, releasing energy in the form of heat. The heat is transferred to a second heavy water system pumped at high pressure through the tubes to steam generators. There its heat is transferred to ordinary water, which boils to become the steam that drives the turbine generator. The importance of detritiation of heavy water from nuclear reactors is well known, as is the implications of detritiation in fusion processes and installations. The nature of the fluids that are processed in detritiation requires the operation of the plant in maximum security conditions in order to protect the working staff and environment. The paper presents how the data acquisition and control system could be made for an experimental heavy water detritiation plant. The plant must be designed to be operated without any working staff in the technological space. The purpose of the security and control system is to ensure for the population irradiation risk, below the prescription limits. The radiological risk is the tritium leakage that can be gas, vapor or liquid.

THE INFORMATIONAL STRUCTURE OF THE SYSTEM

The overall goal is to foster the usage of FLOSS backend platforms and services and to generate new business opportunities for the Open Source developer community. The open source-based set of tools is supposed to have a high socio-economic effect for both, the providers and users of FLOSS.
The informational system is designed first for data and technical parameter acquisition and then for operative supplying of information required for the process control, for exploitation and maintenance of the equipment. The entry information is received as digital or analog signals from transducers through the process interface. As a result of data processing, the computing system elaborates the output information, analog or numerical, that can be used for controlling, adjusting or for displaying on video terminals.
The system has a database for data and programs on the mass memory (one or more magnetic hard drives).

TECHNIQUES DESCRIPTION

Typically, classical control techniques use analog PID (PI, filtered PID) controllers, or on/off controllers for slower processes. Controllers are used in structures like:

- Simple adjusting with imposed reference
- Cascade adjusting
- Combined adjusting by perturbation and reference

These types of controllers have the advantage that they can be easily implemented, but very often their parameters are fixed. That's why classical control structures can be used only for those subsystems of technological installation that have very well defined models, or have modest control requirements.
From all the diverse modern control techniques we have chosen to implement some algorithms by computer programs. This implementation has a great flexibility: it is very easy to modify the dates of the program that contains the adjusting parameters.

HARDWARE AND SOFTWARE ARCHITECTURE

The hardware platform used as a support of the control programs is a personal computer with a Pentium IV processor, 1024 Mb RAM, 256 bits video graphic accelerator, SCSI 160Gb HDD, data acquisition card, one multiplex for 64 serial ports and a network of PLC-s. On this machine we have used Unix operating system. This system offers multitasking facilities for parallel managing aspects regarding data acquisition, data transmission, adjusting and data displaying. Using multitasking we could manage more control loops on the same computer even though some of the processes are controlled only by PLC-s.
The development platform used was Visual C++ because it can compile programs for Unix operating system and it has the following facilities:

- Permits creation of separate threads with different adjusting algorithms for each of the processes.
- Offers communication methods inter-threads for transferring data between data acquisition processes and controlling processes. Communications are made using message boxes and critical sections.
- Permits realization of communication modules in a TCP/IP network. That makes possible implementation of the hierarchical architecture.

- Offers the possibility to create a user-friendly interface for the product, in Unix environment.

The only disadvantage of this platform is that it hasn't any dedicated functions for complicated mathematical calculus needed in adjusting algorithms, so that we had to write them ourselves.

The process computer is placed in the control room of technological installation because it must be protected from the environment of the technological process.

PROCESS OPTIMIZATION

The mathematical problem posed by the general optimization formulation is often remote from that of a practical design problem.

The optimization problem is "given a dynamic system (S) that evolves on a finite time interval $[T_0...T_f]$ and the performance index J(u) it is required a command that minimizes J". In this standard formulation there are no restrictions, but it follows a movement in the states space for minimum energy consumption. Using the computer we have chosen dynamic optimization, that relies on the idea of finding a procedure for generating a "relaxing" array defined by the following condition:

$X1, X2, X3, \ldots Xn$, so that $f(X1) \geq f(X2) \geq f(X3) \geq \ldots$ with the property that if $Xk \rightarrow X$, then $f(Xk) \rightarrow f(X)$, where X is the extreme value of the objective function in a specific domain.

The relaxing arrays generating procedure has two components:

- Choosing the descending direction of the function with the modified Newton method. This is a gradient method that means successive approximations of the function.
- Determining the step on the descending direction.

Newton-type method calculate the Hessian matrix H directly and proceed in a direction of descent using a line search method to locate the minimum after a number of iterations. Calculation H numerically involves a large amount of computation. Quasi-Newton methods avoid this, by using the observed behavior of f(X) to build up curvature information, in order to make an approximation to H using an appropriate updating technique.

CONTROL SYSTEM FUNCTIONS FOR HEAVY WATER DETRITIATION PLANT

The automation system works in a collecting-processing way, interfering with and conducting the controlling processes. The system does the following functions:

- Supervision of technological equipment and devices
- Displaying and printing any damage that may occur, showing the moment of appearance, the kind and the place of the damage
- Displaying and printing the alarms at the moment of the occurrence
- Showing the entries and outputs from the programmable logic controllers for debugging any anomaly or damage
- Controlling the measuring equipment and supervising the process
- On demand, displaying the parameters of the technological process

Some of the subsystems of the technological installation are variable in time due to action of stochastic perturbations. That determines large variations of the parameters of the attached mathematical model, but fortunately the structure of the model remains the same.

The control takes place at the level of the central computer, and also at the level of the PLC-s. We have used a hybrid adjusting method, i.e. some adjusting loops are implemented with classical algorithms (PI, PID), and the others are adaptive.

The system is composed by:

- Classical algorithms running on the PLC-s that also have implemented data acquisition modules, and analog and digital command modules. Programs are written in the PLC's language.
- Controlling algorithms implemented with the process computer. Some of them are typical numerical algorithms, and the others are adaptive control algorithms.

CONCLUSIONS

Implementation of this system in a Romanian heavy water detritiation plant has the following advantages: obtaining a high efficiency and combustible saving, limited efforts for developing a new application in a short period of time, and high performance of the system in solving the demands of applications.

The system uses secure socket connections (SSL) to transmit all sensitive information during the control process. The application has been tested in an integrated system, with several servers running Unix, connected in a network. The system was configured easily, and it has worked very fast because the communication protocol transmits just the information needed. The aim of the system is to realize a network of Intelligent Cells with large capacity about calculation and communication in real time but at low cost.

REFERENCES

Hartescu F.: *RT-ARCH – A New Approach in Real Time Application Design*, IASTED Mini and Microcomputers and Their Applications, Lugano, Switzerland, June 1990

Astrom K., J. Wittenmark: *Adaptive Control*, Addison Wesley Company, 1989

Dumitrache I.: *Intelligent Techniques for Control Applications*, The 9th CSCS Conference, May 1993

Cichocki A., R. Unbehauen, *Neural Netwarks for Optimization and Signal Processing*, John Wiley & Sons, Ltd. 1993

Fleming P.J. *Application of Multiobjective Optimization to Compensator Design for SISO Control Systems*, Electronics Letters, Vol.22

Fleming P.J. *Computer-Aided Control System Design of Regulators using a Multiobjective Optimization Approach*, Proc. IFAC Control Applications of Nonlinear Optimization

Gill P.E., Murray W. *Numerical Linear Algebra and Optimization,* Vol.1, Addison Wesley, 1991

Selic B., P.T. Ward: *The Challenges of Real Time Software Design*, Embedded Systems Programming, Miller Freemans Inc., Oct. 1996

Selic B., G. Gulliksons, P.T. Ward: *Real Time Object Oriented Modelling*, John Wiley & Sons, 1994

W. Wang – Designing Secure Mechanisms for Online Processes

Hauslein A., Page B., Knowledge-based Approaches to Modelling and Simulation Support. Systems Analysis, Modelling Simulation 8 (1991), 4/5, pp. 257-272.

Nussbaumer H., Computer Communication Systems, John Wiley & Sons, 1990.

Selic B., P.T. Ward: The Challenges of Real Time Software Design, Embedded Systems Programming, Miller Freemans Inc., Oct. 1996

Tanenbaum S. A., Computer Networks, Prentice Hall.

Ovidiu Balteanu, Iulia Stefan, Carmen Retevoi - *Safety System in a Heavy Water Detritiation Plant*

Carmen Maria Retevoi - *Virtual Instrumentation for monitoring the purifying installation from Cryogenic Pilot Plant with Remote Data Acquisition*

Carmen Maria Retevoi, Florin Hartescu, … - *Datalogging and Supervisory Control Module for tritium removal plant, with data base and processes analysis* - REV2006 Remote Engineering and Virtual Instrumentation, University of Maribor, Slovenia, 29 - 30 June 2006

USING A GENERIC METHODOLOGY TO DEVELOP A SIMULATION MODEL OF A PETROCHEMICAL PLANT

Martin Albertyn and Paul S Kruger
Department of Industrial and Systems Engineering
University of Pretoria
0002 Pretoria, South Africa
e-mail: malbert@absamail.co.za and paul.kruger@up.ac.za

KEYWORDS

Generic methodology, Simulation model, Petrochemical plant, "Bottleneck" identification technique, Simul8.

ABSTRACT

The key objective is to present a brief overview of a generic simulation modelling methodology and to demonstrate how the generic methodology can be used to develop a simulation model of a petrochemical plant. The generic methodology consists of a "toolbox" of eight methods and techniques and five high-level simulation model building blocks. Some detail about the two "bottleneck" identification techniques of the generic methodology is provided. A system description of a typical petrochemical plant is established and the key characteristics of the plant are identified. The most important aspects of a Simul8 simulation model of the petrochemical plant that was developed with the generic methodology are described. Through scenario analyses the "bottlenecks" in the petrochemical plant are identified and the impact of additional production capacity at one of the "bottlenecks" is determined. The results are discussed and conclusions are presented.

INTRODUCTION

In this paper a petrochemical application of a generic simulation modelling methodology is presented. A synthetic fuel plant is used to demonstrate this application of the generic methodology. The plant produces various chemical products, including petrol and diesel, from coal. A simplified schematic representation of the plant is shown in Figure 1. The plant is situated near the town of Secunda in South Africa. Changes to the original data of the plant are incorporated to protect client confidentiality.

The need for a simulation model of the Synthetic Fuel plant originally arose because the plant management identified the necessity for a decision support tool on a strategic level (Owen 1994). In a plant of this size and complexity it is extremely difficult to predict what the effect of a proposed change is going to be on the operation of the plant. A simulation model can be used to identify problem areas ("bottlenecks") in the plant and to study the effect of proposed scenarios on the plant. Proposed scenarios may include additional production capacity at "bottlenecks", changes in the maintenance strategy, *etc.*

The key objective of this paper is to present a brief overview of the generic simulation modelling methodology and to demonstrate how the generic methodology can be used to develop a simulation model of a petrochemical plant. A simulation model of the Synthetic Fuel plant that was developed with the generic methodology is used to evaluate two alternative scenarios. The first scenario is used to identify the "bottlenecks" in the plant and the second is used to determine how additional production capacity at one of the identified "bottlenecks" impacts on the throughput of the plant.

GENERIC METHODOLOGY

The generic simulation modelling methodology was developed over a 5-year time period from 2000 to 2004 and is detailed in a conference paper by Albertyn and Kruger (16th European Simulation Multiconference 2002) and a doctoral thesis by Albertyn (2004). The generic methodology accommodates the unique simulation modelling problems that are posed by continuous process systems such as petrochemical plants, ammonia manufacturing plants, paint manufacturing plants, *etc.* These systems are usually referred to as stochastic continuous systems and their key characteristics are continuous processes, two types of discrete events (chronological services and stochastic failures) and complex interrelationships. The generic methodology consists of a "toolbox" of eight methods and techniques and five high-level simulation model building blocks (Albertyn 2004).

The "toolbox" of the generic simulation modelling methodology contains the following eight methods and techniques:

a) The variables technique.
b) The iteration time interval (ITI) evaluation method.
c) The event-driven (ED) evaluation method.
d) The Entity-represent-module (ERM) method.
e) The Fraction-comparison (FC) method.
f) The iterative-loop technique.
g) The time "bottleneck" identification technique.
h) The production lost "bottleneck" identification technique.

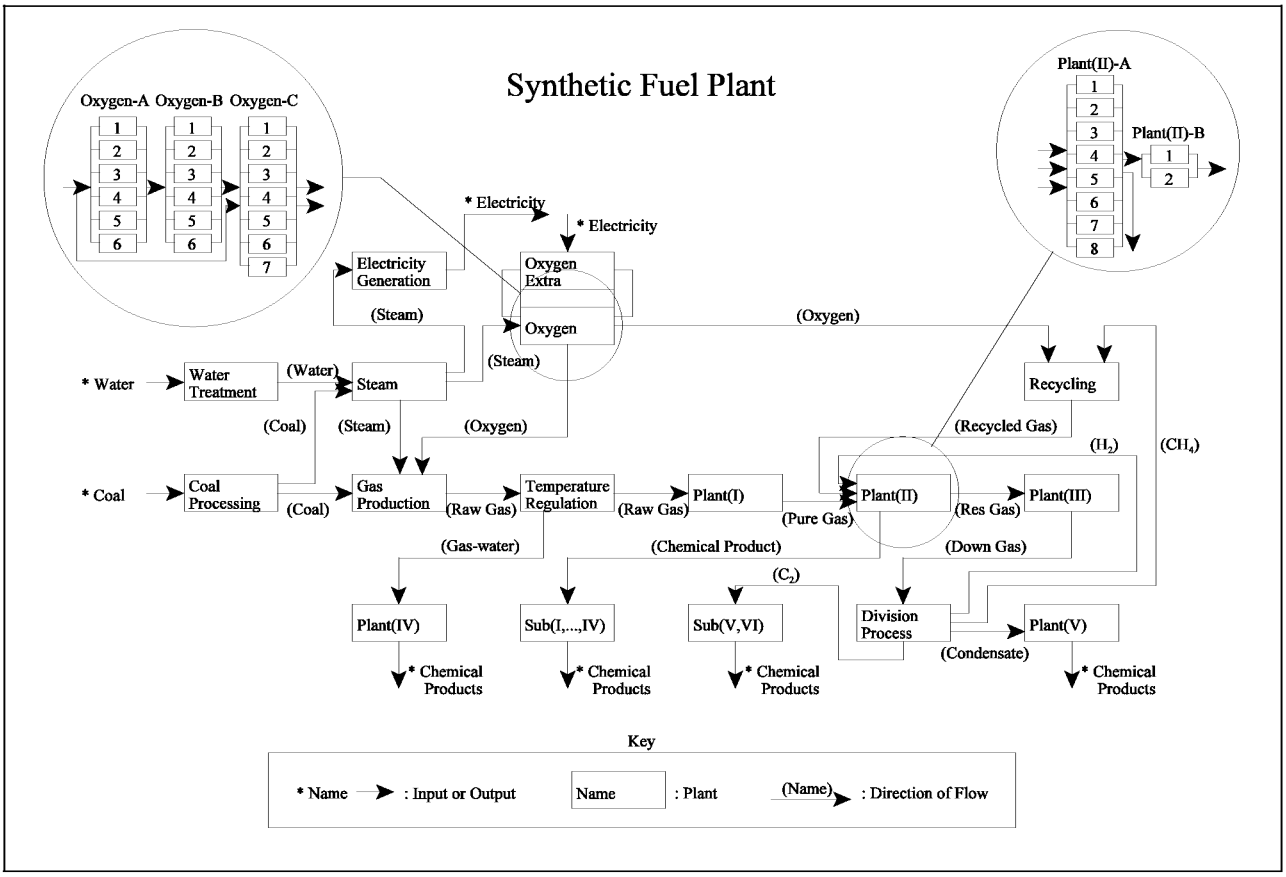


Figure 1: Synthetic Fuel Plant

The generic simulation modelling methodology comprises two parts, namely: an iterative-loop technique part that determines the governing parameters of the system that is under scrutiny before the start of a simulation run and a simulation model part that uses the other seven methods and techniques (embedded in five high-level building blocks) continuously during the simulation run.

The simulation model itself consists of a "virtual" part that deals with the continuous processes and the functioning of the simulation model and a "real" part that deals with the behaviour of the modules. The "virtual" part is represented in the simulation model by the logic engine high-level building block and the "real" part is represented by the four different high-level building blocks of the ERM method.

The four high-level building blocks of the ERM method represent the following four typical elements (*i.e.* the smaller plants) that make up stochastic continuous systems such as petrochemical plants:

a) A smaller plant with a multiple service cycle and failures of the modules.
b) A smaller plant with a service cycle and failures of the modules.
c) A smaller plant with a service cycle of the modules.
d) A smaller plant with failures of the modules.

The five high-level building blocks (*i.e.* the logic engine and the four high-level building blocks of the ERM method) of the generic simulation modelling methodology can be used to construct simulation models of stochastic continuous systems. The generic methodology renders simulation models that exhibit the following characteristics: short development and maintenance times, user-friendliness, short simulation runtimes, compact size, robustness, accuracy and a single software application (Albertyn 2004).

IDENTIFICATION TECHNIQUES

Detailed explanations of the eight methods and techniques of the generic simulation modelling methodology do not fall within the scope of this paper. However, in order to facilitate a better understanding and interpretation of the scenario analyses results that are presented, it is necessary to provide some detail about the two "bottleneck" identification techniques of the generic methodology.

The generic simulation modelling methodology uses the following two "bottleneck" identification techniques:

a) The time that each of the smaller plants is the "bottleneck", as a percentage.
b) The possible production that is lost due to each of the smaller plants, as a percentage.

The time that each of the smaller plants is the "bottleneck" over a chosen period of time, as a percentage, is the period of time that the smaller plant is the "bottleneck" over the chosen period of time, as a constant, divided by the chosen period of time, as a constant, multiplied by 100.

$$\text{"}Btt\text{"}_{PltTim} = ((Tim_{PltBtt})/(Tim_{Tot}))(100) \ (\%) \qquad (1)$$

Where:

a) $\text{"}Btt\text{"}_{PltTim}$ is the time that the smaller plant is the "bottleneck" over the chosen period of time, as a percentage.
b) Tim_{PltBtt} is the period of time that the smaller plant is the "bottleneck" over the chosen period of time, as a constant, in hours.
c) Tim_{Tot} is the chosen period of time, as a constant, in hours.

The production that is lost due to each of the smaller plants over a chosen period of time, as a percentage, is the production that is lost due to the smaller plant over the chosen period of time, as a percentage of the steady state maximum possible production over the chosen period of time, divided by the total production that is lost over the chosen period of time, as a percentage of the steady state maximum possible production over the chosen period of time, multiplied by 100.

$$\text{"}Btt\text{"}_{PltPrdLst} = ((Prd_{PltLst})/(Prd_{SFPltLst}))(100) \ (\%) \qquad (2)$$

Where:

a) $\text{"}Btt\text{"}_{PltPrdLst}$ is the production that is lost due to the smaller plant over the chosen period of time, as a percentage.
b) Prd_{PltLst} is the production that is lost due to the smaller plant over the chosen period of time, as a percentage of the steady state maximum possible production over the chosen period of time.
c) $Prd_{SFPltLst}$ is the total production that is lost over the chosen period of time, as a percentage of the steady state maximum possible production over the chosen period of time.

Even though Equations (1) and (2) use the term "smaller plant", they are equally applicable when the term "smaller plant" is replaced with the term "possible "bottleneck" point" to make provision for instances where some of the smaller plants consist of groupings of different types of modules.

SYSTEM DESCRIPTION

The configuration and process flow of the Synthetic Fuel plant can be derived from Figure 1. For the purpose of the simulation model the total plant is considered to consist of 20 smaller plants and 147 modules. A module can be defined as a grouping of components that has a specific function. The modules are characterised by their production capacities, service schedules and failure characteristics.

Some of the smaller plants consist of groupings of different types of modules. For example, the Oxygen plant consists of three groupings of different types of modules and a total of 19 modules, namely: Oxygen-A with six air turbine and compressor sets, Oxygen-B with six cold boxes and Oxygen-C with seven oxygen turbine and compressor sets (see the detail view of the Oxygen plant in Figure 1). The Oxygen plant actually consists of six parallel lines, each one containing three modules, namely: an air turbine and compressor set, a cold box and an oxygen turbine and compressor set. Such a serial, parallel line within a smaller plant is sometimes referred to as a "train". The seventh oxygen turbine and compressor set represents a reserve production capacity and it was introduced because of the high failure rate of the oxygen turbine and compressor sets. Plant(II) shows another configuration with two groupings of different types of modules and a total of 10 modules, namely: Plant(II)-A with eight chemical reactors and Plant(II)-B with two turbine and compressor sets (see the detail view of Plant(II) in Figure 1).

The following key characteristics of the Synthetic Fuel plant can be identified:

a) It is a continuous process plant.
b) The plant is subject to two types of discrete events, namely:
   i) Chronological events (*i.e.* the services of the modules).
   ii) Stochastic events (*i.e.* the failures of the modules).
c) The plant has complex interrelationships.

The motion of the coal, gases and liquids in the Synthetic Fuel plant is characterised as flow and therefore the process of the plant is characterised as continuous. Even though the coal from the mines is in the solid phase, it is considered as a fluid because it consists of chunks that are moved along on conveyor belts. The services of the modules are strictly chronological events and are characterised by the service cycles of the modules. The failures of the modules are stochastic events and are characterised by the failure characteristics of the modules.

The complex interrelationships of the Synthetic Fuel plant are manifested in both the process flow and process logic of the plant. The complexity of the process flow follows from the presence of feedback-loops to Plant(II)-A from the Division Process and Recycling plants and the division of the output of the Steam and Oxygen plants. Crowe *et al.* (1971) provide some insights into the problems that are posed by the recycling (*i.e.* the feedback-loops) of either heat or matter in chemical plants. The complexity of the process logic of the plant follows from the rules of operation of the plant. For example, one of the rules of operation states that steam will only be supplied to the Electricity Generation plant once the Gas Production and Oxygen plants have been supplied. The continuous nature

of the process of the plant implies that all 147 modules are, in a way, intrinsically interlinked as far as the effect of the service or failure of a module is concerned. Any breakdown in the production capacity at one point because of the service or failure of a module, does have an immediate effect on upstream and downstream operations.

SIMULATION MODEL

For the purpose of the PhD research the five high-level simulation model building blocks of the generic simulation modelling methodology were developed in two different simulation software packages, namely: Arena and Simul8. The high-level building blocks were then used to construct identical simulation models of the Synthetic Fuel plant in both Arena and Simul8. Two simulation models were constructed to illustrate clearly that the generic methodology is not based on, or restricted to, a specific simulation software package.

Both the Arena and Simul8 simulation models of the Synthetic Fuel plant make provision to use either the iteration time interval (ITI) evaluation method option or the event-driven (ED) evaluation method option. The ED evaluation method option evaluates the simulation models only when an event takes place and not every time interval like the ITI evaluation method option. The ITI and ED evaluation methods are compared and their strengths and weaknesses are discussed in a conference paper by Albertyn (2000 Summer Computer Simulation Conference 2000). For the purpose of this paper the ED evaluation method option Simul8 simulation model of the Synthetic Fuel plant is used because of its excellent simulation runtime of only 6,8 minutes for 20 replications of a simulated time period of one year.

The results of the different replications of a simulation run of a stochastic simulation model are usually not identical because of the random (*i.e.* the stochastic) behaviour of the random phenomena like failures. This implies that a simulation run consisting of more than one replication has to be completed in order to obtain a mean result that is representative of the simulated scenario. The determination of the minimum number of replications that would yield a mean result that is representative of the simulated scenario is a determination of minimum sufficient sample size problem. An equation from Crow *et al.* (1960) can be used to determine the minimum sufficient sample size. From Crow *et al.* it follows that the minimum sufficient sample size for a simulation run of the ED evaluation method option Simul8 simulation model of the Synthetic Fuel plant is 13 (Albertyn 2004). The minimum sufficient sample size was calculated from the results of a simulation run comprising 20 replications of a simulated time period of one year and with an allowance for a 0,5% deviation from the real-world mean output throughput value of the Gas Production plant and a 99% confidence interval. The number of replications completed should be more than, or equal to, the calculated minimum sufficient sample size for the answers to be taken as representative of the simulated scenario and it is obvious that this constraint is adhered to in this instance.

Various authors and manuals stress the importance of comprehensive simulation model verification and validation before the results that are generated by a simulation run can be accepted as representative of the simulated scenario (Harrell and Tumay 1999; Kelton *et al.* 1998; Pegden *et al.* 1995; *Simul8®: Manual and Simulation Guide* 1999). According to Pegden *et al.* verification is the process of determining that the simulation model operates as intended and validation is the process of reaching an acceptable level of confidence that the inferences drawn from the simulation model are correct and applicable to the real-world system that is represented. The ED evaluation method option Simul8 simulation model of the Synthetic Fuel plant is verified and validated in the doctoral thesis (Albertyn 2004). For example, as part of the verification process, the number of services and failures generated by the simulation model during a simulation run are compared with the real-world number of services and failures that occur and, as part of the validation process, the mean output throughput value of the simulation model in a known scenario is compared to the real-world mean output throughput value of the Synthetic Fuel plant in the known scenario. The known scenario is the 1993 production year of the Synthetic Fuel plant and the mean raw gas output throughput value of the Gas Production plant is used as the variable of comparison. The mean output throughput value of the Gas Production plant of the ED evaluation method option Simul8 simulation model deviates less than 1% from the mean output throughput value of the Gas Production plant during the 1993 production year (Albertyn 2004). The deviation was calculated from the results of a simulation run comprising 20 replications of a simulated time period of one year. This result (a deviation of less than 1%) indicates that it can be accepted that the ED evaluation method option Simul8 simulation model is a valid representation of the Synthetic Fuel plant.

As part of the validation process the ED evaluation method option Simul8 simulation model of the Synthetic Fuel plant is also subjected to a statistical validity test in the doctoral thesis (Albertyn 2004). An equation from Miller *et al.* (1990) can be used to test the hypothesis that a population mean is statistically equal to a specified value. The term "statistically equal" is used to indicate that even though the population mean and the specified value may not be numerically equal, the difference between the population mean and the specified value may not be significant for a particular level of significance and hence the population mean and the specified value may be considered as equal if viewed from a statistical perspective.

In this instance the mean output throughput value of the Gas Production plant of the ED evaluation method option

Simul8 simulation model (calculated from the results of a simulation run comprising 20 replications of a simulated time period of one year) represents the population mean and the mean output throughput value of the Gas Production plant during the 1993 production year represents the specified value. From Miller *et al.* (1990) it follows that the mean output throughput value of the Gas Production plant of the ED evaluation method option Simul8 simulation model is statistically equal to the mean output throughput value of the Gas Production plant during the 1993 production year, with a 5% level of significance (Albertyn 2004). Therefore it can be accepted (in terms of the mean output throughput of the Gas Production plant) that the ED evaluation method option Simul8 simulation model is a valid representation of the Synthetic Fuel plant, with a 5% level of significance.

SCENARIO ANALYSES

The ED evaluation method option Simul8 simulation model of the Synthetic Fuel plant is used to evaluate the following two alternative scenarios:

a) Scenario I is representative of the Synthetic Fuel plant without the inclusion of the Oxygen Extra plant (*i.e.* the inclusion of an extra oxygen "train" that comprises an air turbine and compressor set, a cold box and an oxygen turbine and compressor set) and is used to identify the problem areas or "bottlenecks" in the plant.
b) Scenario II is representative of the Synthetic Fuel plant with the Oxygen Extra plant incorporated and is used to determine how this addition impacts on the throughput of the plant.

One of the rules of operation of the Synthetic Fuel plant states that only the smaller plants that form part of the main-gas-cycle can act as "bottlenecks" that influence the maximum possible throughput of the Synthetic Fuel plant. There are 10 smaller plants in the main-gas-cycle and they are sometimes referred to as the "heart" of the Synthetic Fuel plant. Two of the 10 smaller plants consist of groupings of different types of modules. The Oxygen plant consists of three groupings of different types of modules and Plant(II) consists of two groupings of different types of modules. The 10 smaller plants of the main-gas-cycle therefore represent 13 possible separate points, any one of which can be the "bottleneck" that determines the maximum possible throughput of the Synthetic Fuel plant at any given moment in time. The 13 possible "bottlenecks" in the main-gas-cycle are the following: Coal Processing, Steam, Gas Production, Temperature Regulation, Oxygen-A, -B and -C, Plant(I), Plant(II)-A and -B, Plant(III), Division Process and Recycling. These "bottlenecks" determine the maximum possible throughput of the Synthetic Fuel plant at any given moment in time.

A simulation run comprising 20 replications of a simulated time period of one year was completed for both Scenario I and II with the ED evaluation method option Simul8 simulation model of the Synthetic Fuel plant. Table 1 provides the Scenario I results for the "bottlenecks" in terms of the time and production lost criteria (see Equations (1) and (2)).

Table 1: Scenario I "Bottlenecks"

No.	"Bottleneck" Name	Tim (%)	PrdLst (%)
1	Coal Processing	0,08	0,09
2	Steam	0,00	0,00
3	Gas Production	1,09	0,51
4	Temperature Regulation	0,00	0,00
5	Oxygen-A	11,17	18,45
6	Oxygen-B	1,32	2,14
7	Oxygen-C	0,19	0,31
8	Plant(I)	27,91	28,20
9	Plant(II)-A	57,53	46,70
10	Plant(II)-B	0,04	0,10
11	Plant(III)	0,26	1,34
12	Division Process	0,41	2,16
13	Recycling	0,00	0,00
Total		**100,00**	**100,00**

Where:

a) No. is the "bottleneck" identification number.
b) Tim is the time that the "bottleneck" is the "bottleneck" (%).
c) PrdLst is the production that is lost due to the "bottleneck" (%).

From Table 1 it follows that the three most important "bottlenecks", in order of decreasing importance, are Plant(II)-A, Plant(I) and Oxygen-A. Both the "bottleneck" identification criteria support this finding. Plant(II)-A is the "bottleneck" for approximately 58% of the time and is responsible for approximately 47% of the production that is lost. Plant(I) is the "bottleneck" for approximately 28% of the time and is responsible for approximately 28% of the production that is lost. These results thoroughly substantiate the perception of the engineering division of the Synthetic Fuel plant that Plant(II)-A and Plant(I) are the "troublemakers". Oxygen-A is the "bottleneck" for only approximately 11% of the time but is responsible for approximately 18% of the production that is lost. This indicates that when Oxygen-A is the "bottleneck", it has a pronounced effect on the throughput of the Synthetic Fuel plant and therefore Oxygen-A is a valid candidate for additional production capacity, even though more production is lost at

Plant(II)-A and Plant(I). Figure 2 provides a graphical representation of the results that are presented in Table 1.

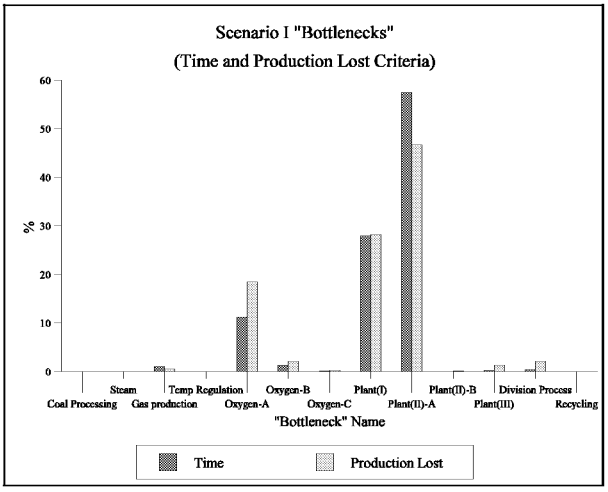


Figure 2: Scenario I "Bottlenecks"

It is interesting to note that these results also substantiate the Pareto principle that states that in every group or collection of entities there exist a vital few and a trivial many (Pegden *et al.* 1995). In this instance the group that comprises the three most important "bottlenecks" (*i.e.* Plant(II)-A, Plant(I) and Oxygen-A) is the "bottleneck" for approximately 97% of the time and is responsible for approximately 93% of the production that is lost, even though the group only represents approximately 23% (3 out of 13) of the possible "bottlenecks".

Before the results of the Scenario I and II simulation runs can be compared, it is necessary to prove that the two scenarios can be differentiated in terms of representing two different outcomes. Pegden *et al.* (1995) and the *Simul8®: Manual and Simulation Guide* (1999) indicate that confidence intervals should be taken into consideration when alternatives are compared. If the confidence intervals for the results of two scenarios overlap, the two scenarios cannot be differentiated in terms of representing two different outcomes. Various sources (Miller *et al.* 1990; Pegden *et al.* 1995; *Simul8®: Manual and Simulation Guide* 1999) detail the theoretical background for the determination of a confidence interval for a population mean. In this instance the 99% confidence intervals for the results of the Scenario I and II simulation runs do not overlap and therefore the two scenarios can be assumed to represent two different outcomes (Albertyn 2004).

Table 2 provides a comparison between the Scenario I and II results for the three most important "bottlenecks" (arranged in order of decreasing importance from top to bottom) in terms of the time and production lost criteria. From Table 2 it follows that the group that comprises the three most important "bottlenecks" (*i.e.* Plant(II)-A, Plant(I) and Oxygen-A) is the "bottleneck" for a total of approximately 97% of the time and is responsible for a total of approximately 93% of the production that is lost in Scenario I. Oxygen-A is the "bottleneck" for approximately 11% of the time and is responsible for approximately 18% of the production that is lost in Scenario I. However, in Scenario II the group that comprises the three most important "bottlenecks" is the "bottleneck" for a total of approximately 98% of the time and is responsible for a total of approximately 95% of the production that is lost while Oxygen-A is the "bottleneck" for less than 1% of the time and is responsible for less than 1% of the production that is lost. These results clearly indicate that Oxygen-A does not qualify as an important "bottleneck" in Scenario II. The results that are presented in Table 2 are graphically depicted in Figure 3 which shows the time criterion comparison on the left-hand side of the graph and the production lost criterion comparison on the right-hand side of the graph.

Table 2: Scenario I and II "Bottlenecks"

No.	"Bottleneck" Name	Scenario I		Scenario II	
		Tim (%)	PrdLst (%)	Tim (%)	PrdLst (%)
9	Plant(II)-A	57,53	46,70	65,75	61,32
8	Plant(I)	27,91	28,20	31,98	33,19
5	Oxygen-A	11,17	18,45	0,21	0,41
Total		**96,61**	**93,35**	**97,94**	**94,92**

Where:

a) No. is the "bottleneck" identification number.
b) Tim is the time that the "bottleneck" is the "bottleneck" (%).
c) PrdLst is the production that is lost due to the "bottleneck" (%).

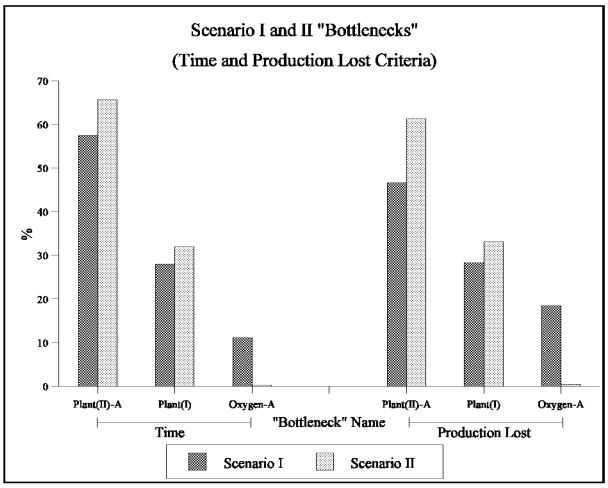


Figure 3: Scenario I and II "Bottlenecks"

The results of the Scenario I and II simulation runs also

indicate that there is an increase of 1,44% in the mean output throughput value of the Gas Production plant from Scenario I to Scenario II. That is equivalent to an increase of approximately five production days in a one year period, if the increase is expressed in terms of Scenario I production days.

CONCLUSIONS

The scenario analyses demonstrate how a simulation model of a petrochemical plant that was developed with the generic simulation modelling methodology is used to evaluate two alternative scenarios. In the first scenario two "bottleneck" identification techniques are used to identify the "bottlenecks" in the plant and in the second the impact of additional production capacity at one of the identified "bottlenecks" is determined. It stands to reason that the simulation model can also be used for the evaluation of a multitude of other scenarios.

The most obvious possible range of application of the generic simulation modelling methodology is found within the petrochemical industry. The coal-to-liquid process (demonstrated by the example in this paper), the classic crude oil refinement process and the gas-to-liquid (GTL) process can all be accommodated by the generic methodology without any difficulty. There are many petrochemical plants (especially crude oil refineries) all over the world that use these processes and each of these plants is a potential client for an application of the generic methodology. The possible range of application of the generic simulation modelling methodology is by no means restricted to only the petrochemical industry. Any plant that exhibits the same characteristics as the Synthetic Fuel plant (*i.e.* any stochastic continuous system) can readily be accommodated by the generic methodology. For example, a plant that manufactures ammonia or paint is obviously also a potential client for an application of the generic methodology.

Unfortunately there is no *Petrochemical Plant Simulation for Dummies*. In *The Goal* (Goldratt and Cox 1992) the character Alex Rogo remarks: "*The complexity in this plant - in **any** manufacturing plant - is mind-boggling if you contemplate it.*" Petrochemical plants are, by the very nature of the processes involved, extremely complex systems. There is no quick and easy way to develop a high quality simulation model of a petrochemical plant. It takes time, commitment and diligence. The aim of the generic simulation modelling methodology is to provide a "tool" that makes the development and use of simulation models of stochastic continuous systems (such as petrochemical plants) easier and more effective. The efficiency of the generic methodology follows from a structured approach and the characteristics of simulation models that are developed with the generic methodology, namely: short development and maintenance times, user-friendliness, short simulation runtimes, compact size, robustness, accuracy and a single software application. The generic methodology, however, is not a magic wand that can effortlessly render high quality simulation models of petrochemical plants.

REFERENCES

16th European Simulation Multiconference (ESM'2002) (2002: Darmstadt, Germany). 2002. *Proceedings*. Ghent, Belgium: [SCS] A Publication of SCS Europe. p.29-36.

2000 Summer Computer Simulation Conference (2000: Vancouver, British Columbia, Canada). 2000. *Proceedings*. San Diego, CA: The Society for Computer Simulation International. p.129-134.

Albertyn, M. 2004. *Generic Simulation Modelling of Stochastic Continuous Systems. Philosophiae Doctor* Thesis. Department of Industrial and Systems Engineering, University of Pretoria, Pretoria. p.140-149, 152-153, 176-177, 210-214.

Crow, E.L.; F.A. Davis; and M.W. Maxfield. 1960. *Statistics Manual*. New York, Dover Publications, Inc. p.48.

Crowe, C.M.; A.E. Hamielec; T.W. Hoffman; A.I. Johnson; D.R. Woods; and P.T. Shannon. 1971. *Chemical Plant Simulation: An Introduction to Computer-Aided Steady-State Process Analysis*. Englewood Cliffs, N.J.: Prentice-Hall, Inc. p.1.

Goldratt, E.M. and Cox, J. 1992. *The Goal*, Second Revised Edition. [*S.l.*]: Creda Communications. p.43.

Harrell, C.; and K. Tumay. 1999. *Simulation Made Easy: A Manager's Guide*, 5th ed. Norcross, Georgia: Engineering & Management Press. p.87-88.

Kelton, W.D.; R.P. Sadowski; and D.A. Sadowski. 1998. *Simulation with Arena*. Boston, Massachusetts: WCB McGraw-Hill. p.444-446.

Miller, I.; J.E. Freund; and R.A. Johnson. 1990. *Probability and Statistics for Engineers*, Fourth Edition. Englewood Cliffs, New Jersey: Prentice-Hall International, Inc. p.197-199, 210-214.

Owen, R. 1994. Strategic maintenance is essential. *Mechanical Technology*, May 1994, p.15, 17.

Pegden, C.D.; R.E. Shannon; and R.P. Sadowski. 1995. *Introduction to Simulation Using SIMAN*, 2nd ed. New York: McGraw-Hill, Inc. p.15-16, 36-38, 129-153.

Simul8®: Manual and Simulation Guide. 1999. Herndon, V.A.: Simul8 Corporation. p.34, 39-48.

AUTHOR BIOGRAPHY

MARTIN ALBERTYN was born on 21 December 1962 in Pretoria, South Africa and studied at the University of Pretoria. He obtained a Batchelor degree in Mechanical Engineering (*Cum Laude*) in 1985 and Honour, Master and PhD degrees in Industrial Engineering in 1992, 1996 and 2005 respectively. He has been involved with the development of simulation models in both the military and industrial environments since 1987.

PAUL KRUGER was born on 2 November 1944 in Johannesburg, South Africa and holds DSc-Engineering and MBA degrees. He is at present a full professor in the Department of Industrial and Systems Engineering at the University of Pretoria and has been intimately involved in the teaching and practice of simulation modelling for more than 30 years.

MILITARY OPTIMIZATION AND RISK ASSESSMENT

USING A SIMULATION MODEL TO DETERMINE THE OPTIMUM NUMBER OF PILOTS IN AN AIR FORCE

Martin Albertyn
Defence Institute,
a Division of Armscor Business (Pty.) Ltd.
PO Box 65014
0001 Pretoria, South Africa
e-mail: martina@sadi.co.za

Stephen D Trichard
Directorate Air Capability Plan
South African Air Force
Private Bag X199
0001 Pretoria, South Africa
e-mail: sct1@mweb.co.za

KEYWORDS

Simulation model, Air force pilots, Simul8.

ABSTRACT

The key objective is to present a simulation model that is used to determine the optimum number of pilots in an air force. Background information about the simulation project and a cursory description of the system under consideration are provided. The Simul8 simulation software package, the basic concepts of the simulation model, the determination of minimum sufficient sample sizes and simulation model verification and validation are discussed. Generic examples of typical results are shown and conclusions about the simulation project are presented.

INTRODUCTION

The *Pilot Operational Personnel Simulation Model* was developed by the Defence Institute for the Directorate Air Capability Plan (DACP) of the South African Air Force (SAAF). The Defence Institute is a division of Armscor Business (Pty.) Ltd. and its primary function is to provide support to the Staff Officers of the South African National Defence Force (SANDF). During 2003 the DACP identified a requirement for a decision support tool to help them to determine the number of pupil pilots that should enter the system every year in order to keep the flying schools and operational squadrons fully staffed with qualified pilots. After discussions between representatives of the DACP and the Defence Institute it was decided that a stochastic simulation model would be the correct decision support tool to utilise in this instance, and henceforth the simulation project was initiated. (It is important to note that the simulation model is primarily concerned with the operational pilots in the SAAF and does not include pilots that are in other postings, with the exception that the simulation model keeps track of the number of operational pilots that moves on to Staff postings during the period of evaluation.)

The simulation model was developed over a period of approximately two-and-a-half years from the latter part of 2003 to 2005. The total expenditure in terms of man-hours over this period was approximately 1700 hours. (This number reflects only the man-hours expended by the Defence Institute and does not include the man-hours expended by the Staff and Operational Officers of the SAAF.) This development history relates closely to a comment from Crowe *et al.* (1971) to the effect that it may take a few man-years to supply answers to complex problems with a simulation model.

The further development and use of the simulation model are ongoing processes and it is anticipated that the simulation model will continue to be enhanced, refined and used as a decision support tool in the foreseeable future. Even though the simulation model was originally established to answer the core question regarding the number of pupil pilots that should enter the system every year, its role as a decision support tool has evolved over time.

Currently the key aspects under consideration are the following:

a) The determination of the optimum number of pupil pilots that should enter the system every year.
b) The determination of the optimum number of pilots at the flying schools and the operational squadrons.
c) The identification of areas of concern (*i.e.* by identifying "bottlenecks" in the system, by monitoring the ratio between pupils and instructors, *etc.*).

The simulation model is used to determine the optimum number of pilots in the SAAF for every scenario that is evaluated. It is important to note that in this instance the term "optimum" does not refer to a formal mathematical optimum but rather to an answer that is good enough (*i.e.* acceptable) in the given circumstances, even though it may not be the absolute best answer possible.

Pegden *et al.* (1995) propose 12 steps that should be taken in every simulation project. A detailed discussion of the steps of a simulation project, as well as the way that they manifested in this particular instance, does not fall within the scope of this paper. However, the system under consideration, the simulation software package that is used, the basic concepts of the simulation model, examples of typical results and conclusions about the simulation project will be discussed.

SYSTEM DESCRIPTION

In order to facilitate a better understanding of the *Pilot Operational Personnel Simulation Model*, it is necessary to provide an overview of the system under consideration. In this instance the system under consideration consists of all the schools and squadrons of the SAAF where operational pilots are trained and used.

The simulation model incorporates the Central Flying School and the three System Groups of the SAAF, namely: the Combat System Group, the Helicopter System Group and the Air Transport and Maritime System Group (see Figure 1 for a simplified schematic representation of the structure of the simulation model). The Air Transport and Maritime System Group consists of a conversion squadron and three different "lines", namely: the Passenger line, the Tactical line and the Maritime line. The pupil pilots enter the system through the Central Flying School and those that qualify as pilots are posted, according to requirements, predisposition and personal preference, to the System Groups. Upon their arrival at the System Groups, the pilots are first converted to the required aircraft type and then those that qualify are moved forward on the career path of each System Group. Each school and squadron comprises a number of sub-groupings of pupil pilots, pilots, copilots, commanders and instructors.

For example, 35 Squadron consists of three sub-groupings, namely: the C-47 copilot group, the C-47 commander group and the C-47 instructor group. Upon arrival at 35 Squadron the pilots are designated C-47 copilots and are trained on the aircraft type by the C-47 instructors. Once the C-47 copilots have completed a tour of duty of approximately two years and amassed enough flying hours, they qualify as C-47 commanders and complete another tour of duty of approximately four years. Therefore a pilot will typically spend approximately six years at 35 Squadron in the roles of copilot and commander while the tour of duty of an instructor at the squadron is approximately three years. The same basic concept, with either simplifications or complications in some instances, applies to all the schools and squadrons.

Along the way and upon completion of the pilot career paths of the different System Groups, pilots with the required predisposition and experience may return to the Central Flying School to be trained as instructors. Once qualified as instructors they complete a tour of duty at the Central Flying School and then return to the System Group that they originally came from to be trained and used as instructors at the schools and squadrons. The whole system therefore actually consists of two parallel career paths, one for the pilots and the other for the instructors. Because of the rigours of flight training, the ratio of pupils (*i.e.* the pupil pilots at the Central Flying School and the pilots that are trained on the aircraft types at the other schools and squadrons) to instructors, is of the utmost importance.

There are also a multitude of other complexities that have to be accommodated by the simulation model, which are not indicated in the simplified diagram that is presented in Figure 1. For example, pilots that fail to qualify as combat pilots at 85 Combat Flying School, may be rerouted from the Combat System Group to the Air Transport and Maritime System Group where they will start their conversion to transport pilots at 41 Squadron. Conversely, C-130 commanders and instructors that have completed their tour of duty at 28 Squadron in the Air Transport and Maritime System Group, may transfer to 60 Squadron in the Combat System Group (*i.e.* move from multi-engine turboprop powered aircraft to multi-engine turbofan powered aircraft). From these examples it is evident that a myriad of possible career paths exists for pilots in the SAAF.

MODEL DETAIL

Background

According to Pegden *et al.* (1995) very few real-world systems are free from the influence of random variation. A simulation model that accommodates the randomness of a system can render a very close approximation of the behaviour of the system and therefore it was decided that a stochastic simulation model would be the correct decision support tool to use in this instance.

Kelton *et al.* (1998) propose that the following three dimensions can be used to classify simulation models: static versus dynamic, discrete versus continuous and deterministic versus stochastic. According to this proposed system of classification the *Pilot Operational Personnel Simulation Model* may be classified as a dynamic, discrete, stochastic simulation model. Such a simulation model is usually, for the sake of simplicity, referred to just as a stochastic simulation model (the dynamic and discrete parts of the classification are usually assumed or taken for granted).

Software Package

The simulation model was developed in the Simul8 simulation software package. Two different simulation software packages were originally considered, namely Simul8 and Arena. (Both Simul8 and Arena are registered trademarks that are usually denoted by Simul8® and Arena® respectively but for the sake of simplicity it will be written simply as Simul8 and Arena in this paper.) The primary reasons why Simul8 was chosen are the following: a relatively low acquisition cost, no annual licencing fees, a comparatively simplistic modelling environment and the inclusion of an internal logic programming language. (Albertyn (2004) provides a cursory comparison between the Arena and Simul8 simulation software packages.)

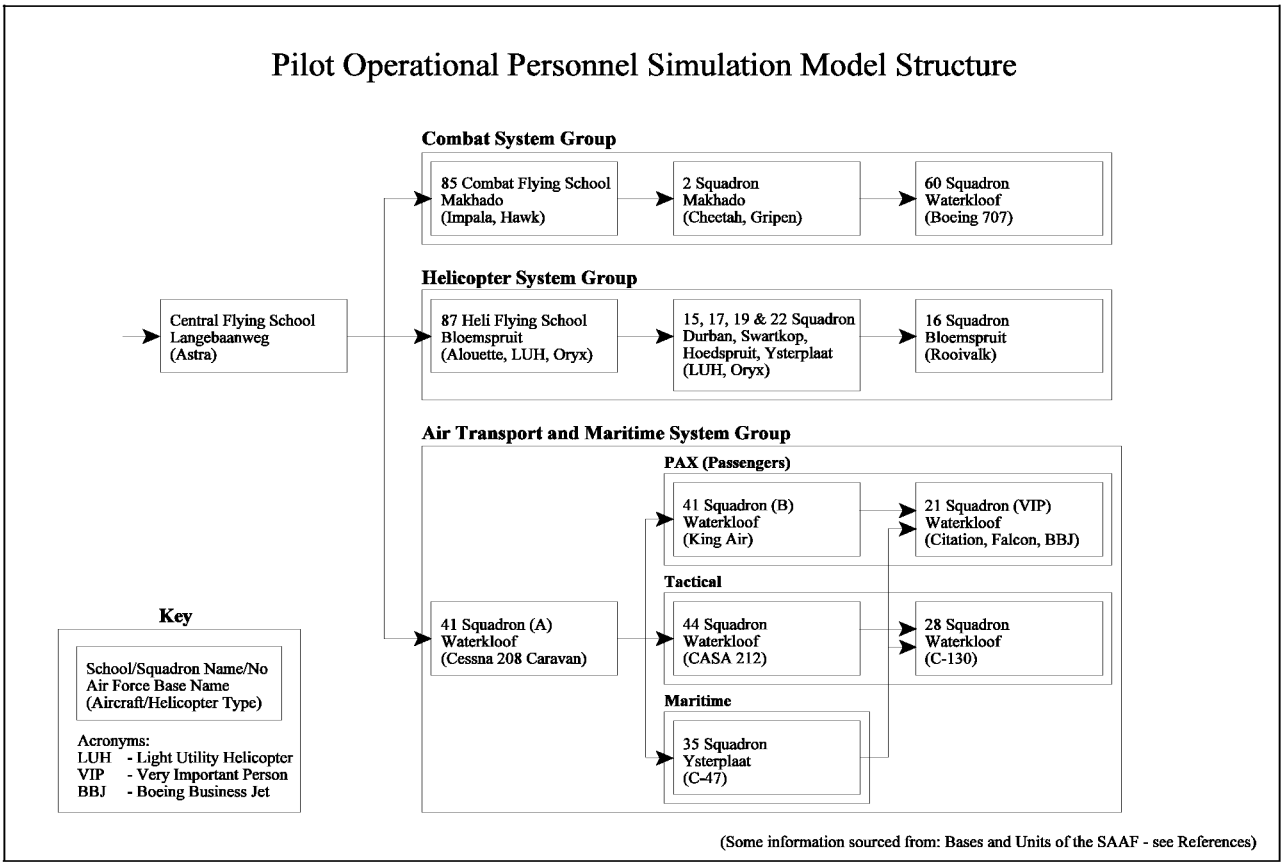


Figure 1: Pilot Operational Personnel Simulation Model Structure

The low acquisition cost and simplistic modelling environment were deemed to be important factors because in the original planning of the simulation project it was envisaged that the Defence Institute would be responsible for the development and maintenance of the simulation model but that the day-to-day use of the simulation model for scenario analysis would be handled by the DACP. The implementation of that plan would obviously have resulted in the acquisition of an additional Simul8 simulation software package for the DACP. That plan has been revised since and the current model of operation is that DI develops, maintains and uses the simulation model for scenario analysis, in close collaboration with the Staff Officers of the DACP. The internal logic programming language of Simul8 is called Visual Logic (VL). The following quotation from the *Simul8®: Manual and Simulation Guide* (1999) explains what VL is and how it is used in a simulation model: "*Visual Logic (VL) is Simul8's logic building environment. In a simulation of significant complexity you will want to add your own rules for deciding how to process work. VL lets you add very detailed logic to control the operation of your simulation.*"

Description

At the start of the simulation project a lot of effort was expended to establish a system description and the boundaries of the system under consideration. More specifically, the major problem was to determine the required resolution or level of detail of the simulation model. The level of detail that is required should be chosen in accordance with the objectives of the simulation model. Enough detail should be included to validate any inferences drawn from the use of the simulation model, without making the simulation model cumbersome by the inclusion of unnecessary trivia. During discussions with representatives of the System Groups, the required level of detail was determined and the expectations of the clients were aligned with the technical possibilities of simulation modelling.

The simulation model consists of the following three hierarchical levels:

a) Level 1 - contains the sub-groupings of pupil pilots, pilots, copilots, commanders and instructors of the flying schools and the operational squadrons.
b) Level 2 - contains the modelling constructs of the sub-groupings.
c) Level 3 - contains the detail to model the attrition of pilots that are lost due to non-renewal of contracts and resignations.

Figure 2 shows the layout of the simulation model in the Simul8 simulation window. The Central Flying School and the Combat System Group are shown in the view that is presented, while scrolling down in that view on a computer screen will reveal the Helicopter System Group

and the Air Transport and Maritime System Group. This view represents the first hierarchical level (*i.e.* Level 1) and each sub-grouping of pupil pilots, pilots, copilots, commanders and instructors is represented by an aircraft or helicopter icon. The layout of the icons in the simulation window conforms very closely with the diagrams of the system description that were developed during the initial stages of the simulation project. According to Elder (1992) the realistic representation of a simulation model in a layout or configuration that is immediately recognisable is fundamental to the successful familiarisation with, orientation to, and acceptance of, the simulation model by clients and users.

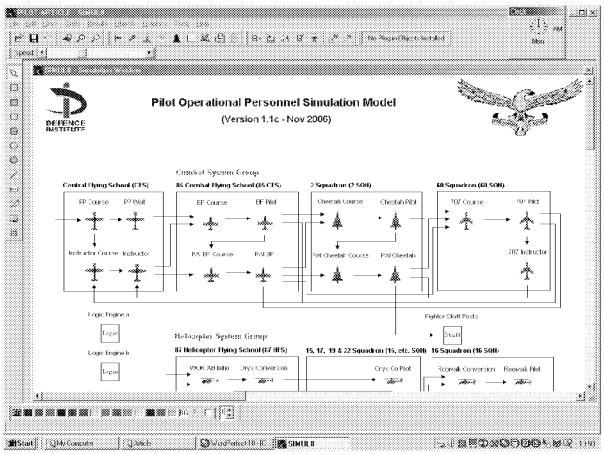


Figure 2: Simul8 Simulation Window

Clicking on an aircraft or helicopter icon will reveal a Simul8 subwindow that contains the modelling constructs of the second hierarchical level (*i.e.* Level 2). Among the constructs of each subwindow is an icon that provides access to yet another subwindow that contains the constructs that model the attrition of pilots that are lost due to non-renewal of contracts and resignations on the third hierarchical level (*i.e.* Level 3).

Each pilot is created as an entity in the simulation model and flows forward according to the appropriate career path, qualifying on specific aircraft types, completing tours of duty, qualifying as an instructor, exiting the system, *etc.* The attributes of each pilot are assigned to different labels that are attached to the entity that represents the pilot. For example, upon successful completion of the instructor course at the Central Flying School, the instructor label of the pilot will be changed from zero to one, thereby signifying that the pilot is qualified as an instructor.

The pilots are routed at decision points in the simulation model with the following two types of decision branches:

a) Rule-based decision branches (for example, only qualified instructors may become instructors at the schools and squadrons).
b) Percentage-based decision branches (for example, for a specific scenario the pupil pilots that qualify as pilots at the Central Flying School may be posted to the System Groups according to a split of 15%, 50% and 35% to the Combat, Helicopter and Air Transport and Maritime System Groups respectively).

The percentage-based decision branches provide the variability or stochastic aspect of the simulation model. According to the *Simul8®: Manual and Simulation Guide* (1999) the destinations to which entities are routed are decided randomly but the exact percentage of entities going to each destination can be specified.

For every scenario that is evaluated the simulation model has to be set up according to the configuration of the system that is modelled in that specific scenario. This involves ensuring that the simulation model parameters represent the scenario that is evaluated, populating the simulation model with all the pilots that are in the system at the start of the simulation run and providing the correct schedule for the addition of pupil pilots to the system during the simulation run. The simulation model parameters determine the process flow and process logic (*i.e.* the rules of operation) of the simulation model. To populate the simulation model with all the pilots that are in the system at the start of the simulation run is no arbitrary task because each pilot has to be placed (*i.e.* created as an entity) at exactly the right place in the simulation model and assigned (to the labels of the entity) the qualifications already obtained by that pilot. For example, at the start of the simulation run a specific pilot may be in the second year of a three-year tour of duty as a Cheetah pilot at 2 Squadron and may be qualified as a pilot on all the aircraft types that lead up to that point in the career of a Combat System Group pilot, namely: Astra, Impala and Cheetah. The schedule for the addition of pupil pilots to the system during the simulation run consists of two intakes per year of pupil pilots for the duration of the simulation run.

Any logical time period can be evaluated by the simulation model but it is currently set up to evaluate a time period of 10 years. That provides insight into the probable behaviour of the system for a planning horizon of 10 years or two consecutive five-year planning cycles into the future.

Minimum Sufficient Sample Size

The results of the different replications of a simulation run of a stochastic simulation model are usually not identical because of the random behaviour of the random phenomena like failures in manufacturing plants, or in this instance, the variability that is introduced into the simulation model by the percentage-based decision branches. This implies that a simulation run consisting of more than one replication has to be completed in order to obtain a mean result that is representative of the simulated scenario. The determination of the minimum number of replications that would yield a mean result that is

representative of the simulated scenario is a determination of minimum sufficient sample size problem. An equation from Crow *et al.* (1960) can be used to determine the minimum sufficient sample size for sample sizes less than 30 and an equation from Miller *et al.* (1990) can be used for sample sizes larger than or equal to 30.

A simulation run comprising 100 replications of a simulated time period of 10 years was completed with the simulation model and the results were evaluated. The results indicate that a sample size of 100 satisfies the requirements for an allowance for a deviation of 5% (*i.e.* the maximum error of the estimate is 5%) from the values of the variables under consideration and a 95% confidence interval (using the equation from Miller *et al.*), while a sample size of 20 satisfies the requirements for an allowance for a deviation of 10% from the values of the variables under consideration and a 90% confidence interval (using the equation from Crow *et al.*). This allows simulation runs comprising 20 replications to be used during the initial stages of scenario analysis when several ideas are bandied about and simulation runs comprising 100 replications to be used in detailed scenario analysis once the scenarios under consideration have been clearly defined.

Runtime

The simulation runtime of a simulation run comprising 100 replications of a simulated time period of 10 years is approximately 65 minutes while the simulation runtime of a simulation run comprising 20 replications of a simulated time period of 10 years is approximately 13 minutes. (In both instances the time that it takes to write the results to preformatted spreadsheets are not taken into account.) The principal features of the hardware configuration of the computer that was used are an 800-megahertz processor and 128 megabytes of RAM (Random Access Memory).

Verification and Validation

Various authors and manuals stress the importance of comprehensive simulation model verification and validation before the results that are generated by a simulation run can be accepted as representative of the simulated scenario (Harrell and Tumay 1999; Kelton *et al.* 1998; Pegden *et al.* 1995; *Simul8®: Manual and Simulation Guide* 1999). According to Pegden *et al.* verification is the process of determining that the simulation model operates as intended and validation is the process of reaching an acceptable level of confidence that the inferences drawn from the simulation model are correct and applicable to the real-world system that is represented. In this instance the simulation model was verified by meticulously checking that the process flow and process logic of the simulation model function as intended (*i.e.* in accordance with the system description that is established for each scenario that is evaluated). The validation of the simulation model is more problematic because the simulation model projects from the present (*i.e.* the current configuration of the system) into the future (*i.e.* the different proposed new configurations of the system in the future) and therefore there is no historical data available against which to validate the outputs of the simulation model. However, this problem was circumvented by analysing the outputs of the simulation model together with representatives of the System Groups to ascertain whether the results realised by the simulation model could reasonably be expected in the given circumstances.

Pilot Attrition

Two types of attrition are modelled in the simulation model, namely the attrition of pilots that are lost due to accidents (modelled as a constant percentage over the period of evaluation) and the attrition of pilots that are lost due to non-renewal of contracts and resignations (modelled as a variable percentage over the period of evaluation). The simulation model originally modelled the attrition of pilots that are lost due to non-renewal of contracts and resignations as a constant percentage over the period of evaluation but was later enhanced to allow the modelling of the attrition as a variable percentage over the period of evaluation. The reason why this attrition needs to be modelled as a variable percentage is because even though the percentage of pilots that do not renew their contracts every year may be fairly constant, the number of pilots whose contracts are due for renewal each year is not constant due to historical reasons (*i.e.* moving from seven-year contracts to thirteen-year contracts recently). The percentage of pilots that resign (*i.e.* that leave the system before the end of their contracted period) each year may also be subject to fluctuation due to external factors such as a new commercial airline recruiting aggressively for experienced pilots in a specific year.

A comparison between the predicted attrition (derived from historical data and knowledgeable estimates) of pilots that are lost due to non-renewal of contracts and resignations and the results for the attrition from the simulation model for a typical scenario and a 10-year period of evaluation is shown in Figure 3. (The numerical values of the y-axis (*i.e.* the vertical axis) are not shown for confidentiality reasons.)

It is obvious that the predicted attrition shows large variation over the period of evaluation and that the simulation model is able to model the predicted attrition relatively closely. (Even though the individual results for some years may show larger deviations between the predicted attrition and the simulation model attrition, the total simulation model attrition over a 10-year period of evaluation typically deviates less than 3% from the total predicted attrition.)

It is important to note that the simulation model is not

used to predict pilot attrition. The pilot attrition is just one of many concepts that needs to be incorporated by the simulation model in order to accurately model the system under consideration.

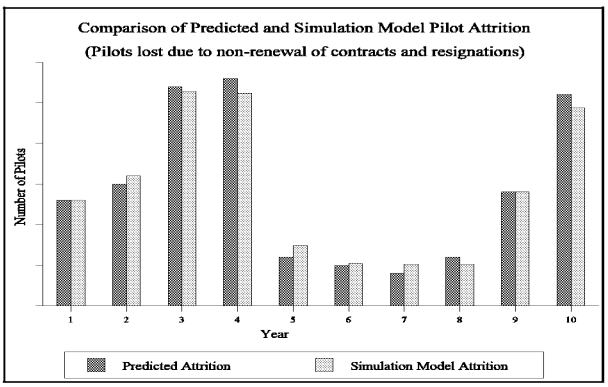


Figure 3: Comparison of Predicted and Simulation Model Pilot Attrition

User-friendliness

Upon the completion of a simulation run the simulation model automatically writes the results of the simulation run to a set of preformatted Microsoft Excel spreadsheets. (Microsoft Excel is a registered trademark and is usually denoted by Microsoft® Excel but for the sake of simplicity it will be written simply as Microsoft Excel in this paper.) This feature greatly enhances the user-friendliness of the simulation model because the user does not have to waste time manipulating blocks of data to find mean values, generate graphs, *etc.* The spreadsheets are set up to represent the required results in the required format and lead to a standardised format for the presentation of the results of different simulation runs.

Size

The size of the Simul8 file of the simulation model is 2619 kilobytes. When viewed in isolation, it is difficult to attach any significance to this numerical value. However, if this simulation model size is compared to the 937 kilobytes simulation model size of a Simul8 simulation model of a petrochemical plant (see Albertyn (2004) for a detailed discussion of the Simul8 simulation model of the petrochemical plant), it can be concluded that the *Pilot Operational Personnel Simulation Model* is a fairly sizeable simulation model.

MODEL APPLICATION

Currently there are three key areas where the simulation model is used as a decision support tool, namely: the determination of the optimum number of pupil pilots that should enter the system every year, the determination of the optimum number of pilots at the flying schools and the operational squadrons and the identification of areas of concern (*i.e.* by identifying "bottlenecks" in the system, by monitoring the ratio between pupils and instructors, *etc.*). A large number of scenarios have already been evaluated and it is foreseen that the simulation model will continue to play a role as a decision support tool in the future.

The results from the different scenarios are used to evaluate the following two basic concepts:

a) The staffing structure of a "design to functional requirements" SAAF.
b) The staffing structure of a "design to budget" SAAF.

The "design to functional requirements" SAAF represents the system that is needed to successfully complete the operations, missions and tasks that fall within the sphere of responsibility of the SAAF and the "design to budget" SAAF represents the system that is possible within the allocated budget. The goal is to meet all the functional requirements within the restrictions of the allocated budget.

In most instances an iterative process is used when a scenario is evaluated with the simulation model. First the simulation model is set up according to the configuration of the system that is modelled in that specific scenario and then a simulation run is completed to provide a baseline result. The baseline result is used as the point of departure for the further adjustment and refinement of the values of the parameters that define the configuration of the system that is modelled in the scenario. This process entails the completion of a series of simulation runs where the result of every simulation run is evaluated and then used to determine how the values of the parameters should be adjusted for the next simulation run. Through this iterative process the optimum number of pilots in the SAAF is determined for every scenario that is evaluated.

Examples of typical results that follow from the simulation model are shown in the following two figures. Figure 4 shows the results for the number of copilots, commanders and instructors at 35 Squadron for a typical scenario and a 10-year period of evaluation. (The numerical values of the y-axis are not shown for confidentiality reasons.)

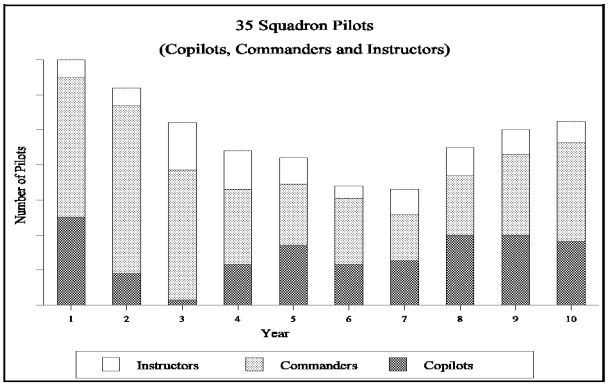


Figure 4: 35 Squadron Pilots

From Figure 4 it is obvious that the total number of pilots

at 35 Squadron first decreases and then increases slightly over the 10-year period of evaluation. This behaviour correlates closely with the future requirements at the squadron. It is also evident that even though the ratio between copilots and instructors shows some variation over the 10-year period of evaluation, the final ratio represents a sound pilot operational personnel situation for the squadron. The same principle applies to the ratio between copilots and commanders.

Figure 5 shows the results for the total number of pilots in the SAAF for a typical scenario and a 10-year period of evaluation. (Once again the numerical values of the y-axis are not shown for confidentiality reasons.)

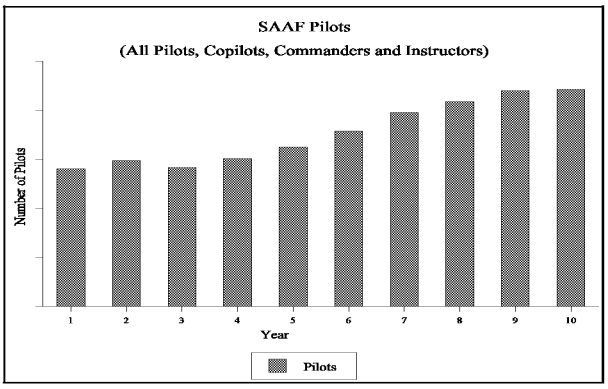


Figure 5: SAAF Pilots

From Figure 5 it is obvious that the general trend for the total number of pilots in the SAAF is upwards and stabilises around years nine and ten for this scenario. This upward trend can be ascribed to the fact that the SAAF has to grow its pilot numbers to accommodate various new acquisitions to the inventory of the SAAF, namely: the Light Utility Helicopter (LUH), the Hawk and Gripen combat aircraft and the A400M transport aircraft. The decline in the total number of pilots in the SAAF during year three can be ascribed to the high number of pilots that are lost due to non-renewal of contracts and resignations in year three (see Figure 3).

It is also interesting to note that the simulation model proved flexible enough to be used as a decision support tool in other areas of concern than the original "design domain". For example, several scenarios were evaluated with the simulation model to provide input to the responsible Staff Officer with regard to a study on the racial transformation of the SAAF.

CONCLUSIONS

At this stage in the life cycle of this simulation project it is the general consensus of all the parties that are involved that the development and use of the *Pilot Operational Personnel Simulation Model* may be classified as a successful simulation project insofar that the original goals have been met and that the simulation model has proved to be a flexible decision support tool.

The success story of this simulation project may be attributed to the following factors:

a) The belief of the director of the DACP in the power of simulation modelling as a decision support tool and a willingness to invest resources in the simulation project.
b) The exposure (through presentations and demonstrations) of the simulation project to various forums in the higher echelons of the SAAF in order to gain recognition for the simulation model and establish its credibility.
c) The involvement of the representatives from the System Groups at the start of the simulation project to establish a system description and during the simulation project to validate the outputs of the simulation model.
d) The close collaboration between the responsible Staff Officer and the Defence Institute personnel that are involved with the simulation project.
e) The fact that the required level of detail of the simulation model was kept within acceptable limits and that the expectations of the clients were aligned with the technical possibilities of simulation modelling.
f) The fact that there is an established knowledge base in the Defence Institute in terms of simulation modelling and the use of Simul8.

The *Simul8®: Manual and Simulation Guide* (1999) indicates that one of the important things to realise about simulation modelling in practice is that some benefit can already be derived just from the process of developing a simulation model, even before the simulation model is used to evaluate alternative scenarios. In this instance the process of establishing a system description of the system under consideration forced the representatives of the System Groups to revisit and re-evaluate the career paths of the pilots and the insights gained from this led to a better understanding of the intricacies of the system for all the parties that are involved.

Kruger (2003) states that: "... *many more* [simulation] *models suffer from too much detail than suffering from not enough detail. An attempt should be made to keep the* [simulation] *model as simple as possible* ...". The importance of the concept to keep it as simple as possible was once again highlighted by this simulation project and even though the concept was applied diligently during the development of the simulation model, the size and complexity of the system under consideration still lead to a fairly sizeable and complex simulation model.

Throughout the world the military budgets of countries are coming under increased pressure, leading to a dire need in the military to increase efficiency. Simulation modelling has been identified as one of the key technology areas of future research by the European Union (Geril 2002). It

therefore stands to reason that simulation modelling qualifies as a best practice technology that can be used by the military in the quest for higher efficiency. It is encouraging to see that the SAAF has taken recognisance of this fact and is using simulation modelling as a decision support tool.

The following quotation from the director of the DACP, Brigadier-General PJ van Zyl, provides the client's perspective on this simulation project:

"*Our strategic planning environment is characterised by considerable financial constraints, emphasising the need to balance resources to a much higher degree than previously required. The number of pilots in the system cannot exceed that which is affordable within the budget allocation, and that requires strict control over the number of candidates recruited for pilot training. The training cycle from recruitment to wings qualification, however, is so long that any change made in the present will only become evident three to four years hence.*

The number of variables in the Air Force pilot system is surprisingly high, and each has the potential to invalidate any correction made in the present. For this reason, it was felt that a simulation model with the ability to represent real life as closely as possible, should be able to assist the Air Force in projecting current trends into the future so that coherent recruitment and training plans could be compiled.

In this respect the simulation model developed by Albertyn & Trichard has proved a major asset in the Air Force's attempts to optimally balance its critical resources. Sufficient time has lapsed since its inception to provide some degree of empirical validation, and an extraordinarily close correlation between original personnel predictions and real life has been observed. This has increased the confidence level in the employment of the model, and has enabled strategic planners to focus on alternative outcomes rather than spending valuable time on manually fine-tuning the large range of variables, as was previously required.

From a strategic perspective, one of the most valuable aspects of the simulation model is its ability to assist multiple what-if analyses. By successively changing the most significant variables, such as resignations and pass rates at the training schools, and running multiple simulations, it is easy to identify potential bottlenecks, staffing levels at advanced training schools, and recruitment targets for candidate pilots. The ability to do worst-case projections and then relate those projections to cost and personnel implications, have proved invaluable to the Air Force. In a relatively short period, this model has become an indispensable strategic planning tool in the SA Air Force, and has made an enormous difference to the accuracy level with which we are able to balance scarce resources."

REFERENCES

Albertyn, M. 2004. *Generic Simulation Modelling of Stochastic Continuous Systems. Philosophiae Doctor* Thesis. Department of Industrial and Systems Engineering, University of Pretoria, Pretoria. p.159-162.

Bases and Units of the SAAF. http://www.saairforce.co.za/-bases.htm (8 September 2006).

Crow, E.L.; F.A. Davis; and M.W. Maxfield. 1960. *Statistics Manual.* New York, Dover Publications, Inc. p.48.

Crowe, C.M.; A.E. Hamielec; T.W. Hoffman; A.I. Johnson; D.R. Woods; and P.T. Shannon. 1971. *Chemical Plant Simulation: An Introduction to Computer-Aided Steady-State Process Analysis.* Englewood Cliffs, N.J.: Prentice-Hall, Inc. p.5.

Elder, M.D. 1992. *Visual Interactive Modelling: Some Guidelines for its Implementation and some Aspects of its Potential Impact on Operational Research. Philosophiae Doctor* Thesis. Department of Management Science, University of Strathclyde, Glasgow. p.150-153.

Geril, P. EUROSIS (European Simulation Society). Personal e-mail (10 December 2002).

Harrell, C.; and K. Tumay. 1999. *Simulation Made Easy: A Manager's Guide*, 5th ed. Norcross, Georgia: Engineering & Management Press. p.87-88.

Kelton, W.D.; R.P. Sadowski; and D.A. Sadowski. 1998. *Simulation with Arena.* Boston, Massachusetts: WCB McGraw-Hill. p.9, 444-446.

Kruger, P.S. 2003. The Art of Simulation Modelling. *South African Journal of Industrial Engineering,* May 2003, vol.14, no.1, p.39-49.

Miller, I.; J.E. Freund; and R.A. Johnson. 1990. *Probability and Statistics for Engineers*, Fourth Edition. Englewood Cliffs, New Jersey: Prentice-Hall International, Inc. p.209.

Pegden, C.D.; R.E. Shannon; and R.P. Sadowski. 1995. *Introduction to Simulation Using SIMAN*, 2nd ed. New York: McGraw-Hill, Inc. p.6, 12-13, 129-153.

Simul8®: Manual and Simulation Guide. 1999. Herndon, V.A.: Simul8 Corporation. p.29, 34, 47, 86.

AUTHOR BIOGRAPHY

MARTIN ALBERTYN was born on 21 December 1962 in Pretoria, South Africa and studied at the University of Pretoria. He obtained a Batchelor degree in Mechanical Engineering (*Cum Laude*) in 1985 and Honour, Master and PhD degrees in Industrial Engineering in 1992, 1996 and 2005 respectively. He has been involved with the development of simulation models in both the military and industrial environments since 1987.

STEVE TRICHARD was born on 30 May 1957 in Riversdale, South Africa and joined the South African Air Force in 1977. He received his pilot's wings in 1978 and became a flying instructor in 1982. He did various operational flying tours and was involved in flying and developmental training for a number of years, completing the Senior Command and Staff Course in 1998. He presently fills the post of Senior Staff Officer Air Mobility Plan in the South African Air Force.

SCENARIO MODELING IN ENVIRONMENTAL MONITORING AND CIVIL PROTECTION APPLICATIONS

Alexander Sabov, Gerhard Schreck, Jörg Krüger
Fraunhofer Institute for Production Systems and Design Technology
Pascalstrasse 8-9
D-10587, Berlin, Germany
E-mail: { alexander.sabov | gerhard.schreck | joerg.krueger } @ipk.fraunhofer.de

KEYWORDS
Scenario modeling, catastrophe management system, civil protection, disaster control, environmental monitoring

ABSTRACT

This paper describes a concept for a scenario modeling environment that allows catastrophe management systems to play out worst-case scenarios. The development focuses on the requirements, modeling techniques, architecture and integration of such a system. An example scenario is described for the flood water domain.

INTRODUCTION

Recent discussions on climate change and the forecast of more intensive extreme weather events make our society and policy more sensitive to the topics of environmental monitoring and disaster protection (IPCC 2007).

Global monitoring for environment and security has been on the research agenda for many years at the national and international level. Identified elements needed to support policy and authorities include:

- More effective and comprehensive systems for monitoring
- Validated models to provide the capability for prediction and scenario analysis
- Tools, protocols and standards for the interoperability of different systems

Process models in the context of integrated information systems will play an important role here in the future (Wyatt 2004).

In the area of industrial systems, process modeling and simulation is a technology which is accepted and used. This includes model-based monitoring and control systems, simulation-based decision support, as well as the simulation-supported training of operators.

High-level control systems dedicated to specific environmental monitoring tasks increasingly include modules for simulation support. The heterogeneity of data and models constitutes a challenging task in the development of such systems (Yagüe et al. 2003; Yagüe et al. 2004).

In the case of integrated environmental monitoring and disaster control systems, it is becoming even more important to address the aspect of heterogeneity. Nevertheless, modeling and simulation are suitable tools for the support of system development, system integration and testing, and system operation.

A main task in this type of application is the provision of suitable input data for different simulation modules. On one hand, there are well-known approaches to the use of historical data to gain the related information profiles (Lisounkin et al. 2002; Lisounkin 2003; Lisounkin et al. 2004). On the other hand, there is a need to model system evolution based on heuristics and the experience of human experts by so-called scenarios. Scenarios are neither predictions nor forecasts, but rather an alternative view of how the future may unfold (Alcamo 2001).

This paper discusses approaches to scenario modeling in the context of integrated systems for environmental monitoring and civil protection. Here, dike monitoring, flood simulation and water quality issues will be managed in an integrated manner. The work constitutes a part of an R&D project conducted with industrial partners well suited to this area. Our research approach is based on the experiences of the authors in relation to simulation and operator training in industrial implementations (Schreck 2002; Lisounkin et al. 2004).

SCENARIO SERVICES

In our approach we assume a catastrophe management system run by operators in a control center. All information on different catastrophes is collected at the center to provide a global view of the situation. Different departments, research institutes and other organizations are connected to the center for the exchange of relevant data and provision of expert components which seamlessly integrate into the global system. Civil protection departments can be informed about new events so that they can take necessary action.

Disaster control exercises must be performed in order to prepare for an emergency. The correct sequence of actions can be verified by playing out worst-case scenarios. The catastrophe management system must support the development, integration and execution of such scenarios. The following requirements must be fulfilled:

- ability to exchange time-consuming steps or unavailable personnel, components or data with a replacement service
- use of historical and fictitious data for scenarios
- concurrent handling of scenario data and real data
- modeling methods for the quick development of fictitious data, replacement services and scenarios

The replacement service can be understood as a simulation of the replaced item. The simulation focuses only on those aspects of the item that are needed in order for the scenario to continue its development. Strong simplifications can be made, or fixed data can be used depending on the purpose of the test. Therefore, the focus must be on methods that allow a fast development of simplified replacement services.

In our concept we consider three different services for scenario support as shown in Figure 1.

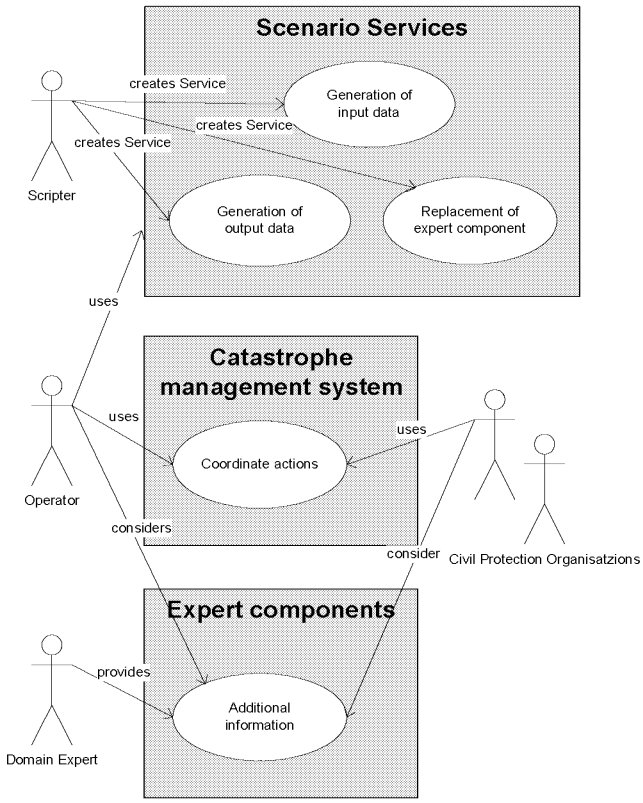


Figure 1: Scenario Support for Environmental Monitoring and Civil Protection

1. Generation of input data for expert components

Many components require input data from other components, humans, sensors and simulations. Characteristic patterns can be found by data-mining the incoming data during operation. In our approach we use a scripting language that allows a scripter to quickly develop scenario scripts that generate characteristic patterns. The script function takes generation parameters for the pattern as input and returns a dataset in the input format of the component.

2. Generation of output data for expert components

Many components generate complex output data. This process may be time-consuming or require a hazardous situation. For scenario purposes it may be enough to obtain simplified output data which does have the accuracy of the expert component but still indicates a hazard and therefore spurs further actions in the complete system. Again, we use a script function with generation parameters to return a dataset in the output format of the component.

3. Providing a replacement service for a component

In some scenario cases the generation of output data is not enough. Especially when testing the complete system, a replacement for a component may be required which interacts with the system as does the real component, using the same type of input and output data. Again, the idea is to provide simplified functionality or data by using lookup tables and interpolation or using historical data.

SYSTEM ARCHITECTURE AND IT-TECHNOLOGIES

The domain of environmental monitoring and civil protection applications consists of heterogeneous databases, model descriptions and applications, in addition to distributed data exchange due to the involvement of different organizations.

The system architecture of the scenario service must fulfill the following requirements:

- web-based communication independent of user platform
- the scripter should only interact with the scripting language to develop and integrate scripts, not with the underlying platform
- the script language should offer support for geographic information system (GIS) functionalities
- the operator should not need knowledge of script development, only of script selection and execution

Figure 2 shows our appropriate system architecture. We make use of an application server (JBoss, http://www.jboss.org) using session beans to handle the business logic, and entity beans for container-managed persistence of the data. The scenario service is based on Java (http://java.sun.com) using XDoclet tags for the generation of classes.

Communication with the scenario service is achieved using web service techniques. A web service description language (http://www.w3.org/TR/wsdl) has been defined describing the interface. The advantage of using web services is the independence of the underlying communication technologies. Communication partners only have a loose coupling, focusing more on a service-oriented rather than an operation-oriented architecture.

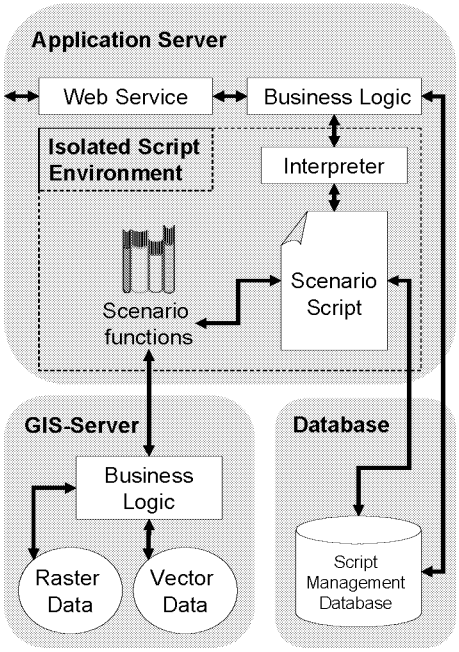


Figure 2: System Architecture

Clients can access the service using remote procedure calls based on SOAP (http://www.w3.org/TR/soap) and XML (http://www.w3.org/XML). The messages are deserialized and processed by the business logic of the script manager.

We use the language Python (http://www.python.org) with its basic elements (variables, loops, classes) for the specification of scenarios. The Python implementation Jython (http://www.jython.org) integrates seamlessly into the Java application. With this approach, scripters can describe their scenarios in the Python language and still interact with the underlying Java platform.

The Python scripting functions have been extended for scenario specific tasks. A library of scenario supporting functions can be used by the scripter for scenario development. When the script contains a function call to such a function, the call is forwarded into the Java domain to be handled. The result is transferred back into the Python environment. This way, a complex function can be implemented in the application server language while the scripter only interacts with the scripting language. Table 1 illustrates the call stack for the scenario.

One common requirement for script language is the support for GIS functionalities. Many applications for civil protection are based on GIS data. When developing a scenario service, access to GIS data may be important for the generation of data. Such functions could be:

- reading/writing values from raster cells
- finding vector objects in a selected area
- reading/writing attributes from vector objects
- combination of layers (union, intersection, …)

Such GIS-functions are typically handled by servers containing GIS engines. Currently we are examining different open-source and commercial tools for how they integrate into the system.

Table 1: Call Stack

Runtime Environment	Action
GIS	provide GIS data
Java Virtual Machine	↑ make GIS Request ↑ ↑ execute function ↑
Python Interpreter	↑ call scenario function ↑ ↑ run script function ↑
Java Virtual Machine	↑ forward message ↑
Client	↑ call web service ↑

SCENARIO SUPPORT FOR THE FLOOD WATER DOMAIN

Flood water is a common hazard in many areas close to rivers. One task of a catastrophe management system for flood water is the constant observation of the water level. The water level is usually measured using tide gauge sensors that have been placed along the river. If the water level exceeds a certain limit, predefined alert phases are activated leading to higher attention and safety actions.

In order to have the best preparation for such hazards, different scenarios which are most likely to occur are defined. For example, a complex flooding model is used to simulate the breach in a dike and to calculate the flood plain after a certain time period. One of the input parameters for the simulator is the hydrograph curve of the water level at the breach point for the next hours. Since it is only a scenario, the current sensor data cannot be used as the hydrograph curve.

One of the purposes of the scenario service is to provide the involved experts a fast way to generate input data. In this case, a hydrograph curve needs to be generated. This requires the following steps:

- analyzing common hydrograph curves for high water situations and finding characteristic trends
- defining generation parameters for hydrograph generation functions
- writing the script
- adding the script and script description into the database, making it available as a service for a catastrophe management system

While the spring high water of the Elbe River in Germany in the year 2006 was mainly caused by a quick melting of snow, in the year 2002 intense rain was the main cause.

The cause strongly influences the trend of the hydrograph curve. The trend mainly depends on:

- the causes of high water
- the local geologic realities
- tributary streams with high water
- breaches in the dike

Figure 3 shows the curve for the city Dresden in spring 2006. The characteristic trend of the curve is a combination of bumps. A script generating such curves would be helpful for the experts in order to quickly feed the simulator with different scenarios.

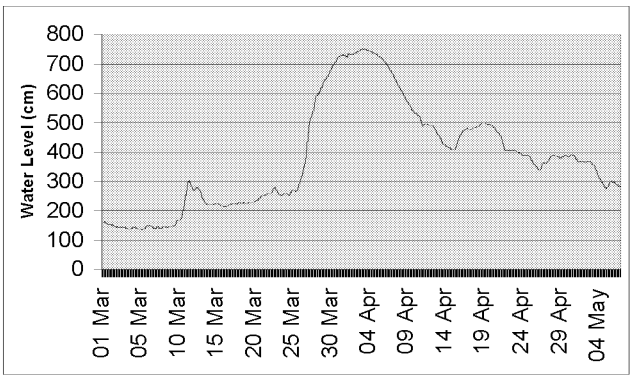


Figure 3: Hydrograph Curve Dresden 2006

As an example for generation of fictitious data we describe a bump with a polynomial interpolation based on three second-order polynomials defined by seven nodes as shown in Figure 4.

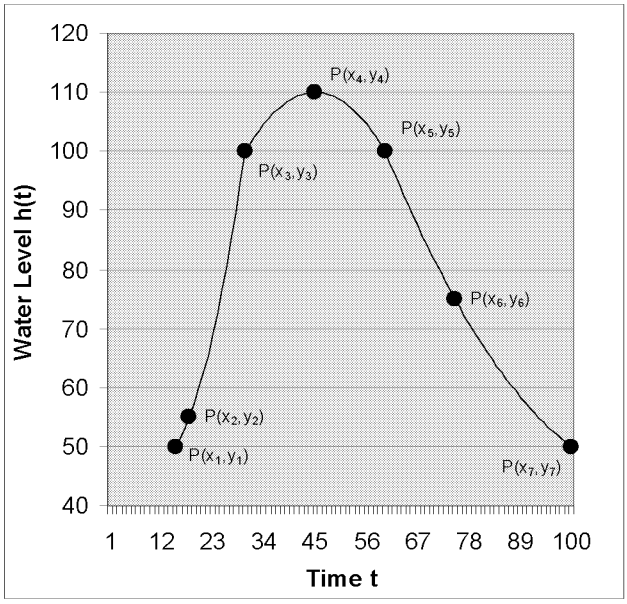


Figure 4: Generation of Tide Gauge Data with Polynomial Interpolation

The interpolation function is implemented in a Python script with the following definition:

def genIDataFloodSim(scriptID, parameter).

For the example above, the call parameters would be:

res=genIDataFloodSim(253,'15 50 18 55 30 100 45 110 60 100 75 75 100 50').

The parameters are defined as a single string using a space separated list in order to support variable parameters in the Python context. The interpolation values are calculated using the following formula:

$$h(t) = \begin{cases} f(t)_1 & x_1 \le t < x_3 \\ f(t)_3 & x_3 \le t < x_5 \\ f(t)_5 & x_5 \le t \le x_7 \end{cases} \qquad (1)$$

$$\text{with} \qquad f(t)_i = a_i \cdot t^2 + b_i \cdot t + c_i$$

$$b_i = \frac{m \cdot (y_{i+2} - y_i) - n \cdot (y_{i+1} - y_i)}{m \cdot (x_{i+2} - x_i) - n \cdot (x_{i+i} - x_i)}$$

$$m = (x_{i+1}^2 - x_i^2), \qquad n = (x_{i+2}^2 - x_i^2)$$

$$a_i = \frac{y_{i+1} - y_i - b_i \cdot (x_{i+1} - x_i)}{x_{i+1}^2 - x_i^2}$$

$$c_i = y_i - a_i \cdot x_i^2 - b_i \cdot x_i$$

APPLICATION

In the previous chapter a script function was described which supports the generation of tide gauge data for the flood water domain. Further script functions related to services for

- dike crashes
- flooding simulation
- identification of risk zones

are in development. The interlinking of such scenario descriptions enables a holistic processing of environmental and civil protection tasks.

The scripts are offered to the catastrophe management system by a scenario management application as a web service. The application allows the browsing, selection and execution of available scripts.

A graphical user interface is not mandatory for a script. The execution can occur by providing generation parameters. Also, user interfaces can only be implemented on the client side due to the usage of web services. A visualization using GIS maps can assist the operator with scripts that require local coordinates. An example for such a user interface is shown in Figure 5.

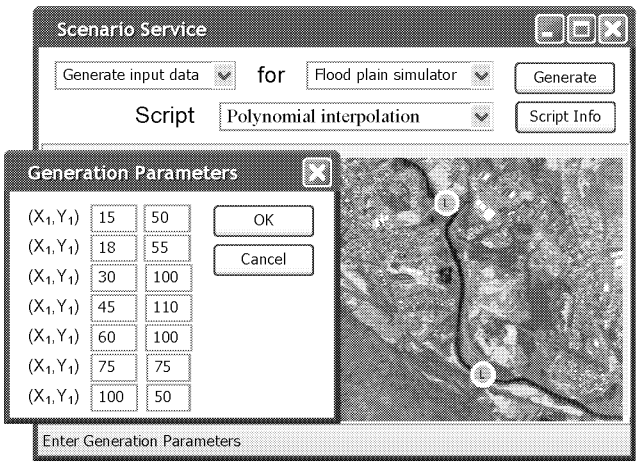


Figure 5: Graphical User Interface for Generation of Tide Gauge Data. GIS-Data by (MassGIS 2007)

CONCLUSION

The presented approach provides an efficient method for scenario modeling in environmental monitoring and civil protection applications. By distinguishing between the system and scripting context, scripters can develop and integrate new scenarios without interfering with the system environment. This leads to faster scenario development compared to catastrophe management systems without an isolated scripting environment. The architecture allows a continuous process chain from scenario modeling to integration and interpretation. The system can be used for a model-based verification of individual control system services as well. For future prospects the following additional features are considered:

- Integration of additional cases of disaster and expert components
- Automated scenario control and access to data traffic
- Extension to a training system for operators

For every local area, different hazards are of great importance for civil protection. Currently, we are considering the following domains for the area of Germany:

- Forest fires
- Flood water
- Water pollution
- Air pollution

ACKNOWLEDGEMENT

The presented architecture was developed within the framework of the R&D project "Grundlagen für ein integriertes Monitoring- und Leitsystem zum Natur- und Umweltkatastrophenschutz (Umweltleitstand)". The aim of the project is to provide solutions for catastrophe management systems that allow a global view on combined catastrophes. The project is supported by public state funds and the European Fund for Regional Development (EFRE) within the scope of the "Programm zur Förderung von Forschung, Innovation und Technologien – ProFIT –" by the Senate of Berlin.

REFERENCES

Alcamo, Joseph. 2001. "Scenarios as Tools for International Environmental Assessments". Environmental Issue Report No. 24, Experts' Corner Report: Prospects and Scenarios No. 5, European Environment Agency. Copenhagen: EEA.

IPCC 2007. "Climate Change 2007: The Physical Science Basis". Summary for Policymaker. Intergovernmental Panel on Climate Change (IPCC). Fourth Assessment Report, February 5th, 2007.

Lisounkin, A. 2003. "Semantic characterization of data series with application to facility control". In *Proceedings of the IASTED International Conference on Signal Processing, Pattern Recognition, and Applications*. June 30 – July 2, 2003, Rhodes, Greece, pp. 113-118.

Lisounkin, A. and Schreck, G. 2002. "Water Consumption Profile Analysis for Facility Control and Operator Training". In *Proceedings of the IASTED International Conference APPLIED SIMULATION AND MODELING* (ASM 2002). June 25-28, 2002, Crete, Greece, Conference Proceedings p. 234 – 239.

Lisounkin, A.; Schreck, G.; Schmidt, H.-W. 2004. "Knowledge Acquisition from Historical Data for Case Oriented Supervisory Control". In *Emerging Solutions for Future Manufacturing Systems*. Edited by Luis M. Camarinha-Matos, Springer, 499 – 506.

Lisounkin, A.; Sabov, A.; Schreck, G. 2004. "On web-based Architectures for Simulation supported Training in Technical Networks". In *Proceedings of the 7th IFAC Symposium on Cost Oriented Automation*. Gattineau/Ottawa, Canada, 24.-26.6.2004

MassGIS 2007. Office of Geographic and Environmental Information (MassGIS), Commonwealth of Massachusetts Executive Office of Environmental Affairs, 2007

Schreck, G. 2002. "Simulation Services for Training of Plant operators". In: *Knowledge and Technology Integration in Production and Services. Balancing Knowledge and Technology in Product and Life Cycle* (V. Marik et al. (Ed)) IFIP, Kluwer Academic Publishers, p. 79-86.

Wyatt, B.K.; Briggs, D.J.; Ryder, P. 2004. "Building a European information capacity for environment and security". European Commission, DG Research, Studies EUR 21109.

Yagüe, A.; Alarcón, P.; Garbajosa, J.; Lisounkin, A.; Schreck, G. 2003. "System with heterogeneous components: Modeling and simulation using a network-based approach". In *Proceedings of the 16th International Conference Software & Systems Engineering and their Applications* (ICSSEA 2003). Paris, France, December 2-4, 2003.

Yagüe, A.; Alarcón, P.; Garbajosa, J.; Lisounkin, A.; Schreck, G. 2004. "Web-based Training System for Forest Fire Office Staff". In *Proceedings of the 6th International Conference on Enterprise Information Systems* (ICEIS 2004). Porto, Portugal, April 14-17, 2004.

Safety/Security Management Model supporting Hazard Identification and Risk Assessment in the Air Traffic Services

Tomas Feglar
KPMG Czech Republic, Ltd.
Pobrezni 1a
186 00 Prague 8
Czech Republic
E-mail: tfeglar@kpmg.cz

KEYWORDS

Aerospace, Risk analysis, Model analysis, System enguneering, Decision

ABSTRACT

Air traffic density in Europe dramatically increased within last decade and air traffic services safety became more sophisticated and complex. Complexity increasing trend is multiplied by automation of subsystems that are involved into air traffic management. Automation uses primarily commercial information and communication technology that is a subject of security relevant initiatives. Safety risk assessment and security risk assessment are influenced by different authorities and currently is difficult to integrate both types of risk assessments. It seriously harmful an efficiency of air traffic management systems and can result gaps in risk mitigation strategy. This article analyzes current situation and introduces a framework allowing harmonization of safety / security risk assessment.

INTRODUCTION

Smooth operation of air traffic system requires a consistent, high level safety in air traffic services (ATS) allowing optimum use of Europe's airspace (Regulation 2004). Systems for air traffic management (ATM) must fulfill a comprehensive set of general safety requirements (Safety) that are further précised by the Eurocontrol Safety Regulatory Requirements (Safety a) particularly ESARR 3 (ATM safety) and ESARR 4 (ATM risk assessment) (Eurocontrol 2000, ESARR 2001).
Last decade all ATM critical subsystems became tightly dependent on information and communication services (ICS). ICS are delivered using commercial of the shelf information and communication technology (ICT). ICT security has relatively long history and the most significant standards applied to these technologies are ISO/IEC 17799 and ISO/IEC 27001 (BC ISI/IEC 2005, BC ISO/IEC 2005). Effective security establishment in accordance with ISO standards needs also risk assessment. Two different approaches to the risk assessments result new risks that cannot be fully covered by new safety regulatory requirement like ESARR 6 (Software in ATM system) (ESARR 2003).

Regulatory requirements are very important but their implementation especially in complex systems like ATM requires sound system engineering approach in accordance with ISO/IEC 15288 (ISO/IEC). This standard defines Risk Management as a process that exists independently on other processes, has its purpose, outcomes and activities. Framework for the Application of Systems Engineering in the Commercial Aircraft Domain follows ISO/IEC 15288 ideas establishing Commercial Aircraft Systems Architecture and splitting Risk Management process into planning phase (Risk Mitigation Plan) and engineering phase (Risk Management) (Commercial 2000). Keeping this in mind we established a framework allowing us harmonization of both approaches (safety motivated risks and information security motivated risk) into one common risk management process.

FRAMEWORK FOR UNDERSTANDING OF CURRENTLY APPLIED RISK ASSESSMENT FOR SAFETY AND SECURITY

Safety Regulation Group published very important guidance on a conduct of hazard identification and risk assessment and a production of safety cases (Safety 2006). Safety cases are very important because they allow air traffic service providers risks' perception tightly coupled with operational scenarios. Each safety case has its life cycle that consists of 7 phases:
Phase 1: Describe the system and its operational requirement
Phase 2: Identify hazards and consequences
Phase 3: Classify the severity of the consequences
Phase 4: Classify the likelihood of the consequences manifested
Phase 5: Evaluate risks (Risk Classification / Tolerability Matrix)
Phase 6: Identify risk mitigation measures (Safety Requirements)
Phase 7: Develop claims, arguments and evidence that the Safety requirements have been met.

Phase 5 produces key decision point for safety protection strategy that distinguished three regions (Table 1):

- Acceptable – the consequence is so unlikely or not severe enough to be of concern.
- Review – the consequence and / or likelihood is of concern; measures to mitigate the risk to ALARP (As Low As Reasonably Practical) should be sought.
- Unacceptable – the likelihood and / or severity of the consequence is intolerable. Major mitigation or redesign of the system may be necessary to reduce the likelihood of severity.

	Probability of Occurance (Likelihood)				
	Extremely improbable	Extremely remote	Remote	Reasonably probable	Frequent
ESSAR Severity	$<10^{-9}$ per hour	10^{-7} to 10^{-9} per hour	10^{-5} to 10^{-7} per hour	10^{-3} to 10^{-5} per hour	1 to 10^{-3} per hour
Accidents	Review	Unacceptable	Unacceptable	Unacceptable	Unacceptable
Major Incidents	Acceptable	Review	Unacceptable	Unacceptable	Unacceptable
Serious Incidents	Acceptable	Acceptable	Review	Unacceptable	Unacceptable
Significant Incidents	Acceptable	Acceptable	Acceptable	Review	Unacceptable
No Effect Immediatelly	Acceptable	Acceptable	Acceptable	Acceptable	Review

Table: Risk Classification / Tolerability Matrix

Risk management decision made on the base of the matrix in the table 1 is associated with particular safety case and applies the Safety Management Systems (SMS) established in accordance with commission regulation EC No. 2096/2005 (Eurocontrol 2000).

Risk analysis and management methods for ICT systems were developed intensively last decade. Various methods are available and some of them are supported by tools - Mehari, CORAS, CRAMM (Mehari, CORAS, CRAMM). Risk management process automation is reasonable because information (and communication) systems include huge amount of components that have to be considered during risk analysis phase.

ICT risk assessment process consists of five phases:
Phase 1: IT assets identification and impact assessment
Phase 2: Threat / Vulnerability Analysis and risk calculation (IT Risk Matrix).
Phase 3: Risk driven countermeasures recommendation
Phase 4: Check ISO/IEC 17799 compliance
Phase 5: Setting countermeasures for implementation

Phase 2 produces key decision point for IT security strategy that distinguished 7 values (Table 2). Tools like CRAMM offer decision makers a set of pre-prepared countermeasures linked with risk values on the base of rules that can be modified by ICT risk manager.

Threat	VL	VL	VL	L	L	L	M	M	M	H	H	H	VH	VH	VH
Weakness	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
Impact															
1	1	1	1	1	1	1	1	1	2	1	2	2	2	2	3
2	1	1	2	1	2	2	2	2	3	2	3	3	3	3	4
3	1	2	2	2	2	3	2	3	3	3	3	4	3	4	4
4	2	2	3	2	3	3	3	3	4	3	4	4	4	4	5
5	2	3	3	3	3	4	3	4	4	4	4	5	4	5	5
6	3	3	4	3	4	4	4	4	5	4	5	5	5	5	6
7	3	4	4	4	4	5	4	5	5	5	5	6	5	6	6
8	4	4	5	4	5	5	5	5	6	5	6	6	6	6	7
9	4	5	5	5	5	6	5	6	6	6	6	7	7	7	7
10	5	5	6	5	6	6	6	6	6	6	7	7	7	7	7

Threat Values: VL - Very Low L - Low M - Medium H - High VH - Very High
Weaknesses Values: L - Low, M - Mdium, H - High

Table : IT Risk Matrix

ICT risk management establishes Information Security Management System (ISMS) in accordance with ISO/IEC 27001.

It is reasonable to assume that smart ATM management must harmonize SMS and ISMS into one consistent

safety/security system that allows overcoming communication difficulties and possible different emergency reactions of ATS/SMS operators and ICT/ISMS operators.

FRAMEWORK FOR BETER COMMUNICATION BETWEEN ATS/SMS OPERATORS AND ICT/ISMS OPERATORS

ATS/SMS Operators have direct responsibility for air traffic management in accordance with safety requirements addressing all systems that are used for these services.
ICT/ISMS Operators have direct responsibility for information and communication management in accordance with security requirements that relate to ICT systems that are critical for ATM systems.
Figure 1 shows two different air traffic management systems perceptions. ATS Operators operate with objects and entities that interact in accordance with particular scenario – aircraft starts or lends, crowds communicate by radio with dispatchers, and so on. Their system perception is operational and risks for them are associated with a scenario not with the system.
ICT Operators operate with ICT systems that control directly RFC systems in particular nodes (Node 2, 3). Appropriate radio frequency setting requires full functionality of all ICT components starting with data (RFC Tables), RFC application, server up to various distribution components (routers, hubs, etc.). ICT Operators' perception is systemic and technical and risks for them are associated with servers, applications, operating systems, data (like Radio Frequency Tables) and network distribution components.
Different risk assessment approaches described in previous paragraph result two different risk management perceptions. First perception is institutionalized by SMS, the second one by ISMS.

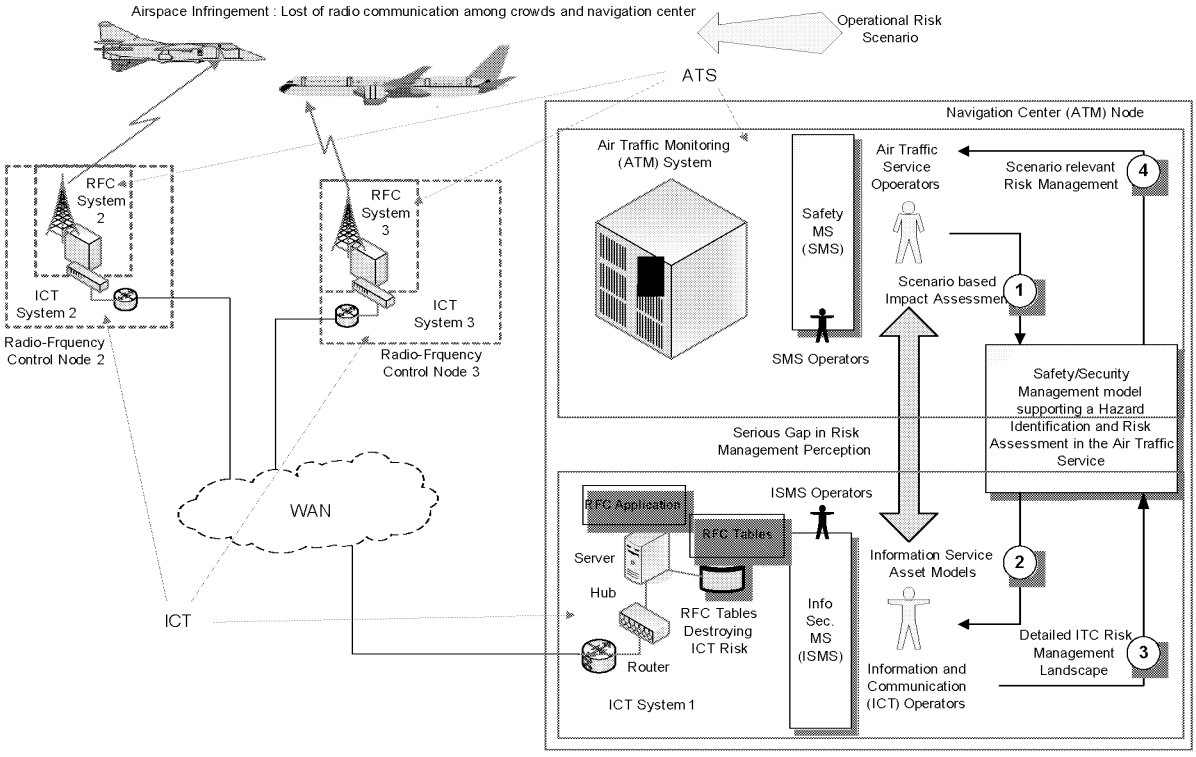


Figure 1: Framework for Understanding of ATS/SMS and ICT/ISMS Operators

The most critical items influencing the gap between perceptions can be summarized as follows:

1. Operational perception captured by scenario is associated with the risks (like Airspace Infringement: in the Figure 1) that are calculated using the Risk Classification / Tolerability matrix.
2. Decisions made on the base of this matrix does not consider risks that relate to ICT
3. ICT risks (like Radio Frequency Tables destroying in the Figure 1) calculation uses IT Risk matrix that is different from the Risk Classification / Tolerability matrix
4. Reasoning that Radio Frequency Table destroying can start Airspace Infringement impacted by the lost of radio communication among crowds and navigation center becomes hidden.
5. SMS designed on the base of the Risk Classification / Tolerability matrix does not consider ICT risks
6. ISMS designed on the base of IT Risk matrix is difficult to optimize because there is no clear linkage to the operational scenario.

CONCLUSION

Topics discussed in this article significantly simplify effort spending during risk assessment of ATM systems. Currently applied safety risk assessment is different from security risk assessment. Both approaches are too different to be tightly synchronized, but reasonable harmonization is possible. This harmonization significantly increases efficiency of each risk assessment life cycle and allows overcoming the gap in operational and systemic risk perceptions.

REFERENCES

BS ISI/IEC 17799:2005: "Information technology – Security techniques – Code of practice for information security management"

BS ISO/IEC 27001:2005: "Information technology – Security techniques – Information security management systems – Requirements".

Commercial Aircraft Domain Working Group: "Framework for the Application of Systems Engineering in the Commercial Aircrtaft Domain"; Draft, Version 1.2a, July 28, 2000.

Commission Regulation (EC) 2096/2005 of 20 December 2005 laying down common requirements for the provision of air navigation services, Official Journal of the European Union, December, 2005

CORAS (http://www2.nr.no/coras/)

CRAMM (http://www.cramm.com)

ESARR: "ESARR 4: Risk Assessment and Mitigation in ATM", 05-04-2001

ESARR: ESARR 6: "Software in ATM systems", April 2003.

Eurocontrol Safety regulatory requirement (ESARR): "ESARR 3: Use of Safety Management Systems by ATM Service Providers", 17-07-2000.

ISO/IEC 15288: "Life Cycle Management – System Life Cycle Process".

Mehari (http://www.clusif.asso.fr/en/clusif/present/)

Regulation (EC) No 549/2004 of the European Parliament and of the Council of 10 March 2004, Official Journal of the European Union, March, 2004.

Safety Regulation Group: "CAP 670 Air Traffic Services Safety Requirements", www.caa.co.uk.

Safety Regulation Group: "CAP 760. Guidance on the Contact of Hazard Identification, Risk Assessment and the Production of Safety Cases", January 2006

TEXTILE SIMULATION

MODELLING OF TEXTILE STRUCTURES

INTER FIBER FRICTIONAL MODEL IN THE COTTON SLIVER

Shahram NOWROUZIEH,
Artan SINOIMERI,
Jean-Yves DREAN
Laboratoire de Physique et Mécanique Textiles,
UMR 7189 CNRS-UHA, ENSISA, 11 rue Alfred Werner, F-68093 Mulhouse Cedex France
E-Mail: shahram.nowrouzieh@uha.fr

Richard FRYDRYCH,
Jean-Paul GOURLOT
CIRAD-CA, TA 70/16
F-34598 Montpellier, France

KEY WORD
Fiber, Friction, statistic model.

ABSTRACT

One of the effective forces in textile process is friction. A new and simply device is developed in order to quantify the friction forces during a quasi-static fibre slippage in a sliver. The device is composed of two identical clamps. One of them is fixed, whereas the second one is moving through a linear guide. A piece of sliver maintained under given pressure in zero gage position with two clamps. The dissociating force of sliver is measured in different speed and normal given load. A statistic model based on these results is proposed. The analysis of variance of data indicates that the effect of speed on maximum frictional force is not significant whereas the effect of normal load on maximum frictional force is significant.

INTRODUCTION

Predicting of yarn tenacity from fiber properties is one of the most important parts of textile research. Mechanistic and statistic models have primarily been used to tackle the problem.
One the first approach is the model (El Mogahzy and Brughton 1989). They carried out multicollinearity between HVI fiber properties and skein break factor (SBF) of yarn. According to the results of the correlation matrix they found high coefficients of correlation between reflectance and elongation, reflectance and finesse, and elongation and finesse. Therefore, the model that best represents the given data and has less severe collinearity is model of the form:

$$SBF = -64.6 + 54.9LU + 35.3FS + 41.9Rd \quad (1)$$

Where LU: Length Uniformity, FS: Fiber Strength and Rd: Reflectance. These properties chose from eight properties which are measured by HVI. The model used by Frydrych (Frydrych 1992) requires the knowledge of the value of twelve variables which have to be determined experimentally. Finally a simple linear regression equation is carried out to predicting yarn tenacity as a function of fiber finesse, yarn twist, fiber breaking stress and strain, and fiber mean length. In these models, the effect of inter fiber frictional force is not taken into account. Nevertheless, the friction forces seem to be very important and affect the yarn spinnability and tenacity. In the present work, a new and simply device is developed in order to quantify the friction forces during a quasi-static fiber slippage in a sliver and then a statistic model is proposed.

METHODS AND MATERIAL

The base of this device is the measurement of the necessary force required to dissociate a sliver under pressure. The device is composed of two identical clamps. One of them is fixed, whereas the second one is moving through a linear guide as shown in Figure 1. A piece of sliver is put down in two clamps in the zero-gage position. The sliver is compressed with two identical loads. The moving one is pulled by the cross head of the traction machine in programmable constant speeds. The inter fiber force and the displacement are measured by force and displacement sensor respectively. In order to avoid an uncontrolled slippage between the sliver and the carriage walls, the channel walls are covered with a thin cotton layer.

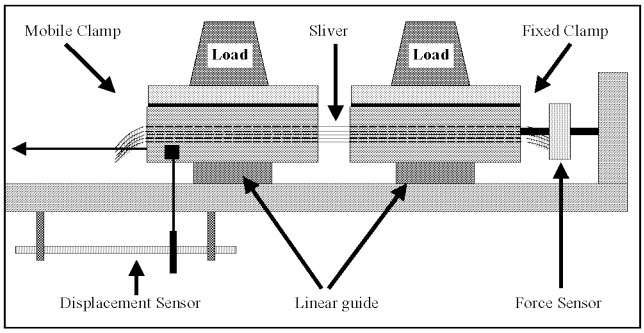


Figure 1: The Device of Measurement the Inter Fiber Friction Force.

One variety of cotton is selected and fiber properties are measured by AFIS, as shown in Table 1.
Where L(w): Fiber length by weight , L(n): Fiber length by number, UQL: Upper Quartile Length, SFC: Short fiber content, IFC: Immature fiber content.

Table 1: The Fiber Properties Measured by AFIS

Nep Cnt/g	L(w) (mm)	SFC(w) %<12.5	UQL(w) (mm)	L(n) (mm)
144	25.67	5.8	31.03	21.19
SFC(n) (mm)	5.0 % (mm)	Fine (mtex)	IFC (%)	Mat. ratio
18.9	35.36	186	5.4	0.95

From this selected cotton, three different count slivers have been prepared and tested with five replications following the conditions described in Table 2.

Table 2: Experimental Table

Sliver count (ktex)	2.65	4.20	5.53
Speed (mm/min)	20	260	500
Load (g)	800 ; 900 ; 1000 ; 1687.6 ; 2187.6 ; 2687.6 ; 3187.6 ; 3687.6		
Five replications for each factorial combination			

RESULTS AND DISCUSSION

A typical shape of the force-displacement curve is shown in Figure 2. This diagram can be divided to four segments. In the segment OA the fibers start to straighten and the curve is approximately a straight line. The point A represents a threshold. After this point the draft forces overcome the inter fiber cohesion and friction forces and the fibers sliding starts. During the slipping period, fibers are pulled out of the clamps. The length of fiber hold in the clamp prescribes the direction of the fiber movement. The fibers which have more than the half of their lengths within the mobile clamp move with it, whereas the rest stays clamped in the fixed clamp. So the fibers will be untaken as a function of the fiber length distribution.

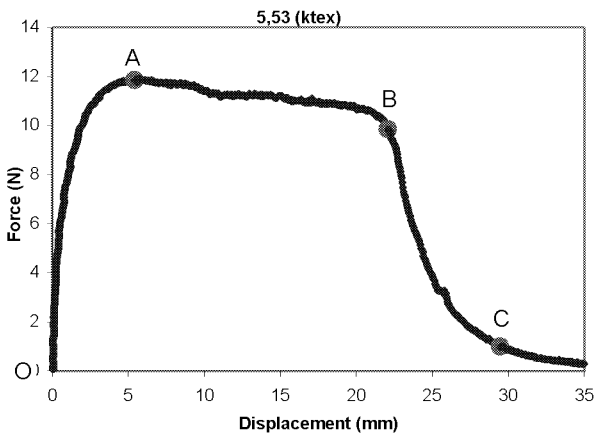


Figure 2: The Force-Displacement Curve for 5.53 ktex under a 1500 g Load

Therefore the friction force decreases down to B point, where there is no more fibers hold by the two clamps. Then the inter fiber force decreases sharply, which is represented by the segment BC of the curve and finally the sliver is broken into two parts.

The optimum point for modeling the friction phenomenon is the point A, where a maximal friction is developed. In this point all the fibers in the sliver cross-section are straighten and hold by the two clamps. With the assumption of a non linear effect of the sliver count and the load applied on the maximum frictional force (MFF), this force may be written as follows:

$$MFF = K \cdot T^{j} \cdot W^{n} \qquad (2)$$

Where K, j and n are constant, T is the sliver count, W is the load.

The normality distribution of data test (Shapiro-Wilkinson, W test) shows that the residual distribution of model is normal with W=0.963 and probability =0.012.

However the MFF increases with speed but it is not significant as indicate in table 3.

Table 3: The Effect of Speed in MFF

Sliver count	Level	Number	Mean	Std Error
2.65 (ktex)	20	40	14.1877 ^{a}	1.0574
	260	40	12.2364 ^{a}	1.0574
	500	40	13.0242 ^{a}	1.0574
4.20 (ktex)	20	40	14.5789 ^{a}	1.0125
	260	40	14.3269 ^{a}	1.0125
	500	40	14.8437 ^{a}	1.0125
5.53 (ktex)	20	40	14.8911 ^{a}	1.0614
	260	40	15.3498 ^{a}	1.0614
	500	40	15.7501 ^{a}	1.0614
The letter *a* at the mean values signify that there is no statistical difference between these results.				

The results of the Table 3 indicate that the model described by the equation (2) may be applied in the speed interval ranging from 20 up to 500 mm/min independently on the speed value.

In order to use a simple linear regression, the equation (2) is converted in a logarithmic form. The simple polynomial obtained is fitted to experimental data as follow:

$$Ln(MFF) = Ln(K) + j \cdot Ln(T) + n \cdot Ln(W) \qquad (3)$$

The estimated model values are assembled in Table 4.

Table 4: The Estimated Value of *ln(k)*, *n* and *j*

Term	Estimate	Std. Error	t Ratio	Prob>|t|	R^2
Ln(k)	-0.1180	0.039	-3.02	0.0030	0.985
n	0.8858	0.009	94.35	0.0001	
j	0.1245	0.019	6.38	0.0001	

As shown in Table 4, the R^2 is very satisfactory and the *Ln(k)*, *n* and *j* are highly statistically significant. These values belong to speed of 20 mm/min.
So, following these results, the best fitting equation for predicting the maximum frictional force (MFF) for the actual cotton is:

$$MFF = 0.888 \times T^{0.1245} \times W^{0.8858} \qquad (4)$$

Where *MMF* and *W* are in Newton and *T* is in ktex. The values of *n* and *j* reconfirm the non linear effect of the sliver count and the load.
The analysis of variance of MFF as shown in Table 5 indicates that the effect of load on the MFF is highly significant.

Table 5: The Analysis of Variance of MFF in Function of Load for Speed of 20 mm/min

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob>F
Model	7	5351.47	764.49	1706.27	<0.0001
Error	112	50.18	0.45	R square: 0.99	
C Total	119	5401.65	45.39		

The means comparisons for each pair of load by using Student's t and Tukey-Kramer test, Figure 3, shows that the only pair which has not significantly effect on the MFF in speed of 20 mm/min is 800-900 g.

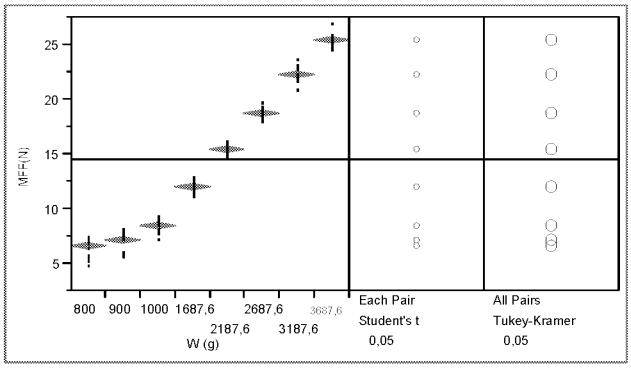


Figure 3: The Effect of Load (w) on the MFF for Speed of 20 mm/min

For testing the model, two pieces of sliver 2.65 ktex are placed together in the channel of two clamps. In fact with doubling the sliver with count of 5.30 ktex from the same cotton is prepared and it is tested under load interval ranging from 1687.6 up to 3687.6 g by step of 500 g. The traction speed is 20 mm/min. The comparison mean of MFF with t-student test shows that there is not different significant between 5.50 (Double) and 5.53 ktex. So we could compare the model MFF double with the experimental results of 5.53 ktex. The curve of mean model MFF, experimental MFF of double and MFF of 5.53 ktex in function of the loads shows in Figure 4. The analysis of variance of experimental and model MFF double indicate the effect of load on the MFF is highly significant as we mentioned before.

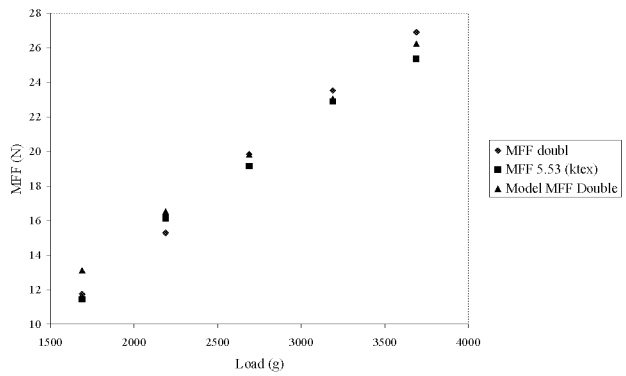


Figure 4: The Mean MFF-Load Curve for Model and Experimental MFF of Double and 5.53 ktex

The mean comparison of MFF (t-student 0.05) as shown in Table 6 indicates that not only the difference between the model MFF double and the experimental MFF double, but also the difference between model MFF double and the experimental MFF 5.53 ktex are not significant for all the loads.

Table 6: The Mean Comparison of the Model MFF Double with the Mean of Experimental MFF Double and 5.53 ktex

Load (g)	1687.6	2187.6	2687.6	3187.6	3687.6
Model MFF double	13.12 *a*	16.51 *a*	19.81 *a*	23.04 *a*	26.22 *a*
Experimental MFF Double	11 .75 *a*	15.27 *a*	19.84 *a*	23.53 *a*	26.90 *a*
Experimental MFF 5.53 ktex	12.39 *a*	15.63 *a*	18.92 *a*	21.88 *a*	26.04 *a*
The letter *a* at the mean values signify that there is no statistical difference between these results.					

These present results are in good accordance with those founded by previous works. Lord (Lord 1955) and El Mogahzy (El Mogahzy and Brughton 1993) have measured the friction force between two thin identical fiber fringes. The Figure 5 shows the Lord's device.

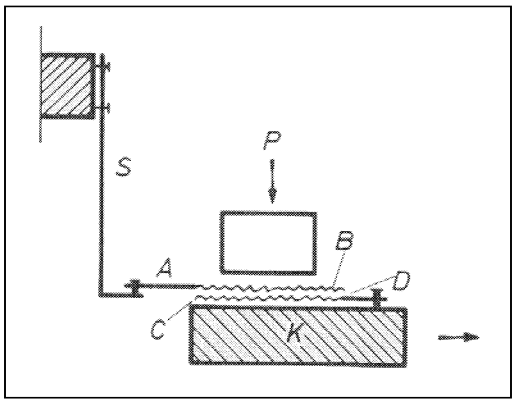


Figure 5: The Lord's Device

Two fringes AB and CD are placed one on top of the other on the movable platform K. Fringe CD and AB has fastened to the platform K and to a calibrated flat spring S respectively. On the top of the fringe AB a load P is placed and then platform K with the fringes and the load is driven at a constant speed by a suitable mechanism. The friction force between two fringes measures by deflecting of the spring S from its position. The calculated *n* for two different cottons with Lord's device is 0.886, the same which is derived in the present work.

A schematic side view of the friction El.Mogahzy's device is shown in Figure 6.

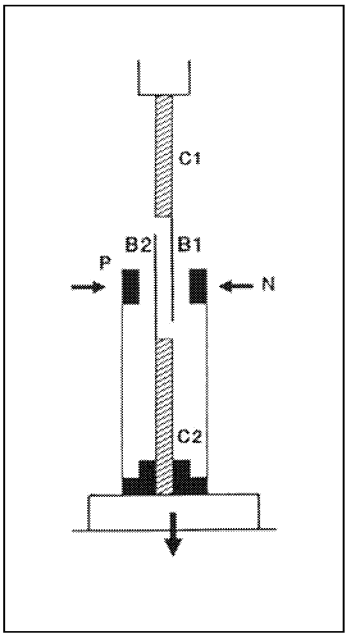


Figure 6: The El Mogahzy's Device

This device has two clamps. The bottom clamp (C2) in which the bottom fibre beard (B2) can be mounted. The top clamp (Cl) is attached to the load cell of the Instron tensile tester. Lateral pressure (P and N) is applied by using a pneumatic system. Both the pressure bar (P) and the bottom clamp (C2) are mounted on the Instron cross head and move with it. The frictional force between two fringes measures by load cell. El Mogahzy reported that the value of *n* for Pima and DPL cotton is between 0.74 and 0.91.

CONCLUSION

A new and simple device is designed in order to evaluate the inter-fiber frictional force. This device is tested for cotton fibers. The maximum frictional force (MFF), developed in different sliver counts, has been measured for different speeds and normal loads. Based on these data, a statistic model is carried out. The model indicates that the effect of the sliver count and the load on the maximum frictional force are non linear and positive. The speed effect on MFF is not significant whereas the Load effect is highly significant on MFF. The test of model indicates that the values of model MFF are close to experimental results and the difference between them is not significant. The power of load *n* is in good agreement with those obtained previously by Lord and El Mogahzy. We could not compare the value of sliver count power with their results because Lord and EL Mogahzy did not explain about the value of sliver count power and its effect on frictional force unfortunately. The perspective of this study is to check out differences between different cotton varieties and to relate these differences with the yarn characteristics and spinnability.

ACKNOWLEDGEMENTS

We are indebted to the French and the Iranian Governments for their financial supporting of the present work.

REFERENCE

El Mogahzy. Y, Broughton. M, 1989, Diagnostic Procedures for Multicollinearity Between HVI Cotton Fiber Properties, Text. Res. J; 59, No. 8, 440-447

El Mogahzy. Y, Broughton. M, 1993, A New Approch for Evaluting the frictional Behavior of Cotton Fibers: Part I: Fundamental Aspect and Mesuring Techniques, Text. Res. J; 63, No. 8, 465-475

Frydrych. I, 1992, A new Approach for predicting Strength Properties of Yarn, Text. Res. J; 62, No. 6, 340-348

Lord. E, 1955, Frictional Force Between Fringes of Fibers, J. Textile Inst. 46, P41-58

FINITE ELEMENT MODELLING OF THE WARP KNITTED STRUCTURE

Argyro Kallivretaki
School of Mechanical Engineering
National Technical University of Athens
Iroon Polytechniou 9, Zografos,
GR-15773 Athens, Greece
E-mail: arkal@mail.ntua.gr

Savvas Vassiliadis
Department of Electronics
TEI Piraeus
P. Ralli & Thivon 250, Egaleo,
GR-12244 Athens, Greece
Email: svas@teipir.gr

Mirela Blaga
Faculty of Textiles and Leather Eng.
Gh. Asachi Technical University
D-trie Mangeron 53,
Iasi, Romania
E-mail: mblaga@tex.tuiasi.ro

Christopher Provatidis
School of Mechanical Engineering
National Technical University of Athens
Iroon Polytechniou 9, Zografos,
GR-15773 Athens, Greece
E-mail: cprovat@central.ntua.gr

KEYWORDS
Computer Aided Engineering (CAE), Numerical Methods, Finite Element Method (FEM), Textile Fabrics

ABSTRACT

The finite element modelling of the Charmeuse warp knitted fabrics is the main subject of the current paper. The proposed model consists of the three-dimensional representation of the warp knitted fabric microstructure and the mechanical analysis of the defined unit cell. An iterative calculation procedure is used for the definition of the geometrical representation of the structure. The parametric modelling is based on the main structural parameters: yarn cross-section, warp density, course density, yarn consumption of the front and the back bar. The flattening of the yarn cross-section has been introduced in the undeformed state of the model for a better approach of the real conditions. The Finite Element Method (FEM) with contact analysis is implemented for the mechanical analysis of the multi-body structural unit. The appropriate contact algorithm is defined for the fast convergence during the solution. Although the complexity of the unit cell is high, the modelling is possible and becomes a tool for the mechanical analysis of the warp knitted fabrics.

INTRODUCTION

In the warp knitting technology the warp yarns are introduced in the longitudinal fabric direction. The characteristic feature of warp knitted structure is the formation of the adjacent loops in the course direction by separate warps. Usually the basic lapping structures are used in combination producing more than one warp sheets. The considered structure, Charmeuse or Locknit, is a combination of tricot and 2×1 plain stitches. Thus two guide bars are used for the manufacturing of the structure under consideration.

The warp knitted fabrics have only a certain range of applications in apparel fabrics and household textiles. However, their continuously spreading use in technical applications increases the research interest related to them. The technological advantages of warp knitted fabrics as formability, contourability, dynamic mechanical properties along with easy and rapid manufacturability are essential reasons for their establishment in composite manufacturing. Nowadays the warp knitted structures produced by high performance fibers made from glass or carbon are used for composite reinforcement. The increasing requirements of industry for high performance technical textiles impose the accurate definition of their properties. Thus the realistic modelling and mechanical analysis of warp knitted structures is of a great importance.

Several research activities have been registered concerning the micromechanical analysis of the basic warp knitted structures. The geometrical representation of the structure comprised obviously the starting point for the researchers. The representations were based mainly on the division of the loop configuration into discernible parts as loop's head and arms, root of the loop and underlapping section. Depending on the approach and complexity, two- and three-dimensional models have been proposed including or ignoring the loop inclination respectively. Convenient geometrical shapes as straight lines and circular parts and in some cases complicated geometries based on empirical assumptions were used for the representation of the defined partitions (Dalidovici, 1940) (Allison, 1958) (Grosberg, 1964) (Korlinski, 1981) (Raz, 1987) (Vékássy, 1960). A different approach proposed by Grosberg based on the elastica theory. According to his approach the loop configuration results from the forces applied to the yarn at the base of the loop (Grosberg, 1964). Some authors proposed algebraic equations for the calculation of the yarn loop length according to the proposed geometrical models. A relative accurate equation for the yarn

consumption has been proposed by Dalidovici for the two guide bars warp knitted structures (Dalidovici, 1940).

The applicative difficulties of the analytical methods in the textile modelling in conjunction to the introduced simplifications and the continuous development of the computer technology led to the numerical modelling and simulation. Particularly the geometrical complexity of the multiple guide bar warp knitted structures impedes their accurate representation using analytical geometrical equations. On the contrary the interpolation techniques for areal and spatial representations provided by the modern Computer Aided Design (CAD) software codes are special facilities for the realistic designing of complex structures. Moreover, the complicated deformation procedure of the structures under consideration precludes the analytical calculation of the mechanical performance. The simultaneous subjection of the constituent yarns in elongation, bending, compression and the presence of multiple contact areas during the deformation impose the use of numerical methods for the analysis. Finite Element Method is known and proved adequate for the prediction of the performance of complex multi-body structures subjected to deformation.

MATERIALS AND METHODS

Used Samples

Twelve samples of Charmeuse structure are examined for this research activity. They have been produced by a 12-filament polyamide flat yarn: PA6 44/12 dtex. The samples are of different structural parameters in order a generalized geometrical modelling to be achieved.

Three-Dimensional Geometrical Representation

Although the current approach focuses on the mechanical simulation of the Charmeuse structure the procedure used for the geometrical modelling is summarized in parallel. The used representation procedure was based on the microscopic observation of the samples, planar and cross-sectional, and the image processing of the received photos. The dimensional characteristics of the structure were correlated and the derived relations introduced the assumptions for the computational geometrical modelling.

Six main structural parameters are used: the wale spacing c (mm), the course spacing w (mm), the yarn loop length of the front B_1 (mm/row) and the back guide bar B_2 (mm/row) and two values of the yarn thickness along the fabric plane D_p (mm) and the fabric thickness D_t (mm). The values of B_1 and B_2 resulted from the measurement of the yarn consumption. The values of c, w, D_p and D_t resulted from the microscopic observation of the samples and the respective image processing. One secondary parameter is also defined, the distance s (mm).

The total loop formation is represented by two sections. The first section, defined as main loop, corresponds to the model shown in Fig. 1a. The second section, defined as connecting part, corresponds to the underlap and the inclined part of the loop's arms.

An iterative procedure is used for the generation of the geometrical model using the main structural parameters. The front and the back loop are firstly designed. The structural characteristics of the double loop are shown in Fig. 1a. Elliptical shape is supposed for the yarn cross-sections. The central axis of each loop is formed by a B-Spline defined by the 1,2,3,4,5 control points. The orientation of the starting and the ending point are defined by the vectors $[D_p/2, -c/2, 0]$ and $[-D_p/2, -c/2, 0]$ respectively. A value is assumed for the parameter s in the range [c/4, 3c/4].

The double loop is then drawn for 3 courses × 3 wales as shown in Fig. 1b. The length of the formed loops is calculated and subtracted from the total yarn length for the calculation of the yarn length corresponding to the connecting parts. The connecting parts are designed with B-Splines defined by four control points and the direction of the Splines at the control points. The Spline direction at the first and forth point results from the double loops. The Spline direction at the other two control points results from an iterative procedure based on the generation of Splines and the selection of the one with the appropriate orientation and length. The appropriate orientation supports the just in touch contact of the yarns precluding the penetration. A mean value of the yarn thickness is used for the calculations at this stage. The appropriate length corresponds to the calculated one. This procedure is used for the designing of the connecting parts of the front (Fig. 1c, Fig. 1d) and the back yarn (Fig. 1f, Fig. 1g). If there is no feasible design constrain for the current value of s the value s is changed and the procedure is repeated. The control points used for the definition of the Splines are shown with the numbers 1,2,3,4 in Fig 1 c,d,f,g.

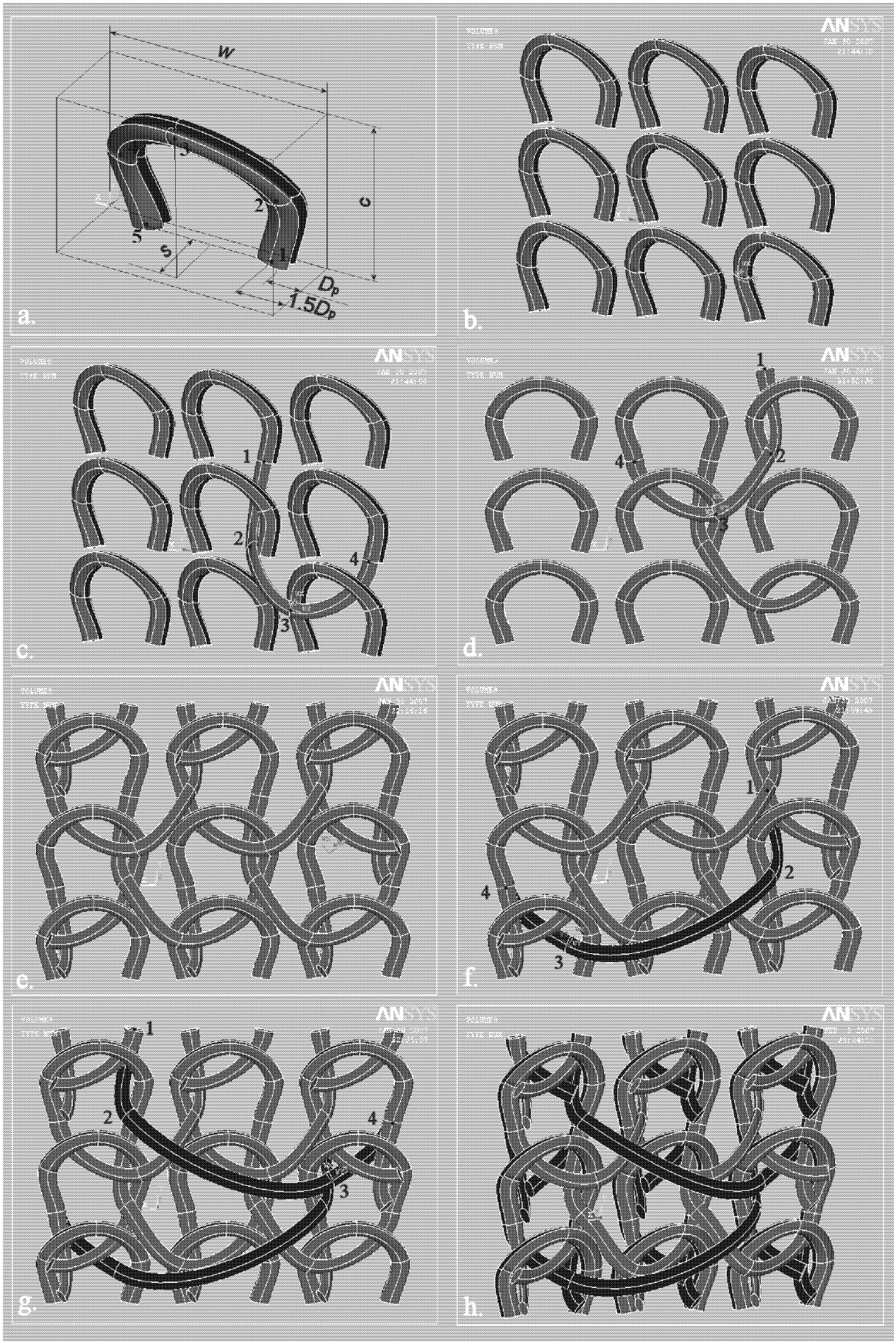


Figure 1. Generation of the 3D structural unit of Charmeuse.

Mechanical Simulation

The modelling and the mechanical simulation of the investigated structure are performed in ANSYS software. The unit cell used for the mechanical simulation consists of three courses and three wales (Fig. 1h). Some of the parts of minor importance are eliminated and their influence in the deformation procedure is foreseen applying the appropriate boundary conditions. Thus the unit cell is simplified, the number of the necessary finite elements is reduced, the computational complexity is lower and consequently the solution is faster.

The proposed approach is based on the generation of a mesh with approximately 13500 hexahedron finite elements (Fig. 2). Excluding the standard signification of the mesh pattern for the load transferring, it is also important for the definition of the contact and target elements. These elements are responsible for the load transfer between the contact areas of the multi body system, the friction forces application and the definition of the contact state (sticking or sliding). The contact and target elements are placed as an overlay on the solid elements in the areas defined as contact-target (contact pair). Thus the appropriate mesh of the bodies in these areas supports the convergence during solution. A mesh including coincident nodes generated in the contact-target area is practically approved as well-defined mesh.

Convergence difficulties are usually presented in the multiple contact models. Thus the selection of the contact algorithm is essential for the regular solution procedure. ANSYS FEM code provides several contact algorithms and specific options in order the convergence to be improved. For the current simulations the surface-to-surface contact is selected. The selected one is suitable for the contact analysis of equally deformable bodies (with the same elastic properties). It permits also large scale sliding between the bodies. Two contact algorithms are used: the Augmented Lagrangian and the Internal Multipoint Constraint (MPC) algorithm in conjunction with bonded contact. The second one is implemented for the contact occurs between the front and the back loop. The first one is implemented for all the others contact pairs.

The initial gap closing and initial penetration reduction of the defined contact pairs are useful options for the convergence success. The initial gap cause large displacements indicating unconstrained motion in the analysis. The initial penetration, on the other hand, incurs overestimation of the contact forces resulting in non-convergences or in breaking-away of the components in contact. The activation of the mentioned options succeeds the adjustment of just touching contact pairs.

The definition of loading and boundary conditions is a certainly demanding phase for the mechanical simulation. Especially the cases of multi-body models require the sufficient constraint of the bodies to preclude the undesirable infinite displacement of them. The undesirable over-constrain of the model, on the other hand, is a usual defect in contact models. The subjection of the proposed models in tensile deformation is achieved with the imposition of displacement on the nodes belonging in one area of the apparent unit cell and the calculation of the reaction forces presented. The nodes belonging in the across area of the apparent unit cell are constrained. Obviously, for the walewise and coursewise tensile test the mentioned areas correspond to the A1, A2 and A3, A4 of Fig. 2 respectively. The restriction U_x=0 is applied on the nodes belonging in A3, A4 and the restriction U_y=0 is applied on the nodes belonging in A1, A2 for the walewise and coursewise tensile test respectively.

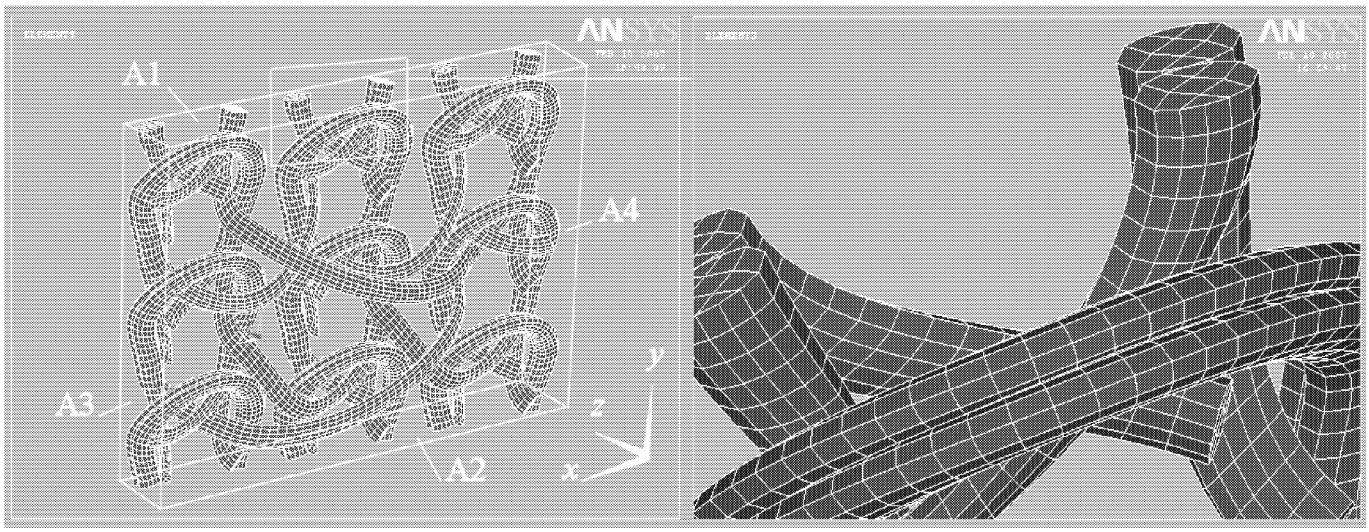


Figure 2. Meshed unit cell

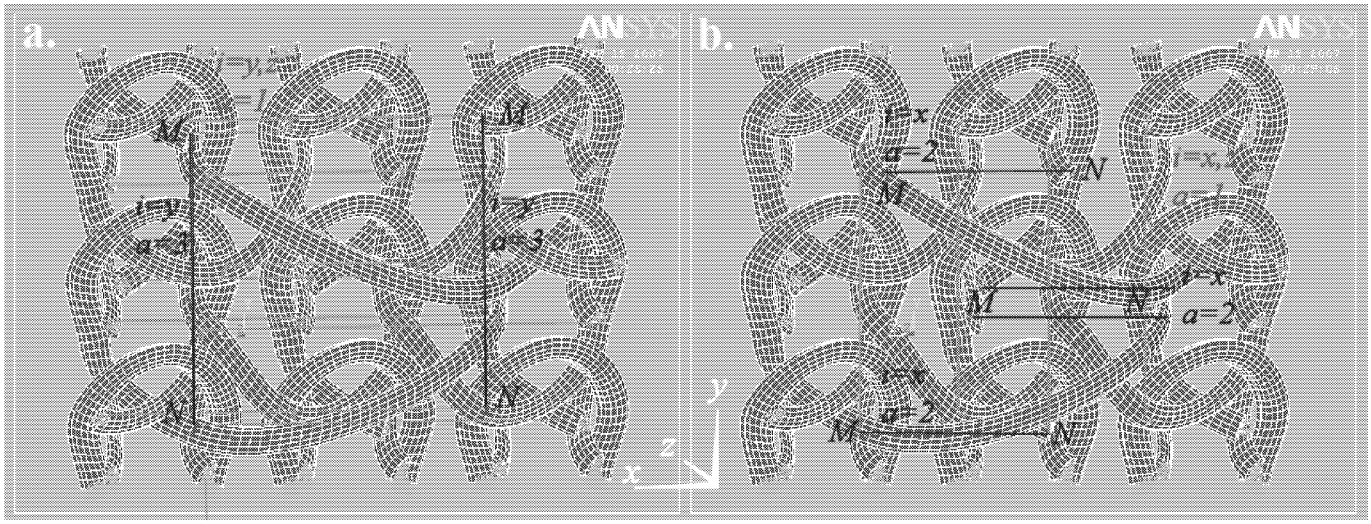


Figure 3. Constraint equations of tensile test (a) along wale direction, (b) along course direction

In the current model the subtraction of the selected parts of the bodies demands additional application of boundary conditions for the even load transferring between the bodies. This subtraction is necessary for the Charmeuse structure since a complete unit cell presenting symmetric and antisymmetric features is unreachable. The additional boundary conditions are applied in the nodes belonging in the areas formed by the subtraction of the selected volumes. The applied boundary conditions are constraint equations (CE) preserving the motion control of the nodes belonging in the mentioned areas according to the respective nodes of the neighbor course or wale. The CE correspond to equations of the type: $U_i(nodeM) = a \cdot U_i(nodeN)$, where $i=x,y,z$ and $a = 1, 2, 3$ depending on each case. Fig. 3 shows schematically the application of the constraint equations.

The simulated deformations are highly nonlinear due to the multiple contact pairs and the large displacements occur. Thus a full Newton-Raphson solution procedure is implemented for faster convergence. The advantage of this algorithm consists in the update of stiffness matrix for every computational iteration. The control of the iterations number is important for the convergence of the solution procedure with the less CPU time consumption. Two options are activated to reduce the number of the necessary iterations. The predictor option extrapolates the DOF solution using the previous history in order to take a better guess at the next solution. The line search option attempts to improve the Newton-Raphson solution by scaling the solution vector by a scalar value termed the line search parameter.

RESULTS AND DISCUSSION

For the evaluation of the modelling method the coursewise and walewise tensile deformation are simulated for a sample in the KES-F tensile testing mode. Moreover the structural unit is designed for the given sample according to the proposed procedure and the tensile tests are simulated. Particularly the accuracy of the predicted performance is a combination of the accuracy of the geometrical model, the success of the meshing, the appropriate material properties introduced in the yarns and the sufficiency of the applied boundary conditions.

Transverse isotropic properties are introduced for the yarn material type. The transverse isotropic properties describe the different elastic properties along the yarn axis direction and the radius direction approaching the non-homogenous behaviour of the yarn. The definition of the elastic properties is based on the typical elastic properties of PA6.

Some of the results received from the walewise and coursewise tensile test of the unit cell are shown in Fig. 4. For the walewise tensile test the displacement is imposed on the nodes of the area A1 and the nodes of the area A2 are constrained along y axis. While for the coursewise tensile test the displacement is imposed on the nodes of the area A3 and the nodes of the area A4 are constrained along x axis. The deformed and undeformed shapes of the unit cell are given in Fig. 4a,b. As it is shown the underlaps present the larger displacements. More representational are the Fig. 4c,d where the displacement vector summary is given for each deformation. Low values of strain (Fig. 4e,f) are presented in the yarns compared to the elongation of the unit cell. Thus the elongation of the unit cell, in the first stage of tensile deformation, is mainly caused by the sliding and the straightening of the yarns rather than the elongation of them.

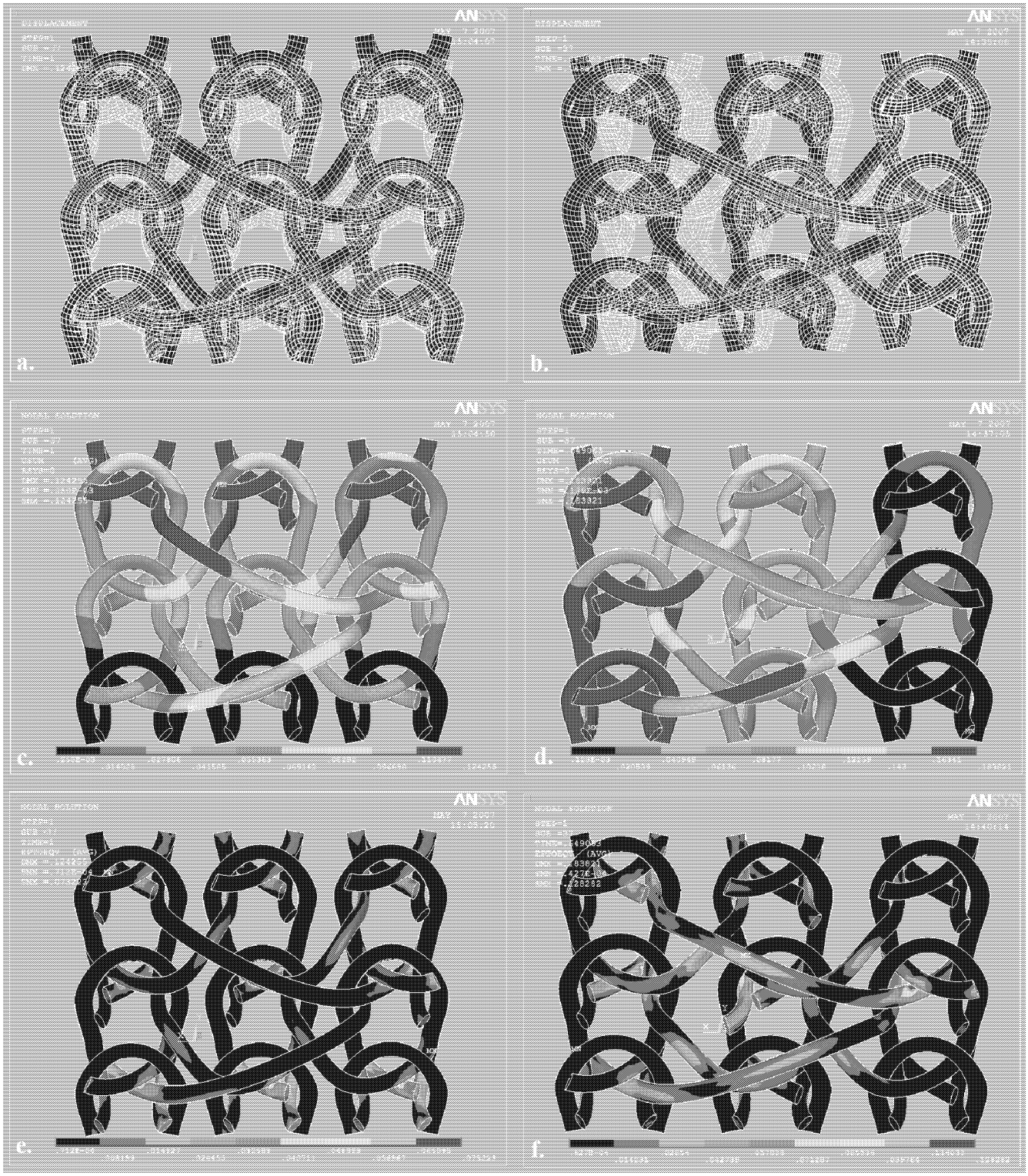


Figure 4. Results from the deformation of the unit cell. (a) Deformed and undeformed shape from the walewise tensile test, (b) Deformed and undeformed shape from the coursewise tensile test, (c) Displacement vector summary from the walewise tensile test, (d) Displacement vector summary from the coursewise tensile test, (e) Von Mises total strain from the walewise tensile test, (f) Von Mises total strain from the coursewise tensile test.

CONCLUSIONS

The advanced numerical methods in modelling and simulation offer special facilities for the design and mechanical analysis of multi-body textile structures. The convenient and rapid use of interpolation methods for geometrical representations in conjunction with parametric design and iterative procedures accomplish the requirements for rapid and accurate modelling of the complex warp knitted structures. The abilities offered by

the developed F.E.M. software codes permit the mechanical analysis of demanding deformation procedures. The convergence difficulties arise from the nonlinearities and the large deflection effects are faced with the selection of the appropriate algorithm. The time needed for the solution procedure is controlled by the proposed options for the minimization of the required CPU sources. Conclusively the micromechanical analysis of complex textile structures like the warp knitted fabrics using numerical methods is considered sufficient tool for the textile design procedure.

REFERENCES

Allison, G.L. 1958. "Warp Knitting Calculation Made Easy". *Skinner's Silk Rayon Rec*, 3, 281-285.

Blaga, M. 1999. "Research upon Yarn Consumption in Warp Knitted Fabric". *Revista Romana de Textile – Pielarie*, 3-4, 81-86.

Dalidovici, A.S. 1940. "Basic Knitting". Moscow.

Documentation for ANSYS

Goktepe, O. and Harlock, S.C. 2002. "Three-Dimensional Computer Modeling of Warp Knitted Structures". *Textile Research Journal*, 72 (3), 266-272.

Goktepe, O. 2001. "Use of Non-Uniform Rational B-Splines for Three-Dimensional Computer Simulation of Warp Knitted Structures",. *Turkish Journal of Engineering and Environmental Sciences*, 25 (4), 369-378.

Grosberg, P. 1964. "The Geometrical Properties of Simple Warp-knit Fabrics". *Journal of Textile Institute*, 55, T18-T30.

Kopias, K. and Pinar, A. 2004. "Warp-knitted Interlock Stitches – New Stitch group. Concept of Formation and Structure". *Fibres & Textiles in Eastern Europe*, 12 (1), 45.

Kopias, K. and Pinar, A. 2004. "Comparative Analysis of Structural Parameters of Warp knitted fabrics with Interlock and Traditional stitches". *Fibres & Textiles in Eastern Europe*, 12 (2), 52-56.

Korlinski, W. 1981. "Podstawy Dziewiarstwa", *Wydawnictwa Naukowo-Techniczne*, Warszawa.

Raz, S. 1987. "Warp Knitting Production", *Veglar Melliand Textilberichte GmbH*, Heidelberg.

Vékássy, A. 1960. "Konfekcioipar". *Tankonyvkiado*, Budapesta.

INTRODUCTION TO MODELLING OF VIRTUAL GARMENT DESIGN IN 3D

Agnieszka Cichocka
Technical University of Lodz
Faculty of Textile Engineering and Marketing
Department of Clothing Science and Technology

Ul. Żwirki 36, 90-942 Łódź, Poland

agnieszka.cichocka@ensait.fr

Pascal Bruniaux
Vladan Koncar
Laboratoire GEnie et Matériaux TEXtiles
(GEMTEX), UPRES EA2161
École Nationale des Arts et Industries Textiles (ENSAIT)
9 rue de l'Ermitage, BP 30329, 59056 ROUBAIX Cedex 01, France

pascal.bruniaux@ensait.fr,
vladan.koncar@ensait.fr

KEYWORDS
Modelling of virtual garment, pattern development process, adaptive model, identification

ABSTRACT

This paper presents an introduction to modelling of virtual garment design process in 3D which is our first approach in this domain. Our global project on virtual clothing design, besides the conception of virtual adaptive mannequin, is devoted to creation and modelling of garment in 3D. Starting from ideas of mass customized, e-commerce and the need of numerical innovations in garment industry we present a model of virtual garment and methodology enabling to conceive the virtual clothing directly on an adaptive mannequin morphotype in 3D. Short introduction of description the overall model of garment under constraints is presented. To explain overall methodology, the basic pattern of trousers is given. This global model of garment creation in 3D is composed of 3 components, including model of human body, model of ease and model of garment. The essential part is the ease model, necessary for proposed process of garment modelling. After describing each element influencing this process, the detailed presentation of ease model in relation with garment model is proposed. This combination of previously mentioned models may be considered as 2 other sub-models. The first one is linked with the front pattern position on the body and the second is leading to the back pattern of trousers position with appropriate values of ease. In order to execute the procedure of identification of the correct ease values and consequently their right position on the human body, the algorithm of identification is proposed. These two sub-models are strongly connected as in the feedback effect coming from interactions of front and back of trousers patterns. This phenomenon appears during modelling and depends on the structure of proposed model of ease. The relatively significant number of parameters requires the use of specific strategy of identification technique.
Finally the superposition of virtual and real pattern was done in order to visualise the right results.

INTRODUCTION

The criteria of virtual garment creation requires to take into account many aspects like model of virtual mannequin, model of garment, model of fabric, which are essential for garments industry. The diagram below in Figure 1 presents our global model of garment creation process without taking into account the fabric model.

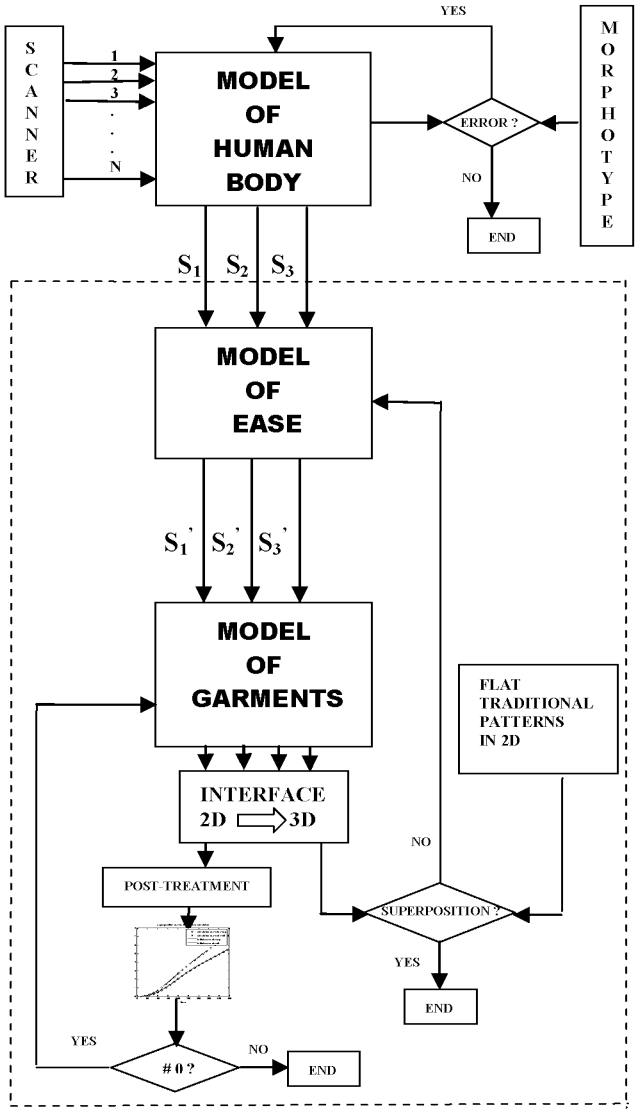


Figure 1. **Global model**

It does not mean that we belittle signification of model fabric in garment creation process, but we assume that it will

influence on the process subsequently. In this study it is not a subject of our analysis.
Proposed model represents a set of MIMO (multi input multi output) systems because of complexity of the virtual garment creation process. Moreover the feedback loops are located in order to translate the internal phenomena with the structure of model and accordingly to its particular elements.
In the next part more detailed description of each element is presented.

MODEL OF HUMAN BODY

To start a modelling of the human body the data from body scan are crucial in order to respect the human morphology. After scanning the human beings, a 3D shape can be imported from any 3D body scan device utilized in the garment industry. The data directly issued from a body scan are presented in the form of cloud points. In order to adapt the human data to 3D, another translation is required. This means that a cloud of points must be interpreted in order to obtain interpolation of human data and at once understood for CAD environment. In our case the RAPIDFORM software was used to transform the cloud of points by triangulations process to final superficial form of virtual mannequin.
Next step consists in the building of the parameterized mannequin morphotype. It is indispensable to base it on a definite model of surface representation using many curves as a result from the sections of a scan human body. Before the processes of an adaptive mannequin making, the final shapes of parametric body must be clearly determined and all related data correctly structured. The concept of parameterized model morphotype chosen, it is very important to set their parameters. They influence the selection and the position of these curves and relations between themselves and data. Moreover, this model should link perfectly a methodology design digitized in a CAD to be able to use it.
The presented mannequin model is controlled by a variety of parameters which are directly related to the curves controlling the mannequin surface without the upper limbs (Cichocka and al. 2006). All these parameters influencing the future shape have been identified and validated during the construction of the model. Those parameters are:

- Direct parameters (driving parameters) that allow easy and interactive changes and
- Indirect parameters (estimated parameters) that must be adapted and perfectly associated to the anthropometrical data acquired through various system inputs.

To validate a virtual mannequin morphotype created during the modelling process, the comparison process is required with the "initial" mannequin shape called morhotype in diagram on Figure 1. After accomplish this practice we are looking for errors related to differentiation between the compared mannequins. This procedure is seen on the previous diagram as the feedback loop. Presented mode of control makes us possible to correct the posture of mannequin morphotype form and adapt it properly to original model.

Finally we created an entirely parametric virtual mannequin which was segmented in various sections (S_1, S_2, and S3) which are related to different kind of garments and its position on the human body. Owing to this process we were identified 3 segments:

- Superior members linked with sleeves
- Inferior members linked with trousers legs.
- Bodice linked with waistcoat.

Those segments fulfilled at once the role of outputs for model of human body and also the inputs for next element of global model i.e. the ease model.
Subsequently according to the huge range of garment, the classification of the garment became required. In order to categorize the garment, the association of particular segments is indispensable from the point of view the chosen model of garment
In the figure below an example of final model of mannequin adaptive morhotype is presented.

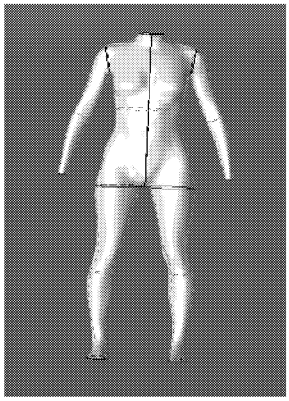


Figure 2. **Model of adaptive mannequin morphotype**

More detailed study about the methodology of conception an adaptive mannequin morphotype was presented in our previous work (Cichocka and al. 2006).

MODEL OF EASE

Whilst the conception of virtual adaptive human body is succeeded the next stage could begin. The issue from previous phase available under the contours form will make the base for proposed ease model. This means that the core of the design of virtual garment process will be considered. It was highlighted on the global model that the ease model depends and remains in a close relationship with the garment model which is normal. For each alteration on the ease model the direct influence on the garment model is observed. Accordingly to classification of garment the proper contours of human body must be chosen. It means that the established segments offer the different lines determining the human morphology. Moreover the garment conception imposes very strong constraints in the choice of contours. Additionally, in majority of cases the garments are composed of a combination of different contours containing at least two proposed segments. For example the garments intended to upper part of human body like the t-shirts, have to consist of

the bodice and superiors members segments. The more complex example, like a smoking jacket, includes three segments mentioned in previous section. In current project we focus on the trousers model assigned to lower part of body. In this case the bodice and inferiors members segments will be a subject of our assumptions.
According to preceding digressions, the segments represent not only the contours of human body but also they have to match up to the analysed kind of garment.
Management of this model of ease was accomplished with the significant numbers of ease parameters. These control parameters are placed on those human body contours, rightly selected to fit to the trousers model. Also, the repartition of controlling points was crucial because of the opportunity of future administration of ease model. Because the human body represents a very complex set of shapes, it was essential to put more points on the curvatures places on the human body than on the places which are "flats" and straights. On the figure 3, we can find the example of the control point's arrangement on the contours the trousers model.

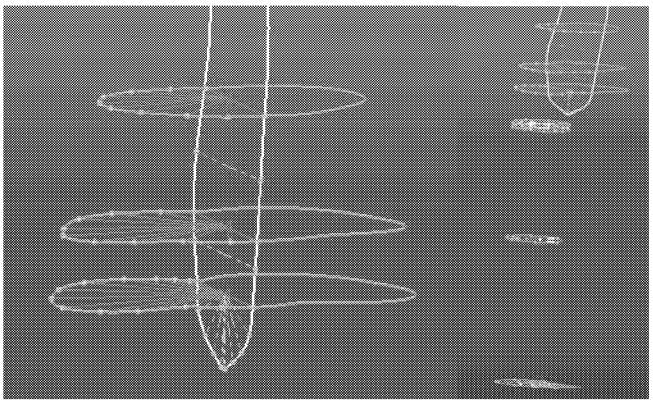


Figure 3. **Control point's arrangement on the trousers contours**

While the problems to using the contours for upper part of trousers with waistband and hips were quite easy, the leg of trousers design was a difficult task. The procedure of scanning of human body requires assuming a particular pose. In general the legs and the arms must be spread to an appropriate angle relating to scanner device chosen. We have compared two methods of measurement the human body corresponding to the needs of the garment industry. Traditional manual method and advanced high-speed automatic scanner were compared in order to verify the accuracy of the last one. It measures the changes of skin in 13 properly chosen points on human body, linked with anthropometrical surveys and simultaneously changed their position on the body. For each new position of human body the measures are taken identically as in previous cases in order to assure measurement recurrences. After an analysis of these two ways, it has been obvious that even if the time required to take the measurements using the body scanner was shorter, the significant differences between the results issues from the body scanner and traditional method appeared. Moreover, the more significant accuracy of measurements obtained with scanner device shows that accordingly to the skin deformation, the small modification of arm position introduces the changes to the results of measurement. Furthermore the traditional pattern making process takes into account the appropriate posture of human body which finally influences on the pattern shape. Consequently there is no possibility to use the original posture of virtual human body on the legs level. Therefore, knowing that during garment conception, the legs of the basic trousers fall in straight-line, starting from the thigh to the bottom of trousers, we have to follow it and to keep this condition. Then the copies of thigh contours were distributed relatively to the required levels of knee and ankle which was presented on Figure 4.

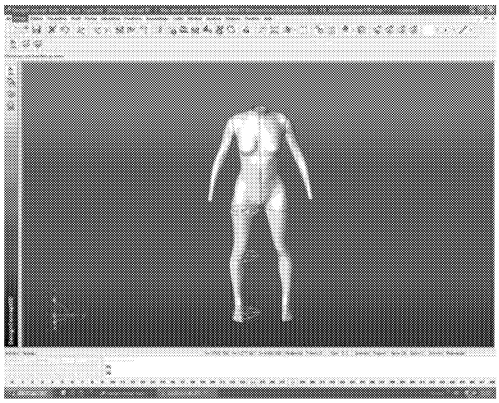


Figure 4. **Conception of legs trousers**

Finally, in order to have a correct course of virtual garment design, it was necessary to throw additional contours. Besides the addition of these curves and the controlling points, it was necessary to put them in the places which are essential to model of basic trousers. In our case the supplementary contours are situated between the waistband line and hip line and below thigh line. Also the zone of buttocks needed the complementary contours for the same reasons as we explained for the increased number of points in the curvatures places on the human body. The proposed model of ease for the basic lady trousers model is presented on the Figure 5.

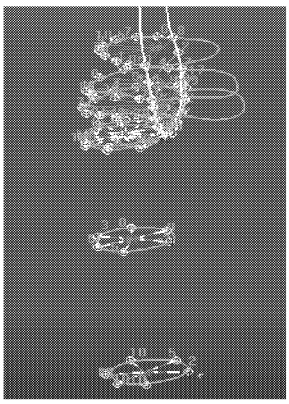


Figure 5. **Model of trousers ease**

This model of ease is represented by a set of lines which will support the future garment. Accordingly to the morphology of human body, our proposed model of ease constitutes the form of skeleton supporting the proposed model of trousers.

MODEL OF GARMENTS

In clothing domain, it is common that the prototyping and pattern design process is in fact an intuitive process which depends on the designer's competence and experience. For that reason, it is very difficult to simulate his work which could be in fact definite by the expression "to know how to make his proper profession". Moreover, once the garment design is finished, the pattern making process is required. Generally, in garment industry this work is done: firstly in 2D using usually the CAD software, secondly, by another person called pattern maker. We hardly need to mention that this process is very laborious and arduous. Besides, often there are many problems in communication, between designer and pattern maker, because of differences appearing during the sketch reading process. Proposed methodology using software which work in 3D, could aid in solving this misunderstanding between them by suggesting an attitude for direct pattern making process during garment design course.

The last element of our vision of virtual garment creation contains the model of garment. Garment designer needs the knowledge about morphology of human body and pattern making process but also about the garment position on the human body. Consequently the core of our model was highlighted in order to show the inseparable relationship between ease model and garment model acting mutually on each other.

Starting from the ease model which determines the skeleton of garment (S_1', S_2', S_3') then the seam lines are required. These seams must focus and reflect in the model of garment. It means that in this stage, the look of trousers is definite. Supporting on the ease model and seams, the resultant surfaces take into account the garment division into 2 elements, analogous to garment conception: front and back of the trousers presented on the figure 6.

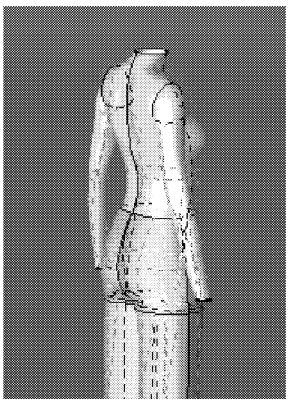


Figure 6. **Model of virtual garment**

Respecting the procedure of pattern making, firstly one front and one back were created only. The all pieces of trousers will be available at the end of ease identification process like symmetrical reflection in a mirror the front and back for another half of body. Next, using interface enabling the passage from 3D to 2D, the post-treatment is executed. This process in general consists of the triangulation process of resultant surfaces. After the definition of triangulation parameters appropriate to the model, further analysis can be continued. During this action, an analysis of fabric deformation in warp and weft directions is carrying out, as showed in Figure 7.

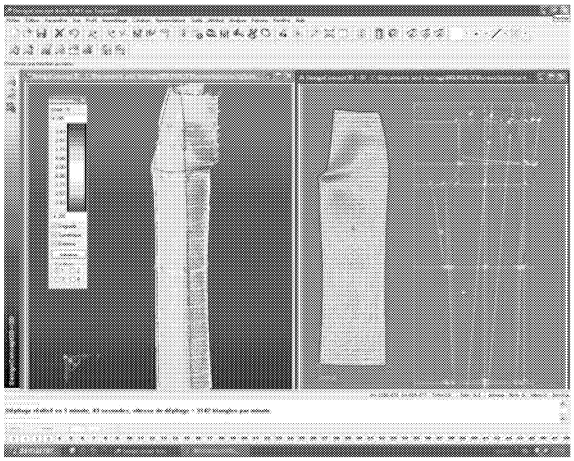


Figure 7. **Post-treatment of trousers surface - front of trousers,**

The different colours illustrate the stretched and compressed zones. Owing to this analysis and the stretching-stress curves of the chosen fabric, we are able to predict if the draft is required and where is the best position to put it. In order to verify our methodology, the point of reference was needed. For that, it was indispensable to place the 2D flat patterns of trousers. Accomplishing a traditional and industrial procedure of pattern trousers making process for a given person, we are obligated to put it in 2D, using DC3D software. Also it was very important to take into account that in general, during the pattern making process the seam values are contained in overall pattern surface, in order to enabling the assembling of garment pieces. While our virtual 3D trousers supported on mannequin, represents a pattern surface without the seam values. Our pieces of trousers are assembled by putting their edges together, that is indicated by the outside and inside seam lines. It means that we do not consider the influence of seam presence in the garment and it is except in this simulation. To allow the evaluation of our process by superposition of patterns, we were required to adapt the industrial patterns reducing the overall surface of patterns by the appropriate seam values. Depending on the results of the assessment of the differences between the patterns the process could end relating to good results or to be continued in opposite case. The example of an evaluation process is presented in the Figure 9 and Figure 10.

In the next part the identification scheme will be presented in order to detail our procedure and the identification algorithm of appropriate model of garment.

IDENTIFICATION SCHEME

In the course of identification process we noticed a property that the global, superficial model of garment may be decomposed into two sub-models which are strongly linked together. To specify, we mean, the zones which are close to the interface of front and back of the trousers. The overall procedure of identification the models of garment and ease,

is defined in 2 main stages. The first consists of the ease identification and managing of the model of trousers front. Consequently the second is devoted to the ease values identification and to control the model of the trousers back.

PROCEDURE AND IDENTIFICATION ALGORITHM

The model of the garment was judiciously selected in order to avoid the problems of identification. The model division into submodels is strategic, not only because it enables an identification of vector parameters as a whole by assembling parameters of submodels, but also because of specificity of garment conception process. In our case, the identification was carried out in several stages. The first was to identify the parameters of submodel related to the front of trousers model. The second corresponds to the identification of the parameters of the submodel related to trousers back.

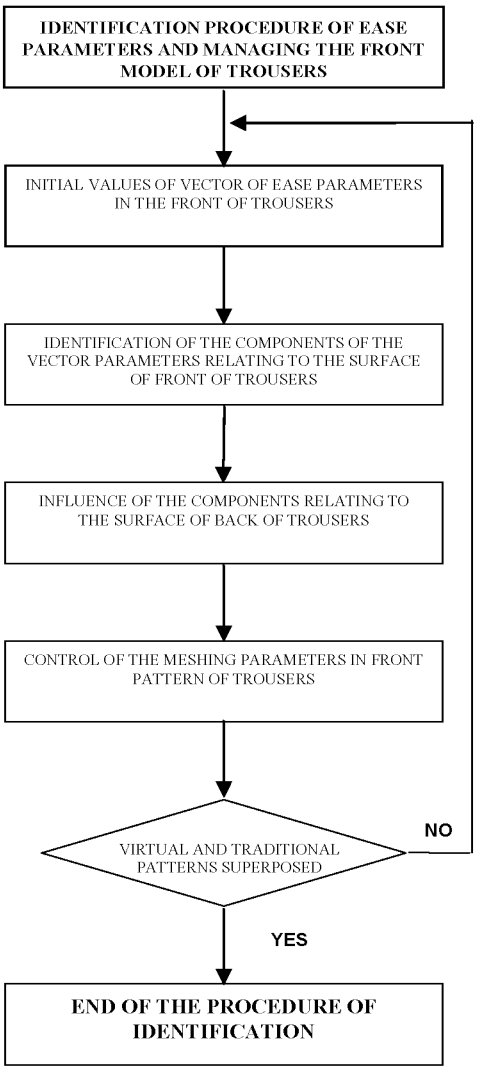


Figure 8. **Procedure of identification**

The flow chart of Figure 9 recalls the methodology described by the various points:

- Analysis of the front of garment model in order to set approximates values of the ease.
- Introduction of the initial vector of ease parameters necessary to start the identification. Only submodel of front of trousers is used.
- Identification the components of the vector parameters relating to the surface of trousers front. In this case, we control the surface appearance in front of trousers. The incorrect values of ease could cause the surface errors like too much divergence of slopes between different lines of garment.
- Identification of the components of the vector parameters relating to the back surface trousers. In this case, we control the surface appearance in back of trousers. Because of complex interrelations between those elements of trousers, each change of the ease values on the front, in particular, those close to the interface front-back, may influence on the surface of the trousers back. Consequently, further actions can not be accomplished.
- Identification and introduction of the triangulation components. In this stage is recommended to increase the number of triangles in order to guaranty to appropriated and smoothed shape of pattern edges.
- Transferring and flattening of the 3D patterns on 2D space. This step enables us to verify the virtual patterns and make the decision of achievement of identification process or continuity.
- Presentation of the results obtained by the superposition of the industrial and simulated pattern (fronts of trousers)
- Acceptation or negation of the results. If negation, it is necessary to run the procedure again as in the first point.

SIMULATION RESULTS

Once the correct values of ease and proper model of front trousers are founded, we continue the analysis for the back part of trousers. Obviously, in case of the backs of trousers we are executing the same algorithm of identification. In the Figure 9 we can find the results of virtual and traditional patterns superposed for the fronts and backs.

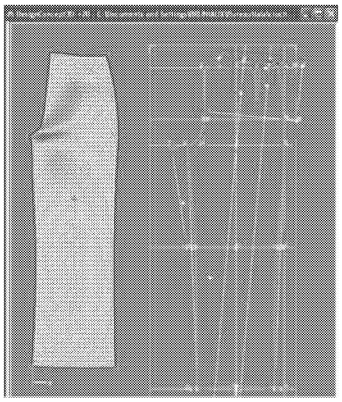


Figure 9. **Superposition of virtual and industrial patterns**

CONCLUSION

This article presents a new approach of garment design process modelling in which appear three interconnected sub-models. The feedback loops incorporate to the global model owing to the part highlights, showed the strong interaction between front and back of trousers. The interest of our model and proposed methodology of garment design is that it could be applicable in almost any software working in 3D. Moreover during the garment design process we are able to visualise the results in the form of 2D patterns immediately. Also it was necessary to set up an identification strategy specific to the problem. Identification procedure showed the sensibility of global model and emphasized the necessity of knowledge in garment conception.

In this method the divergence in the mannequin position and falling of trousers has been seen. It means that in general, during the measurement process of human body its silhouette is straight with the legs connected together whilst during the scanning process the legs are put separately. This is why in our case; we cannot cover the mannequin surface, describing the essential segments of human body for the model of trousers: the inferior's members and bodice. However, in garment conception, the legs of the basic trousers falling in straight-line from the hips, by thigh, to the bottom of trousers, which was taking into account by vertical distribution the copies of contour. This distribution was relative to the required levels of knee and ankle in order to ensure the correct and predictable shape of trousers.

To accomplish the process of garment design and give more realistic image transfer, obviously the stage of fabric simulation is required in order to verify the correct drape of trousers and finally validate the overall process. Nevertheless, we assume that the designer having the information about the fabric assigned to chosen trousers, in the first time, determine the position of garment on the body in a similar way as on the sketches, so our theory takes it into account. In the future work we think about an automatic graduation of patterns. We consider also conceiving the set of other segments in order to cover of human body, like upper limbs.

REFERNCES

Cichocka A., Bruniaux P. "Adaptive model of the human body – methodology of design of the mannequin morhpotype" Industrial Simulation Conference 2006, Italy

Chi L., Kennon R. "Body scanning of dynamic posture" International Journal of Clothing Science and Technology, vol. 18, No. 3, pp. 166-178, 2006

Boudjemaï F. "Reconstruction de surfaces d'objets 3D a partir de nuages de points par réseaux de neurones 3D-SOM", thèse de USTL Lille, 2006

AGNIESZKA CICHOCKA is preparing her PhD at ENSAIT Roubaix in France and at the Technical University of Lodz in Poland since 2004. She is doing her research in the field of dynamical modelling and simulation of fabrics for a design process. She participates to industrial projects with Lectra CAD company specialized in garment industry.

PASCAL BRUNIAUX was born in Denain, France in 1959 and went to the University of Lille, where he studied Automatic control and obtained his PhD in 1988. He worked at HEI engineering institute and in 1990 he came at ENSAIT, the textile engineering institute as Associate Professor. He is teaching in the field of mass customization and textile design process using virtual reality tools. He is doing his research in the field of fabric drape modeling and new approach to garment design using 3D CAD tools.

VLADAN KONCAR was born in Belgrade, Serbia in 1962 and went to the University of Lille, where he studied Automatic control of complex systems and obtained his PhD in 1991. He worked at the University of Lille as Assistant Professor. In 1991 he came at ENSAIT, the textile engineering institute as Associate Professor. He is currently Full Professor. He is teaching in the field of virtual reality, computer networks, communicative and intelligent textile structures and system theory. He is doing his research in the field of intelligent textile sensors and actuators, flexible screens and fabric drape modeling for design purposes.

INTERACTIVE VISUALISATION OF SIMULATED NONWOVEN FABRICS

Carlo Moritz Göllner
VICOMTech
Mikeletegi Pasealekua 57
E-20009 San Sebastián, Spain
email: mgoellner@vicomtech.es

Stefan Wundrak
Fraunhofer IGD
Fraunhoferstraße 5
D-64283 Darmstadt, Germany
email: stefan.wundrak@igd.fhg.de

Álvaro Segura
VICOMTech
Mikeletegi Pasealekua 57
E-20009 San Sebastián, Spain
email: asegura@vicomtech.es

Aitor Moreno
VICOMTech
Mikeletegi Pasealekua 57
E-20009 San Sebastián, Spain
email: amoreno@vicomtech.es

André Stork
Fraunhofer IGD
Fraunhoferstraße 5
D-64283 Darmstadt, Germany
email: andre.stork@igd.fhg.de

KEYWORDS

Computer Graphics, Application, Scientific Visualisation

ABSTRACT

The computational simulation of the production of spunbond nonwoven fabrics is a new and emerging field of research. Visualising the outcome of these simulations, large sets of thin randomly oriented fibres, has not yet been scientifically covered. We apply a set of visualisation techniques for the analysis of the simulation data using established methods such as surface reconstruction, parallel coordinates, and colour coding of specific physical fibre properties. Furthermore, we propose a model for realistic rendering of nonwoven fabrics to create a virtual preview of the simulated nonwoven fabric. This is a topic that has not been addressed before. Our model for realistic rendering is based solely on the output of the simulation and relies on existing techniques for realistic rendering of fibrous materials. We use modern programmable graphics hardware to ensure interactive frame rates.

INTRODUCTION

Nonwoven fabrics, short nonwovens, are manufactured mats made of randomly oriented synthetic fibres and are an important component of technical, hygienic, medical, and cosmetic products. Machine manufacturers cooperate with researchers to simulate and optimise the production process of nonwovens by combining computational fluid dynamics and fibre dynamics.

We focus on the outcome of the simulated production of spunbond nonwovens (Hietel and Wegener 2005, Marheineke and Wegener 2006). During the main production process, semi molten plastic fibres are injected into an airflow field where they are at first dragged and thinned. Then, fibres start to crimple due to air turbulences. Finally, the fibres deposit on a moving conveyor belt for removal and further post-processing such as dyeing and calendering (figure 1).

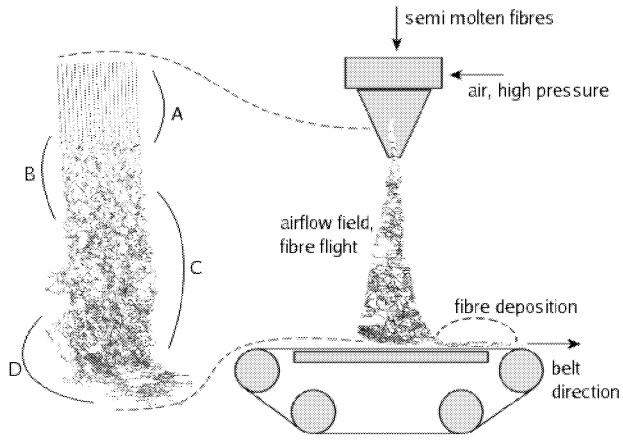


Figure 1: Production process of spunbond nonwovens.

The simulation of this process divides into three components:

1. Simulation of the turbulent airflow field using computational fluid dynamics. The outcome is an airflow field defined on an unstructured grid.

2. Simulation of fibre dynamics in the airflow field, in the following denoted as fibre flight. The outcome is a set of fibres approximated by time varying polylines in R^3. Each node of a polyline contains several physical parameters such as velocity or outer forces.

3. Simulation of the fibre deposition onto the conveyor belt. The result is a mat of fibres approximated by stationary polylines in R^2. The fibres form a flat mat since the simulation method omits fibre-to-fibre collision, yet.

The main points of interest regarding the fibre flight are the time variant geometry, physical parameters, and the overall distribution of fibres within the airflow field. Moreover, four

main zones of fibre behaviour should be located during visualisation: dragging zone (A), transition zone (B), crimpling zone (C), and deposition zone (D) (figure 1).

The main points of interest regarding the fibre deposition are the density distribution and structural properties such as anisotropy of fibre orientations, which is an important factor for tensile strength.

We refer to Laramee et al. (2004) for CFD visualisation techniques related to the airflow field.

RELATED WORK

Interactive Scientific Visualisation

Established techniques are suitable to visualise the fibre flight, for instance colour coding for the visualisation of scalar parameters (Schumann and Müller 2000, Rogowitz et al. 1996). Iso-contouring is a technique to visualise value boundaries of scalar parameters. However, iso-contours are applied to grid-aligned values, a data structure not offered by the fibre flight datasets. Glyph visualisation of vector values indicate the direction and magnitude of vectors. Pseudo-colouring may be used to map the components of a vector to RGB colour channels. However, interpreting this technique intuitively is difficult (Ware 2000).

The fibre flight data contains a set of physical parameters. Many visualisation approaches for such multidimensional datasets apply a projection into lower dimensions, for instance scatter plots (Cleveland 1993) or parallel coordinates (Inselberg and Dimsdale 1990) for 2D projection.

Interaction techniques assist the active process of visualisation and exploration: gaining an overview, zooming, filtering, acquiring detail-on-demand, and relating different parts of the data (Card et al. 1999). Filtering is a promising technique for visualising simulated nonwovens due to the mass of chaotically oriented polylines in the datasets. Linking and brushing is an interaction technique for filtering (Martin and Ward 1995). One selects parts of a dataset – brushing – and other correlated parts of the dataset are highlighted – linking. Alternatively, uncorrelated parts of the dataset may be masked out.

Realistic Rendering of Fibrous Materials

No direct work on realistic rendering of nonwovens has been published, yet. Thus, it is appropriate to identify existing rendering techniques that are applicable to nonwovens as well in order to create a virtual preview of the simulated nonwovens. Hair and cloth fibres are similar to nonwoven fibres in aspects such as diameter and translucency. Real-time shading algorithms in this area divide into two groups, either simulating the reflection behaviour of individual fibres or a complete piece of fabric.

The fibre deposition is a direct model of fibres of a nonwoven fabric. Thus, relying on techniques for shading of individual fibres is reasonable. In this field, the approaches of Kajiya and Kay (1989) and Marschner et al. (2003) are promising ones. Kajiya and Kay extend the Phong shading model to infinitely thin cylinders, whereas Marschner et al. use a model of rough dielectric cylinders to approximate specular reflections of hair fibres. Both models may be evaluated in real-time using graphics hardware (Nguyen and Donnelly 2005, Zoeckler et al. 1996). Zinke et al. (2004) extend Marschner's approach to a near-field model that allows close up views as well.

In addition to the shading model, one has to take occlusion and absorption of light into account, i.e. self-shadowing. Lokovic and Veach (2000) presented deep shadow maps, an extension of shadow mapping (Williams 1978) for volumetric objects. This approach has been improved for real-time execution using graphics hardware (Kim and Neumann 2001, Koster et al. 2004, Nguyen and Donnelly 2005).

The fibres of nonwovens are translucent and thus have to be drawn back to front to compute the attenuation integral using the over-operator. Therefore, a technique for depth-sorting is needed. CPU based sorting algorithms for polyline primitives exist, based on view-aligned bins (Kim and Neumann 2002) or pre-sorted variants of the geometry (Zoeckler et al. 1996). Graphics hardware can do back-to-front sorting using a multi pass technique called depth peeling (Everitt 2001).

Antialiasing is necessary if geometry covers fractional pixel during rendering, which is the case for thin nonwoven fibres. Common antialiasing techniques employ the fractional pixel coverage in a blending model. OpenGL graphics hardware supports accelerated antialiasing of line segments. McNamara et al. (2000) introduced a high quality antialiasing scheme for lines using polygonal geometry and edge filter kernels. However, the overall improvement over hardware accelerated antialiasing is the less significant the thinner lines are.

APPLICATIONS

Interactive Scientific Visualisation of Simulation Results

Fibre Flight - Time Variant Geometry, Physical Parameters

The obvious visualisation method for the time variant geometry of the fibre flight is to render it as an animation in R^3. We visualise the vectorial physical parameters of the fibres by mapping the norm to colour on the fibre geometry using different colour maps such as the spectrum map (figure 2). Experiments with glyph representations for the vectors showed that glyphs collide with the line geometry and result in a cluttered view.

The time step frequency is too small to capture the fast movement of the fibres by means of a smooth animation. Linear interpolation could smooth the animation. However, there are no clear dependencies between the points of two consecutive time steps. Moreover, the simulation of the fibre flight has a stochastic background, thus representing the fibre movement exactly is not necessary. Therefore, linear interpolation for a smooth animation is not applied.

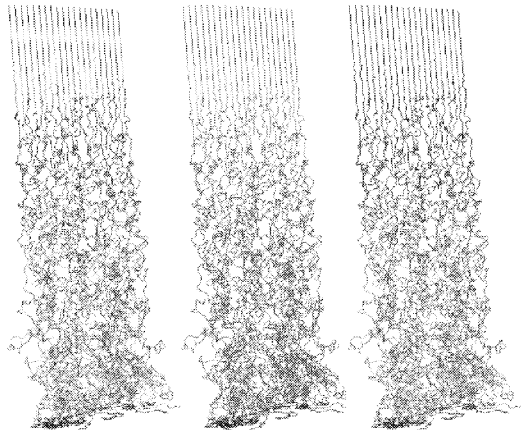


Figure 2: Colour Mapping of Relative Velocity, Fibre Velocity, and Outer Forces

Fibre Flight - Exploration of Zones of Fibre Behaviour

The four characteristic zones of fibre behaviour can be derived from three physical properties: the fibre velocity v, the relative velocity of fibres to the airflow v_r, and the tangent vectors t as reference direction for the related forces. We construct three scalars based on these vectors and map them to a parallel coordinate system:

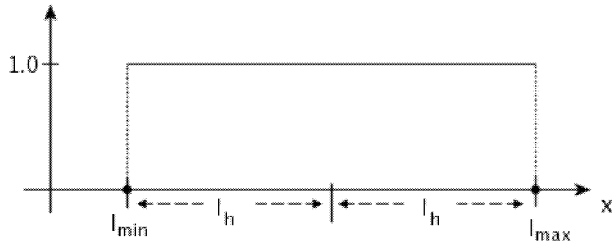
C_1 Scalar projection of normalised relative velocity onto the tangent: $p_{vr} = \frac{v_r}{|v_r|} \cdot \frac{t}{|t|}$. Negative values denote an orientation against the tangent direction, thus the air flows faster than the fibre moves, causing acceleration in tangent direction. Positive values denote the opposite.

C_2 Normalised relative velocity: $n_{vr} = |v_r|$ The value denotes the difference between the velocity of fibres and the airflow. A high C_2 value in combination with a C_1 value near to -1 denotes a strong acceleration in tangent direction.

C_3 Normalised fibre velocity: $n_v = |v|$ The value is high in the dragging zone and overall decreases towards the deposition zone on the conveyor belt. It is thus suitable for spatial segmentation.

We let the user interactively define an interval on interest for the values C_1 to C_3 to extract the four characteristic zones of the fibre flight. We use brushing and linking to filter geometry parts that do not fit the user's selection.

We use the GPU to accelerate both selection highlighting in the parallel coordinate system, and filtering of geometry. Shader code evaluates if a data tuple of a vertex fits into the current selection of value intervals (figure 3). OpenGL alpha testing is used for filtering during rasterisation, based on appropriately assigned alpha values. We filter all available time steps at once to provide a global view on the fibre flight dataset and its zones of fibre behaviour.



$$included = 1.0 - step(I_h + \varepsilon,\ abs(x - I_{min} - I_h))$$

Figure 3: Shader Pseudo Code for Interval Testing

Fibre Flight - Distribution of Fibres in Space

The user is interested in the distribution of fibres in space during the fibre flight. For that purpose, we compute a bounding surface of the fibre flight geometry. At first, we obtain sets of intersection points of the fibre geometry with planes parallel to the deposition plane, followed by the construction of a convex hull for every point set, using the Gift Wrap algorithm (Jarvis 1973).

Subsequently, we create a triangulated surface by connecting points of each two neighbouring convex hulls. The connection of such contour lines is a standard problem of graphic data processing (Lederbuch 1999). Our contour lines are convex and circular, thus we use an effective and fast heuristic by Ganapathy and Dennehy (1982) as connection method.

The bottleneck of our approach is the computation of intersections between line segments and planes. The points of a fibre are approximately sorted downwards the up-direction, neglecting short upturns due to turbulences. Hence, binary search can be used to reduce the overall number of line segments incorporated into intersection calculations. This results in a distortion of the surface, depending on how aggressive one applies the binary search.

Finding an optimal factor to adjust the binary search is difficult. We leave this decision up to the user who can balance between accuracy and performance based on factors such as the amount of turbulence in the dataset.

Fibre Deposition - Density Distribution

The geometry of the fibre deposition datasets is static, opposed to the fibre flight. The density distribution is the first point of interest.

It is common practice to visualise 2D density matrices as height fields using redundant mapping of density to both

height and colour (Rheingans and Landreth 1995). Although a density matrix is not given directly, we construct one by rasterising the fibre deposition geometry using Bresenham's algorithm. The resulting density matrix is visualised as a height field (figure 4).

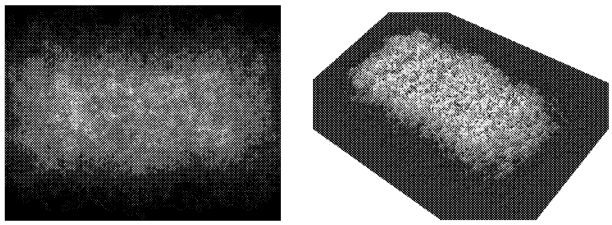


Figure 4: Visualisation of a Density Matrix (190x128)

Fibre Deposition - Predominant Direction Trends

The anisotropy of fibre directions is the second point of interest for the fibre deposition. It is difficult to discover due to the mass of chaotically oriented fibres in the fibre deposition dataset. We assist the discovery of anisotropy using interactive colour mapping and filtering of tangent directions, similar to a technique presented by Guse and Roth (2000).

All tangent vectors of the fibre deposition are coplanar. Thus, we can use the angle between a tangent and a suitable basis vector of the deposition plane in a pixel shader for lookup into a RGBA texture.

The user defines an angle of interest to discover anisotropy in the fibre deposition interactively (figure 5). Using the alpha lookup texture as shown in the lower right corners in the figure, one can reveal anisotropy against belt direction caused by dragging forces of the conveyor belt (angle 180°). Geometry that does not fit the angle of interest is filtered out using alpha blending.

REALISTIC RENDERING OF NONWOVENS IN REAL-TIME

In addition to our techniques for scientific visualisation, we use a model for realistic rendering of nonwovens to create a virtual preview of simulated nonwovens based on the geometry of fibre deposition datasets. Having the structural properties of nonwovens in mind (figure 6), we have chosen appropriate existing techniques covering four aspects: Shading (A), self-shadowing (B), depth-sorting (C), and antialiasing (D).

A The fibre shading model of Kajiya and Kay includes one specular reflection term only whereas Marschner's shading model simulates reflection properties of hair based on a model of dielectric cylinders, resulting in three specular reflection terms: R for direct surface reflections, TT for transmission, and TRT for reflections at the inner surface of the dielectric cylinder. In addition to that, Marschner includes a

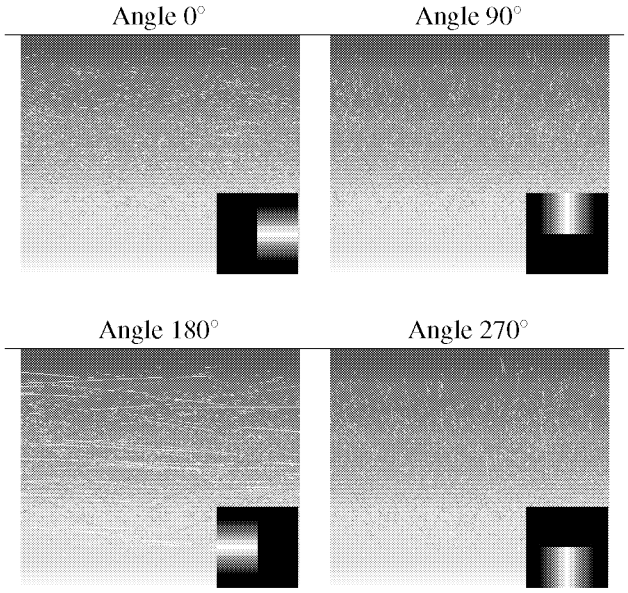


Figure 5: Filtering of Tangent Directions

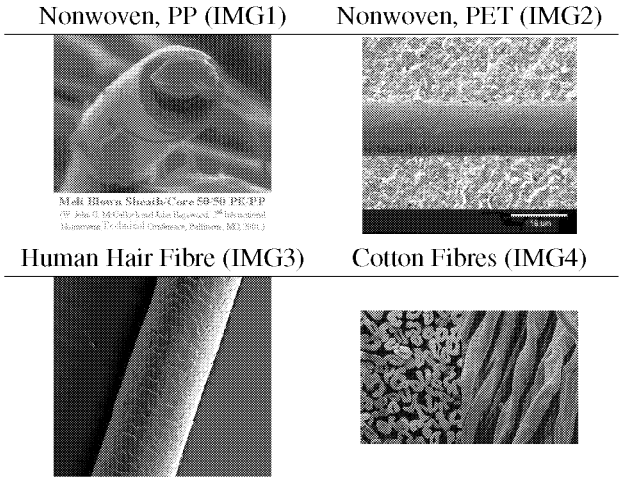


Figure 6: Electron Microscopy of Several Fibre Types

term to take tilted scales α on the hair surface and elliptical profiles into account (figure 7).

The 4D scattering function depends on four angles of incidence and can be split up into a sum of three terms for R, TT, and TRT taking into account the scattering behaviour of dielectric cylinders and Bravai's Law (η'). Each term consists of 2D components for longitudinal (M) and azimuthal scattering (N).

$$\begin{aligned} S(\phi_i, \phi_r, \theta_i, \theta_r) = & M_R(\theta_h) \cdot N_R(\eta'(\eta, \theta_d), \phi)/cos^2(\theta_d) + \\ & M_{TT}(\theta_h) \cdot N_{TT}(\eta'(\eta, \theta_d), \phi)/cos^2(\theta_d) + \\ & M_{TRT}(\theta_h) \cdot N_{TRT}(\eta'(\eta, \theta_d), \phi)/cos^2(\theta_d) \end{aligned}$$

Fibres of nonwovens consist of translucent material and resemble circular dielectric cylinders without prominent surface features. Hence, we can use a simplified model of

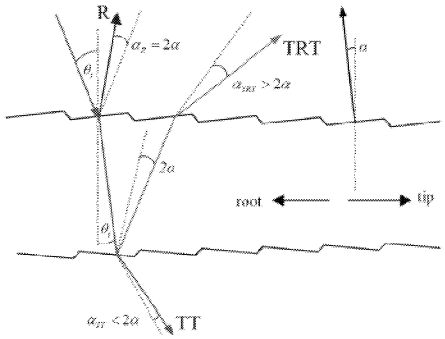


Figure 7: Scattering From a Hair Fibre: R, TT, and TRT (Image Source: Zinke et al. (2004))

Marschner's shading technique by neglecting terms for elliptical profiles and tilted surface scales. These terms are part of the N and M components of the scattering function.

The M and N terms are still too complex to be evaluated directly in graphics hardware. However, they depend on two variables only and may thus be precomputed and stored in 2D floating point lookup textures (True et al. 2004). Nguyen and Donnelly (2005) outline this approach that we adopt.

The shading model alone does not simulate fibre to fibre transmission of light, resulting in an overall too dark appearance. Ray tracing can solve this problem, but it is not suitable for real-time rendering (Moon and Marschner 2006). Thus, we follow the approach of Nguyen and Donnelly (2005) and add the diffuse term as defined by Kajiya and Kay (1989) to increase overall brightness.

B Self-shadowing gives important visual cues, especially concerning the density distribution of nonwovens. We choose the well described opacity shadow mapping approach by Kim and Neumann (2001) which has proven its potential for real-time evaluation using graphics hardware (Nguyen and Donnelly 2005). Figure 8 illustrates the usage of a set of slices to create a piecewise linear approximation of the fractional visibility function Ω.

The fibre deposition datasets of current simulations do not have height information since the collisions of fibres is not yet part of the FIDYST simulation tool (Hietel and Wegener 2005). Therefore, we created synthesised datasets of sufficient thickness and density for testing to overcome this temporary limitation.

C To ensure the correct evaluation of the attenuation integral, we use the bin-sorting technique proposed by Kim and Neumann for depth sorting that has a complexity of $O(n)$ (Kim and Neumann 2002). The view frustum is separated into bins and each line segment is sorted into a bin based on the distance to the camera (figure 9). This is the only CPU based technique in our portfolio.

D The fibres of nonwovens have a diameter ranging from 5 to 30 μm, thus they cover only pixel fractions when rendered

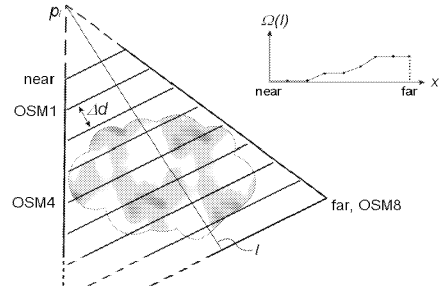


Figure 8: Opacity Shadow Mapping: Approximation of Ω

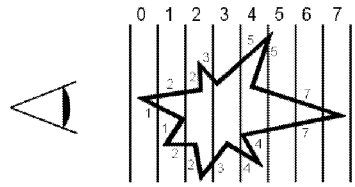


Figure 9: Approximate Depth-Sorting of Line Segments

as polylines, given the typical size of a simulated nonwoven of 13x10 cm. Antialiasing techniques have to be applied.

We choose to rely on standard hardware-accelerated line antialiasing (GL_LINE_SMOOTH) since the advantage of more sophisticated approaches is low for very thin lines such as the fibres of nonwovens.

RESULTS

Scientific Visualisation of the Simulation Results

The interrelations between three physical parameters are characteristic for four zones of fibre behaviour during the fibre flight. We created a solution for an interactive exploration of these zones in real-time using parallel coordinates combined with linking-and-brushing. Figure 10 shows that we can extract three of the four zones from the fibre flight dataset, whereas our method could not extract the deposition zone effectively.

Our hardware accelerated method filters a fibre flight dataset containing $1.2 \cdot 10^6$ line segments at 37 FPS on our test system (see table 2 for test system specifications). Advanced simulation datasets also include obstacles in the airflow that have a strong influence on the fibre flight. Our approach helps to identify the zones of fibre behaviour in these complex setups, opposed to figure 10 where the transition from one zone to another is rather uniform.

Figure 11 shows the extracted surfaces to visualise the fibres' area of distribution in space during the fibre flight. The polyline geometry alone is not suitable to depict this because of the difficult shape perception. Moreover, the surface gives valuable context information in combination with filtering using parallel coordinates. The surfaces also indicate a slight bend of fibres in conveyor belt direction that increases with

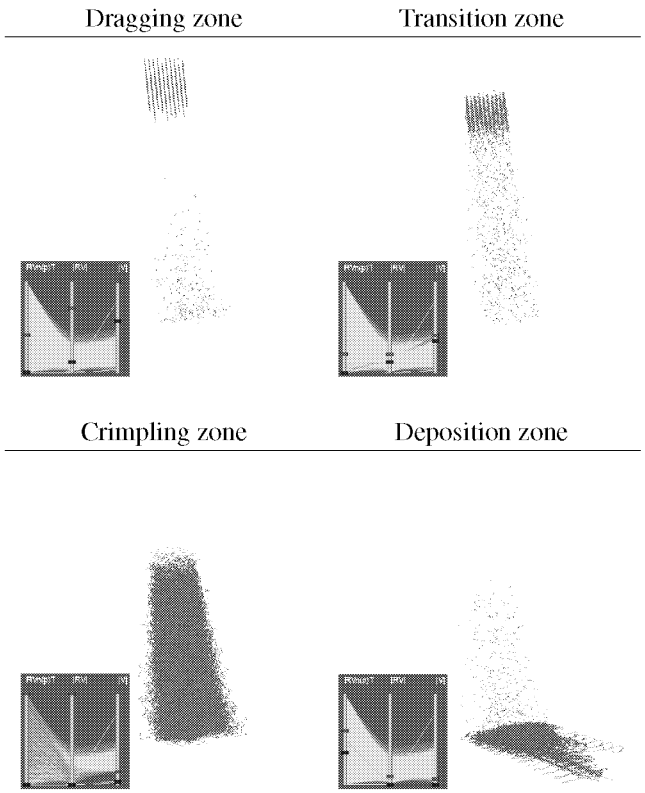


Figure 10: Exploration of Zones of Fibre Behaviour Using Parallel Coordinates

the belt speed. Extracting this feature by using parallel coordinates only is difficult.

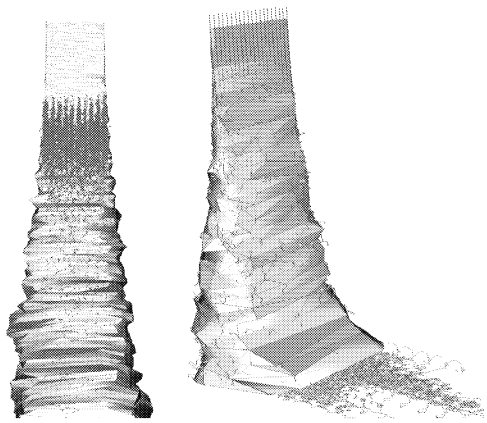


Figure 11: Reconstructed Bounding Surfaces of a Fibre Flight Dataset

As visualisation techniques for fibre deposition datasets, we use height fields to visualise the density distribution (figure 4). This way, the user can judge the uniformness of the density distribution, which is an important property of nonwovens. We reveal other structural properties such as anisotropy using interactive filtering of tangent directions (figure 5). This technique gives a first hint on the tensile strength of the simulated nonwoven fabric.

Finally, we combine the visualisation methods for all three simulation aspects (figure 12). The visualisation techniques for the fibre flight and the fibre deposition are integrated into our visualisation framework (IFX) that already includes CFD visualisation techniques for the airflow field. Thus, we have created a solution for the engineer and simulation expert that offers interactive tools for the visualisation and exploration of all three dataset types of simulated spunbond nonwovens.

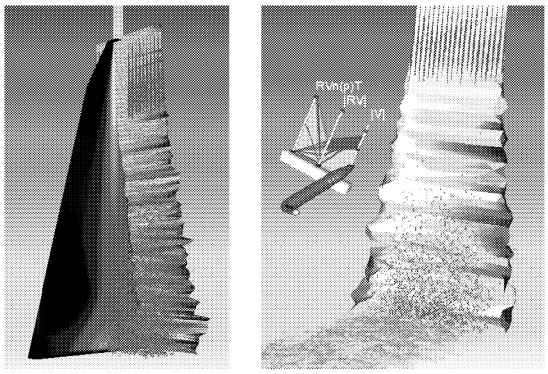


Figure 12: Combined Visualisation of Fibre Flight, Fibre Deposition, and Airflow Field Data

Realistic Rendering of Nonwovens

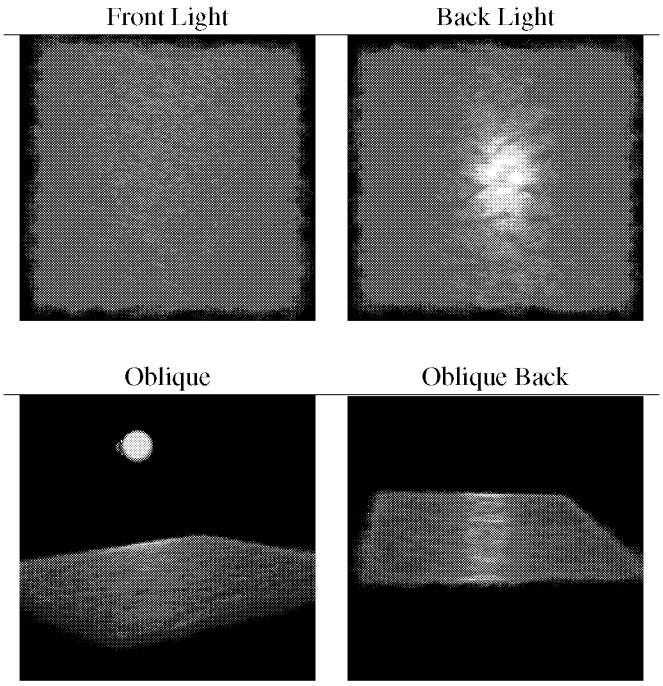


Figure 13: Realistic Rendering of Nonwovens at Various Lighting Situations

The results of our model for realistic rendering of nonwoven fabrics are shown in figure 13. We achieve a realistic virtual preview of nonwoven fabrics on perpendicular incidence of light. In a back lighting scenario, self-shadowing depicts the density distribution, an important property of nonwo-

vens. However, we experience shadow aliasing on oblique incidence in form of striping artefacts, since sixteen opacity shadow maps are not sufficient in this case to sample the flat structure of the fabric sheet.

Figure 14 shows a comparison of the fibre shading model with a real real nonwoven fabric under comparable lighting conditions. In particular, the shading model can reproduce characteristic circular highlights. Note, that the fabric sample on the left in figure 14 shows calendering patterns that shall be disregarded.

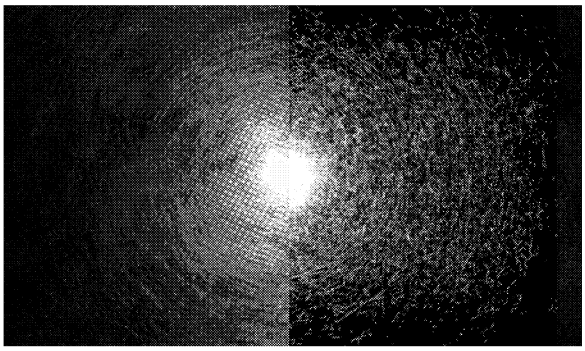


Figure 14: Comparison of a Real Nonwoven Fabric (left) and a Rendered Virtual Preview (right)

On our test system, we measured 8 FPS for a still scene in a render loop that incorporates fibre shading, opacity shadow map lookups, and antialiasing. The test dataset contains $2 \cdot 10^5$ line segments. Render passes for the construction of opacity shadow maps are added when the light source moves. The CPU-based depth-sorting is active in case of a moving camera (table 1).

Scene Movement	FPS	
static light source and camera	8	(1.00)
moving light source	5.5	(0.69)
moving camera	5	(0.63)
moving light source and camera	3	(0.38)

Table 1: Realistic Rendering of Nonwovens: Performance

CPU, RAM	Pentium-IV 3 GHz, 1 GB
Graphics Board	NVIDIA GeForce 6800GT
Window	1024x768
Opacity Shadow Maps	16x 512x512

Table 2: Test System Specifications

SUMMARY AND CONCLUSIONS

In this work, we have presented several techniques for the visualisation and exploration of simulated spunbond nonwoven fabrics. Previously, no work about visualisation specific to this domain has been published. Prior to now, visualisation has been done using general-purpose tools such as (GNUPlot) which do not support the interactive analysis of simulated nonwovens. We have identified, applied, and extended existing visualisation and rendering techniques to support engineers and simulation experts in the nonwoven industry. Most of our techniques use modern programmable graphics hardware to ensure interactive frame rates also for larger datasets and classification tasks.

The domain of simulated nonwovens represents an interesting class of datasets since the underlying chaotically oriented polylines are an uncommon form of data in the field of numerical simulations.

The creation of a virtual preview of the simulated nonwoven fabric is not only an interesting topic since it is a new application field for techniques of realistic rendering. It is also a mean of introducing nonwoven engineers to the field of numerical simulation. The engineers are used to inspect real samples of nonwovens in different lighting situations. Thus, a realistic interactive virtual preview is a suitable first step to introduce the engineers to work on simulated nonwovens in the product optimisation process.

FUTURE WORK

The computational simulation of nonwovens is still in its beginning – future simulations will produce results up to two orders of magnitude larger than our current datasets. Thus, an upcoming field of work in this area will be the handling of large datasets.

Also, we will investigate if our techniques are applicable to other forms of simulated nonwovens, beside spunbond nonwovens. Research from our partners in this area will lead to new challenges.

ACKNOWLEDGMENT

We want to thank our partners from the Fraunhofer ITWM in Kaiserslautern for the provided datasets and their help with interpreting the visualisation of the simulation results.

REFERENCES

Card S.K.; Mackinlay J.D.; and Shneiderman B., 1999. *Readings in information visualization: using vision to think.* Morgan Kaufmann Publishers Inc.

Cleveland W.S., 1993. *Visualizing Data.* Hobart Press.

Everitt C., 2001. *Interactive Order-Independent Transparency.* NVIDIA Corporation.

Ganapathy S. and Dennehy T.G., 1982. *A new general triangulation method for planar contours.* In *SIGGRAPH '82: Proc. of the 9th annual conf. on Computer graphics and interactive techniques.* ACM Press, 69–75.

Guse R. and Roth M., 2000. *RGPSview: a tool for RGPS data visualization. Proc of the IEEE Geoscience and Remote Sensing Symposium IGARSS 2000*, 7, 3006–3008.

Hietel D. and Wegener R., 2005. *Simulation of spinning and laydown processes. Technical Textiles*, 3.

Inselberg A. and Dimsdale B., 1990. *Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry.* In *IEEE Visualization*. 361–378.

Jarvis R.A., 1973. *On the Identification of the Convex Hull of a Finite Set of Points in the Plane. Inf Process Lett*, 2, no. 1, 18–21.

Kajiya J.T. and Kay T.L., 1989. *Rendering fur with three dimensional textures.* In *Proc. of the 16th annual conf. on Computer graphics and interactive techniques.* ACM Press, 271–280.

Kim T.Y. and Neumann U., 2001. *Opacity Shadow Maps.* In *Proc. of the 12th Eurographics Workshop on Rendering Techniques.* Springer-Verlag, 177–182.

Kim T.Y. and Neumann U., 2002. *Interactive multiresolution hair modeling and editing.* In *Proc. of the 29th annual conf. on Computer graphics and interactive techniques.* ACM Press, 620–629.

Koster M.; Haber J.; and Seidel H.P., 2004. *Real-Time Rendering of Human Hair Using Programmable Graphics Hardware.* In *Computer Graphics International*. 248–256.

Laramee R.; Hauser H.; Doleisch H.; Post F.; Vrolijk B.; and Weiskopf D., 2004. *The State of the Art in Flow Visualization: Dense and Texture-Based Techniques.*

Lederbuch P., 1999. *Reconstruction of Surfaces from Planar Contours.* In *Proc. of Conf. Computer Engineering and Informatics CEI*. 218–223.

Lokovic T. and Veach E., 2000. *Deep shadow maps.* In *Proc. of the 27th annual conf. on Computer graphics and interactive techniques.* ACM Press/Addison-Wesley Publishing Co., 385–392.

Marheineke N. and Wegener R., 2006. *Fiber Dynamics in Turbulent Flows - General Modeling Framework. SIAM Journal on Applied Mathematics*, 66, no. 5, 1703–1726.

Marschner S.R.; Jensen H.W.; Cammarano M.; Worley S.; and Hanrahan P., 2003. *Light scattering from human hair fibers. ACM Trans Graph*, 22, no. 3, 780–791.

Martin A.R. and Ward M.O., 1995. *High Dimensional Brushing for Interactive Exploration of Multivariate Data.* In *VIS '95: Proc. of the 6th conf. on Visualization '95.* IEEE Computer Society, 271.

McNamara R.; McCormack J.; and Jouppi N.P., 2000. *Prefiltered antialiased lines using half-plane distance functions.* In *Proc. of the ACM SIGGRAPH/EUROGRAPHICS workshop on Graphics hardware.* ACM Press, 77–85.

Moon J.T. and Marschner S.R., 2006. *Simulating multiple scattering in hair using a photon mapping approach. ACM Trans Graph*, 25, no. 3, 1067–1074.

Nguyen H. and Donnelly W., 2005. *Hair Animation and Rendering in the Nalu Demo.* In M. Pharr (Ed.), *GPU Gems 2*, Addison Wesley.

Rheingans P. and Landreth C., 1995. *Perceptual Principles for Effective Visualizations.* In *G. Grinstein and H. Levkowitz, Perceptual Issues in Visualization.* Springer, 59–73.

Rogowitz B.E.; Treinish L.A.; and Bryson S., 1996. *How not to lie with visualization. Comput Phys*, 10, no. 3, 268–273.

Schumann H. and Müller W., 2000. *Visualisierung - Grundlagen und allgemeine Methoden.* Springer.

True T.; Grantham B.; Kuehne B.; and Shreiner D. (Eds.), 2004. *SIGGRAPH 04 Course Program #16 on Performance OpenGL.*

Ware C., 2000. *Information visualization: perception for design.* Morgan Kaufmann Publishers Inc.

Williams L., 1978. *Casting curved shadows on curved surfaces.* In *Proc. of the 5th annual conf. on Computer graphics and interactive techniques.* ACM Press, 270–274.

Zinke A.; Sobottka G.; and Weber A., 2004. *Photo-Realistic Rendering of Blond Hair.* In *Vision, Modeling and Visualization 2004.* Akademische Verlaggesellschaft Aka GmbH, Berlin, vol. 0, 191–198.

Zoeckler M.; Stalling D.; and Hege H.C., 1996. *Interactive visualization of 3D-vector fields using illuminated stream lines.* In *Proc. of the 7th conf. on Visualization '96.* IEEE Computer Society Press, 107–ff.

HYPERLINKS

(GNUPlot)	GNUPlot, http://www.gnuplot.info
(IFX)	Fraunhofer IGD - IFX visualisation framework, http://www.igd.fhg.de/igd-a2/ifx
(IMG1)	Nonwoven fibre (PP), http://www.fluent.com/solutions/examples/x225.htm
(IMG2)	Nonwoven fibre (PET), http://www.swicofil.com/pes.html
(IMG3)	Human hair fibre, http://www.ualberta.ca/~mingchen/thair.htm
(IMG4)	Cotton fibres, http://www.e4s.org.uk/textilesonline/content/6library/report1/textile_fibres/cotton.htm

SMART TEXTILES MODELLING AND DEVELOPMENT

TEXTRONIC DEVICE FOR MEASURING THE BREATHING RHYTHM

Janusz Zięba and Michal Frydrysiak
Department of Automation of Textile Processes
Technical University of Lodz
Żwirki 116,
Poland
E-mail: jzieba@p.lodz.pl, michal.frydrysiak@p.lodz.pl

KEY WORDS: health care, dynamic modeling, electronics, physiology, software engineering.

ABSTRACT

This article presents the prototype of a textronics sensor intended for measuring the breathing frequency; its design and the working-principle are described. The sensor is composed of two optical waveguides, the sender and the receiver. The characteristic of a textile spring, which is an integral component of the sensor, is presented as a force vs. elongation dependency. Experimental test results of the breathing rhythm are presented for human groups of different age, obtained during different physical effort. A mathematical model of the mechanical sensor's part is described, together with a scheme of the model calculation by the Matlab Simulink program, as well as selected runs of force and elongation for different degrees of physical effort are presented. The sensor described may be used in medical applications integrated with special protective clothing. The authors focused on the design description of a non-invasive system of health monitoring, which extends the traditional clothing functions, and at the same creates a modern textronics product. Special attention was devoted to application of optical waveguide-technique in textile engineering, connected with great expectations considering an easy transmission of measuring signals inside fibrous structures. This latter is an important technological problem, which stimulates the development of new, unconventional manufacturing methods of textronics textile products. Such prototype, as the described above, of an interactive textile product, which presents auto-adaptive features, will be called as a textronics product.

TEXTRONICS

The term textronics refers to a synergic connection of textile science and engineering, electronics and informatics (figure 1) with the use of knowledge from the area of automatic and metrology. The synergy of such a connection is displayed in creation of a new quality, in which the component elements enhance mutually their action. This is achieved by physical integration of microelectronic with textile and clothing structures. The aim of textronics is obtaining multifunctional, intelligent products of complex internal structure, but of uniform functional features. As textronic systems are called measuring and control systems, which include fibrous sensors and/or actuators, or other elements inserted in the fibrous structures, as well as electronic micro-circuits connected with these structures or formed with the use of them.

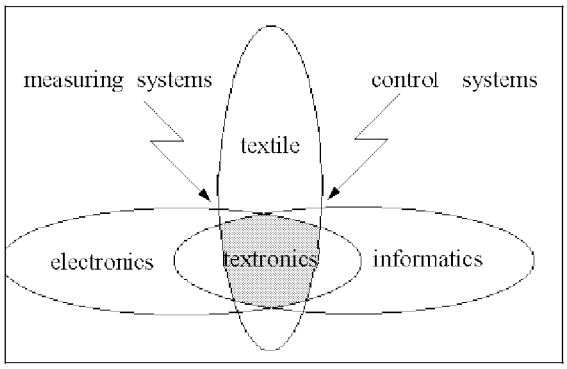


Figure 1: The area of textronics

Humans together with their clothing are the objects of measurement or control in standard textronic systems (Gniotek 2006, Gniotek and Krucińska 2004, Gniotek at al. 2003). At present, one of the basic problems of textronics is the identification of the properties of a human as the object of measuring and control, and this is the reason that only the model of the sensor was presented in this work. The breathing, which were used as input function to design the sensor, were determined by a real set of values of the breathing rhythms.

HUMAN PHYSIOLOGY – THE MECHANICS OF BREATHING

Oxygen is necessary for the vital processes of a human's organism, and it is taken up from air by the breathing process of rhythmic and unconscious character. A very important problem concerned with the breathing frequency measurement is recognition of the mechanics of this process, and the human's physiology connected with it. During

the breathing cycle, the volume of the chest changes as the result of increase in its three basic dimensions: the front-back, the transversal, and the vertical dimensions. The front-back dimension increases as the sternum moves away from the backbone by the action of the muscles which raise the sternum. The increase in the transversal dimension is the result of raising the medium ribs arched to the bottom, whereas the increase in the vertical dimension of the chest cavity is the lowering of the diaphragm. During contraction, the diaphragm presses on the organs positioned in the abdominal cavity, and therefore rhythmic protrusions of the abdominal walls are observed during breathing. The exhalation is a passive action under normal conditions. At the last stage of this process, the inhalation muscles are loosened which results in lowering the ribs. The abdominal pressure also takes part in the exhalation, and pressing on the entrails causes the diaphragm to raise-up. While testing a resting, healthy human we may observe 14 –16 breathing cycles per minute, which is related to 8 litres of air. This means that at each inhalation a volume of air within the range of 350 to 500 ml flow into the lungs. This air is called breathing air, whereas the process is called minute lungs' ventilation. The breathing air value changes in dependence on the human's activity, and for example may be larger than the above-mentioned 'resting' value by more than ten or even more than twenty during intensive action (Sylwanowicz et al. 1980, Traczyk 1989). In this work, we present a textile sensor prototype, which is based on the above-mentioned phenomena and used for measuring a physiological feature, namely the breathing frequency. Taken into account that clothing directly contact with the human body, we took an attempt to design a three-dimensional textile-clothing interface, which should add to the clothing a new function – the interactive adaptation to external stimuli, independently on the clothing's functionality and aesthetics (Tao 2001). The most important function of the textronic sensor presented herein is the increase in human's safety, thanks the possibility of early detection of threats and their proper counteraction. The lack of oxygen raises the level of adrenaline and noradrenaline, which causes pulse acceleration and an increase in the blood arterial pressure. This means that a progress in hypoxia may even cause the hazard of death. The dangerous situations connected with the professional activity of soldiers, policemen, fire-fighters, and lifeguards, as well as common people, who for example spent the period of convalescence in their own houses, create the area of the potential application of the described textronic sensor. The basis of realisation this function is a broad-considered monitoring of working conditions and of the state of health of the worker. This monitoring should be completed by control systems of different kind, including automatic control systems, information storage and transmission systems. This kind of activity, focused on human's health protection is called e-medicine and telemedicine, and recently is going trough dynamic booming (Marsh 2004). The creation of a specific contact between the physician and the patient is possible only thanks to application of modern technology. The device, described by us, allows normal life of patients, which should be monitored, without the need to stay all the time in a hospital.

DESIGN OF THE TEXTRONIC SENSOR FOR MEASURING BREATHING RHYTMS

The sensor prototype designed by us has the shape of a band wrapped tightly around the chest, which is seen in figure 2. This figure shows also the position of the optical waveguides together with the electronic system provided for counting the number of breaths per minute. The design presented is characterised by an easy assembly without disturbing the use-comfort (Zięba and Frydrysiak 2006).

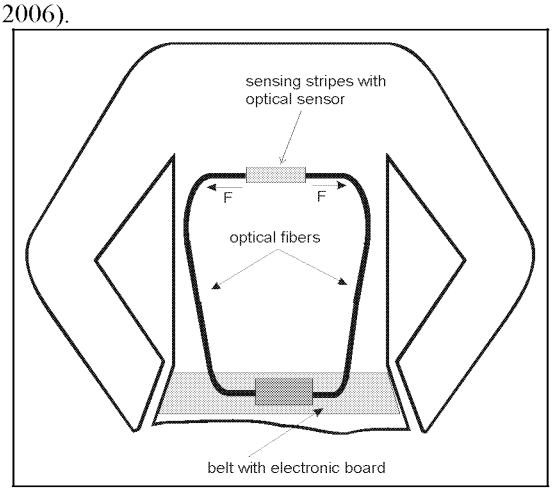


Figure 2: Scheme of the textronic sensor's displacement in the textile structure

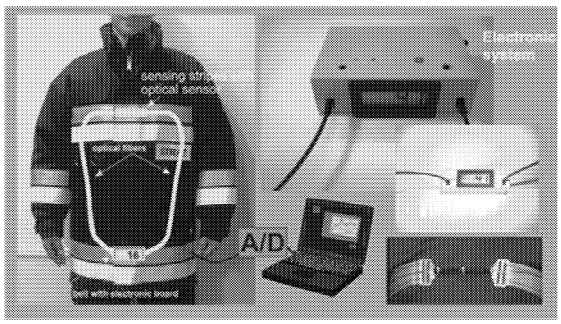


Figure 3: The prototype of implementation breathing frequency sensor in fireman uniform

Design of the breathing rhythm sensor

The device for measuring the breathing rhythm is composed of a sensor head, the sender and the

receiver light waveguides, and the electronic system. The most essential part of the sensor is its measuring head presented in figure 4.

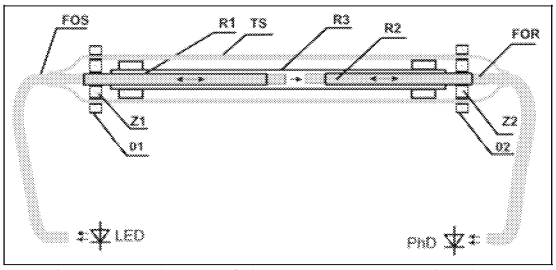


Figure 4: Scheme of the sensor's measuring head

The head includes two optical fibres, the sender (FOS) and the receiver light waveguids. They are positioned horizontal in one line, and each of them is surrounded by a sleeve (R1, R2), which are slidable placed in a common sleeve (R3); the sleeves in turn are fastened by the clamps (O1 and O2) manufactured of a material of good adhesion to the clothing. Each of the sleeves' ends (R1 and R2) is jammed by clips (Z1, Z2) which are mutually connected by a spring. The second end of the optical fibre, of the sender light waveguide (FOR) is placed directly opposite to the electroluminescence diode (LED) of the electronic circuit whereas the second end of the optical fibre, of the receiver light waveguide (FOR) is placed opposite to the photodiode (PhD) connected with the circuit, which in turn is connected to an amplifier, which output (Out) is connected to an electric current frequency recorder. The sensor's working principle consists on recording the changes in the distance between the light waveguide ends in the sensor's head forced by the chest's movements. The view of the electronic device intended for breathing counting is presented in figure 3.

Testing the device for breath counting

Within the scope of our investigation, we determined the static and dynamic characteristics of the sensor and its component elements. The elastic properties of the spring and the complete sensor were determined experimentally with use of an Instron tensile tester, and are presented in figures 5. The spring constant was assessed experimentally on the basis of a series of recorded hysteresis of the force vs. elongation runs. Our further tests were devoted to the determination of the static characteristic of the complete sensor.

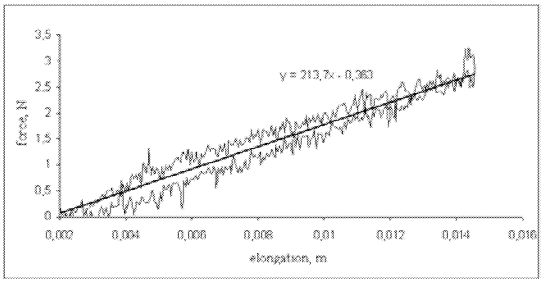


Figure 5: Force vs. elongation dependence of the textile spring

The nonlinearity of the sensor's static characteristic is caused by the nonlinearity of the electronic amplifier within the range of low voltage supply. This phenomenon does not influence the measure of the breathing frequency.

Breathing rhythm recording checked the working correctness of the sensor. Examples of breathing rhythms recorded during some breaths for different persons, of different age, which rested and during their medium activity, are presented in figures 6 and 7 by runs of the photodiode voltage.

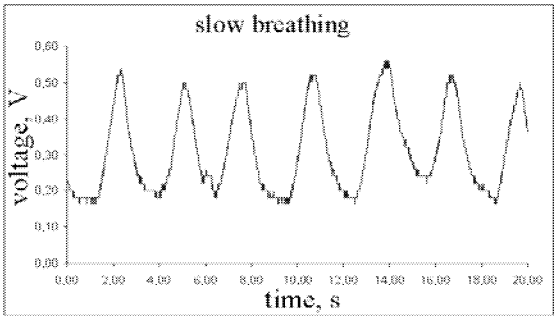


Figure 6: Example of a breathing rhythm run: slow breathing

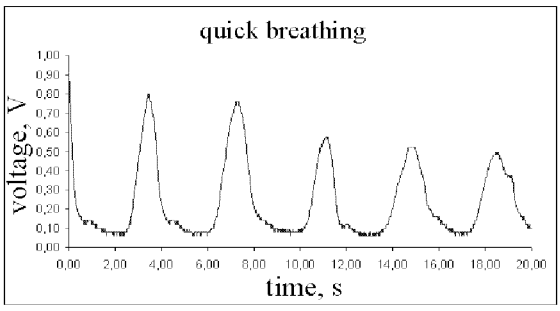


Figure 7: Example of a breathing rhythm run: quick breathing.

During research test was built laboratory stand to one side input forced the end of optical fibers. The researches perform for different frequency forced.

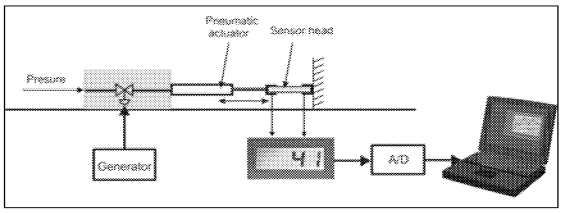


Figure 8: Scheme of testing laboratory stand

MODEL OF THE MECHANICAL PART OF THE SENSOR

Two sensor models were analysis. For the first model variant authors accepted, that the sending light waveguide is motionless, whereas the receiving waveguide slides forced by the chests' movements.

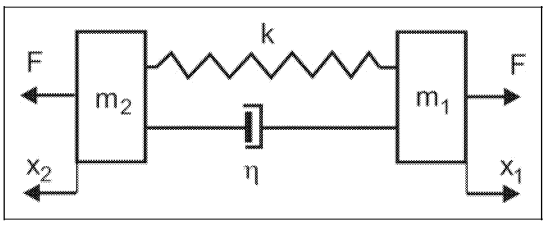


Figure 9: Mechanical model of the breathing sensor

For the second variant authors accepted a model with both the waveguides (the sending and the receiving) sliding coaxial; this variant is presented in figure 9. The whole sensor's displacement consists of components dislocation.

$$x_1 + x_2 = x \qquad (1)$$

During the first phase of the model elaboration we identified the object, i.e. the sensor's head as a two-parameter Kelvin-Voight object.

In order to describe precise the mechanical features of the object, detailed investigation of the system previously described was carried out. The sensor mass including its component elements, and those of the both light waveguides (the sensing and the receiving) was measured, as well as the constant ***k*** of the textile spring was assessed. The force component required to fill the lungs with air was accepted as the single input quantity of the model, whereas the photodiode voltage as the single output quantity. The dynamic properties of the sensor depends mainly on its mechanical part, as the optoelectronic phenomena proceeds very rapid.

The dynamic of the sensor for the may be described following:

$$m_1 \frac{d^2 x_1}{dt^2} - \eta\left(\frac{dx_1}{dt} + \frac{dx_2}{dt}\right) - k(x_1 + x_2) = F \qquad (2)$$

$$m_2 \frac{d^2 x_2}{dt^2} - \eta\left(\frac{dx_1}{dt} + \frac{dx_2}{dt}\right) - k(x_1 + x_2) = F$$

where:

m_1, m_2 – masses of the light waveguides,

x_1, x_2 – displacement end of the light waveguides,

η - coefficient of friction of the sleeve,

k – constant of the textile spring,

F – component force.

A mathematical model of the sensor's measuring head was designed on the basis of the system of equations (2) with use of the Matlab-Simulink program, figure 10. The real forces recorded by a strain gauge during the breathing cycle were used as the system forcing. The results obtained in the form of model runs are presented in figure 11.

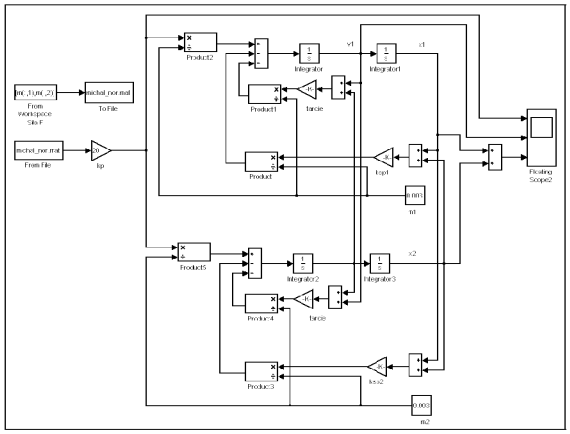


Figure 10: Simulation model of mechanical part of sensor created In Matlab Simulink program

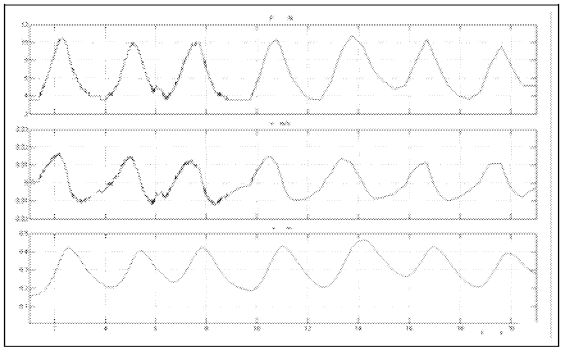


Figure 11: Force component course of breathing F, Speed changes of breathing v, chest movement x.

Conclusions

The construction of the textronics breathing rhythm sensor was working correctly. The advantage of this sensor is its textile structure which enables easy integration with the textile-clothing product. The results of simulation tests are close to those of experimental results. The easiness of servicing of the device described, and its small dimensions and mass are decisive considering its universality.

Acknowledgement

This work was supported by the Grant No 8 T10C 019 17.

Biography

JANUSZ ZIĘBA was born in Poland, and graduated with a M.Sc. degree in industrial automatic from the Technical University of Łódź, Faculty of Electrical Engineering. He attained a Ph.D. degree at the Textile Faculty in 1983, and at present is Senior Lecturer. His main scientific fields of interest are research into electrical, electronic, and control engineering and textronic. He is author and co-author of circa 60 publications.

MICHAŁ FRYDRYSIAK was born in 1979 in Łódź. He graduated from the Technical University of Łódź, Textile Faculty with a M.Sc. degree in 2003. The M.Sc. thesis he finalised in the Department of Automation of Textile Processes where he also commenced his post graduate studies into the field of modeling and simulating textronic systems under the scientific supervision of Prof. Krzysztof Gniotek Ph.D., D.Sc. His knowledge into textronic he broadened during studies at the University of Ghent, Belgium, and the University of Liberec, Czech Republic. Since November 2006, M. Frydrysiak M.Sc., Eng. has been working at the Department of Automation of Textile Processes.

REFERENCE

Catrysse M., Puersa R., Hertleer C., Van Langenhove L., van Egmondc H., Matthys D." *Towards the integration of textile sensors in a wireless monitoring suit*" Sensors and Actuators A 114 (2004) 302–311

Deflin E., Weill A., Koncar V." *Communicating Clothes: Optical Fibre Fabric for a New Flexible Display"* Proceedings Avantex Symposium 2002

Giergiel J. Warszawa 1990."Tłumienie drgań" PWN.

Gniotek K., Stempień Z., Zięba J. 'Textronic, a new field of knowledge , Przegląd Włókienniczy nr 2, Poland 2003

Gniotek K., Krucińska I.: *The Basic Problem of Textronics*, Fibres & Textiles in Eastern Europe, January/March 2004. Vol. 12 No 1(45)

Hung V.V, Ramin S. E.: Dynamic systems: Modeling and analysis, McGRAW-HILL COMPANIES, INC, New York, 1998

Kirstein T, Cottet D., Grzyb J., Tröster G. "T*extiles for Signal Transmission in Wearables*" Electronics Laboratory Swiss Federal Institute of Technology, 2003

Marsh A. "*3G Medicine – Removing the wires of telemedicine", www.e-care-ist.net 2004,*

Mazzoldi A., D. De Rossi, *Smart textiles for wearable motioncapture system*, proceedings world textile conference, 1-3 July 2002, p65-71

Sylwanowicz W., i inni. „Anatomia i fizjologia Człowieka", PZWL Poland 1980

Tao X. " *Smart fibres, fabric and clothing"* The textile Institute, Smart Textile Composites integrated with fibre optic sensors p174-199, 2001

Traczyk W.Z. „Fizjologia człowieka w zarysie, wydanie IV poprawione i uzupełnione" PZWL Poland 1989

Zięba J., Frydrysiak M.,"*Textronic – Electrical and electronical textiles sensors for breathing frequency measurement"* Fibers & Textiles in Eastern Europe No 5(59) 2006, Vol.14, p.43-47

Vanhala J. „*Wearable Technology Development in Finland" Tempere University of Technology*, Wearable electronic and smart textiles Leeds, UK 11,07, 2004

TEXTILE MAGNETIC CORE

Janusz Zięba
Department of Automation of Textile Processes
Technical University of Łódź
Żeromskiego 116, 90-924 Łódź
Poland
E-mail: janusz.zieba@p.lodz.pl

KEYWORDS: dynamic modeling, electrical engineering, software engineering.

ABSTRACT

This work describes textile magnetic elements which are component parts of textronic clothing used for monitoring selected physiological parameters of humans. Textile magnetic circuits consist of elementary monofilament magnetic fibres. Textile magnetic coils composed from a textile casing, electro-conductive yarn, and magnetic fibres are presented. Textile magnetic elements are mainly textile cores, which are the basic elements of each textile electromagnetic device, such as sensors, actuators and transformer. We present macroscopic magneto-mechanical and magnetic models which in the future will be the basis for mathematical description and simulation procedures of magnetic fibres and textile magnetic cores. The results of an analysis of the transversal and longitudinal magnetic fibres are presented. The mathematical problem of designing textile magnetic cores with consideration of the magnetic fibres' interlacement is also described. Calculation models realised by the Matlab-Simulink program are presented.

INTRODUCTION

Multifunctional fibres also called as hybrid fibres, recently world-wide manufactured, independently on their natural textile features are characterised by new properties which increase their application possibilities in textronic products [Gniotek K, Stempień Z., Zięba J.][Zięba J.]. These properties may be electroconductive, magnetic, piezoelectric, characterised by transmission radiation in general, and light in particular figure 1.

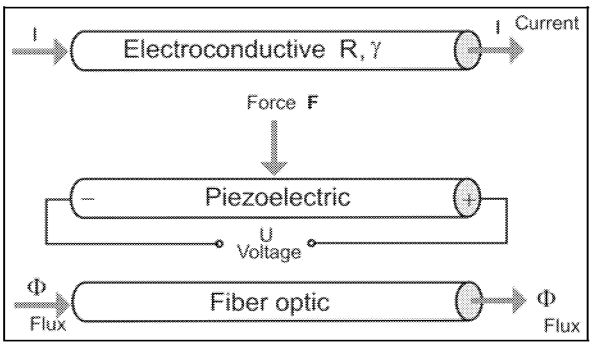


Figure 1: Symbolic presentation of a multifunctional fibre [Zięba J.]

Multifunctional fibres, yarns, and nonwovens are currently elements of textile electro-engineering and electronics, and their additional properties are dominant in relation to the mechanical properties. A specific kind of multifunctional fibres, the magnetic fibres, which are characterised by distinct magnetic properties are the subject of the elaboration presented.

A magnetic fibre may be symbolically presented as shown in figure 2, where the magnetic properties are presented in analytical form by equation $B = \mu H$, and in graphic form by the magnetisation dependency, with the magnetic flux Φ flown along the fibre axis.

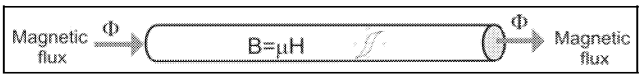


Figure 2: Symbolic presentation of a magnetic fibre [Zięba J.], [Rubacha M., Zięba J.]

Magnetic fibres create new possibilities for designing textile magnetic elements, and at the same electronic circuits and devices consisting parts of intelligent clothing, for example such which can be used for monitoring selected physiological parameters of humans.

The example of this kind of construction could be sensors, which measure breath frequency or pulse. Magnetic nonwoven can be made from magnetic fibers, which potentially could be an component of textiles sensor or actuator. Properties of magnetic nonwoven could find its application in shielding of magnetic field. Magnetic nonwoven could be easily integrated with clothing structure.

STRUCTURE AND MANUFACTURE OF MAGNETIC FIBRES

Magnetic fibres are manufactured by introducing ferromagnetic nano-particle powders into the fibre matter during fibre production [Łaszkiewicz and al., 2003], [Łaszkiewicz and al., 2004], [. This results in an implementation of the ferromagnetic into the fibre which becomes a macroscopic, monolithic material called a composite, as the polymer (the matrix) together with the powder filler make a discontinuous phase. Structure fibres presented in figure 4 and figure 5.

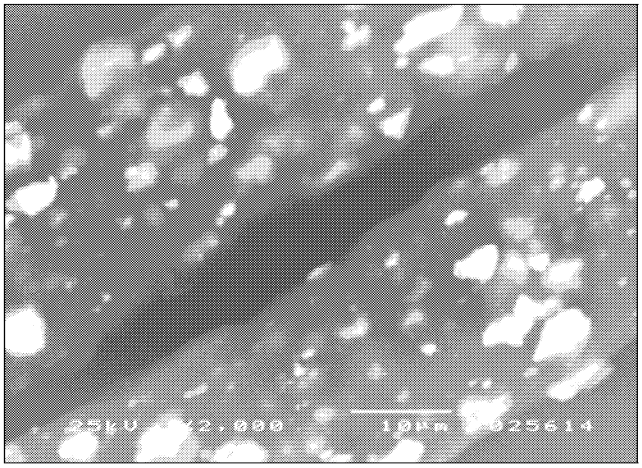


Figure 4: Longitudinal view of the magnetic fibres we obtained. [Rubacha M., Zięba J.]

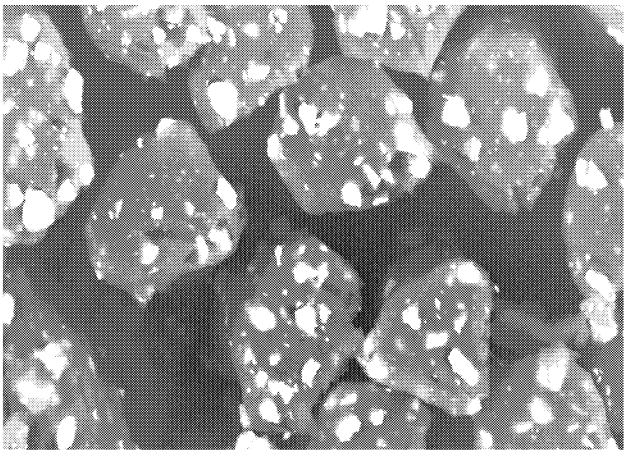


Figure 5: Cross-sections of magnetic fibres obtained by us. obtained. [Rubacha M., Zięba J.]

These composites are also called as electronic composites, as they have many applications in structures of microelectronic and nanoelectronic micro-mechanical systems (MEMS) and bio-micro-mechanical systems (BioMEMS) [Taya M. 2005]. A model may represent the elementary structure of magnetic fibres, where the path of the magnetic flux of the magnetic field Φ goes through the magnetic granules and the polymer figure 6. This is related to the magnetic reluctance R_m and the polymer reluctance R_p, and shown by equations (1):

$$R_p = \frac{l_p}{\mu_p \cdot S}, \qquad R_m = \frac{l_m}{\mu_m \cdot S} \qquad (1)$$

where:
R_p – the reluctance of the polymer,
l_p – the length of the way through the polymer between the ferromagnetic grains,
μ_p – the permeability of the polymer,
R_m – the eluctance of the ferromagnetic.
l_m – the ength of the way through the ferromagnetic,
μ_m – the permeability of the ferromagnetic,
S – the elementary area of the ferromagnetic cross-section.

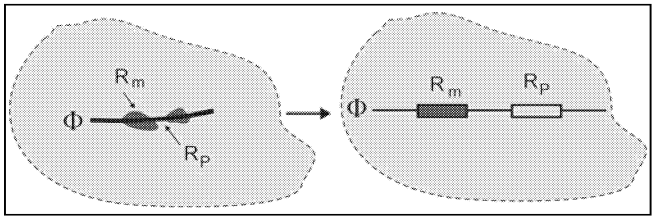


Figure 6: Model of the magnetic fibre and its equivalent circuit.[Zięba], [Rubacha M., Zięba J.]

The implementation of magnetic micro-powder into the fibres' polymer matrix causes that the fibres achieve magnetic properties. Magnetic fibres filled with magnetic materials are characterised by lowered mechanical tensile strength. The non-uniformity of the structure and the low mechanical strength of ferromagnetic materials contribute to this feature [Rubacha M., Zięba J. 2006].

MACROSCOPIC MODEL OF A TEXTILE MAGNETIC CORE

The elasticity of a magnetic fibre and a textile magnetic core influences both the mechanical as well as the magnetic fibre's properties. A macroscopic two-parameter structural model of a magnetic fibre or magnetic textile core, which considers mechanical and magnetic phenomena, is presented in figure 7.

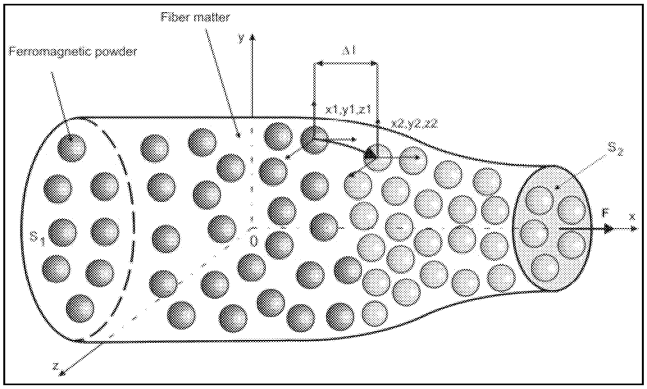


Figure 7: The segment of stretched fibre

The deformation of the magnetic fibre causes deformation of the crystalline net of the ferromagnetic, and at the same influences the changes of magnetic permeability Δμ. The tensions and deformations of the fibre cause a change of the geometric fibre parameters.
Figure 7 presents the move of granulated product particles (magnetic powder) together with matrix, where new and initial position of particles is marked. Thus, the spatial distribution of granulated product is changed. The mechanical properties of magnetic fibres are mapped in rheology model in figure 8. The analysis of phenomena has conducted up to fiber's microstructure in surrounding of magnetic grain. The model takes into account the mechanical proprieties of polymer matrix and modificator particles. Polymer matrix differs in springy properties from granulated product particles. Magnetic materials is characterized by large stiffness and small elasticity. The consequence of this interpretation is established rheology model.

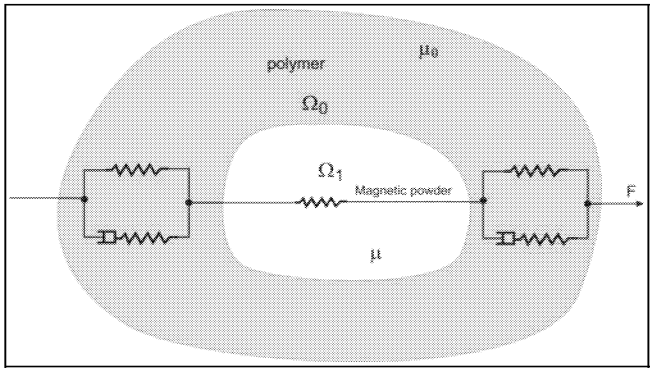
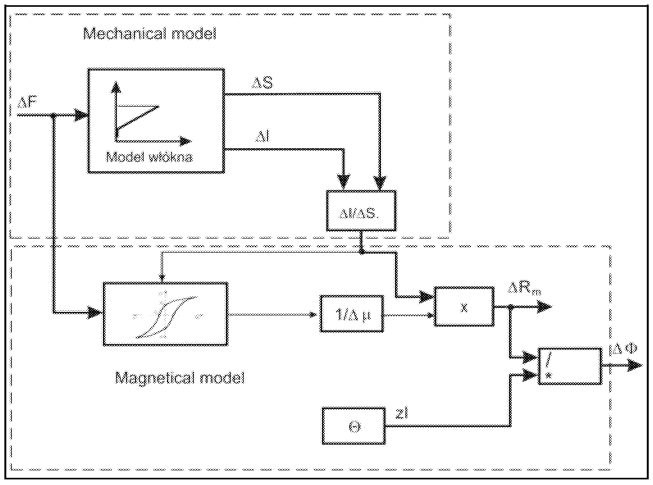


Figure 8: Rheology model of the magnetic fibre

Rheology model of fiber consists of magnetic powder and polymer matrix.
The changes in the fibre permeability and in the geometrical dimensions cause together a change in the fibres reluctance ΔR_m figure 9.



Figures 9: Magnetical and mechanical model of fiber; ΔF – change of force acting on the fibre; Δl – change of the fibre's length, ΔS – change of the cross-section area, Δμ - change of the magnetic permeability, ΔR_m - change o the fibre's reluctance [Zięba], [Rubacha M., Zięba J.]

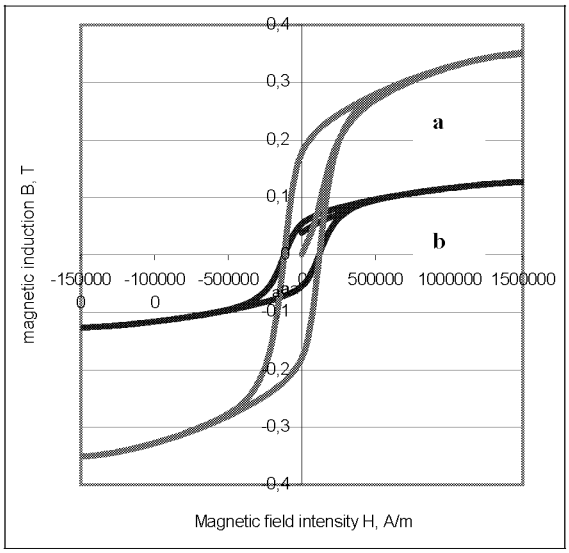


Figure 10: Magnetic loops of magnetising a) barium ferrite (hard magnetic), and b) fibres including 50 wt% (23 vol%) of barium ferrite, [Rubacha M., Zięba J.]

The fibres' magnetic properties are determined by the magnetic loops of magnetising presented in figure 10 for various percentage coefficients (by volume) of the magnetic material incorporated into the polymer matrix [Rubacha M., Zięba J. 2006].

ANALYSIS OF A TEXTILE MAGNRTIC CIRCUIT

An elementary textile magnetic circuit is a circuit, which consists of a monofilament magnetic fibre figure 11a with a coil wound around the fibre. A substitute scheme of the elementary circuit is presented in figure 11b.

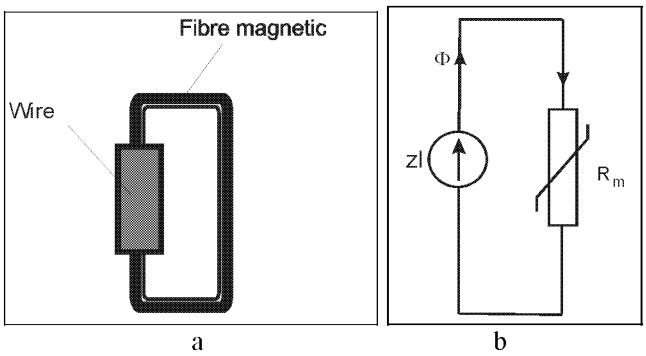


Figure 11: An elementary monofilament textile magnetic core (a) and its substitute circuit (b).

The elementary magnetic circuit is described by the following equation: (2)

$$z \cdot I = H \cdot l \qquad (2)$$

where:
z – the number of the coil windings,
I – the electrical current,
H – the magnetic field strength, and
l – the length of the magnetic circuit.

The textile magnetic coil with magnetic core, and metalic wires or electroconductive yarn as the winding [Zięba J. 2005], is presented in figure 12.

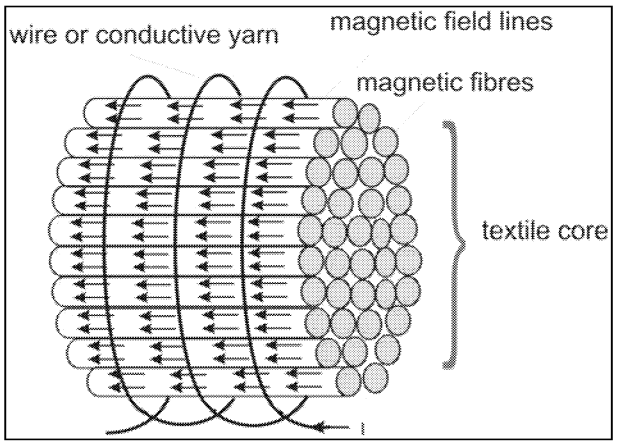
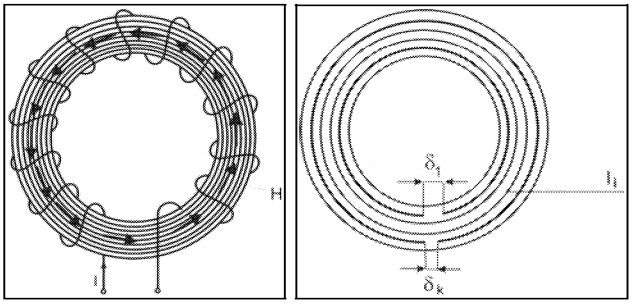


Figure 12: Schematic view of the design of a textile coil with magnetic core.

The representation of equation (1) is the calculation scheme presented in figure 14.
The shape of the magnetic core depends on the construction type of the electromagnetic textile device figure 13a.



a b

Figure 13: Textile magnetic cores; a) toroidal coil, b) magnetic coil with air-gap.

A textile magnetic core composed of magnetic fibres may be considered as a laminated core-assembly. A textile core after being formed in the shape of a torus or other form requires interlacement of the magnetic fibres in order to avoid formation of micro air-gaps. In laminated metal-sheet core-assemblies, the core's front surfaces are grinded in order to avoid micro-gaps, which unfortunately is not possible in textile cores, and therefore the fibres must be interlaced. As the magnetic fibres may have different lengths, a part of these fibres would not form a closed magnetic circuit, as the both ends of the particular fibres would not fit together or overlay mutually. This case is illustrated by figure 13b, where the core includes some fibres, which do not fit, and form micro air-gaps of various width (δ_1, , δ_2). This problem concerns not only magnetic cores of toroidal shape but also those of a typical rectangular form. The substitute scheme circuit of the magnetic core is presented in figure 14. The designifcali on substitute scheme circuit below:

where:

S – the substitute cross-section of the textile magnetic core,

S_i – the substitute cross-section of a single magnetic fibre,

B – the magnetic induction.

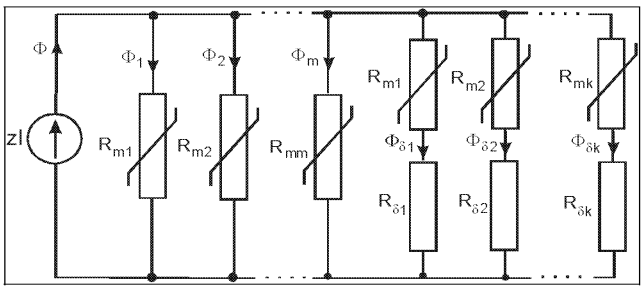


Figure 14: Substitute scheme circuit of the magnetic core of a textile coil.

The magnetic coil presented in figure 14b is described by equation (3):

$$z \cdot I = \sum_{i=1}^{m} H_i \cdot l_i + \sum_{i}^{k} H_{0i} \cdot \delta_i \qquad (3)$$

where:

H_{oi} – the magnetic field strength in the air-gap,

δ_i – the width of the air-gap,

R_{mm} – the magnetic reluctance of the m-th fibre,

$R_{\delta k}$ – the magnetic reluctance of the k-th fibre,

Φ_m – the magnetic flux of the m-th fibre,

$\Phi_{\delta k}$ – the magnetic flux of the k-th air-gap, and

Φ - the substitute magnetic flux of the textile magnetic core.

$$S = \sum_{i=1}^{m} S_i, \quad \Phi = \sum_{i=1}^{m} \Phi_i, \quad \Phi = B \cdot \sum_{i=1}^{m} S_i, \quad R_m = \Theta \sum_{i=1}^{m} \frac{1}{\Phi_i},$$

$$R_{mi} = \frac{l_i}{\mu \cdot S_i}, \quad R_m = \frac{1}{\mu} \sum_{i=1}^{m} \frac{l_i}{S_i}$$

In order to calculate the magnetic core's parameters it is necessary to known the core and fibre cross-sections, and therefore it was necessary to introduce the substitute cross-section S_i of the fibre. This causes that the substitute circuit of the textile magnetic core is composed from two circuits, one without a gap, and the second with an air-gap.

The average cross-section of a magnetic fibre with an $F_{12}BaO_{19}$ filler equals about 3.97×10^{-8} mm^2, and the cross-section of the textile magnetic core depends on the number of fibres.

The representation of the physical model is the simulation model realised by the Matlab-Simulink program presented in figure 15.

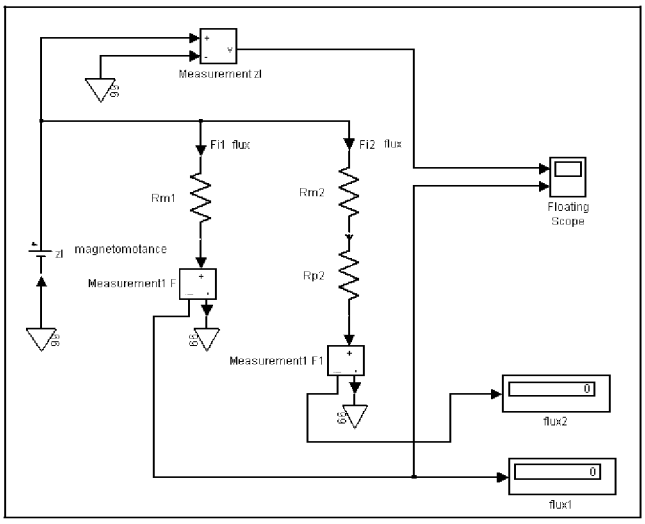


Figure 15: A simulation model of the magnetic circuit.

CONCLUSIONS

Magnetic fibres are fibres determined by their mechanical and magnetic properties, and are characterised by the following features:

- The magnetic properties depend on the kind of the implemented magnetic filler its and percentage content by volume in the fibre matter.
- Textile cores require an appropriate arrangement of the magnetic fibres in the coil casing, in order to avoid air-gap formation.
- It was proved that magnetic fibres can be elements of magnetic cores, which as parts of textile transmitters, gauges, and actuators may be broadly applied in textronic, intelligent clothing products.
- An analysis of the magnetic circuits enables the determination of its magnetic quantities and parameters.

References

Laszkiewicz, B., Rubacha, M., Czarnecki, P., Niekraszewicz, B.: Cellulose Fibres As Chemical Sensors, Conference Proceedings, 9th International Meeting on Chemical Sensors, Boston, USA, July 7-10, 2002, p. 393.

Laszkiewicz B., Czarnecki P., Kulpiński P., Niekraszewicz B., Rubacha M., Present status and future potential of smart fibers and fabric, VIII Int. Scien. Confer. November 2004, Łódź, Proceedings,
Gniotek K., Stempień Z. Zięba J., 'Textronic – a new branch of knowledge (in Polish)', Przegląd Włókienniczy, Nr 2, 2003.
Zięba J., Polish Patent Application ('Method of obtaining a magnetic coil destined especially for placing into textile products (in Polish)'.
Zięba J., 'Magnetic fibres (in Polish)', Proceedings of the 6th Seminar for 'Textronic, 'Measurement and autumation in textile engineering'. Konopnica, 2005.
Zięba J., 'Modelling of magnetic fibres (in Polish)', Proceedings of the 7th Seminar for 'Textronic, "Measurement and automation in textile engineering', Konopnica, 2006.
Rubacha M., Zięba J., Magnetic Textile Elements, Fibres & Textiles in Eastern Europe, Vol. 14, No 5(59) 2006, p. 48-52.
Taya M., Electronic composites, Modeling Characterization Processing, and MEMS application, Cambridge University Press, Cambridge 2005,

Acknowledgement
This work was supported by the Grant No 8 T10C 019 17.

Biography

JANUSZ ZIĘBA was born in Poland, and graduated with a M.Sc. degree in industrial automatic from the Technical University of Łódź, Faculty of Electrical Engineering. He attained a Ph.D. degree at the Textile Faculty in 1983, and at present is Senior Lecturer. His main scientific fields of interest are research into electrical, electronic, and control engineering and textronic. He is author and co-author of circa 60 publications.

TEXTRONIC MODEL OF CLOTHING CONSIDERING CHANGES IN ITS THERMOINSULATING PROPERTIES

Michał Frydrysiak, Stefan Ziegler, Krzysztof Gniotek
Department for Automation of Textile Processes
Technical University of Łódź
Żeromskiego 116, Łódź, Poland
E-mail: michal.frydrysiak@p.lodz.pl, kgniotek@p.lodz.pl

KEY WORDS: textronics, thermoregulation, temperature measurement, homeostasis, heat transfer model.

ABSTRACT

This work presents a model of the human – clothing system and describes the mechanisms of heat exchange which takes place in this configuration. Over the first stage of investigation, the authors have focused on the elaboration of a homeostasis model, i.e. the human's thermoregulation mechanism. This was the basis for further detailed considerations into the phenomenon of heat exchange between the human's skin and the textronics clothing layer positioned directly to the skin, which means that clothing with a complex internal structure but of uniform functional properties was considered. Therefore the detailed recognition of the phenomena and processes, which occur between the particular elements of the clothing system, will enable the design of more complex and detailed textronics models over further investigations. This work also presents quality-models of the objects considered, especially those concerned with the thermoregulation properties of humans and textiles being in contact with the human's skin. We assume that textronics elements implemented into textile structures may cause some consequences, such as for example local temperature changes which may influence the clothing users. Therefore the creation of detailed textronics models is important, which would enable the answer on the following question: does the incorporation of textronics elements into the textile clothing will not disturb the users comfort and cause essential temperature changes on the clothing surface and below the clothing.

HOMEOSTASIS – THE HUMAN'S THERMOREGULATION

A human's being maintains its organism's internal temperature independently on the changes of the environmental temperature. This process is called homeothermy, and its action consists in adaptation of the quantity of heat produced in the organism, as the result of metabolic transformations, to the quantity of heat transmitted between the organism and the environment, in order to achieve a thermal balance of the human's body system under changing environmental conditions (Traczyk 1989, Kozlowski 1999). The body's internal temperature is the result of the temperature of various internal organs, and under thermo-neutral environmental conditions, i.e. at a temperature of 21^0C, humidity of 50%, and air flowing with a velocity of 0.1 m/s equals about 36.6^0C. On the basis of literature dealing with physiology (Traczyk 1989, Korczak 1969), within the limits of the available knowledge, we identified the features of the object discussed, this is of the human's being. The input parameters, the disturbances, and the physiological parameters, which influence the skin temperature changes, were determined. It should be emphasised that the phenomenon discussed is complex and the object itself essentially non-linear. The changes occurring at the object's output, which means the skin temperature are measurable but cannot be controlled, as we have no possibility to interfere into the human's physiological processes (the homeostasis cannot be controlled). Therefore the investigation into textronic thermoregulation processes, where the possibility occurs of compensating the influence of disturbances in the layer between skin and clothing, is a great challenge. Figure 1 presents a quality-model of the human's thermoregulation (Frydrysiak, Gniotek 2004).

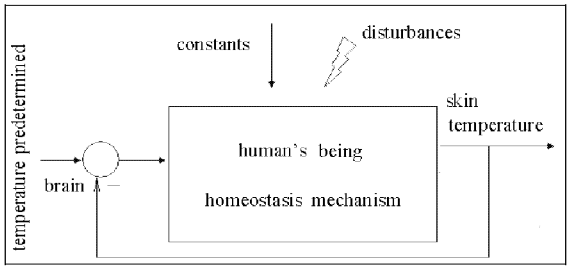


Figure 1: Quality-model of human's thermoregulation

The constants considered in Figure 1 are especially the dimensions and the mass of the human's body, as well as the thickness of the adipose tissue and the amount of perspired sweat, which may be approximately estimated for the particular case. The authors assumed in their work that the constants

mentioned above will be equal for all the model variants, and that the input quantities are independent.
On the basis of the above-mentioned information, a mathematical model of the human's being was created with the use of the Matlab Simulink program. The design assumptions provided the division of the human's body into 15 segments. In order to perform calculations with the possible greatest accuracy, the particular parts of the body were simplified and compared to simple geometric solids presented in figure 2 (Frydrysiak, Gniotek 2006). The model developed presents heat exchange by conduction, convection, radiation, and evaporation for a naked humanoid which is the initial heat source for further model considerations.

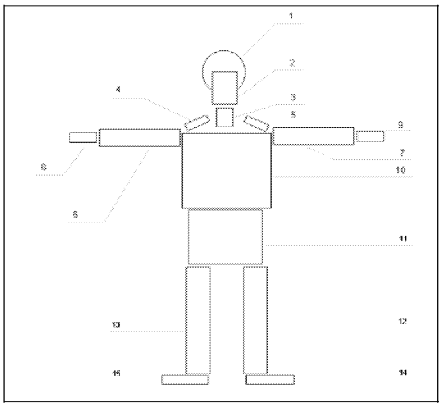


Figure 2: Scheme of the human's body division into geometrical solids, where: 1 head; 2 face; 3 neck; 4,5 shoulders; 6,7 arms; 8,9 hands; 10 thorax; 11 abdomen; 12,13; legs; 14,15 feet

From the point of view of heat exchange, we may consider textile products as permeable barriers of small thickness, and a specially formed surface (Żurek 1977). We assumed that clothing does not contact directly the human's skin with its whole surface, which results in the need to consider the existence of air channels in the layer between skin and clothing in which occur the phenomena of convection (q_{conv}) and heat radiation (q_{rad}); the simplified block scheme of the human – clothing system is presented in figure 3.

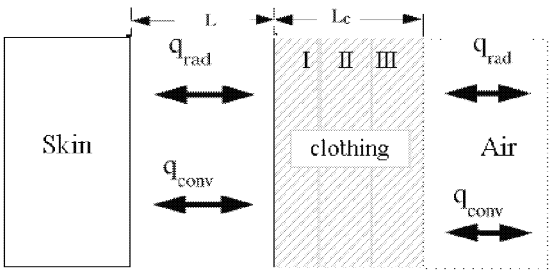


Figure 3: Simplified block scheme of the human – clothing system

where L is dimension of air-gap and L_c is width of clothing packet.
We also assumed that the particular layers of the clothing packet fit tightly one to the other. Only the implementation of an electronic element between the layers I and II causes that a local increase in the textile packet layer's thickness occurs. In turn, the thickness change results in a change of the conductivity. Determination of the preliminary assumptions necessary for designing the model is presented in figure 4.

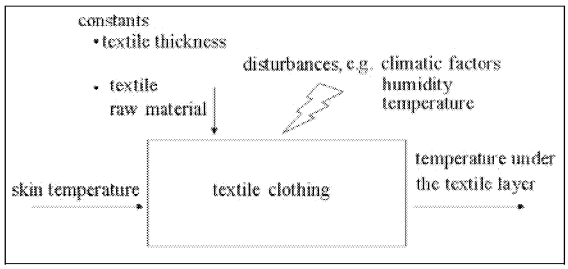


Figure 4: Quality model of the textiles' thermal properties.

On the basis of solutions presented previously, the authors propose a human – clothing quality-model which will be helpful for designing the model in the Matlab Simulink program.

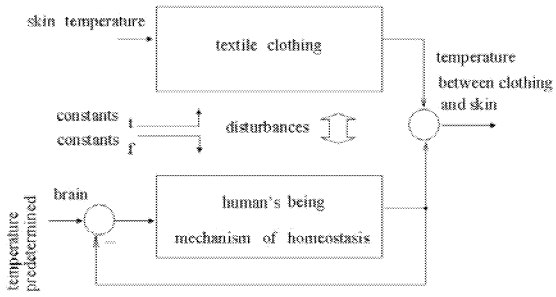


Figure 5: Combination of two quality models with constants ***t*** and ***f*** which are characteristic for textiles and the physiology of humans

EXPERIMENT

Within the scope of preliminary investigations, the basic properties of the component layers and the clothing packet were determined, and the coefficients of thermal conductivity & thermal resistance were assessed. The clothing packet is thermal resistant oriented. The following basic set of materials dedicated to applying in fire-fighter clothing were tested: I the outside layer – Delta T art. 8022 Nomex woven fabric with an area mass of 196 g/m^2 and a thickness of d = 0.46 mm; II the layer protecting against moisture – Fireblocker N2LY Gore Tex woven fabric with an area mass of 150 g/m^2 and a thickness of d=0.73 mm; III stitched cotton nonwoven, non-combustible finished, quilted with a polyester-cotton lining also non-combustible

finished, with a total area mass of 120 g/m^2 and a thickness of d = 2.25 mm. The tests were carried out with the use of a conductometer of the two-plate-Poensgen type. The temperature distribution on the sample surface was measured with the use of a DM53 radiation pyrometer equipped with a sensor from Cole Parmer Infrared Co. A heater was used as the heat source. It should be stressed that the following three factors influence the temperature under the clothing: the phenomenon of conduction between the textile layers, the presence of an electronic integrated circuit between the packet's layers, and the presence of feeding wires.

Table 1: Average values of the overall heat-transfer coefficient and thermal resistance of the textile packet tested

Nr.	kind of sample	$\bar{\lambda}$ W/m°K	$\bar{R}$ m^2/W°K
1	1	0,0273	0,0171
2	2	0,0257	0,0287
3	3	0,0251*	0,0904
4	1+2+3	0,0337*	0,1076

where *is the equivalent coefficient λ for an inhomogeneous layer

An assumption of summation of the influence of all three factors for designing the quality model was accepted, whereas the mutual influences between them were omitted.

MODEL OF THE TEXTRONIC CLOTHING PACKET

We assume that the textile packet tested is composed of three component layers which fit closely together as it is presented in figure 6,

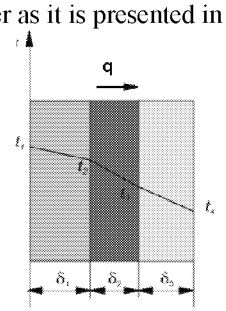


Figure 6: Scheme of the mutual positions of the clothing packet's layer

where:

$\delta_{1,2,3}$ –the thickness of the particular textile layers,

$\lambda_{1,2,3}$ – the thermal conductivity related to the particular textile layers,

t_1 – the temperature under the clothing,

$t_{2,3}$ – the temperature at the contact of the particular component layers, and

t_4 – the temperature on the surface of the textile packet.

We assume that the heat flux is the same for all three layers. On the basis of the Fourier law, we can write the equation which describes the heat conduction through the textile layers:

$$q = \frac{t_1 - t_4}{\frac{\delta_1}{\lambda_1} + \frac{\delta_2}{\lambda_2} + \frac{\delta_3}{\lambda_3}} \quad \text{,in W} \qquad (1)$$

where:

δ_i / λ_I – is the textile resistance of the i-th layer.

We can also calculate the unknown temperatures t_2 and t_3 according to the equations (2):

$$t_2 = t_1 - q\frac{\delta_1}{\lambda_1} \quad \text{, in } °C$$

$$t_3 = t_2 - q\frac{\delta_2}{\lambda_2} = t_1 - q\left(\frac{\delta_1}{\lambda_1} + \frac{\delta_2}{\lambda_2}\right) \quad \text{, in } °C \qquad (2)$$

$$t_3 = t_4 - q\frac{\delta_3}{\lambda_3} \quad \text{, in } °C$$

This simplify scheme described only textile pack. The more detailed model of textile clothing structures have to take into consideration thermal resistance of air between skin and fabric layer and heat transfer mechanism to surroundings, it is shows figure 3. The basic thermodynamic low are described this phenomenon by equations:

$$q_r = \varepsilon C_0\left[\left(\frac{273+t_4}{100}\right)^4 - \left(\frac{273+t_o}{100}\right)^4\right] \text{, in W} \qquad (3)$$

$$q_k = \alpha_k(t_4 - t_o) \quad \text{, in W} \qquad (4)$$

$$q = q_k + q_r \quad \text{, in W} \qquad (5)$$

where q_r is heat of radiation and q_k is heat of convection.The analyses of heat similarity criteria as a $Nu=(G_rP_r)^n$ (where Gr-Grashof number, Pr-Prandtl number, Nu-Nusselt number), could calculate heat convection coefficient α_k and formulate with earlier equations (1-5) formula (6).

During heat inflow from outside layer on radiation transfer mechanism isn't take into consideration

$$q = f[C(G_rP_r)^n + \varepsilon C_o\left[\left(\frac{273+t_4}{100}\right)^4 - \left(\frac{273+t_o}{100}\right)^4\right] \qquad (6)$$

temperatures of radiation body, equation 6. Heat flux q and temperature of clothing surface t_4 could be solved by equation number (1) and (6), where C and n are the constant. Farther consideration, authors make only for textile pack, be expressed by equation number (1). The calculations were carried out assuming that the heat source was placed outside the clothing, which means at the side of the layer III. The calculation model realised by the Matlab Simulink program is presented in figure 7. The temperature drop across the layers I-III was calculated for the packet with and without implementation of an electronic system in the form

of a thin-layer epoxy resin plate with the following parameters:

$$\delta_{pt} = 10^{-3} m,$$
$$\lambda \cong 0{,}2 \ \frac{W}{m \cdot K},$$
$$R_{pt} = 5 \cdot 10^{-3} \ \frac{m^2}{W \cdot K},$$

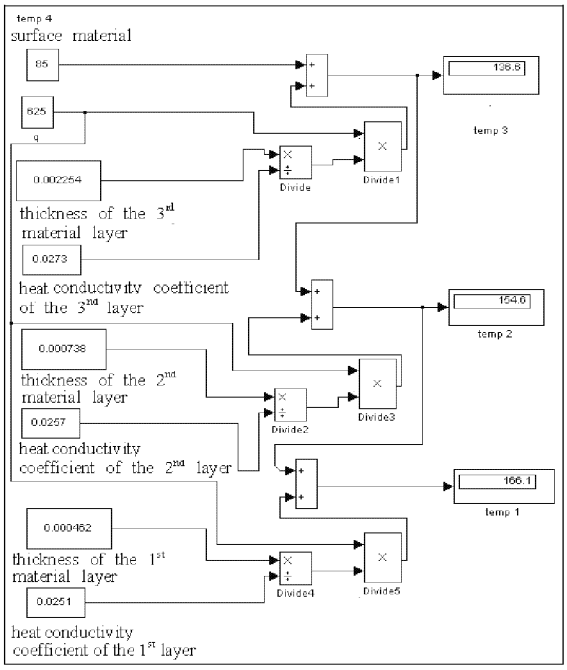


Figure 7: Calculation scheme of the temperature drop across the layers of the clothing packet

A digital LME temperature sensor and a J-type thermo element were attached to the plate. The calculation and measurement results are presented in Tables 2 and 3. The calculations were carried out on the assumption that an insignificant change of the heat flux occurred. Temperatures necessary to calculations ware temperature of out side layer t_4 and surroundings temperature t_o.

Table 2: The average temperature drop across the layers I-III of the textile packet without implementation of the electronic element

temperature difference $\Delta t_{14} = q \cdot R_{14}$	
calculated	measured
67,2 °C	80,3 °C

Table 3: The average temperature drop across the layers I-III of the txtile packet with implementation of the electronic element.

temperature difference $\Delta t_{14} = q \cdot R_{14}$	
calculated	measured
75,0 °C	81,0 °C

The temperature drop across the implemented plate itself was also calculated and is shown in Table 4.

Table 4: The average temperature drop across the implemented electronic plate

temperature drop across the plate $\Delta t_{pl} = q \cdot R_{pl}$	
calculated	measured
3,1 °C	1,5 °C

Conclusions

Such a model, as the type described by us may be a good tool to verify new generation clothing and to set designing guidelines.

The authors would like to state that the considerations presented herein have a cognitive character and consist of an introduction to farther, more detailed investigations. The results obtained look promising, whereas the estimation of the significance of the differences between calculation and measurement results will be the subject of our further works.

Biography

MICHAŁ FRYDRYSIAK graduated from the Technical University of Łódź, Textile Faculty with a M.Sc. degree in 2003. The M.Sc. thesis he finalized in the Department of Automation of Textile Processes where he also commenced his post graduate studies into the field of modelling and simulating textronic systems under the scientific supervision of Prof. Krzysztof Gniotek Ph.D., D.Sc. His knowledge into textronic he broadened during studies at the University of Ghent, Belgium, and the University of Liberec, Czech Republic. Since November 2006, M. Frydrysiak M.Sc., Eng. has been working at the Department of Automation of Textile Processes.

STEFAN ZIEGLER is a research worker in the Department for Automation of Textile Processes at the Technical University of Łódź. He is a specialist into the field of measuring the parameters of textile processes, especially those concerned with temperature.

KRZYSZTOF GNIOTEK, Prof., PhD, DSc, Head of Department for Automation of Textile Processes Technical University of Lodz. He is interested in measuremet science and systems, automation and process control, methodology of experimental research and textronics as a synergy connection of textile technology, electronics and computer science. Author and co-author of 3 monographs, 74 papers and reports and 44 patents.

REFERENCES

Bakinowska K., i inni, Pomiary Cieplne Część I, Wydawnictwo Naukowo-Techniczne, Warszawa1995

Frydrysiak M., Gniotek K. "MODELLING OF THE PHENOMENON OF HUMAN THERMOREGULATION" IMTEX conference 2004.

Frydrysiak M., Gniotek K. "A computer 15 node model of human thermoregulation", IX International Conference, Faculty of Engineering and Marketing of Textile 2006

Gniotek K., Stempień Z., Zięba J. 'Textronic, a new field

of knowledge , Przegląd Włókienniczy nr 2, Poland 2003.
Gniotek K., Krucińska I.: The Basic Problem of Textronics, Fibres & Textiles in Eastern Europe, January/March 2004. Vol. 12 No 1(45)
Gniotek K.: " Indentification Methodology of Some Textile Object Properties", Polish Academy of Sciences, Textile Commision, Lodz, 2004, chapter 9- Textronics
Kozlowski S., Nazar K., Wprowadzenie do fizjologii klinicznej , PZWL Warszawa 1999
Michalski L., Eckersdorf K., Pomiary Temperatury, Wydawnictwo Naukowo-Techniczne Warszawa
Traczyk W.Z. „Human phisiology script, IV edition " PZWL Poland 1989
Żurek W., Kopias K. Struktura Płaskich wyrobów Włókienniczych, WNT Warszawa 1977

MODELLING FOR EVALUATION AND PREDICTION

USERFRIENDLY PERMEABILITY PREDICTING SOFTWARE FOR TECHNICAL TEXTILES

B. Verleye*, G. Morren**, S.V. Lomov*, H. Sol**, D. Roose*
K.U.Leuven;**Vrije Universiteit Brussel
email: bart.verleye@cs.kuleuven.be

KEYWORDS
Composites, Permeability, CFD, RTM, LCM

Abstract

For the simulation of the injection stage of the manufacturing process of composite materials, the permeability value of the textile is an important input parameter. The permeability of textiles can be determined experimentally, however, this is a time and resource consuming process, and no standard procedure is yet available. In this paper we present an easy to use software package to simulate the flow through a textile model, and predict the permeability. Results are presented for a woven fabric and are compared with experimental results.

INTRODUCTION

Liquid Composite Moulding (LCM) is a rapidly developing manufacturing process. It involves: laying up of a textile reinforcement in a mould cavity of a desired 3D shape; injection of a liquid resin; polymerization (thermosets) or solidification (thermoplasts) of the resin (Parnas, 2000). The permeability of a textile is a key characteristic for composite manufacturing and is of particular importance for the simulation of the injection stage of LCM. The evaluation of textile permeability gained importance due to the often encountered problems of non-uniform impregnation, void and dry spot formation.
The permeability is a geometric characteristic related to the structural features of the textile at several length scales. Textile is a porous medium, so permeability can be defined using Darcy's law

$$\langle \vec{u} \rangle = -Re\underline{\underline{K}} \cdot \langle \nabla p \rangle \,, \tag{1}$$

with $\vec{u} = \vec{u}(x, y, z)$ the fluid velocity,Re the Reynolds number, $p = p(x, y, z)$ the pressure, $\langle\rangle$ volume averaging and $\underline{\underline{K}}$ the permeability tensor of the porous medium. Equation (1) is a homogenized equation, the information of the internal geometry of the reinforcement being hidden in K. Finite element or finite difference Darcy solvers require the input of K. Unfortunately, measurements of textile permeability are time and resource consuming (Hoes, 2003), hence reliable numerical prediction of K is required for the Darcy solvers.

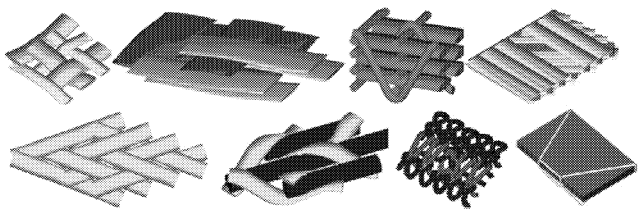


Figure 1: Examples of models of textile reinforcements: 2D woven; 2D woven laminate; 3D woven; UD laminate; 2-axial braid; 3-axial braid; weft-knitted; Non Crimp Fabric

TEXTILE MODELLING

A key task in permeability modelling is the characterisation of the reinforcement. For the creation of a single layer model of the reinforcement, we use the WiseTex software (Verpoest and Lomov, 2005). The software package WiseTex implements a generalised description of the internal structure of textile reinforcements on the unit cell level. The description integrates mechanical models of the relaxed and deformed state of 2D and 3D woven (Lomov et al., 2003, 2000, 2001), two- and three-axial braided (Lomov et al., 2002b), weft-knitted (Moesen et al., 2003) and non-crimp warp-knit stitched (Lomov et al., 2002a) fabrics (NCF) and laminates (Lomov et al., 2002c) (Fig. 1). All these models, including the models of deformed fabrics, use a unified description format of the geometry of the reinforcement unit cell. This format allows the calculation of physical and mechanical parameters of the fibres near an arbitrary point in the unit cell. The reader is referred to (Verpoest and Lomov, 2005) for more details.

MATHEMATICAL APPROACH

To compute the permeability via the law of Darcy (1), we solve the Stokes equations

$$\begin{cases} -Re\nabla p + \Delta\vec{u} = \vec{g} \\ \nabla \cdot \vec{u} = 0 \end{cases} , \tag{2}$$

in a unit cell of a textile modell (Fig. 2). Periodic boundary conditions are imposed in XYZ-direction for the velocity and pressure. The design engineer can

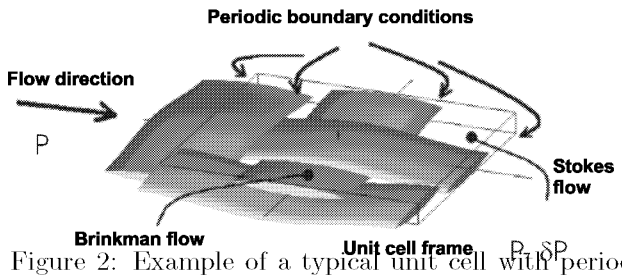


Figure 2: Example of a typical unit cell with periodic boundary conditions

choose to put wall conditions in one or two directions. As driving force, a global force $\vec{g}$ or a constant pressure drop ∇p_0 can be imposed.
If one wants to take the micro flow inside the yarns into account, a penalty term K_{yarn}^{-1} is added to the stokes equations which results in the Brinkman Brinkman (1947) equations:

$$\begin{cases} -Re\nabla p + \Delta\vec{u} - K_{yarn}^{-1}\vec{u} = \vec{g} \\ \nabla \cdot \vec{u} = 0 \end{cases} . \quad (3)$$

The software package FlowTex, is based on the freely available finite difference Navier-Stokes solver, NaSt3DGP, developed at the institute for Numerical Simulation of the University of Bonn (Griebel, 2004; Griebel et al., 1998). The Stokes solver uses the PETSc library (Balay et al., 2004, 2001, 1997), to solve the resulting system of discretised equations.
As the solver is based on a finite difference discretisation, no meshing problems arise in the preprocessing. That allows the design engineer to neglect the mathematics of the solver, and to compute the permeability with one button (Figure 3). The more advanced user can also use the software on Unix clusters (without GUI), which allows to perform many experiments at once.

EXPERIMENTAL SETUP

The experimental validation is performed with a highly automated central injection rig, called PIERS set-up. This PIERS (Permeability Identification using Electrical Resistance Sensors) set-up consists of a mold cavity with two sensor plates, each containing 60 electrical sensors (Fig. 4). After placing the reinforcement and closing the mold, the test fluid can be injected. This is done centrally in the reinforcement through a hole in the middle of the lower sensor plate. While the flow front propagates through the reinforcement, the fluid flow makes contact with the electrical sensors. Since an electrical conductive fluid is used, the wetting of these DC-resistance sensors will change their electrical resistance. This variation is registered and hence an arrival time for the sensors can be stored.

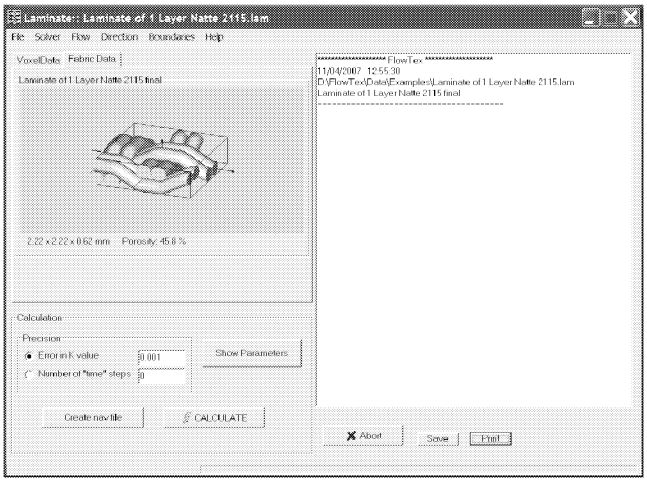


Figure 3: The GUI of FlowTex

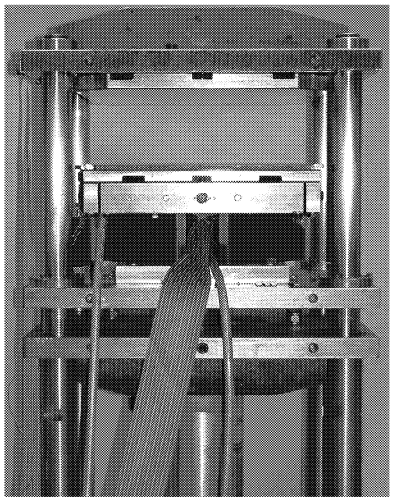


Figure 4: The PIERS set-up

Model	Permeability (mm^2)
Single layer	3.44 E−04
Random nesting	3.24 E−04 ± 1.03 E−04
Maximum nesting	1.54 E−04
Experimental	2.7 E−04 ± 10%

Table 1: Results for the Monofilament Fabric with different nesting

From the stored date, the experimentally determined permeability is computed with the inverse method (Morren et al., 2006).

VALIDATION

The Monofilament fabric Natte 2115 is a realistic structure which is close to actual textile reinforcements, and for which permeability is experimentally validated. The fabric consists of fibres which are coated with a PVC coating, so micro flows are eliminated. The full description of the Monofilament Fabric Natte 2115 test-fabric can be found in (Hoes, 2003; Hoes et al., 2001). Table 1 compares the results of calculations on the single layer model, the two layer model with maximum nesting, the two layer model with random nesting, with the experimental result. The permeability of the model with random nesting is the average of 15 calculations on models with different random nesting.
The experiments on the Natte fabric where performed with several layers of Natte in order to fit into the experimental mold. The layers are compressed in the mould cavity, which leads to a random nesting. The random nesting parameters are however not fitted to the laying up process. In table 1 we see that the experimental value is in between the single layer and the maximum nesting value, which is an acceptable result. For the Natte structure calculations with a fine discretisation ($\Delta x = 0.02mm$) took about one minute on the K.U.Leuven high performance cluster.

CONCLUSIONS

In this paper we presented the permeability predicting software FlowTex. Textile models are designed with WiseTex and then imported into FlowTex. A fast finite difference solver is used to solve the Stokes equations and compute the permeability. The software is efficient and userfriendly for non CFD-experts.

ACKNOWLEDGEMENTS

This research is part of the IWT-GBOU-project (Flemish government): Predictive tools for permeability, mechanical and electro-magnetic properties of fibrous assemblies: modelling, simulations and experimental verification
The authors thank Prof. Griebel, M. Klitz and R. Croce of the University of Bonn for the interesting discussions and the collaboration on the adaptation of the NaSt3D-code Griebel (2004).
This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its author(s).

References

Balay S.; Buschelman K.; Eijkhout V.; Gropp W.D.; Kaushik D.; Knepley M.G.; McInnes L.C.; Smith B.F.; and Zhang H., 2004. *PETSc Users Manual.* Tech. Rep. ANL-95/11 - Revision 2.1.5, Argonne National Laboratory.

Balay S.; Buschelman K.; Gropp W.D.; Kaushik D.; Knepley M.G.; McInnes L.C.; Smith B.F.; and Zhang H., 2001. *PETSc Web page.* Http://www.mcs.anl.gov/petsc.

Balay S.; Gropp W.D.; McInnes L.C.; and Smith B.F., 1997. *Efficient Management of Parallelism in Object Oriented Numerical Software Libraries.* In E. Arge; A.M. Bruaset; and H.P. Langtangen (Eds.), *Modern Software Tools in Scientific Computing.* Birkhäuser Press, 163–202.

Brinkman H., 1947. *On the permeability of media consisting of closely packed porous particles. Applied Scientific Research*, 1, no. 1, 81.

Griebel M., 2004. *http://wissrech.iam.uni-bonn.de/ researchprojectsNaSt3DGP.*

Griebel M.; Dornseifer T.; and Neunhoeffer T., 1998. *Numerical Simulation in Fluid Dynamics, a Practical Introduction.* SIAM,Philadelphia.

Hoes K., 2003. *Development of a new sensor-based setup for experimental permeability identification of fibrous media.* Ph.D. thesis, Vrije Universiteit Brussel.

Hoes K.; Dinesku D.; Vanhuele M.; Sol H.; Parnas R.; Belov E.; and Lomov S., 2001. *Statistical distribution of permeability values of different porous materials.* In H. Sol and J. Degrieck (Eds.), *10th European Conference on Composite Materials (ECCM-10).*

Lomov S.; Belov E.; Bischoff T.; Ghosh S.; Chi T.T.; and Verpoest I., 2002a. *Carbon composites based on multiaxial multiply stitched preforms. Part 1: Geometry of the preform. Composites part A*, 33, no. 9, 1171–1183.

Lomov S.; Chi T.T.; Verpoest I.; Peeters T.; Roose D.; Boisse P.; and Gasser A., 2003. *Mathematical modelling of internal geometry and deformability of woven preforms. International Journal of Forming Processes*, 6, no. 3-4, 413–442.

Lomov S.; Gusakov A.; Huysmansa G.; Prodromou A.; and Verpoest I., 2000. *Textile geometry preprocessor for meso-mechanical models of woven composites. Composites Science and Technology*, 60, 2083–2095.

Lomov S.; Huysmans G.; Luo Y.; Parnas R.; Prodromou A.; Verpoest I.; and Phelan F., 2001. *Textile Composites Models: Integrating Strategies. Composites Part A*, 32, no. 10, 1379–1394.

Lomov S.; Nakai A.; Parnas R.; Ghosh S.B.; and Verpoest I., 2002b. *Experimental and theoretical characterisation of the geometry of flat two- and three-axial braids. Textile Research Journal*, 72, no. 8, 706–712.

Lomov S.; Verpoest I.; Peeters T.; Roose D.; and Zako M., 2002c. *Nesting in textile laminates: geometrical modelling of the laminate. Composites Science and Technology*, 63, no. 7, 993–1007.

Moesen M.; Lomov S.; and Verpoest I., 2003. *Modelling of the geometry of weft-knit fabrics.* In *Tech-Textil Symposium.* Frankfurt, CD–Edition.

Morren G.; Gu J.; Sol H.; Verleye B.; and Lomov S., 2006. *Stereolithography specimen for calibrating the permeability measurements of fibrous reinforcements for RTM. Advanced composites letters*, 119–125.

Parnas R., 2000. *Liquid Composite Molding.* Hanser Publishers, Munich.

Verpoest I. and Lomov S., 2005. *Virtual textile composites software Wisetex: integration with micro-mechanical, permeability and structural analysis. Composites Science and Technology*, 65, no. 15-16, 2563–2574.

MODELLING THE AGILITY IN THE APPAREL INDUSTRY

Antonela Curteza and Adela Florea
Technical University of Iasi
B -dul D. Mangeron, Nr. 53,
70050, Iasi, Romania
E-mail: acurteza@tex.tuiasi.ro

KEYWORDS
agility, design, modeling, quantification, apparel firms

ABSTRACT

The enterprises' agility, considered as being the capacity of delivering a response "just in time" to the market changes, corresponds to the nowadays reality of globalisation and to the sine qua non condition of the companies' existence. therefore, the modelling of the enterprises' behaviour towards turbulences represents a permanent preoccupation for the theoreticians as well as for the managers. In order to meet the demands described above, the collective involved in this project proposed a way to model and assess the agility. this action implied the resolution of two issues: the establishment of the subsystems that define agility and the identification and differentiation of the specific indicators for each subsystem.

The instrument chosen was the fuzzy clips because the subsystems were expressed both numerically and attributively and used for the fuzzy technical modelling. The quantification of the indicators by means of the fuzzy technique has raised a number of difficulties, which constitute the problematic of the present paper.

INTRODUCTION

A preliminary study of the foreign companies' web sites as well as of the specialized literature (see attached bibliography) provided the following information:

- the "agility" concept is relatively new, dating only from 2001 and it refers to the capacity of the companies to react in an environment characterised by variable turbulences;

The thematic is new for the textiles and apparel industry and, in our projects, it was oriented [Curteza, A] towards:

- underlining the way in which an agile enterprise, with non-automated processes, reacts in an environment with variable turbulences, and not only for extreme turbulences;
- finding new solutions in the case of maximum, extreme turbulences, in order to stimulate the capacity of innovation.

DIMENSIONS (SUBSYSTEMS) INVOLVED IN APPRECIATING THE AGILITY OF AN ENTERPRISE AGILE FIRM.

The essential concept of agile fabrication is the integration of the organization, the people and the technology in a coordinated interdependent system that responds quickly to changes. In this section, we reproduce from literature the concept of firm's agility and the essential elements that lead to it. In Figure 1, a number of dimensions of the agility are identified, as well as some characteristics ("pressures") of the environment in which the firm develops its activity.

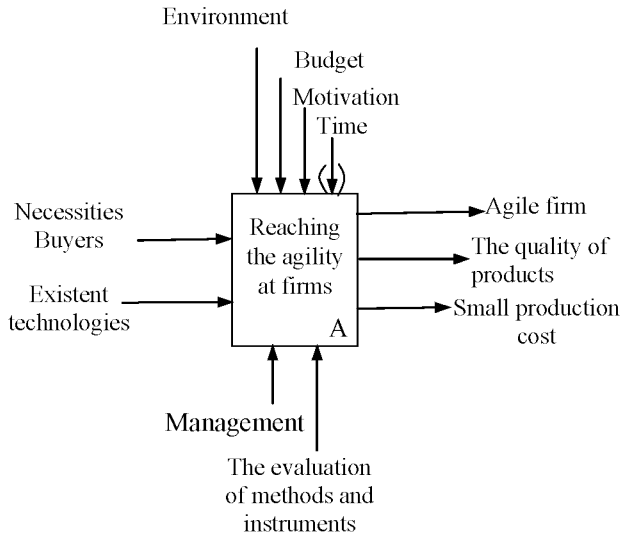


Figure 1: Reaching the first firms agility level (A-0 level), [after Siew-Pang Wong and Larry Whitman, Attaining Agility at the Enterprise Level.]

According to these authors, the quality of products and services is not a parameter of the agility, but a parallel result to agility, like the low production cost. The management and the evaluation capacity are parameters that have a direct influence on the achievement of agility [Wong, S.,P., Whitman, L.]. In fact, the different vision of diverse corporations and companies, and the non-unitary definition of this concept in the specialty literature creates the image of a theoretic concept, and not practically applicable.

THEORETICALLY RESULTS OBTAINED CONCERNING AGILITY QUANTIFICATION AND SIMULATION

The obtained theoretical results show that:

- agility is considered a multidimensional measure, because many of its characteristics cannot be reduced to just one characteristic, furthermore, they are irreducible one at another;
- the modelling of the enterprise agility has been made only by means of four subsystems;
- the modelling using fuzzy logic is the only one noticed;
- there were observed various theories and random applications of the concept in automated industries (please refer to the attached bibliography);
- the importance of expert systems has been revealed for the collection of basic principles and for the evaluation of its applications, as being an important asset in the assessment of the production subsystem agility.

DIFFICULTIES WHEN QUANTIFYING THE APPAREL FIRMS' AGILITY

The necessity of simulating the agility in the apparel industry brought about theoretical limitations and specific difficulties regarding the quantification of the apparel firms' agility, presented as below.

Theoretical limitations:

- the thematic approach was done only from a technical point of view, the reduced production cost being considered a parallel result to agility;
- the overwhelming influence of design and research on a company's agility has not been discussed;
- the aspects of engineering knowledge are considered essential for agility, due to the sustained activity on the market of knowledge processing programs, a limitation which insists upon cognitive aspects of agility; the nature of the inclination is commercial.
- mentions are made neither about the lack of waiting domain, nor about the correlation between response times and degree of context surprise, which leads to quite vague definite demands.

Problems about quantification of apparel firms' agility consist in:

- the necessity of establishing a set of parameters specific for the apparel industry. In Table 1 there are presented measured parameters mentioned in the specialty literature [Nikos C. Tsourveloudis], as well as parameters proposed by the authors of the present study for evaluating the agility of the design infrastructure;
- lack of a one-to-one correspondence between the agility factors and the physical characteristics of the enterprise. As it can be seen in Table 1, the defining factors for enterprise agility are expressed by means of various elementary physical features. The result consists in an unsubstantial behaviour of certain parameters involved in the quantification of agility;

Table 1: Proposed measurable parameters

Subsystem	Parameters
Production	"Changing effort","Versatility" "Range of adjustment or adjustability" "Substitutability", "Operation Commonality" "Job variety","Part variety", "Part commonality"
Market	"Re-configurability","Modularity index","Expansion ability","The market for reachable volumes "
Personal	"Training level","Rotation (flexibility) of the workspace"
Information	"Interoperability", "Networking"
Design	"Complexity and size of the apparel design team", "Number of collections proposed by a firm/year","Dimension and complexity of the collection", Value of the information used for developing the collections","Level of creativity/innovation – the quality of the ideas","Level of originality of products", "Sales volume of the previous collections", "Number of constant consumers", "Level of satisfying the consumers' demands"

- the percentage variation of the indicators involved in the quantification of agility. The challenge when quantifying of the agility measurements is given by the fact that the involved parameters are not homogenous in percentage, the variation depending mainly on the development or maintenance of each enterprise on the market strategy;
- different temporal spans for collecting primary information. The information is obtained by processing the elementary data recorded in the registry or in computers. A part of this data precedes the fabrication process, a second part is being created during the fabrication and another part is obtained after the fabrication of the apparel product, following the lot to the client or remaining in the factory's archive. This implies a post-factum correlated processing of all data; the obtained information can be practically used for evaluating the agility of another manufacture cycle;
- different types of information that must be correlated. Even though many quantification methods are known, the application in apparel industry is restricted by the particularities of the products. So, the apparel industry presents a larger number of characteristics than the products of other industries, which can be evaluated by subjective analysis. It is the case of both aesthetical and comfort characteristics and some characteristics that manifest during the use of the product. Moreover, because the manufacturing systems are managed by human beings, it is possible that most of the data and information to be delivered by the persons working at

the executive level; they convey the information in words, sentences and phrases. This is why it is necessary to record and use human knowledge and perceptions when it comes to agility and its factors.

MODELLING THE DESIGN SUBSYSTEM, A STEP TOWARDS SIMULATING THE AGILITY

In the apparel industry, a definitive role is held by the Design subsystem. Why design? Because design matters. So, in the heart of innovation lies creativity and the ability to put together technical know-how and an understanding of the user's needs; in other words, creativity and innovation are the engines behind design, and design is a strong component of an agile company [Bruce, M., Bessant, J.].

The modelling of the design subsystem by applying the fuzzy theories

The engineering of the fabrication systems doesn't dispose of enough analytical solutions or strictly mathematical forms, failing to treat in a unitary way the variable dimensions of the agility. This is why, for the purpose of not wasting data and the information source in defining the total agility, various methods are applied to use the non-numerical dates, like the fuzzy theories. The fuzzy technique was used in order to model the total agility, taking into account the fact that the indicators of the analyzed subsystems are expressed both attributively and numerically. The used software was Fuzzy Clips [Teodorescu H., N.].

The parameters for the design subsystem

At first, the agility can be characterized by taking into account The Design Atlas Tool [Bruce and Bessant, 2002], using a series of parameters (characteristics) that form a base.

An analysis of these characteristics shows that they are referring to the following aspects of agility:

- apparel design team (one direct parameter: Complexity and size of the apparel design team, and two indirect parameters: Level of creativity/innovation – the quality of the ideas and the Level of product originality);
- apparel collections (characteristic parameters: Number of collections proposed by a firm/year, Dimension and complexity of the collection, Value of the information used for developing the collections, Level of creativity/innovation – the quality of the ideas and Level of product originality);
- proposed method for covering the market (characteristic parameters: Sales volume of the previous collections, Number of constant consumers and the Level of satisfaction of the consumers' demands).

Instead of the traditional ways of solving the problem, which refer especially to the organizational and structural agility of the enterprise, the agility of the teams or of the work teams from the enterprise, the proposed way focuses on agility on the market.

The quantification and fuzzification of primary parameters in design

- *Complexity and size of the design team*

This characteristic includes the three enumerated parameters that relate to the team, and also characterizes the capacity of the team to be agile. The parameters can be evaluated by linguistic degrees such as: the complexity (of the team): big, medium, acceptable, small. The attribution of the linguistic degrees mentioned above and their probable evaluation by using belonging functions remains at the expert's latitude.

- *The number of collections proposed by a firm per year*

This parameter can be kept numerical in nature, but is brought down to an interval of fixed values, through normalization. For example, normalization can be made by dividing it by the maximum known number of possible collections. Another alternative is transforming the number of collections – which can be very high – into a smaller number, through logarithms:

$$C_{2\,\text{collections/year}} = \log(n_{\text{colections/year}}) \qquad (1)$$

A third alternative is to apply a function with limitation, for example a typical sigmoid function:

$$C_2 = -0.5 + \frac{1}{1+e^{-an}} \qquad (2)$$

The final alternative is to fuzzify this parameter (table 2):

Table 2: The connection between the numerical interval and the linguistic degree

Numerical interval	0-2	2-5	6-8	9 –15	>15
Linguistic degree	Very small	Small	Medium	Big	Very big

The correspondence of belonging functions can be chosen by triangulating, with the exception of the functions for very small and very big (figures 2 and 3).

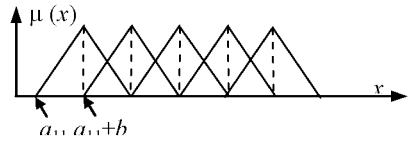


Figure 2: The graphic of five triangular belonging functions, over positioned two by two

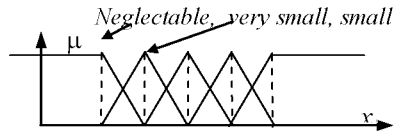


Figure 3: The graphic of five belonging functions from which three are triangular and two of go up and go down type, over positioned two by two

- *Dimension and complexity of the collection*

Dimension and complexity can be specified through attributes like: full, almost full, partial, not full enough and separated. The estimation of the dimension and of the complexity can be made numerically; another alternative is that the estimation of the collection's dimension and complexity is made by an expert, through fuzzy degrees.

- *The value of the information used in developing the collections*

The value of the information is difficult to be estimated. The evaluation made by linguistic degrees and fuzzy values is justified. A realistic evaluation alternative is:
"The information used is complete enough, with a trust degree of 0,6 AND
The information used is not complete enough, with a trust degree of 0,3".

- *Level of creativity/innovation*

The creativity level also cannot be numerically evaluated. The evaluation with linguistic degrees and with fuzzy values is justified as above like that:
"Creativity is high, with a trust degree of 0,8 AND
The information used is medium, with a trust degree of 0,3".

- *The level of product originality*

This evaluation can be done through 3-5 linguistic degrees, as chosen by the human expert.

- *The volume of sales of the previous collections*

This parameter is purely numerical and might be identified with the number of proposed collections per year. The parameter can be expanded on each collection and then an average can be calculated.

- *The number of constant consumers*

This parameter, although it is typically numeric, can be very well interpreted by using linguistic degrees and by fuzzification. Attributes often used for such a parameter are: number of constant consumers: negligible, very low, low, very big.

- *The level of satisfying the consumers' demands*

This is a parameter that can very easily be fuzzified. Furthermore, it is a difficult parameter; the task of being treated through binary or probability logics is almost unrealistic. Linguistic "natural" attributes are: very low, low, medium, big, very big or similar (eventually unsatisfying or totally satisfying the demands, which correspond to the limit of binary logics).

RESULTS OBTAINED BY MODELATING THE DESIGN SUBSYSTEM

The above information shows that the system of parameters used in characterizing the agility is big, each parameter having a relatively big number of linguistic degrees (five or more) after linguistic evaluation and fuzzification. This leads to a complex system of fuzzy rules, with a number of rules of about 50 - 100, hard to manipulate and hard to describe even by human experts.

The difficulty comes from using a low number of linguistic degrees for conclusions (5, 7 or 9), low compared to the number of rules. The result (the conclusion) will be the same for many sets of premises (with different linguistic degrees of the premises).

Therefore:

-in order to eliminate the confusions that could occur and to maintain the integrity of the information, we recommend giving up the method used until now, that is considering agility a scalar, and to consider agility a vector (of linguistic degrees, or fuzzy, a belonging functions vector), with all its components,

-we also propose the grouping (clustering) and setting the hierarchy of these parameters, to create coherent rules.

METHODS USED FOR SELECTING THE AGILITY'S PARAMETERS AND THEIR TRANSPOSAL INTO COMPARABLE UNITS

Two concepts were applied when selecting the involved parameters: the integrating concept and the selective concept, which means selection of some features of vital importance for the agility.

In the situation of the selective concept, when the target of the examination consists of a series of measurable characteristics, representative for a sub-system of agility, the analysis is easy, numerous practical methods of solving the problem already being known. For example, in the case of selecting only one representative, measurable characteristic, the easiest method was to record graphically, usually in chronological order the variation of the values for the measurable characteristic. When selecting more independent measurable characteristics, the statistical processing of the data and the mathematical modelling was used.

As a result, in the practical quantification of the data, the following numeric methods have been used: multi-criteria evaluations by calculating some statistic indicators (the arithmetical mean, the geometrical mean, the harmonic mean, the polygonal area, evaluation through the relative complex indicators, evaluation through point attributing, evaluation through optional indicators, evaluation using equal and differential ranks, evaluation by applying the informational entropy method). It has been ascertained that:

- the engineering of the manufacturing systems doesn't dispose of enough analytical solutions or rigorous mathematical forms, even in the easiest possible cases;
- the algebraic formula fail at the reunification of the variable dimensions of agility, in correlation with the human perception of agility .

The mathematical modelling implies both processes of phenomena investigation and optimization. The latter one is based on variables selection, on choosing the issues that must be experimented, on the statistical analysis of the data, on obtaining the mathematical models and the result

interpretation. Mathematical modelling is characterized by a high degree of uncertainty in the apparel industry, given by: modelling is difficult to be applied, once because of the multitude of factors that are mutually-conditioned and also because the evaluation by partially-subjective methods of certain parameters of the sub-systems.

For the success of modelling-related problems, a sequential approach of the sub-systems is required. So as not to waste the informational deposit obtained by using linguistic values - such as: low, medium, high – methods for processing non-numeric data were applied. The data collecting has been achieved by means of interview and investigation. The correct selection of the sample, the way of processing the results and the test for concordance between opinions were also extremely important. The results are processed by means of statistical and mathematical methods. Thus, the expert-sequential systems are considered as a fast instrument for modelling and assessing the textile and apparel companies' agility.

By using experts when applying this method it has been ascertained that:

- the success of the application of this method depends on the qualification and skills of the expert group;
- the mathematical processes required for the validation of the results are numerous and require, in order to facilitate the work, the usage of certain statistics software.

In the case of immeasurable characteristics, when the integrator concept is used, it is taken into consideration the fact that it is being operated with both measurable and attributive data, which can be divided into detectable and undetectable (it can be measured only at the end of the operation cycle for the specific characteristic); for example, calculating the degree of occupation of equipment, the average tact for a technological line, the real productivity of the same operation and so on.

The work methodology, in these cases, involved three stages:

-transforming the undetectable characteristic into a detectable characteristic;

-analysis and synthesis of the previously-known data, in order to draw a correlation between the qualitative obtained results and the estimation of the levels influenced by the process variables;

-making certain experiments which lead to the identification and hierarchy of the causes and solutions for the problem.

The following ways of transforming an undetectable characteristic into a measurable characteristic are known:

- measuring the effect induced by the characteristic during the development of the process;
- obtaining data by estimation, directing the decisions before the development of the manufacturing process.

The option for one of the above-stated methods has to be made in correlation with the importance of the undetectable characteristic and with the investigation target. In the case of characteristics that cannot be transformed into detectable ones, if the target can be achieved by reducing them to a subsystem based on similar characteristics resulted from similar processes, the used methods are the statistical and modelling ones. In the case of characteristics that cannot be transformed into detectable ones and contain data based on new information or information that is drawn at the end of the process, the used methods require multidisciplinary examinations and are approached by the "gurus" of the domain.

CONCLUSIONS

In this research we brought a number of clarifications about the general concept of enterprises' agility, which permit a better grounded theoretical analysis of the capacity of an enterprise to be agile.

When it comes to the agility dimensioning for the apparel firms, we clarified some characteristics that allow the determination of the agility, by considering the information from the specialty literature as a point of reference.

The essential contribution is represented by the inclusion of the *Design infrastructure* in the total agility of an enterprise and by the way in which this influences the estimation of the total agility.

The fuzzy technique was used in order to model the total agility, taking into account the fact that the indicators of the analyzed subsystems are expressed both attributively and numerically.

The paper also revealed the main difficulties related to the quantification of the agility of apparel enterprises.

REFERENCES

Bruce, M., Bessant, J., Design in Business .Strategic Innovation Through Design, *Prentice Hall, Pearson Education Limited,* UK, 2002;

Curteza, A., Florea, A., The Apparel Firms Agility, *AGIL texDESIGN, 1st Internation Workshop DESIN-INNOVATION-DEVELOPMENT*, July, 2006, Romania;

Tsourveloudis, N., C,. Knowledge-Based Measurement of Enterprise Agility, *P1: KMX/JXJ P2: KMX/GXL QC: KMX/GXL T1: GXL KB016A-03 KB016/Leondes KB016-LEONDES-v1.cls* April 13, 2004;

Wong, S.,P., Whitman, L., Attaining Agility At The Enterprise Level, *Proceedings of The 4th Annual International Conference on Industrial Engineering Theory, Applications and Practice*, November 17-20, 1999, San Antonio, Texas, USA.

LATE PAPER

Secondary nucleation study of sucrose crystallization: experiments and modelling

S Ouiazzane[(1)*], B.Messnaoui[(1)], J Wouters[(2)], T.Bounahmidi[(1)]

(1) Laboratoire d'Analyse et Synthèse des Procédés Industriels (LASPI), Ecole Mohammadia d'Ingénieurs, B.P: 765 Université Mohammed V Agdal -Rabat Maroc
(2) Faculté Universitaire de N.D de la Paix, Département de Chimie, Laboratoire de Chimie Biologique Structurale (CBS), Rue de Bruxelles, 61 Namur B-5000, Belgique.

* Courriel: ouizzanesanae@yahoo.fr. Tel :0021260896490

Abstract:

This work is devoted to study the effect of impurity on the the quality of sucrose crystallization in bach crystallizer. The liquid phase concentration in crystallizer and a supersaturation profiles was measured by using a refractometer. The logiciel LUCIA "Laboratory Universal Computer Analysis Image" was used to determine the Crystal Size Distribution (CSD) by analysing the obtained image by optical microscope and Scanning Electron Microsopic (SEM). A non-linear optimization method is applied to estimate the growth and nucleation rate parameters from batch cooling crystallization experiments. The calculated results are in agreement with those observed and determined by LUCIA. The effect of the impurties on the mean particle size of the final product is discussed.

Keywords: *Nucleation; secondary nucleation; crystallization; sucrose; modeling*

Introduction

Crystallization is a solid–liquid separation process, where molecules is transferred from a solute dissolved in a liquid phase to a solid phase. The molecule is governed by nucleation and growth rates. This is known as crystallization kinetics. In sugar manufacturing, crystallization is a crucial step upon which the rate of sugar extraction depends. This extraction rate depends primarily on syrup quality and the efficiency of the purification unit. For a given syrup quality, optimization of extraction and processing rates in a sugar factory requires efficient monitoring of the crystallization unit. This in turn, hinges on a understanding of the physicochemical phenomena involved.

Crystallization in sucrose solutions involves both nucleation and crystal growth. According to the classical theories of nucleation [1] there are three different elementary nucleation processes: primary

homogeneous, primary heterogeneous and secondary nucleation. It has been reported that primary homogeneous nucleation occurs spontaneously, while primary heterogeneous nucleation only occurs in the presence of a foreign nucleus. Secondary nucleation requires the presence of a solute particle interface. Secondary nucleation is the main pathway followed industrial crystallization processes. The formation of nuclei is unwanted in many crystallizations, especially sugar crystallization, because it widens the range of the crystal size distribution. A broader size distribution range means the product of the crystallizer is out of specification and the flow properties of the product are unsatisfactory. It is to the industry's benefit to avoid the occurrence of nucleation. Nucleation in the presence of other crystals does not occur below the secondary nucleation threshold (SNT). Thus, the position of the SNT should be identified and the variables that affect the nucleation above the SNT should be investigated. The nucleation can be a major determinant of quality in sugar-based products.The kinetics of nucleation and crystal growth require supersaturation, which is obtained by temperature changes and can be affected by impurities.

For a better understanding the role of oversursaturation and the nucleation phenomenon, several models have been developed in the literature [1,2,3]. All models are based on the population balance equation. These works are treated the crystallization of sucrose without studying the effect of impurities on the nucleation and growth rates. Indeed, the effects of impurities in the secondary nucleation have little attention in the literature.

The objectif of the present work is aiming to investigate the effect of impurities on the growth and nucleation rates. This is achieved by comparing the measured concentration profile in solution and the profile predicted by a dynamic mathematical model.

2. Experiment

The sugar contains 99% of sucrose, was provided by the Belgian company of sugar refinery (Tirelemeont). Glucose (99%) and fructose (99%) were supplied by Sigma. Distilled water was used for the preparation of the treated solutions. The experimental apparatus used in this study consisted of a Crystallizer with contenance 2 L, the solution is agitated using an agitator provided with a propeller Mixel TT equipped with 2 against-blades, the output engine is 0,37 kw. The stirring velocity is selected so as to ensure a good homogeneity of the solution; the speed is kept constant at 250 rpm. The temperature in the crystallizer was controlled by means

of a heating bath equipped with an external temperature Pt100O sensor which was set in the solution, the accuracy is ± 0.01°C. For water-sucrose mixture, a saturated solution was prepared by dissolving 761.15 g of sucrose in 233.82 g of distilled water at 70°C. These mass are measured by using the KERN FOB 1.5K0.5 balance with accuracy of 0.0001 g. In the case of impur solutions, 3% of each impurity was introduced. Thus, the solution containing the fructose as impurity, a saturated solution was prepared by dissolving 691.80 g of sucrose in solution contains 233.82g of water + 22.19 g of fructose. For the solution containing the glucose as impurity, a saturated solution was prepared by dissolving 754.52 g of sucrose in solution contains 233.82g of water + 22.19 g of glucose. For the solution containing the fructose combined with glucose as impurity, a saturated solution was prepared by dissolving 685.60g of sucrose in solution contains 233.82 g of water + 22.19 g of (fructose + glucose).

Initially the saturated solution is prepared by dissolving sucrose in distilled water. The temperature was kept constant at 70 °C for 30 min while agitating. The experiment continues by lowering the temperature inside the solution following a controlled temperature linear profile. The terminal temperature reached is 40°C. The kinetic measurements was obtained using the samples as the time function. Samples of the soluion are periodically removed from the crystallizer and placed under the microscope for image capture. Solution concentration was measured with an Abbe5 refractometer, the accuracy is 0,1. Each data point presented is an average of three experiments.

The microscopic analysis was carried out by a binocular Nikon microscope provided with an objective of a maximum enlargement of 30*50*2.5. The lighting of the sample is ensured by optical fibres providing a cold light favorable to the maintenance of the sample in its state initial. A camera Nikon DXM 1200 F is connected to a system of frame grabbing on PC. The images are acquired and are analyzed by the software of images analysis LUCIA "Laboratory Universal Computer Analysis Image". The accuracy of the method does not exceed 1% in the concentrations of sugars.

3. Model formulation and solution

For a perfectly mixed batch crystallizer of constant volume, in which crystal breakage and agglomeration are assumed to be negligible, the population balances equation PBE is a partial differential equation in time, t and crystal size, L:

$$\frac{\partial n(L,t)}{\partial t} = -G\frac{\partial n(L,t)}{\partial L} \qquad (1)$$

where n is the population density and G is the size-independent crystal growth rate. Subject to the boundary and initial conditions:

$$n(0,t) = n^o = \frac{B_0}{G} \quad (2)$$

$$n(L,0) = 0 \quad (3)$$

where n^o is the population density at the initial size of nuclei, L=0, B_0 is the nucleation rate.

The solution phase concentration in the crystallizer is described by a mass balance of the crystallizing solute:

$$\frac{dC}{dt} = -3\rho_c k_v h \int_0^\infty G\, n(L,t)\, L^2\, dL \quad (4)$$

in which C is the concentration (mass of solute per total mass of the mother liquor), ρ_c is the crystal density, k_v is the volume shape factor converting L^3 into crystal volume and h converts solvent mass to slurry volume.

In this work, the temperature $T(t)$ recorded experimentally will be used instead of the energy balance equation. This is done to avoid the error in estimating the overall heat transfer coefficient and to avoid estimating the heat of crystallization and its dependency on concentration.

The driving force for the nucleation and growth rates is the difference between the bulk concentration and the saturation concentration calculated at the bulk temperature. The kinetic expressions for nucleation and growth are taken to be empirical power-laws in supersaturation. The nucleation kinetics are described by the following empirical expression.

$$B_0 = k_b S^b \quad (5)$$

The growth kinetics are described by :

$$G = k_g S^g \quad (6)$$

where the saturation S is defined as:

$$S = \frac{C - C_{sat}(T(t))}{C_{sat}(T(t))} \quad (7)$$

$C_{sat}(T(t))$ is the concentration of saturation of sucrose in solution. The saturation concentration of sucrose in water is done as:

$$C_{sat}(T(t)) = 64.47 + 0{,}10336(T(t)) + 14{,}24\ 10^{-4}\ (T(t))^2 - 60{,}2\ 10^{-7}\ (T(t))^3 \quad (8)$$

For the saturation limit of sucrose in solution phase with the presence of impurity, the $C^*_{sat}(T(t))$ is obtained from:

$$K_{sat} = \frac{C^*_{sat}(T(t))}{C_{sat}(T(t))} \quad (9)$$

With these equations the model is completely defined. The parameters k_b, b, k_g, and g are inferred from the experimental data.

The PBE Eq. (1) is solved using the method of moments. The ith moment of the PBE is defined as:

$$\mu_j = \int_0^\infty n(L,t)\, L^j\, dL \quad (10)$$

The first four moments are required for the general batch crystallizer model:

$$\frac{d\mu_0}{dt} = B_0 \quad (11)$$

$$\frac{d\mu_j}{dt} = jG\mu_{j-1} \quad \text{for j = 1, 2, 3, 4} \quad (12)$$

The initial conditions are:

$$\mu_j(0) = 0 \quad \text{for } j = 1, 2, 3, 4. \quad (13)$$

$$C(0) = C_0 \quad (14)$$

The Eqs. (4), (11) and (12) will be solved using a fourth-order Runge–Kutta.

4. Parameters estimation

The calculation of the best parameters in the representation of experimental data was performed either by the Marquardt–Levenberg's algorithm for current minimization process (W. Marquardt, 1963). The minimization is carried out on sums of squares of deviations between calculated and experimental quantities. The objective function in terms of the least square estimations takes the form:

$$\phi(k_g, g, k_b, b) = \sum_{i=1}^{Ndata} \left(C_i^{cal} - C_i^{\exp}\right)^2 \quad (15)$$

The approach uses a nonlinear optimization routine to generate successive guesses of the kinetic parameters. For each guess, the dynamic model equations describing the process are solved and the resulting measurement predictions are compared with the data according to an optimization criterion. Generally, the criterion is a measure of the rediction errors. The model equations are solved implicitly each time the optimizer provides a parameter guess.The iterative optimization process ends when the specified tolerance has been achieved.

5. Results and discussion

In the parameters optimization process, the concentration phase in the crystallizer, at each times, is simultaneously calculated for the studied system by using a fourth-order Runge–Kutta. Four kinetic parameters (k_g, g, k_b, b) are fitted from the experimental data relative to the measured concentration phase in the crystallizer v.s times.

Table 1: Optimum kinetics parameters for crystallization of sucrose systems: 1: Sucrose-water, 2: Sucrose-glucose- water; 3: Sucrose - fructose-water, 4: Sucrose-glucose-fructose-water

	g	*b*	*kg* (m/s) (g/g water)$^{-B}$	*kb* (No./g/s) (g/g water)$^{-b}$
System 1	*1.830*	*0.3915*	*0.041*	*162.84*
System 2	*1.929*	*0.3557*	*0.0.42*	*166.73*
System 3	*2.748*	*0.1674*	*0.007*	*176.01*
System 4	*2.851*	*0.1448*	*0.004*	*182.74*

Figs. 1 and 2 illustrate the calculated concentration and saturation profiles in the crystallizer plotted against time, respectively. For comparison, we include the experimental data obtained by refractometric analysis. These figures show the quality of the adjustment of the kinetic parameters. It can be seen that the deviation standard for concentration is less than 0.1 % for the sucrose-water sytem,

0.03 % for the sucrose-glucose-water system, 0.04% for the sucrose-fructose-water system and 0.05% for the sucrose-glucose-fructose-water system. Concerning the effect of impurities on the crystallization of sucrose, one concludes from these two figures that the addition of impurities permits a reduction in the solubility of sucrose in the solution phase of the crystallizer, as mentioned in the literature [7]. This suggests that these impurities are in competition with sucrose. Consequently, an increase in solution phase supersaturation was obtained in the crystallizer and the time of crystallization of sucrose was decreased .Figs. 3 and 4 show the profiles of nucleation rate and growth rate plotted against time for the four systems studied. It appears from these figures that the effect of impurities allow an increase in the nucleation rate B_0 and a decrease in the growth rate G of the sucrose crystals, which is more pronounced in the presence of fructose.

The crystals formed in these experiments exhibit the occurrence of a secondary nucleation phenomenon. Figure 5 was taken using a field emission scanning electron microscope (SEM). It illustrates an example of the needle-like crystal(s) formed after a peak of superrsaturation. The observation of the aspect of the crystals before and after the peak gives more insight. Before the peak, crystals show regular surfaces. After the peak (fig. 5), crystals show imperfect surfaces with many fine that seem to have grown on the surface. From the SEM image, it can be observed that the crystal adopts long needle shaped structures: it shows that secondary nucleation has occurred.

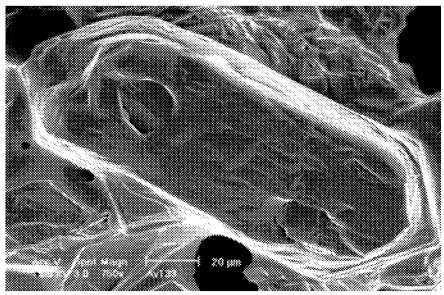


Fig. 5: (SEM) typical after a peak of sursaturation

The table 2 present the estimated mean dimension sizes at different times for the four studied system. For comparison, the experimental data were included in this table. Indeed, one observes that the calculated results are in agreement with the maximal measured sizes. By comparing the obtained sizes of the particle for the sucrose-water and sucrose-glucose-water systems, one notes that the presence of glucose causes a reduction in length of the obtaiend final product.

Sucrose -water			Sucrose-glucose water			Sucrose -fructose-water			Sucrose -glucose-fructose-water		
Time min	Max µm	Cal. µm	Time min	Max µm	Cal. µm	Time min	Max µm	Cal. µm	Time min	Max µm	Cal. µm
30	20	12.72	30	50	71.30	30	50	76.66	30	-	89
58	-	40.10	60	-	116.68	60	-	103.67	49	100	79.34
65	50	74.69	66	200	151.42	66	130	122.63	64	-	119.24
80	200	193	80	-	230.00	80	180	170.50	90	-	166
130	-	420.02	150	-	366.94	150	-	284.51	150	-	277.65

Table 2: Crystal average size as function of times for different system

The same effect was observed by comparing the sucrose-water and sucrose-glucose-fructose-water systems. For the comparison of particle dimensions between the sucrose-glucose-water and the sucrose-fructose-water systems, one observes an increase in the particle length with the time, but one remarks that the effect of glucose is important than fructose, i.e., the addition of fructose permit to produce the smallest particle size of the final product compared to the glucose.

Conclusion

Secondary nucleation and growth of sucrose crystallization was investigated at supersaturated conditions. The liquid phase concentration containing in batch crystallizer was measured as function of time by using a refractometer.The parameters of the growth and nucleation rates have been successfully estimated for sucrose crystallization by using the non-linear optimization method and for a linear cooling profile. The obtained kinetic parameters are physically realistic and comparable with the other parameters in the literature. The estimated results are in agreement with the experimental data. It has been shown that the presence of impurities in liquid phase in bach crystallizer of sucrose permit to produce the smallest mean particle sizes of the final product.

References

[1] J. Garside, Industrial Crystallization from Solutions, Chem. Eng. Sci., 40, N° 1, 3-26 (1985)

[2] A. Mersmann, J. Garside, J. Nývlt .Measurement of Crystal Growth Rates, European Federation of Chemical Engineering, Working Party on Crystallization (1990).

[3] V. Maurandi,.; G.Vaccari,; G.Mantovani,; A.Rossi, Kinetic, Variations in Sucrose Crystallization for Different Inter-Crystal Distances. Zuckerindustrie, 114 (7), 562-566 (1989).

[4] R.Farrel , and Y. Tsai, Modeling, Simulation and Kinetic Parameter Estimation in Batch Crystallization Processes. AIChE J., 40(4), (1994), 586.

[5] P.Quintana, E.Bolanos, I.Salcedo, Mathematical modeling and kinetic parameter estimation in batch crystallization, AIChE J, 50, (2004) 1407–1417 .

[6] H.A. Mohameed, B. Abu-Jdayi, M.Al Khateeb, Effect of cooling rate on unseeded batch crystallization of KCl, Chemical Engineering and Processing 41 (2002) 297-302.

[7] Y. Abed, N.Gabas, M. L. Delia, & T. Bounahmidi, Measurement of liquid–solid phase equilibrium in ternary systems of water–sucrose–glucose and water–sucrose fructose and predictions with UNIFAC. Fluid Phase Equilibria, (1992). 73, 175–184.

AUTHOR LISTING

AUTHOR LISTING

AUTHOR LISTING