Spectrum Analysis of GMA Welter in Various Working Modes

Gökhan Gökmen, Yelda Karatepe

Marmara University Technical Education Faculty, Department of Electricity, Goztepe Campus, Kadıkoy, Istanbul, 34722, Turkey
gokhang@marmara.edu.tr, ykaratepe@marmara.edu.tr

Tahir Çetin Akıncı

Kırklareli University, Engineering Faculty, Department of Electrical and Electronics Engineering, Kavaklı Campus, 39100, Kavaklı, Kırklareli, Turkey
cetinakinci@kirklareli.edu.tr

Memduh Kurtulmuş

Marmara University Technical Education Faculty, Department of Electricity, Goztepe Campus, Kadıkoy, Istanbul, 34722, Turkey
memduhk@marmara.edu.tr

Abstract: In this study, a current drawn by a welter at initial, stable-state and finish modes is examined using spectral analysis. The current shunt measurement method is utilized in order to measure the current drawn by the welter. The study involves the examination of welding stages of a material with the electrode of a welter. First, the current drawn by the welter is measured in the initial mode of the welding process. Then the current value during the stable-state mode of the welding process is measured. Finally, the current drawn at the finishing mode of the welding process is measured. Fast Fourier Transform (FFT) of all these measured current values are calculated and spectral analysis is performed using these transforms. During the study, it is observed that current drawn by the welter during these three modes of welding are different from each other. For each mode, frequency domain analysis of the measured current is performed.

Keywords: Welters; high frequency inverter; working modes; current shunt measurement; spectral analysis

1 Introduction

GMA (Gas Metal Arc) welding is widely used in many industrial applications for metal joining. The GMA welding machine produces an electrical arc between a metal electrode and the weld pool, with shielding from externally supplied gas, which may be an inert, active or a mixture gas. The occurred arc and its heat melts the metal surface and the metal electrode, then molten metal of the electrode is transferred to the work where it becomes the deposited weld metal [1, 2, 3, 4, 5, 6].

All electric arc welding machines work similarly. They dissolve electrodes for connection [8, 9, 10, 11]. In practice, there is a lot of work on the electric arc welding machines [12, 13].

According to the application of power electronics technology, in the GMA welding machine area, higher quality, less spatter generation and more automation are required. These requirements can be met by using the high-frequency inverter arc welding machine that are in this study.

Generally, the GMA Welding Machines widely used in the industrial area can be classified into MIG (Metal Inert Gas), MAG (Metal Active Gas) and GMAW (Gas Metal Arc Welding) welding machine types, according to the utilized shielding gas which prevents oxidizing molten pool or globule. Especially among these GMA welding machines, the CO_2 GMA welding machine is widely used because the price of shielding gas utilized is less expensive. But it has a major disadvantage: it generates more spatter during the welding procedure. The spatter which is generated during welding procedure is the small article radiated to space, nozzle and base metal. So additional work to remove this spatter is needed [14]. The machine with separate wire feeder used in this study is a professional type water-cooled semi-automatic welding machine designed for heavy industry. It is suitable for high quality welding seams for constructional low-carbon and stainless steels, aluminum and their alloys.

It is important to determine the spectral properties of the current signals of welter for current harmonic analysis, the optimization of welding process and time, calculating of the electromagnetic field and the determination of radiation distribution. In this study, the spectral properties of the current signals of welter for each working mode are analyzed respectively.

One of the most effective ways is the current shunt measurement for monitoring, acquiring and measuring applications of currents. It is also suitable for measuring high currents such as the welter primer current.

2 Current Measurement Method

There are basically three methods of monitoring current. Which of these three is used will depend on a number of factors both intrinsic and extrinsic to the application. These requirements may sometimes also be conflicting. Therefore, a careful balancing of requirements to select the optimum method is required. These three basic methods are resistive, optical and magnetic current measurement. The optically isolated resistive method has a medium current range and high isolation, but it has low accuracy and a medium-range cost. Magnetic measurement methods, such as traditional or Hall Effect current transformer, have high current range and medium accuracy, but their cost is very high. The resistive measurement method has high current range and accuracy, and also its cost is very low [15].

In this study, the resistive measurement method was chosen. The measurement of the instantaneous primer current of the welter was carried out by a shunt current measurement. In this method, the current to be measured passes into a resistor and it leads to voltage of shunt resistor. This voltage is proportional to measurement current. The shunt voltage is very low, so it must be amplified by an operational amplifier. A sense resistor must be placed between load and source [16, 17, 18].

As a shunt resistor, the MP2060 (0.005 Ω) power film resistor was preferred [11, 12, 13, 14] and the INA 146 difference amplifier (Burr-Brown Corporation) was chosen as gain amplifier. The connection scheme of the shunt current measurement method is shown Fig. 1 [18, 19, 20, 21].

The calculation of the output voltage of the INA 146 amplifier is given below [18, 19, 20, 21].

$$V_o = I.R_s.[0.1(1 + R_{G2} / R_{G1})] \tag{1}$$

In this equation, V_o represents the output voltage of the amplifier in V, I represents the welding current in A, R_s represents the shunt resistor value in Ω, R_{G1} and R_{G2} represent the gain adjustment resistor in Ω. $[0.1(1 + R_{G2} / R_{G1})]$ refers to the gain value of the amplifier. If R_{G1} is 100 Ω and R_{G2} is 49.9 kΩ, the gain is 50 [6, 7, 10]. The relationship between the welding current and output voltage is given below:

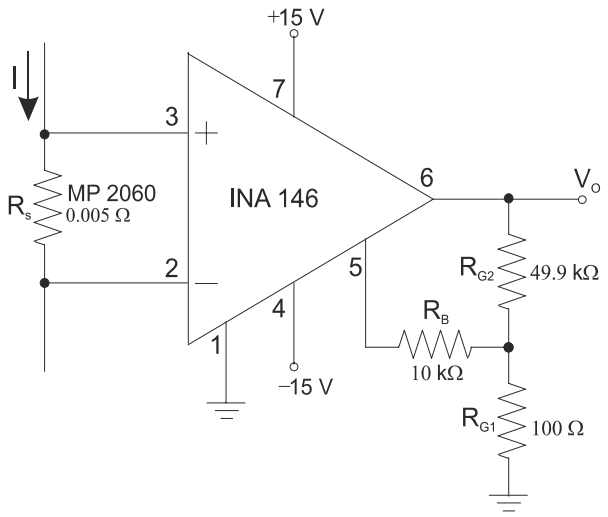


Figure 1
Shunt current measurement [19]

$$I = \frac{V_o}{R_s.[0.1(1+R_{G2}/R_{G1})]} = 4.V_o \tag{2}$$

If the output voltage is selected ± 14.475 V (it is almost the maximum amplifier output voltage), the maximum welding current value is calculated as ±57.9 A. This value is accepted as ±1 pu for convenience and subsequent measured current values are specified according to that value [18, 19].

3 Power Spectrum Density

A common approach for getting information about the frequency properties of a random signal is to transform the signal into frequency domain using Discrete Fourier Transform. For data with N-samples, the transformation at $m\Delta f$ frequency is defined as the following equation [18, 22, 23, 24, 25, 26, 27].

$$X(m\Delta f) = \sum_{k=0}^{N-1} x(k\Delta t)^{-j2\pi km/N} \tag{3}$$

In this equation, Δf and Δt represent frequency and resolution at the sampling time, respectively. In this context, specific power spectral density of time series *x(t),*' which is N-sample long, is given as Sxx(f) in Eq. 4.

$$S_{xx}(f) = \frac{1}{N}\left|X(m\Delta f)\right|^2, f = m\Delta f \tag{4}$$

The cross power spectral density approximation, which is defined between two time series like x(t) and y(t), can be given in a similar way. The statistical accuracy of the approximation given in Eq. 4 increases with the increasing number of discrete data or increasing number of data block containing a sufficient amount of data.

4 Data Acquisition Systems

For data acquisition, a shunt resistor was connected to the primer coil of the welding machine. The output voltage of the amplifier was transmitted to the computer at a sampling rate of 0.005 seconds via an Advantech 1716L Multifunction PCI card, and data analysis was performed using MATLAB-Simulink. This data acquisition system is shown in Fig. 2 [11, 12].

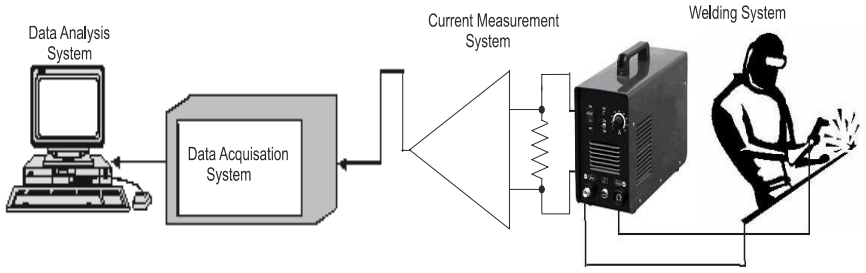


Figure 2
Data acquisition systems [18, 19]

The voltage values of the amplifier are saved separately for three different operation modes of the welding machine. Respectively, these modes are the initial mode, the stable-state mode, which corresponds to the moment reached after a specified time has passed from the starting point, and the finishing mode. As far as these three modes are analyzed, it is seen that the maximum current drawn is 0.92 pu (53.2681 A) and it is drawn at the initial mode. For the other two modes, the peak value of current is observed as 0.6 pu (34.7401 A). The peak current value decreased and is fixed around 0.025 pu (1.44 A) right after the end of the welding [18, 19].

In this study, the welding process is applied to *ST 37* type material by "rutile" basis electrode using the Metal Active Gas (MAG) method. Some specific properties of this welter can be listed as below [18, 19]:

- Frequency: 50/60 Hz
- Number of phases: 3~ AC.
- Primer: 100 VA, Voltage: 440 V-220 V-240 V
- Primer Current: 17/29 A
- Secondary: 55V (DC), 100 %:200 A - 35%: 315 A

5 Data Analysis and Feature Extraction

Data obtained through experimental studies are examined at time-current plane. After the analysis performed, it is observed that the instant alteration of current is composed of high-frequency components. In order to obtain the properties of the high-frequency components, the spectral method is utilized.

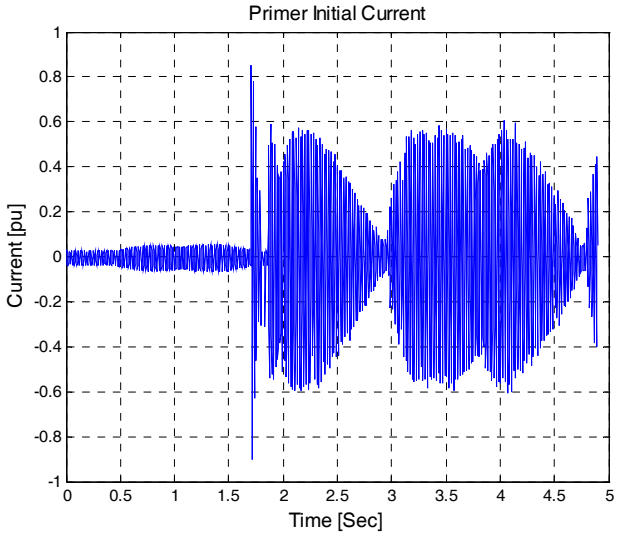


Figure 3
Welding current at the initial mode [13, 14]

In Fig. 3, the current-time graph for the primer current is given. It can be seen that the high current drawn after 1.7 seconds lasts until 3 seconds. At 3 seconds, it drops down to the minimum value and then increases again. When this current-time graph of welding machine is observed, it can be said that it is a characteristic graph for the primer current [18, 19].

When the current-time graph generated for primer initial current of welding machine is expressed at time-frequency plane (Fig. 4), it can be seen that in a frequency band of 45-55 Hz at intervals there are high amplitudes of around 5 seconds. Since mains frequency is 50 Hz, this frequency operation interval can be said to be feasible. In Fig. 4, a high-frequency region is observed at the region labeled as 1 between 1.5-2.2 seconds. This region corresponds to the high current arising after 1.7 seconds at current-time graph. Moreover, high frequency regions are also observed in regions labeled as 2 and 3. These regions are related to the instants when high currents are drawn, as can be observed from current-time graph.

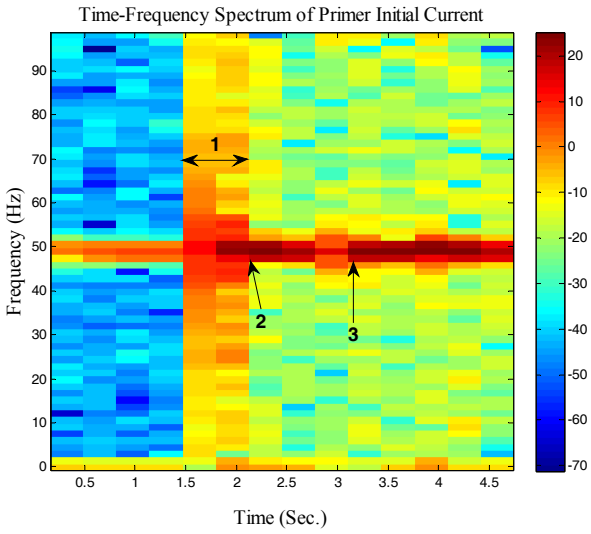


Figure 4
Time- Frequency spectrum of primer initial current

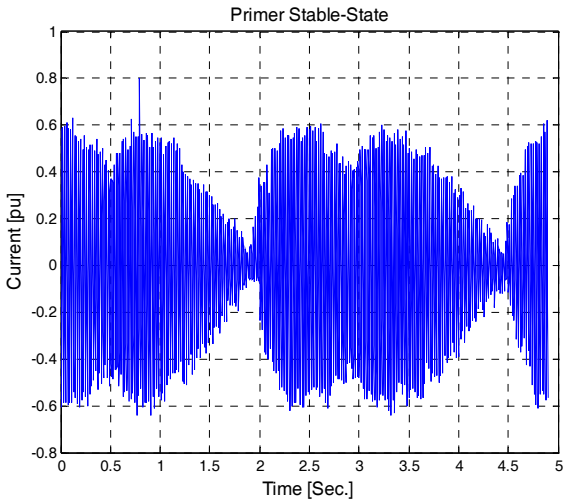


Figure 5
Current-Time graph of stable-state mode

The primer current for stable-state instant is depicted in Fig. 5. Here, high current can be observed starting from the starting instant. Even though it decreases at 1.7 seconds, it exhibits an increasing characteristic from that point on. When this current-time graph of the welding machine is observed, it can be said that it is a characteristic graph for the primer stable- state mode.

When the current-time graph generated for primer initial current of welding machine is expressed at time-frequency plane (Fig. 6).

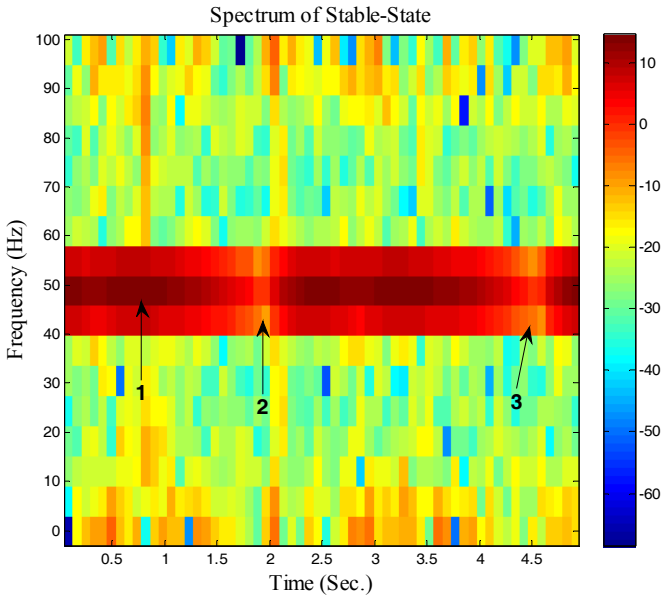


Figure 6

Time-Frequency spectrum of finish-mode

it can be seen that in a frequency band of 45-55 Hz, at intervals there are high amplitudes of around 5 seconds. Here, the region labeled as 1 shows the region at which the welding machine generates the first stable arc. Moreover, this part is the part that builds up the high frequency components. This area corresponds to 10's scale at the color bar on the graph. At the region labeled as 2, there is a lower frequency region between 1.5-2 seconds. It can be seen that the region labeled as 3 is the region at which low arcs occur; low frequency and current are drawn. Lower currents are drawn at regions labeled as 2 and 3 when compared to the region labeled as 1.

In Fig. 7, the current-time graph, which is damped at 0.5 seconds and exhibits an increasing characteristic until 3 seconds, is depicted. This characteristic can be shown as the finish-mode characteristic of the welding machine. In this study, it is observed that the welding machine draws high currents until 2.9 seconds and then it does not perform the welding operation until 5 seconds.

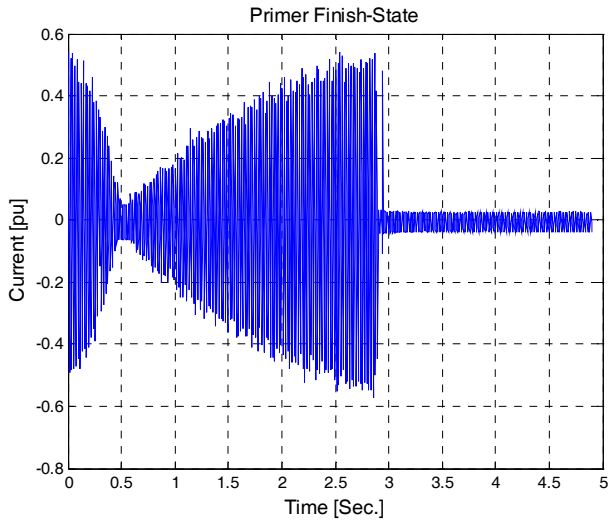


Figure 7
Current-time graph of finish-mode

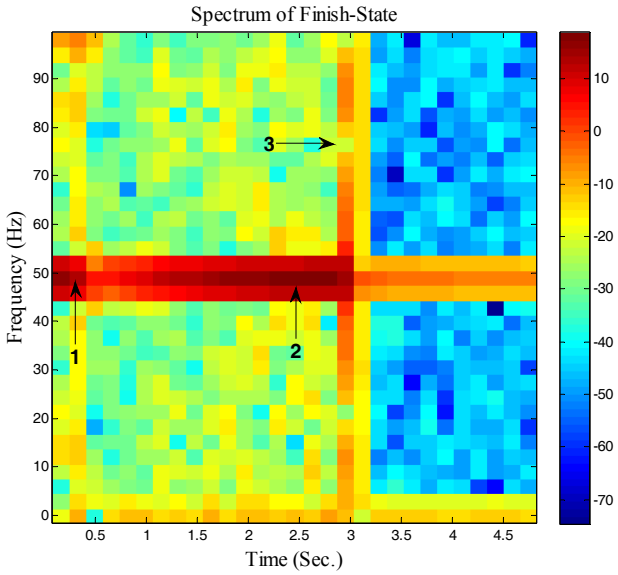


Figure 8
Time-Frequency spectrum of finish-mode

In Fig. 8, the spectrum graph of the Finish mode, which belongs to the current-time graph shown in Fig. 7, is depicted. As can be seen from the spectrum, while region 1 and 2 can be expressed with high-frequency components, region 3 can be expressed with low-frequency components.

Conclusions

In this study, a data acquisition system utilizing a current shunt measurement method for the measurement of the current drawn by welding machine is used, and the data acquired is analyzed using the MATLAB-Simulink package program. After the analysis, the current characteristics of the initial, stable-state and finish modes are determined and the time-frequency spectrums of these characteristics are analyzed.

The extraction of the properties of currents drawn by the welding machine at initial, stable-state and finish modes and performing their spectral analysis are used to specify the current-frequency properties for these different modes. The findings of the analysis result show that the device has a frequency around the fundamental frequency of 50 Hz. In this study on the electric arc welding machine, the spectral analysis method, the source initial mode, operating mode and ending mode frequencies are determined. In addition, this also raises the properties of the different modes. The acquired information contains meaningful results related to the operation states of the welding machine.

References

[1] Correia, D. S., Gonçalves, C. V., Sebastião S. C., Ferraresi V. A: GMAW Welding Optimization Using Genetic Algorithms, Journal of the Brazilian Society of Mechanical Sciences and Engineering, Vol. 26, No. 1, pp. 28-33, 2004

[2] Deruntz, B: Assessing the Benefits of Surface Tension Transfer Welding to Industry, Journal of Industrial Technology, Vol. 19, No. 4, August 2003

[3] Jones, L. A., Eagar, T. W., Lang, J. H: Metal Transfer Control in gas Metal Arc Welding, Tenth Symposium on Energy Engineering Sciences-Argonne National Laboratory, Argonne, IL, May 1992

[4] Funderburk, R. S: Key Concepts in Welding Engineering´-Welding Innovation, Vol. 16, No. 1 and 2, 1999

[5] O'Brien, R: Welding Handbook, Welding Processes. 8th Edition, Miami: American Welding Society, ISBN 0-87171, pp. 110-116, 1991

[6] Ravisankar, V, Balasubramanian, V., Muralidharan, C: Selection of Welding Process to Fabricate Butt Joints of High Strength Aluminum Alloys Using Analytic Hierarchic Process, Science Direct, Material & Design. Vol. 27, No. 5, pp. 373-380, 2006

[7] Ruth, K: Welding Basics: An Introduction to Practical & Ornamental Welding- Creative Publishing International Inc. Minnesota, ISBN 1-58923-139-2, pp. 7-11, 2004

[8] Lancaster, J. F: The Physics of Welding, Phys. Technology, Vol. 15, pp. 73-79, 1984

[9] http://www.ehow.com/a bout_4661176_electric-welding-machines.html. access date 15.06.2010.

[10] Palanco, S., Klassen, M., Skupin, J: Spectroscopic Diagnostics on CW-Laser Welding Plasmas of Alumi-Num Alloys, Spectrochim - Acta B, At. Spectrosc., Vol. 56, No. 6, pp. 651-659, Jun. 2001

[11] Karabegović, I, Hrnjica, B: Simulation of Industrial Robots for Laser Welding of Load Bearing Construction. -Mechanika. -Kaunas: Technologija, Nr. 2(76), pp. 50-54, 2009

[12] Harry, J. E: Measurement of Electrical Parameter of ac Arcs- IEEE Transactions on Industry and General Ap-plications, Vol. IGA-5, No. 5, pp. 624-632, September/October 1969

[13] Hao, X., Song, G: Spectral Analysis of The Plasma in Low-Power Laser/Arc Hybrid Welding of Magnesium Alloy- IEEE Transactions on Plasma Science, Vol. 37, Issue 1, pp. 76-82, 2009

[14] Lanchester, L: The Physics of Welding, 2nd ed. Pergamon Press, ISBN: 13 978-0080340760, pp. 340, 1986

[15] Bode, P. T: AN39 Current Measurement Applications Handbook, ZETEX Semiconductors, Application Notes, Issue 5, pp. 1-42, 2008

[16] http://focus.ti.com/analog/docs/mirosite.tsp?sectionId=560 microsite.tsp?sectionId=560&tabId=2182&micrositeId=7, access date 15.06.2010.

[17] Koon, W: Current Sensing for Energy Metering, Technical Article, Analog Devices, Inc., pp. 2-9, 2010

[18] Akinci, T. C: Time-Frequency Analysis of the Current Measurement by Hall Effect Sensors Electric Arc Welding Machine. Mechanika, ISBN: 1392-1207, Vol. 5, No. 85, pp. 66-70, 2010

[19] Akıncı T. C, Nogay H. S, Gokmen G: Determination of Optimum Operation Cases in Arc Welding Machine Using Neural Network, Journal of Mechanical Science and Technology, Vol. 25, No. 4, pp. 1003-1010, 2011

[20] Caddock Electronics, Inc: MP2060 Kool-Pak Clip Mount Power Film Resistor, Data Sheet 28_IL128.1004, p. 1, 2004

[21] Texas Instruments: INA199A1-A3EVM, User's Guide SBOU085, pp. 4-11, 2010

[22] Burr-Brown Corporation: INA146 High-Voltage, Programmable Gain Difference Amplifier, Data Sheet PDS-1491A, pp. 1-11, 1999

[23] Vaseghi, S. V: Advanced Signal Processing and Digital Noise Reduction, 3rd ed. John Wiley & Sons Inc, ISBN: 0-470-09494-X, p. 449, 2006

[24] Seker, S: Determination of Air-Gap Eccentricity in Electric Motors Using Coherence Analysis, IEEE Power Engineering Review, Vol. 20, No. 7, pp. 48-50, 2000

[25] Taskin, S., Seker, S., Karahan, M., Akinci, T. C: Spectral Analysis for Current and Temperature Measurements in Power Cables- Electric Power Components and Systems, Vol. 37, Issue 7, pp. 415-426, April 2009

[26] Dutoit, T., Marques, F: Applied Signal Processing- A Matlab-based Proof of Concept, Springer Science +Business Media, ISBN: 978-0-387-74534-3, pp. 149-179, 2009

[27] Arfib, D., Keiler, F., Zölzer, U: Time Frequency Processing. In: DAFX:Digital Audio Effects U. Zölzer, Ed. Hoboken (NJ), John Wiley & Sons, 2002

Novel Degree-based Molecular Descriptors with Increased Discriminating Power

Tomislav Došlić

Faculty of Civil Engineering, University of Zagreb
Kačićeva 26, Zagreb, Croatia
doslic@grad.hr

Tamás Réti

Óbuda University
Bécsi út 96/B, H-1034 Budapest, Hungary
reti.tamas@bgk.uni-obuda.hu

Abstract: In the present study we investigate some general problems concerning the degeneracy of widely used topological indices (graph invariants), and we propose a novel family of molecular descriptors characterized by a decreased degeneracy level. A special feature of topological indices of novel type is that they take into account the degrees of vertices on increasing distances from a single vertex. According to the comparative tests performed on samples of isospectral graphs and of graphs of small diameter, the new descriptors are judged to be more efficient for discriminating between topological structures of molecular graphs than several traditional molecular indices.

Keywords: Zagreb indices; pseudo-regular graphs; QSAR/QSPR studies

1 Introduction

A promising trend in theoretical and structural chemistry is the employment of graph invariants (topological indices) for the characterization of the combinatorial structure of carbon-based chemical compounds and the prediction of their physico-chemical properties. Topological invariant is a real number derived from the structure of a graph in such a way that it does not depend on the labeling of vertices. Hundreds topological invariants (indices) have been invented so far, and numerous reviews have been published on their applications in the QSAR/QSPR studies [1-10].

One of the main difficulties when using topological indices for discriminating and prediction purposes is their degeneracy, i.e., the fact that two (or more) non-isomorphic graphs have the same value of a topological index. The degeneracies are unavoidable; however, it makes sense to search for indices whose degeneracy is as low as reasonably possible [11, 12].

In the present study we investigate the discriminating potential (application limits) of traditional degree-based descriptors, especially, how to decrease the degeneracy by an appropriate modification or generalization of their structure, and finally we propose a set of novel topological invariants having improved discriminating potential.

2 Definitions, Basic Notions

All graphs considered in this study are finite, simple and connected graphs (without loops and multiple edges). We use the standard terminology; for the concepts not defined here, we refer the reader to any of standard graph theory monographs such as, e.g., [13] or [14]. For a connected graph G, V(G) and E(G) denote the set of vertices and edges, and $|V(G)|$ and $|E(G)|$ the numbers of vertices and edges, respectively.

An edge of G connecting vertices u and v is denoted by (u,v). The diameter of a graph G (written by diam(G)) is defined as the greatest distance between any pairs of vertices in G. The degree of vertex u, denoted by d(u), is the number of edges incident to u. We denote by $\Delta=\Delta(G)$ and $\delta=\delta(G)$ the maximum and the minimum degrees, respectively, of vertices of G. A graph is called regular (R-regular), if all its vertices have the same degree R. To avoid trivialities we always assume that $|V(G)| \geq 3$, and $d(u)\geq 1$. A connected graph with maximum vertex degree at most 4 is said to be a "chemical graph".

Consider the topological descriptor X(G) defined in the general form

$$X(G) = F(Z_1(G), Z_2(G), ..., Z_J(G))$$

where F is a J-variable, non-negative real function, $Z_1(G)$, $Z_2(G)$,...,$Z_j(G)$,... $Z_J(G)$ are appropriately selected topological invariants given as

$$Z_j(G) = \sum_r \sum_{s\leq r} E(r,s) f_j(r,s),$$

where $f_j(x,y)$ are real symmetric functions for $1\leq j\leq J$, and quantities E(r,s) denote the number of edges in G with end-vertices of degree r and s. (The number E(r,s) are sometimes denoted by $m_{r,s}$).

The descriptors represented by X(G) are called *the generalized edge-additive topological indices*.

It follows that if for graphs Hu and Hv the equalities $E_{Hu}(r,s)=E_{Hv}(r,s)$ are fulfilled, then $Z_j(Hu)=Z_j(Hv)$ and $X(Hu)=X(Hv)$ hold, independently of the type of functions $f_j(x,y)$ and the J-variable function F. This means that the topological descriptors of the type X(G) are not suitable for discriminating between graphs Hu and Hv.

By specializing functions F and $f_j(x,y)$ one can obtain several indices (molecular descriptors) from recent literature [3-10]. In a particular case, by selecting topological parameters $Z_1(G)$ and $Z_2(G)$ as

$$Z_1(G)=\sum_r\sum_{s\le r}E(r,s)\left(\frac{r}{s}+\frac{s}{r}\right) \quad \text{and} \quad Z_2(G)=|V(G)|=\sum_r\sum_{s\le r}E(r,s)\left(\frac{1}{r}+\frac{1}{s}\right),$$

we can construct the topological index $X_E(G)=Z_1(G)/Z_2(G)$ introduced in a recent paper [10]. It has been verified that for index $X_E(G)$ the following identity is fulfilled:

$$X_E(G)=\frac{Z_1(G)}{Z_2(G)}=\frac{1}{|V(G)|}\sum_r\sum_{s\le r}E(r,s)\left(\frac{r}{s}+\frac{s}{r}\right)=\frac{1}{|V(G)|}\sum_{u\in V(G)}m(u)$$

In the above formula m(u) stands for the average degree of the vertices adjacent to vertex u in G.

In certain cases the discriminating ability of topological descriptors of the type X(G) (for example $X_E(G)$) is strongly limited. This is demonstrated in the following example. In Fig. 1 a pair of isospectral graphs, G_1 and G_2, are shown [15]. For them the equality $X_E(G_1)= X_E(G_2)$ holds; consequently, they cannot be distinguished by the topological index $X_E(G)$.

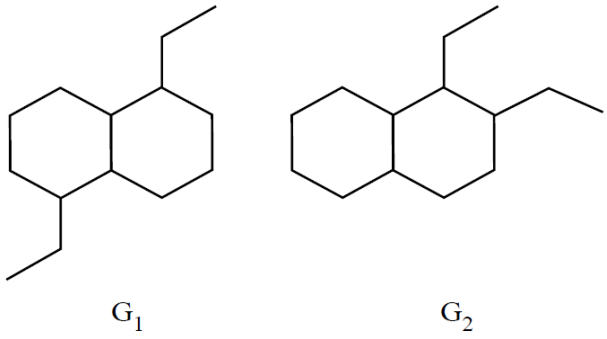


Figure 1
A pair of isospectral chemical graphs that cannot be distinguished by indices of the type X(G)

A possible solution to improve the discriminativity of a topological index is to modify it so as to include more information encoded in the graph adjacency matrix. For this purpose, it seems logical to take into account the degrees of vertices on increasing distances from a single vertex [16-18]. Hence, for i ≥1 we define the quantitites

$$Q_i(u) = \sum_{v \in N_i(u)} d(v)$$

where $N_i(u)$ denotes the set of all vertices at distance i from vertex u. If $N_i(u)$ is empty, we set $Q_i(u) = 0$ by definition. Obviously, $Q_i(u)$ is equal to zero for all i that exceed the diameter of G. Now we define

$$m_i(u) = \frac{Q_i(u)}{n_i(u)}$$

where $n_i(u)$ is the cardinality of $N_i(u)$. It is assumed that $m_i(u) = 0$ if $N_i(u)$ is empty. It is easy to see that relations $\delta \le m_i(u) \le \Delta$ and $\sum n_i(u) = |V(G)| - 1$ hold for any vertex u. By averaging $m_i(u)$ over all vertices of G we obtain global topological indices

$$\langle m_i(G) \rangle = \frac{1}{|V(G)|} \sum_{u \in V(G)} m_i(u),$$

and topological parameters defined as

$$Q_i(G) = \sum_{u \in V(G)} Q_i(u) = \sum_{u \in V(G)} m_i(u) n_i(u) = \sum_{u \in V(G)} \sum_{v \in N_i(u)} d(v)$$

for $1 \le i \le \mathrm{diam}(G)$. From the previous considerations if follows that the topological index $X_E(G)$ now appears as a special case $X_E(G) = \langle m_1(G) \rangle$.

3 Some Theoretical Considerations

The Zagreb indices belong to the family of the widely used molecular descriptors. In what follows we analyse some correspondances between the quantities $Q_i(u)$, $m_i(u)$, $n_i(u)$ and the Zagreb indices.

Recall that the first Zagreb index $M_1(G)$ and the second Zagreb index $M_2(G)$ of a graph G are defined as

$$M_1 = M_1(G) = \sum_{u \in V(G)} d^2(u) = \sum_r \sum_{s \le r} E(r,s)(r+s)$$

$$M_2 = M_2(G) = \sum_{(u,v)\in E(G)} d(u)d(v) = \sum_r \sum_{s\le r} E(r,s)rs$$

We refer the reader to surveys [4, 5, 19-21] for more information on Zagreb indices.

Proposition 1 ([10]): Let [d(G)] denote the average degree of a connected graph G. Then $\langle m_1(G)\rangle \ge [d(G)]$ holds with equality if and only, if G is regular. ■

Proposition 2 ([22]): Let G be a connected graph. Then

$$M_1(G) = \sum_{u\in V(G)} m_1(u)d(u) \quad \text{and} \quad 2M_2(G) = \sum_{u\in V(G)} m_1(u)d^2(u)$$ ■

Corollary 2.1 Because $\delta \le m_1(u) \le \Delta$ this imples that

$$2|E(G)|\delta \le Q_1(G) = M_1(G) \le 2|E(G)|\Delta$$

$$\delta M_1(G) \le 2M_2(G) \le \Delta M_1(G)$$

Lemma 1 ([19]): Let G be a connected graph. Then

$$Q_1(u) = d(u)m_1(u) \le 2|E(G)| - d(u) - (|V(G)| - 1 - d(u))\delta$$ ■

Lemma 2 Let G be a connected graph. Then

$$Q_1(u) = d(u)m_1(u) \le 2|E(G)| - |V(G)| + 1$$

Proof. Because $\delta \ge 1$ and $d(u) \le |V(G)| - 1$, from Lemma 1 it follows the claim. ■

From Lemma 2 the following proposition yields:

Proposition 3 Let G be a connected graph. Then for k=1,2,...positive integers

$$\sum_{u\in V(G)} Q_1^k(u) = \sum_{u\in V(G)} (d(u)m_1(u))^k \le |V(G)|(2|E(G)| - |V(G)| + 1)^k$$

with equality if G is a complete graph K_n or a star graph S_n on $n\ge 3$ vertices. ■

Corollary 3.1 ([31]): As a particular case, for k=1 we have

$$M_1(G) = \sum_{u\in V(G)} d(u)m_1(u) \le |V(G)|(2|E(G)| - |V(G)| + 1)$$

with equality if G is a complete graph K_n or a star graph S_n on $n\ge 3$ vertices.

Proposition 4 ([23]): Let G be a connected graph. Then

$$Q_1(u) = d(u)m_1(u) = \sum_{v\in N_1(u)} d(v) \ge d(u) + n_2(u)$$

and

$$d(u)Q_1(u) = d^2(u)m_1(u) \geq d^2(u) + d(u)n_2(u)$$

with equality if and only if G is a triangle- and quadrangle-free graph. ■

Corollary 4.1 Consider the Gordon-Scantlebury index S(G) of a graph G [1, 24]. This is a widely-used molecular descriptor of the type X(G) which can be calculated as

$$S(G) = \frac{1}{2}\sum_{u\in V(G)} d(u)\big(d(u)-1\big) = \frac{1}{2}\big\{M_1(G) - 2|E(G)|\big\}$$

From the prevous considerations it follows that

$$\sum_{u\in V(G)} n_2(u) \leq \sum_{u\in V(G)} Q_1(u) - 2|E(G)| = M_1(G) - 2|E(G)| = 2S(G)$$

and

$$2M_2(G) = \sum_{u\in V(G)} d^2(u)m_1(u) \geq M_1(G) + \sum_{u\in V(G)} d(u)n_2(u)$$

with equality if and only if G is a triangle- and quadrangle-free graph.

Proposition 5 Let G be a connected graph. Then

$$Q_2(G) = \sum_{u\in V(G)} Q_2(u) = \sum_{u\in V(G)} m_2(u)n_2(u) = \sum_{u\in V(G)} d(u)n_2(u)$$

Proof. It is based on the following identity:

$$\sum_{u\in V(G)} Q_2(u) = \sum_{u\in V(G)}\sum_{v\in N_2(u)} d(v) = \sum_{u\in V(G)} d(u)n_2(u) \qquad ■$$

Proposition 6 Let G be a connected graph. Then

$$\sum_{i\geq 2} Q_i(G) = 2|E(G)|\big(|V(G)|-1\big) - M_1(G)$$

Proof. For any vertex u we have

$$2|E(G)| - d(u) = \sum_{v\in N_1(u)} d(v) + \sum_{i\geq 2}\sum_{v\in N_i(u)} d(v) = \sum_{v\in N_1(u)} d(v) + \sum_{i\geq 2} Q_i(u)$$

The claim now follows by summing over all vertices. ■

Corollary 6.1 Let $G \neq K_n$ where K_n denotes the complete graph on n-vertices. Then

$$M_1(G) + Q_2(G) \leq 2|E(G)|\big(|V(G)|-1\big)$$

with equality if and only if, diam(G) =2.

Proposition 7 Let G be a connected graph. Then

$$\sum_{u\in V(G)} m_1^2(u) \geq M_1(G)$$

with equality if and only if G is regular.

Proof. From the Cauchy-Schwarz inequality one obtains

$$Q_1(G) = \sum_{u\in V(G)} m_1(u)d(u) \leq \sqrt{\sum_{u\in V(G)} m_1^2(u)} \sqrt{\sum_{u\in V(G)} d^2(u)}$$

Since $M_1(G)= Q_1(G)$, we have

$$M_1(G) \leq \sqrt{\sum_{u\in V(G)} m_1^2(u)} \sqrt{M_1(G)}$$

and the claim follows. ■

A connected graph is called pseudo-regular [25, 26] if there exists a positive constant p=p(G) such that each vertex of G has the average neighbor degree number equal to p, i.e., $m_1(u)=p(G)$ for any vertex u in G. Of course, every regular graph is also pseudo-regular. Moreover, it is obvious that $\langle m_1(G)\rangle = p(G)$ for any pseudo-regular graph.

In Fig. 2 two infinite sequences of pseudo-regular graphs denoted by $G_A(k)$ and $G_B(k)$ are shown [27]. It is interesting to note that $m_1(u)=3$ holds for any vertex u of graphs $G_A(k)$ and $G_B(k)$, where $k\geq 3$.

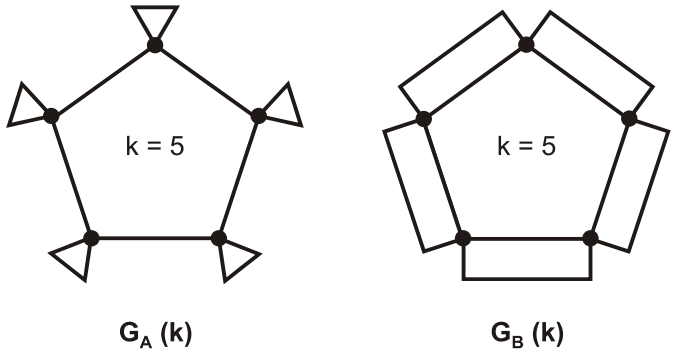


Figure 2
Pseudo-regular chemical graphs $G_A(5)$ and $G_B(5)$ that are not regular

Moreover, for graphs $G_A(k)$ and $G_B(k)$ the identities E(2,2,)=k, E(4,2)=2k, E(4,4)=k, $|E(G)| = 4k$ are fulfilled. This implies that graphs $G_A(k)$ and $G_B(k)$ cannot be distinguished by topological indices of the type X(G).

Proposition 8 Let G be a pseudo-regular graph. Then

$$p(G) = \langle m_1(G) \rangle = \frac{2M_2(G)}{M_1(G)}$$

Proof. From Proposition 2 it follows directly that if G is pseudo-regular then $M_1(G) = 2|E(G)|p$ and $M_2(G) = |E(G)|p^2$. ■

Proposition 9 Let G be a connected graph. Then

$$\sqrt{\frac{\sum_{u \in V(G)} Q_1^2(u)}{M_1(G)}} = \sqrt{\frac{\sum_{u \in V(G)} d^2(u) m_1^2(u)}{\sum_{u \in V(G)} d^2(u)}} \geq \frac{2M_2(G)}{M_1(G)}$$

with equality if and only if G is pseudo-regular.

Proof: We start from the Chebyshev inequality ([28], p 43). By specializing $b_j = a_j$ and $\sum_{j=1}^{J} w_j = 1$ we obtain the inequality

$$\sum_{j=1}^{J} w_j a_j^2 \geq \left(\sum_{j=1}^{J} w_j a_j \right)^2$$

with equality if and only if $a_1 = a_2, \ldots = a_J$. Now we denote by u_j the j-th vertex of G and define $a_j = m_1(u_j)$ and $w_j = d^2(u_j)/M_1(G)$ for $1 \leq j \leq J = |V(G)|$. We have

$$\frac{\sum_{u \in V(G)} d^2(u) m_1^2(u)}{\sum_{u \in V(G)} d^2(u)} \geq \left(\frac{1}{M_1(G)} \sum_{u \in V(G)} d^2(u) m_1(u) \right)^2 = \left(\frac{2M_2(G)}{M_1(G)} \right)^2$$

Now the claim follows with equality if and only if G is pseudoregular. ■

It is interesting to note that the left-hand side of the inequality of Proposition 9 is a sharp lower bound of the spectral radius of G ([25]).

Proposition 10 Let G be a connected graph. Then

$$\sqrt{\frac{\sum_{u \in V(G)} (d^2(u) + Q_1(u))^2}{\sum_{u \in V(G)} d^2(u)}} = \sqrt{\frac{\sum_{u \in V(G)} (d^2(u) + d(u) m_1(u))^2}{M_1(G)}} \geq 2\sqrt{\frac{M_1 G)}{|V(G)|}}$$

with equality if and only if G is regular.

Proof. We compute the variance of quantities defined as $b_j = d^2(u_j) + d(u_j)m_1(u_j)$ for $1 \leq j \leq J = |V(G)|$. We have

$$\mathrm{Var}(b) = \frac{1}{J} \sum_{j=1}^{J} b_j^2 - \left(\frac{1}{J} \sum_{j=1}^{J} b_j \right)^2 \geq 0$$

Consequently,

$$\frac{1}{|V(G)|}\sum_{u\in V(G)}(d^2(u)+d(u)m_1(u))^2 \geq \left(\frac{1}{|V(G)|}\sum_{u\in V(G)}(d^2(u)+d(u)m_1(u))\right)^2 = \frac{4M_1^2G)}{|V(G)|^2}$$

and this further implies

$$\frac{1}{M_1(G)}\sum_{u\in V(G)}(d^2(u)+d(u)m_1(u))^2 \geq \frac{4M_1G)}{|V(G)|}.$$

From there the claim follows, with equality if and only if G is regular. ■

The left-hand side of the inequality of Proposition 10 represents a sharp lower bound on the Laplacian spectral radius of G [29].

4 Possibilities of Increasing the Discriminativity

Comparing the topological indices $\langle m_1(G)\rangle$ and $\langle m_2(G)\rangle$ it is clear that $\langle m_2(G)\rangle$ should be more discriminative than $\langle m_1(G)\rangle$. Considering the pseudo-regular graphs $G_A(k)$ and $G_B(k)$ in Fig. 2, from the previous considerations it follows that $\langle m_1(G_A(k))\rangle = \langle m_1(G_B(k))\rangle = 3$ for $k \geq 3$, and $\langle m_2(G_A(k))\rangle = 28/9$ and $\langle m_2(G_A(k))\rangle = 32/9$ hold for $k \geq 5$.

For isospectral graphs G_1 and G_2 depicted in Fig. 1 it can be verified that $\langle m_1(G_1)\rangle = \langle m_1(G_2)\rangle = 16/7$, moreover $\langle m_2(G_1)\rangle = 177/70 \approx 2.52857$ and $\langle m_2(G_2)\rangle = 2117/840 \approx 2.52024$. As we can observe the numerical values of $\langle m_2(G_1)\rangle$ and $\langle m_2(G_2)\rangle$ are very close.

In the following, we will analyse the situations where employing the topological descriptor $\langle m_2(G)\rangle$ does not result in an improvement of disriminating performance. One such situation is, obviously, when G is regular; another one is when the graph is, in a sense, "small".

Proposition 11 Let G be a connected graph of diameter 2. Then the descriptors $\langle m_1(G)\rangle$ and $\langle m_2(G)\rangle$ are algebraically dependent quantities.

Proof. If the diameter of G is equal to 2, we have $n_1(u) = d(u)$ and $n_2(u) = |V(G)| - 1 - d(u)$ for any vertex u. By Proposition 6, one obtains

$$2|E(G)| - d(u) = \sum_{v\in N_1(u)} d(v) + \sum_{v\in N_2(u)} d(v) = Q_1(u) + Q_2(u)$$

This implies that

$$Q_2(u) = 2|E(G)| - d(u) - Q_1(u) = 2|E(G)| - d(u)\big(1 + m_1(u)\big)$$

From there it follows that

$$\langle m_2(G)\rangle = \frac{1}{|V(G)|}\sum_{u\in V(G)}\frac{Q_2(u)}{n_2(u)} = \frac{1}{|V(G)|}\sum_{u\in V(G)}\frac{2|E(G)| - d(u)\big(1+m_1(u)\big)}{|V(G)|-1-d(u)}$$

Hence, if two graphs of diameter 2 have identical d(u) and $m_1(u)$ for all vertices, we cannot discriminate between them based solely on the values of $\langle m_1(G)\rangle$ and $\langle m_2(G)\rangle$. ■

For an illustration, look at the polyhedral graphs of diameter 2 shown in Fig. 3.

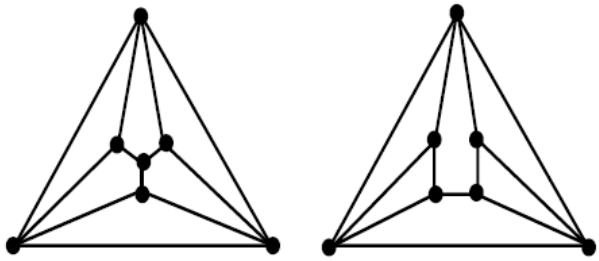


Figure 3
Graphs of diameter 2 that cannot be distinguished by $\langle m_1(G)\rangle$ and $\langle m_2(G)\rangle$

It can be easily verified that graphs G_b and G_c are characterized by the following identical topological parameters: E(3,3)=3, E(4,3)=6, E(4,4)=3, moreover, $\langle m_1(G_b)\rangle = \langle m_1(G_c)\rangle = 7/2$, $\langle m_2(G_b)\rangle = \langle m_2(G_c)\rangle = 23/7 = 3.285714$. It should be noted that for G_b and G_c the corresponding Wiener indices (W) are also identical, namely $W(G_b)=W(G_c)=30$ [1, 11].

The above examples demonstrate that there exist several molecular graphs having the same $\langle m_1(G)\rangle$ index. Moreover, in certain cases, indices $\langle m_i(G)\rangle$ for small values of i (i=1,2) still suffer from degeneracy and narrow numerical range.

As we have already mentioned the occurrence of degeneracy can be decreased by taking into account the degrees of neighboring vertices in $N_i(u)$, that is, the degrees of all vertices at distance i≥1 from u. Based on this concept, Randić and Plavšić proposed a descriptor of the following type [17]:

$$AVS(G) = \sum_{u\in V(G)} d(u) + \sum_{u\in V(G)}\sum_{i\geq 1} Q_i(u)P_i(u)$$

In the above formula, constants $P_i(u)$ are appropriately selected positive weights. In general, the weight is a strictly decreasing positive function of i. In the chemical literature, when $P_i = 1/2^i$ for i=1,2, ..., the AVS(G) index is called "the augmented valence sum" [17]. This is a useful measure of complexity of chemical graphs. It is interesting to note, that if $P_i(u)=1$ for any $i\geq 1$ and for any vertex u, then it follows from the Proposition 6

$$AVS(G) = \sum_{u\in V(G)} d(u) + \sum_{u\in V(G)} \sum_{i\geq 1} Q_i(u) = 2|E(G)| + \sum_{i\geq 1} Q_i(G) = 2|E(G)||V(G)|$$

Moreover, if $P_i(u)=1/m_i(u)$ for $i\geq 1$, then we get the equality

$$AVS(G) = 2|E(G)| + \sum_{u\in V(G)} \sum_{i\geq 1} n_i(u) = 2|E(G)| + |V(G)|\left(|V(G)| - 1\right)$$

Finally, if $P_i(u)=1/n_i(u)$ for $i\geq 1$, then one obtains

$$AVS(G) = 2|E(G)| + \sum_{u\in V(G)} \sum_{i\geq 1} m_i(u) = 2|E(G)| + |V(G)| \sum_{i\geq 1} \langle m_i(G)\rangle$$

as a particular case. The descriptor AVS(G) has a better discriminating ability than most other traditional topological indices [17]. The only drawback to the computation of AVS(G) descriptors is that in every cases, it is necessary to determine the corresponding graph distance matrix.

5 A Novel Set of Molecular Descriptors Based on Dissimilarity Functions

An alternative approach to improve the discriminating ability is to try to combine the information captured by $\langle m_1(G)\rangle$ and $\langle m_2(G)\rangle$. That would amount to quantifying the change in the average degree when one passes from distance 1 to distance 2 from a given vertex. Starting with this concept, we selected a topological quantity of the type

$$\frac{1}{|V(G)|} \sum_{u\in V(G)} D\left(m_1(u), m_2(u)\right).$$

In the expression above the non-negative function D(x,y) is a measure of dissimilarity of its arguments. There are several ways to meaningfully choose D(x, y). As a particular case, we consider here the dissimilarity function defined as D(x,y)= min(x,y)/max(x,y). In that way we obtain a new topological index

$$T(G) = \frac{1}{|V(G)|} \sum_{u\in V(G)} \frac{\min(m_1(u), m_2(u))}{\max(m_1(u), m_2(u))} \leq 1$$

It can be verified that T(G) discriminates between chemical graphs G_1 and G_2 in Fig. 1. Indeed, $T(G_1)=9553/11760=0.81233$ and $T(G_2)=19391/23520= 0.82447$. The result is even more interesting when we take into account the fact that G_1 and G_2 are isospectral [15]. A imilareffect appears on a pair of isospectral graphs depicted in Fig. 4. [30].

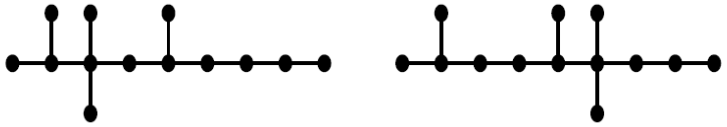


Figure 4
A pair of isospectral trees that cannot be distinguished by indices of the type X(G), but are discriminated by index T(G)

Moreover, for graphs of diameter 2 in Fig. 3 we have $T(G_b) = 0.8641002$ and $T(G_c)= 0.8868275$. Hence we have reasons to consider T(G) as a valuable addition to the repertoire of topological indices discriminating among isospectral and/or graphs of small diameter.

Concluding Remarks

We conclude the paper with some remarks on the properties of the topological descriptor T(G). According to our comparative studies performed on isospectral graphs and graphs of small diameter, it was found that the topological index T(G) has a quite low degeneracy. This is a favorable property when considering the efficiency of discrimination among real chemical graphs. Additionally, a practical advantage is that descriptors T(G) are simply computed. For this purpose, it is enough to determine the degree-distribution of the first and second order neighboring vertices (i.e. degrees of vertices at distances i= 1 and 2 from a given vertex u).

It is obvious from the definition that T(G) is equal to one for regular graphs. That suggests that T(G) could be used as a kind of measure of non-regularity of a graph. For a path P_n on n vertices we easily obtain $T(P_n) = 1 – 1/n$, in accordance with our perception of a path as a quite regular tree. Computing the value of T(G) for a star graph S_n on n vertices, however, we face a serious problem since $m_2(u)$ is equal to zero for the central vertex u. It follows that $T(S_n) = 1/n$, a result difficult to reconcile with the fact that the star is also a fairly regular tree. On the other hand, the drastic changes in the average degree of neighbors on distances 1 and 2 in the star graph are well captured by the index.

Generally, T(G) is likely to have problems whenever G has a well-connected vertices, i.e., vertices adjacent to all other vertices. However, that is not a serious problem in practical applications, since well-connected vertices are necessarily of a high degree, while chemically interesting and relevant graphs contain vertices of degree at most four.

Acknowledgment

Partial support of the Ministry of Science, Education and Sport of the Republic of Croatia (Grants No. 037-0000000-2779 and 177-0000000-0884 and a bilateral cooperation project) is gratefully acknowledged. This work was partially supported by the Hungarian National Office for Research and Technology (NKTH) as a part of a Bilateral Cooperation Program (under contract no. HR-38/2008).

References

[1] N. Trinajstić, Chemical Graph Theory, 2nd revised ed. CRC Press, Boca Raton, USA, 1992

[2] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, 2nd ed., Wiley–VCH, Weinheim, 2009

[3] M. Randić, The Connectivity Index 25 Years after, Journal of Molecular Graphics and Modelling, 20 (2001) 19-35

[4] S. Nikolić, G. Kovačević, A. Miličević, N. Trinajstić, The Zagreb Indices 30 Years After, Croat. Chem. Acta, 76 (2003) 113-124

[5] I. Gutman, K. Ch. Das, The First Zagreb Index 30 Years after, MATCH Commun. Math. Comput. Chem. 50 (2004) 83-92

[6] B. Zhou, N. Trinajstić, On a Novel Connectivity Index, J Math Chem 46, (2009) 1252-1270

[7] B. Furtula, A. Graovac, D. Vukičević, Atom-bond Connectivity Index of Trees, Discrete Applied Mathematics, 157 (2009) 2828-2835

[8] D. Vukičević, B. Furtula, Topological Index Based on the Ratios of Geometrical and Arithmetical Means of End-vertex Degrees of Edges, J Math Chem 46 (2009) 1369-1376

[9] D. Vukičević, M. Gašperov, Bond Additive Modeling 1. Adriatic Indices, Croat. Chem. Acta, 83 (2010) 243-260

[10] T. Došlić, T. Réti, D. Vukičević, On the Vertex Degree Indices of Connected Graphs, Chem. Phys. Lett. 512 (2011) 283-286

[11] I. Gutman, Y-N. Yeh, S-L. Lee, Y-L. Luo, Some Recent Results in the Theory of the Wiener Number, Indian Journal of Chemistry, 32A (1993) 651-661

[12] I. Gutman, Y-L. Luo, Y-N. Yeh, S-L. Lee, The Mean Isomer Degeneracy of the Wiener Index, Journal of the Chinese Chemical Society, 40 (1993) 195-198

[13] F. Buckley, F. Harary, Distance in Graphs, Addison–Wesley, Redwood, 1990

[14] D. B. West, Introduction to Graph Theory, Prentice-Hall, Upper Saddle River, NJ, 1996

[15] K. Balasubramanian, C. S. Basak, Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters, J. Chem. Inf. Comput. Sci. 38 (1998) 367-373

[16] M. Randić, On Complexity of Transitive Graphs Representing Degenerate Rearrangements, Croat. Chem. Acta, 74 (2001) 683-705

[17] M. Randić, D. Plavšić, On the Concept of Molecular Complexity, Croat. Chem. Acta, 75 (2002) 107-116

[18] I. Lukovits, S. Nikolić, N. Trinajstić, On Relationships between Vertex-degrees, Path-numbers and Graph Valence-shells in Trees, Chem. Phys. Lett. 354 (2002) 417-422

[19] K. Ch. Das, I. Gutman, Some Properties of the Second Zagreb Index, MATCH Commun. Math. Comput. Chem. 52, (2004) 103-112

[20] A. Ilić, D. Stefanović, On Comparing Zagreb indices, MATCH Commun. Math. Comput. Chem. 62 (2009) 681-687

[21] D. Vukičević, I. Gutman, B. Furtula, V. Andova, D. Dimitrov, Some Observations on Comparing Zagreb Indices, MATCH Commun. Math. Comput. Chem. 66 (2011) 627-645

[22] K. Ch. Das, Maximizing the Sum of Squares of Degrees, Discrete Math. 285 (2004) 57-66

[23] S. Yamaguchi, Estimating the Zagreb Indices and the Spectral Radius of Triangle- and Quadrangle-free Connected Graphs, Chem. Phys. Lett. 458 (2008) 396-398

[24] M. Gordon, G. R. Scantlebury, Non-Random Polycondensation: Statistical Theory of the Substitution Effect, Trans. Faraday Soc. 60 (1964) 604-621

[25] A. Yu, M. Lu, F. Tian, On the Spectral Radius of Graphs, Linear Algebra and its Applications 387 (2004) 41-49

[26] T. Réti, I. Gutman, D. Vukičević, On Zagreb Indices of Pseudo-Regular Graphs, Journal of Mathematical Nanoscience, 1 (2012) 1-12

[27] H. Hargitai, personal communication

[28] G. H. Hardy, J. E. Littlewood, G. Pólya, Inequalities, Cambridge University Press, Cambridge, 1952

[29] G.-X. Tian, T.-Z. Huang, B. Zhou, A Note on Sum of Powers of the Laplacian Eigenvalues of Bipartite Graphs, Linear Algebra and its Applications, 430 (2009) 2503-2510

[30] E. R. van Dam, W. H. Haemens, Which Graphs Are Determined by Their Spectrum? Linear Algebra and its Applications, 373 (2003) 241-272

[31] B. Zhou, Zagreb Indices, MATCH Commun. Math. Comput. Chem. 52 (2004) 113-118

Shaft Sensor-less FOC Control of an Induction Motor Using Neural Estimators

Peter Girovský, Jaroslav Timko, Jaroslava Žilková

Department of Electrical Engineering and Mechatronics, Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic
peter.girovsky@tuke.sk, jaroslav.timko@tuke.sk, jaroslava.zilkova@tuke.sk

Abstract: The paper deals with a shaft sensor-less field oriented control structure for an induction motor based on neural network estimators. The first part presents the theoretical knowledge. The second part presents the simulation and results of designing neural estimators for observing the magnetic flux and the motor angular speed for induction motor field oriented control in MATLAB-Simulink. Controllers for simulation of shaft sensor-less field oriented control have been designed by state space method. An achieved simulation result of the neural angular speed estimator has been verified by system of AC converter – induction motor by Real-Time system.

Keywords: induction motor; neural network; sensor-less control; vector control

1 Introduction

Motors play important roles in industrial production and in many other applications. In their early days, DC motors had the advantage of precise speed control when utilized for the purpose of accurate driving. However, DC motors have the disadvantage of brush erosion, maintenance requirements, environmental effects, complex structures and power limits. On the other hand, induction motors are robust, small in size, low in cost, and almost maintenance-free.

Hasse [9] and Blaschke [10] developed a field oriented control theory to simplify the structure of IM speed control used to drive the DC motor. In recent years, the field oriented control theory has become more feasible due to progress in the development of electronics techniques and high-speed microprocessors. Nonlinear control problems can often be solved if full state information is available; in the IM case, the rotor states are immeasurable and often it is too costly to monitor the angular speed of the rotor.

In most applications, speed sensors are necessary in the speed control loop. On the other hand, there are applications where lower performance is required, cost reduction and high reliability are necessary, or a hostile environment does not

allow for using speed sensors. In these fields, speed sensor-less IM control can be usefully applied. Many different solutions for the estimation of states variables or model parameters have been proposed recently, for example, estimators utilizing motor construction properties, estimators based on the drive dynamic model or estimators based on artificial intelligence [7, 8, 13, 15, 16].

Sensor-less controllers have been proposed which depend on adaptive control and observer theory, on optimal observer design by applying Kalman filter theory [11, 12], on sliding mode control [2, 3], and on using artificial intelligence methods [1, 4, 5, 6, 14].

At present, requirements on the dynamic precision are not too strict and virtual or soft sensors are alternatively successfully utilized. Estimators based on artificial intelligence are divided into the following groups:

- systems based on the fuzzy logic,
- systems based on neural networks,
- systems based on hybrid systems,
- systems based on evolutionary algorithms (genetic algorithms).

2 Simulation Design of a Neural Estimator for Field Oriented Control of Induction Motor

The neural modelling can perform estimation of the induction motor angular speed or of other non-measurable variables on the neural networks base.

Nowadays, there are field oriented controlled drives based on different solutions and performances which are commonly used in industry. With field-oriented techniques, the decoupling of flux and torque control commands of the IM is guaranteed, and the induction motor can be controlled linearly, like a separately excited DC motor. The DC motor like performance can be obtained by preserving a fixed and orthogonal orientation between the field and armature fields in the induction motor by orientation of the stator current with respect to the rotor flux in order to attain independently controlled flux and torque. Using the field oriented control principle, the stator current component i_{d1} is aligned in the direction of the rotor flux vector and the stator current component i_{q1} is aligned in the direction perpendicular to it. The rotor flux orientation in the squirrel-cage rotor IM cannot be directly measured, but it can be obtained from terminal variables.

After using transformation of coordinates d, q to the rotating system x-y, the electric torque is proportional to the i_{1y} component and the relation between the rotor flux and i_{1x} component is given by the first order linear transfer function with $T_2 = L_2/R_2$ time constant.

From this fact and for the considered flux control, the stator current and voltage components were chosen as input signals for the reconstruction of the induction motor speed. The developed estimators were trained according to selected training patterns from the direct field oriented control of the induction motor. Block diagram of the control scheme is presented in Figure 1.

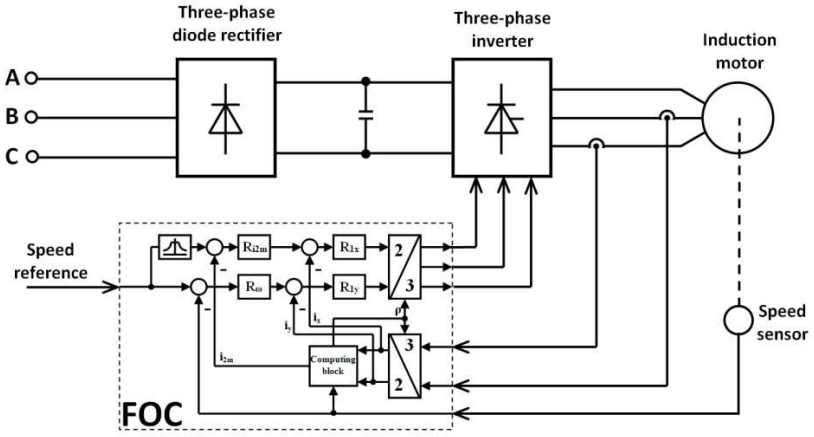


Figure 1
Basic field oriented control scheme

2.1 Induction Motor FOC Simulation Design

Field oriented control simulation design had been made for induction motor with the following parameters: P_n=0,75 kW; U_n= 220V/380V; I_n= 3,8 A/2,2 A; n_n= 1380 rpm; p=2; s=0,08; J=5,4.10^{-3} kgm^2

In the design of state control by method of the poles determine for two input variables and one output based on the following equations:

$$\sigma T_1 \frac{di_{1x}}{dt} + i_{1x} = \frac{K_T u_{1x}}{R_1} - (1-\sigma) T_1 \frac{di_{2m}}{dt} + \sigma T_1 \omega_{2m} i_{1y} \tag{1}$$

$$\sigma T_1 \frac{di_{1y}}{dt} + i_{1y} = \frac{K_T u_{1y}}{R_1} - (1-\sigma) T_1 \omega_{2m} i_{2m} - \sigma T_1 \omega_{2m} i_{1x} \tag{2}$$

$$T_2 \frac{di_{2m}}{dt} + i_{2m} = i_{1x} \tag{3}$$

$$\frac{i_{1y}}{T_2 i_{2m}} + \omega = \omega_{2m} \tag{4}$$

$$\frac{J}{p} \frac{d\omega}{dt} = \frac{3p}{2} \frac{L_h}{1+\sigma_2} i_{2m} i_{1y} - m_z \tag{5}$$

Define the state variables: $i_{2m}=x_1$; $i_{1x}=x_2$; $\omega=x_3$; $i_{1y}=x_4$; $m_z=z$; $u_1=u_{1x}/K_T$; $u_2=u_{1y}/K_T$

Then, written can be the state equation for induction motor:

$$\dot{\underline{x}} = \begin{bmatrix} -a_1 x_1 + a_1 x_2 \\ f_2(\underline{x}) \\ a_3 x_1 x_4 \\ f_4(\underline{x}) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ b & 0 \\ 0 & 0 \\ 0 & b \end{bmatrix} . \underline{u} + \begin{bmatrix} 0 \\ 0 \\ -e \\ 0 \end{bmatrix} . z = a(x) + B.u + e.z \tag{6}$$

$$y = x_3 = c(x)$$

The constants and functions used in the state equation (6):

$$a_1 = \frac{1}{T_2}; \quad a_2 = \frac{1}{\sigma T_1}; \quad a_3 = \frac{3p^2}{2J}\frac{L_h}{1+\sigma_2}; \quad a_4 = \frac{1-\sigma}{\sigma}; \quad \sigma_2 = \frac{L_{2\sigma}}{L_h}; \quad \sigma = 1 - \frac{L_h^2}{L_1.L_2};$$

$$b = \frac{K_T}{\sigma L_1}; \; e = \frac{p}{J}$$

$$f_2(\underline{x}) = a_1.a_4.x_1 - (a_2 + a_1.a_4).x_2 + x_3.x_4 + a_1.\frac{x_4^2}{x_1}$$

$$f_4(\underline{x}) = -(a_2 + a_1.a_3).x_4 - a_4.x_1.x_3 - x_2.x_3 - a_1.\frac{x_2.x_4}{x_1}$$

Nonlinear function $f_2(x)$, $f_4(x)$ in the control scheme shown in Fig. 2 compensating for introduction of control u, so as to simplify the state equation:

$$u = \frac{1}{b}\begin{bmatrix} -f_2(x) + v_2 - r.x_2 \\ -f_4(x) + v_4 - r.x_4 \end{bmatrix}$$

$$\dot{\underline{x}} = \begin{bmatrix} -a_1 & a_1 & 0 & 0 \\ 0 & -r_2 & 0 & 0 \\ 0 & 0 & 0 & a_3 \\ 0 & 0 & 0 & -r_4 \end{bmatrix} . \underline{x} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} . \begin{bmatrix} v_2 \\ v_4 \end{bmatrix}$$

2.1.1 Current Sub-Circuit (R_{1x}, R_{1y})

$$\begin{bmatrix} \dot{x}_2 \\ \dot{v}_2 \end{bmatrix} = \begin{bmatrix} -r_2 & 1 \\ -K_2 & 0 \end{bmatrix} . \begin{bmatrix} x_2 \\ v_2 \end{bmatrix} + \begin{bmatrix} 0 \\ K_2 \end{bmatrix} . w_2$$

$$v_2 = k_2 \int (w_2 - x_2).dt \; ; \; \dot{v}_2 = K_2.(w_2 - x_2)$$

The characteristic polynomial of system $P_{(\lambda)}$:

$$P_{(\lambda)} = \det\left(\lambda.\underline{I} - \underline{A}\right) = \det\begin{bmatrix} \lambda + r_2 & -1 \\ \mathrm{K}_2 & \lambda \end{bmatrix} = \lambda^2 + r_2.\lambda + K_2$$

For current controller select the damping $d=0.85$, regulation time $t_r=0.05s$ and determine the desired characteristic polynomial $P_{(s)}$:

$$P_{(s)} = s^2 + 146.s + 7354$$

By comparing the characteristic polynomial and the desired characteristic polynomial we obtain controller constants K_2, K_4 and r_2, r_4 where $K_2 = K_4$ and $r_2 = r_4$.

2.1.2 Superior Circuit of Magnetizing Current (R_{i2m})

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{v}_2 \\ \dot{v}_1 \end{bmatrix} = \begin{bmatrix} -a_1 & a_1 & 0 & 0 \\ 0 & -r_{12} & 1 & 0 \\ -K_2.r_{11} & -K_2\left(1+r_{12}\right) & -K_2.d_{12} & K_2 \\ -K_1 & 0 & 0 & 0 \end{bmatrix}.\begin{bmatrix} x_1 \\ x_2 \\ v_2 \\ v_1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ K_2 \end{bmatrix}.w_1$$

$$w_2 = v_1 - r_{11}.x_1 - r_{12}.x_2 - d_{12}.v_2 \; ; \; \dot{v}_1 = K_1.\left(w_1 - x_1\right)$$

The characteristic polynomial of system $P_{(\lambda)}$:

$$P_{(\lambda)} = \det\left(\lambda.\underline{I} - \underline{A}\right) = \lambda^4 + \lambda^3\left(a_1 + r_{12} + K_2.d_{12}\right) + \lambda^2\left(a_1.r_{12} + a_1.K_2.d_{12} + K_2 + K_2.r_{12}\right) +$$

$$+ \lambda.K_2.a_1.\left(r_2.d_{12} + 1 + r_{12} + r_{11}\right) + a_1.K_1.K_2$$

Select the damping $d=0.85$, regulation time $t_r=0.1s$ and determine the desired characteristic polynomial $P_{(s)}$:

$$P_{(s)} = s^4 + 214.s^3 + 1793s^2 + 162020.s + 9079700$$

By comparing of the characteristic polynomial and the desired characteristic polynomial we obtain controller constants K_1, r_{11}, r_{12}, d_{12}.

2.1.3 Superior Circuit of Speed (R_ω)

$$\begin{bmatrix} \dot{x}_3 \\ \dot{x}_4 \\ \dot{v}_4 \\ \dot{v}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & a_3 & 0 \\ 0 & 0 & 1 & -r_4 \\ -K_4.r_{33} & -K_4\left(1+r_{34}\right) & -K_3.d_{34} & K_4 \\ -K_3 & 0 & 0 & 0 \end{bmatrix}.\begin{bmatrix} x_3 \\ x_4 \\ v_4 \\ v_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ K_3 \end{bmatrix}.w_3$$

$$w_4 = v_3 - r_{33}.x_3 - r_{34}.x_4 - d_{34}.v_4 \; ; \; \dot{v}_3 = K_3.(w_3 - x_3)$$

The characteristic polynomial of system $P_{(\lambda)}$:

$$P_{(\lambda)} = \det(\lambda.\underline{I} - \underline{A}) = \lambda^4 + \lambda^3(r_{34} + K_4.d_{34}) + \lambda^2.K_4(r_4.d_{34} + 1 + r_{34}) + \lambda.(a_3.K_4.r_{33}) + a_3.K_3.K_4$$

Select the damping $d=0.85$, regulation time $t_r=0.1s$ and determine the desired characteristic polynomial $P_{(s)}$:

$$P_{(s)} = s^4 + 214.s^3 + 1793s^2 + 162020.s + 9079700$$

By comparing of the characteristic polynomial and the desired characteristic polynomial obtained are controller constants K_3, r_{33}, r_{34}, d_{34}.

Constants calculated for all circuits of controller are used by the field oriented control scheme shown in Figure 2.

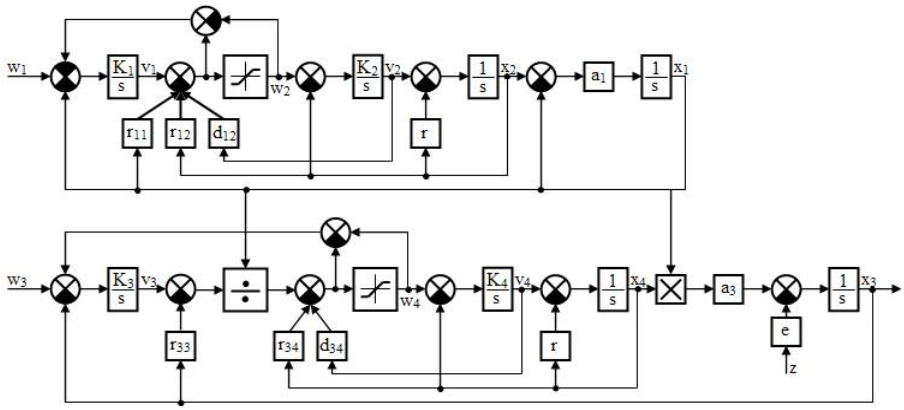


Figure 2
Simulation field oriented control scheme

2.2 Magnetising Current Neural Estimator

If for vector control the *x*-th component of the stator current vector is considered as a basis of current-creating component, then the magnetising current i_{2m} estimator will process current-creating component of the stator current.

As mentioned above, the magnetising current i_{2m} neural estimator bases its estimation of the current-creating component of stator current i_{1x}. Dependence between currents i_{2m} and i_{1x} is linear, and hence the estimator can be made up of a feed-forward neural network without any hidden layer. For the activating function, the *purelin* linear function can be used. The input data vector consists of values of the stator current i_{1x} in step *(k)* and step *(k-1)*, respectively, and also the preceding value of magnetising current i_{2m} in step *(k-1)*. A basic diagram of such neural estimator is shown in Figure 3.

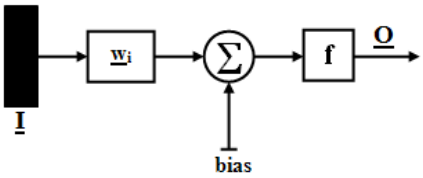


Figure 3
Basic diagram of magnetising current i_{2m} neural estimator

Here, $\underline{O}$ stands for output values vector, $\underline{I}$ is the input data vector, and w_i presents weights of individual connections of neurons.

$$\underline{O} = f\left[\sum \underline{I}.\underline{w}_i + bias\right] \tag{7}$$

Substituting the input matrix to equation (7), we will obtain the equation for the magnetising current neural estimator in the following form:

$$i_{2m}(k) = purelin\left(\begin{bmatrix} i_{1x}(k) \\ i_{1x}(k-1) \\ i_{2m}(k-1) \end{bmatrix} \underline{w}_i + bias\right) \tag{8}$$

where current $i_{2m}(k)$ is the output variable and the input variables are $i_{1x}(k)$, $i_{1x}(k-1)$ and $i_{2m}(k-1)$.

2.3 Speed Neural Estimator

If for the basis of torque-creating component we establish the *y*-th component of the vector, then the speed estimator will estimate this torque creating component from the stator voltage and current.

As was already mentioned above, the angular speed ω neural estimator bases its estimation on the torque component of stator voltage u_{1y} and current i_{1y}. The relation between the input and output quantities is not represented by a simple linear dependency, and this is the reason why for the estimation a cascade neural network with one hidden layer consisted of eight neurons will be used. As an activating function for the hidden layer used, there was the *tansig* nonlinear function and for the output layer used was a *purelin* linear function. The input data vector is represented by values of stator voltage u_{1y} and stator current i_{1y} in steps *(k)* and *(k-1)*, as well as by value of magnetising current i_{2m} in steps *(k)* and *(k-1)*. Basic diagram of such a neural estimator is shown in Figure 4.

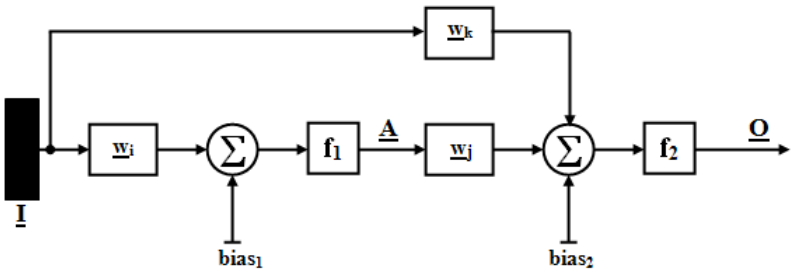


Figure 4
Basic diagram of ω motor angular speed neural estimator

In the figure, $\underline{O}$ is the output values vector, $\underline{I}$ presents a vector of input variables and w_i, w_j, w_k are weights of individual connections of neurons.

$$\begin{aligned} \underline{A} &= f_1\left[\sum \underline{I}.\underline{w}_i + bias_1\right] \\ \underline{O} &= f_2\left[\sum \left(\underline{A}.\underline{w}_j + \underline{I}.\underline{w}_k\right) + bias_2\right] \end{aligned} \tag{9}$$

Post substituting the input matrix to equation (9) the neural speed estimator can be described by the following equation:

$$\omega(k) = purelin\left(\begin{bmatrix} u_{1y}(k) \\ u_{1y}(k-1) \\ i_{1y}(k) \\ i_{1y}(k-1) \\ i_{2m}(k) \\ i_{2m}(k-1) \end{bmatrix} \underline{w}_j + tansig\left(\begin{bmatrix} u_{1y}(k) \\ u_{1y}(k-1) \\ i_{1y}(k) \\ i_{1y}(k-1) \\ i_{2m}(k) \\ i_{2m}(k-1) \end{bmatrix} \underline{w}_i + bias1\right) \underline{w}_k + bias2\right) \tag{10}$$

where the output quantity is $\omega(k)$ angular speed value and where the input are values $u_{1y}(k)$, $u_{1y}(k-1)$, $i_{1y}(k)$, $i_{1y}(k-1)$, $i_{2m}(k)$ and $i_{2m}(k-1)$.

3 Simulation Results

In the following, we show the simulation results of sensor-less vector control of an induction motor when applying neural estimators of the speed and magnetising current, respectively.

The principal diagram of the vector control with connected neural estimators of the magnetising current and speed is shown in Figure 5.

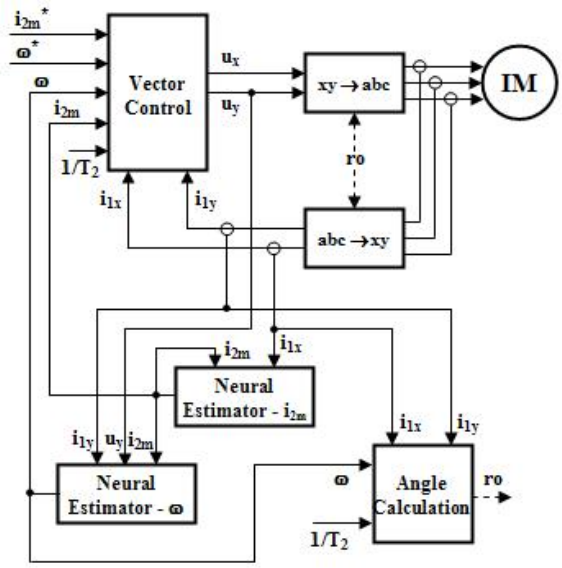


Figure 5
Basic diagram of vector control with neural estimators

Simulation, design and training of neural estimators were performed for the induction motor with parameters: P_n=0,75 kW; U_n= 220V/380V; I_n= 3,8 A/2,2 A; n_n= 1380 rpm; p=2; s=0,08; J=5,4.10^{-3} kgm^2

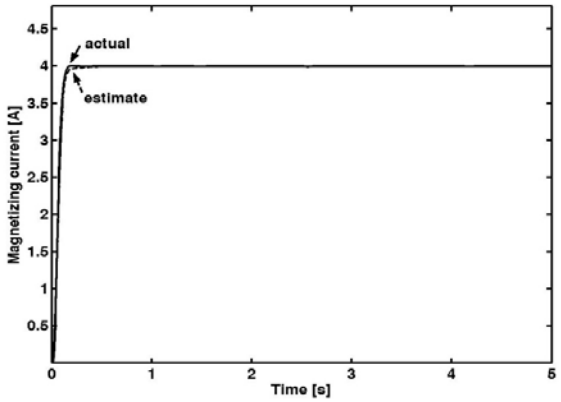


Figure 6
Comparison of the estimated versus actual magnetising current

Figures 6 and 7 show a comparison of real and observed values of the magnetizing current and the angular speed. A dashed line shows there is the required angular speed value during starting, reversing and loading transients. In time of 2s the motor was loaded by the rated load torque.

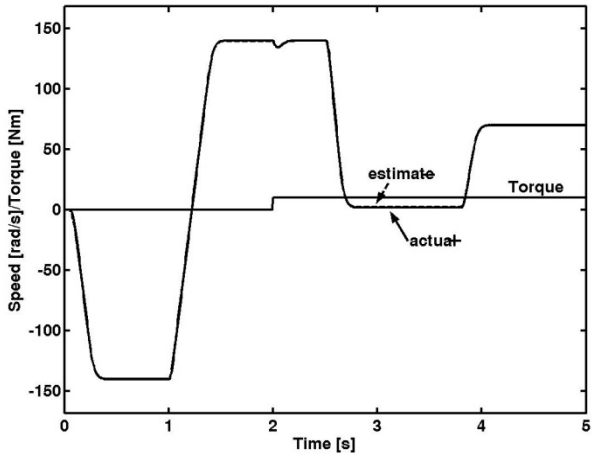


Figure 7
Comparison of the estimated versus actual speed of the IM

The waveforms shown in Figures 6 and 7 are valid for the case of no feedback to control from the neural observers but led directly from the motor mathematical model.

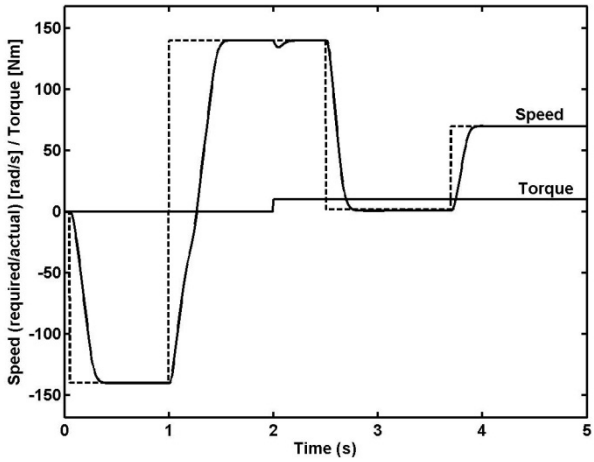


Figure 8
Transients of desired versus real angular speed and the motor load torque

Shown in Figure 8 is a simulated response of the induction motor angular speed (in solid line) at conditions identical with the previous one, shown in Figure 7. In this case, and the same as in any following ones, the feedback to control was introduced from neural observers of the magnetising current and angular speed.

3.1 Experimental Verification

For verification of simulation results, an experimental Real-Time system based on RT-LAB system was used. The principal scheme of the whole system is show in Figure 9.

The experimental system consists of the SIMOVERT MASTERDRIVES Vector Control and an induction motor with the same parameters as those of the motor used for simulation. Used as the load there was a dynamo with resistor and the base of this experimental system consists of Real-time system with NI PCI-6025E.

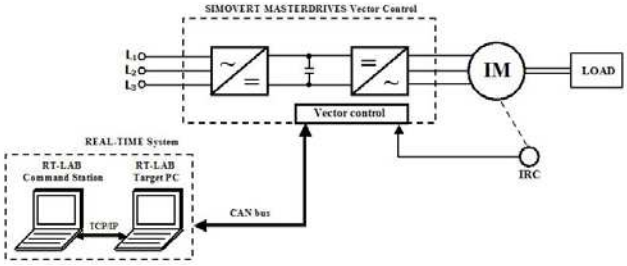


Figure 9
Principal scheme of Real-Time system

3.2 Neural Estimation for Experimental Verification

Regarding different ways of vector control in SIMOVERT MASTERDRIVES, a Vector Control (system in rotary coordinates *d-q*), used for design of speed neural estimator, was the input stator voltage in step *(k)* – $\boldsymbol{u}(k)$, in the step *(k-1)* - $\boldsymbol{u}(k-1)$ and value of current components *d,q* in step (k) – $\boldsymbol{i}_d(k)$, $\boldsymbol{i}_q(k)$ and in step *(k-1)* - $\boldsymbol{i}_d(k-1)$, $\boldsymbol{i}_q(k-1)$. For off-line training using the Levenberg-Marquardt algorithm 126013 samples in aggregate were used. The output vector for training is represented by value of the rotor speed $\hat{o}(k)$ in step *(k)*.

For the speed neural estimator we used a cascade neural network with one hidden layer having six input neurons and six hidden neurons. For the hidden layer activating function used was the *tansig* nonlinear function and for the output layer we used *purelin* linear function. Using them we obtained an equation for neural estimator of speed in the following form:

$$\hat{o}(k) = purelin\left(\begin{bmatrix} u(k) \\ u(k-1) \\ i_d(k) \\ i_d(k-1) \\ i_q(k) \\ i_q(k-1) \end{bmatrix} \underline{w}_j + \boldsymbol{tansig}\left(\begin{bmatrix} u(k) \\ u(k-1) \\ i_d(k) \\ i_d(k-1) \\ i_q(k) \\ i_q(k-1) \end{bmatrix} \underline{w}_i + bias1\right) \underline{w}_k + bias2\right) \tag{11}$$

4 Experimental Results

Presented in the following are the simulation results of sensor-less vector control of an induction motor when using neural estimators of speed. The principal diagram of vector control effected with the use of neural estimator of speed is shown in Figure 10.

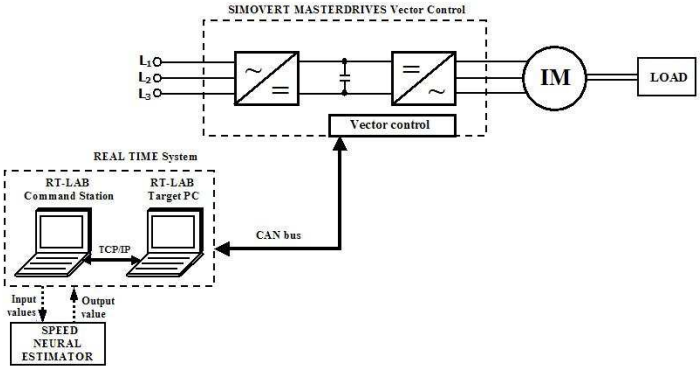
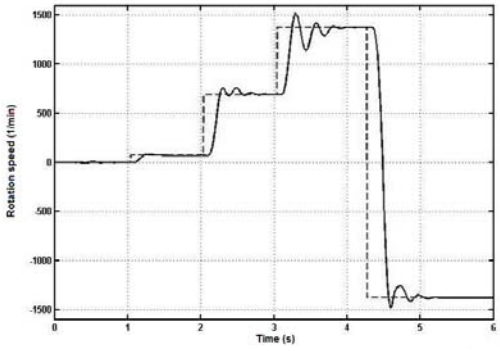
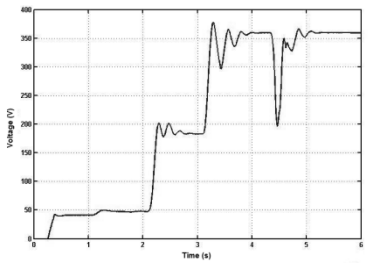


Figure 10

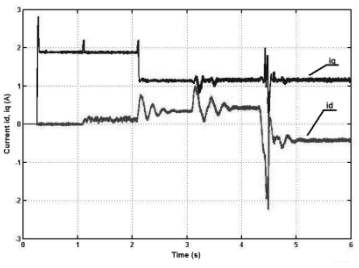
Principal scheme of Real-Time system with neural estimator



a) speed



b) voltage



c) current i_d, i_q

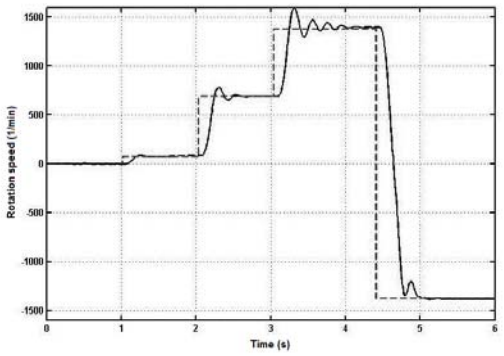
Figure 11

Time courses of desired versus actual rotor speed without load and relevant voltage and current

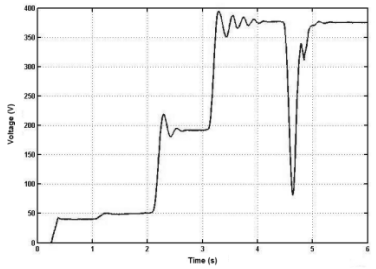
In Figures 11 and 12 are shown courses of desired (dash line) and actual rotor speeds of induction motor in the vector control using the scheme according to Figure 10.

At time 1 second, the required value of rotor speed changed from 0% to 5% of nominal speed, at time 2 seconds, from 5% to 50%; at time 3 seconds, from 50% to 100% of nominal rotor speed and at time 4 seconds, the induction motor reversed.

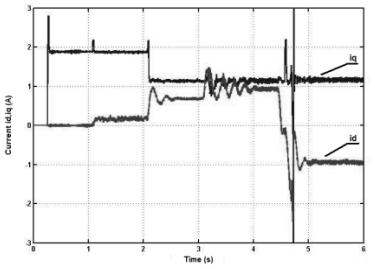
For verification we used an experimental real time system. The results obtained, illustrated by respective waveforms, validate the possibility of utilising artificial neural networks in sensor-less vector control of the induction motor. The drive features better adaptability and robustness in comparison with a drive without estimator.



a) speed



b) voltage



c) current i_d, i_q

Figure 12

Time courses of desired versus actual rotor speed with load and relevant voltage and current

Conclusions

The paper is concerned with designing induction motor neural estimators. Based on easily measurable quantities, such as components of stator current and voltage, we designed estimators of the motor speed and magnetising current, utilizing feed-forward and cascade neural networks. Both these networks were trained off-line

using the Levenberg-Marquardt algorithm, which is a modification of the traditional back-propagation training algorithm.

The results arrived at, illustrated by respective waveforms, validate the possibility of utilising artificial neural networks in the sensor-less vector controlling of an induction motor, while also taking advantage of their advantageous properties, such as adaptability and robustness.

At the end of this paper presented are research results. For applied verification used was experimental Real-Time system.

Acknowledgement

The authors wish to thank for the support to the R&D operational program Centre of excellence of power electronics systems and materials for their components II. No. OPVaV-2009/2.1/02-SORO, ITMS 26220120046 funded by European regional development fund (ERDF).

References

[1] Vas P., "Artificial-Intelligence-based Electrical Machines and Drives", Oxford University Press, Oxford, 1999

[2] Vittek J., Dodds S. J., Makyš P., Lehocký P., "An Observer Design for Forced Dynamics Control of AC Drives", Transcom 2007, Žilina, Slovensko, pp. 207-210, 2007

[3] Vittek J., Bris P., Štulrajter M., Pácha M., "Chattering Free Sliding Mode Control Law for Position Control of the Drive Employing Induction Motor", Power Engineering Conference 2008, AUPEC '08, Australasian Universities, pp. 1-6, 14-17 Dec., 2008

[4] Kuchar M., Brandštetter P., Kaduch M., "Sensor-less Induction Motor Drive with Neural Network". IEEE, Annu. Power Elec. Specialists Conf.m pp. 3301-3305, 2004

[5] Jovankovič J., Žalman M., "Application of the Virtual Sensors Based on the Artificial Neural Networks", EDPE'03, International conference, Slovakia, 2003, pp. 486-490

[6] Bensalem Y., Abboud W., Sbita L., Abdelkrim M. N., "A Sensor-less Neural Network Speed Control of Induction Motor Drive", Int. Journal of Signal System Control and Engineering Application 1 (2): pp. 150-158, 2008

[7] Jadlovská A., Kabakov N., Sarnovský J., "Predictive Control Design Based on Neural Model of a Non-linear System", Acta Polytechnica Hungarica, Vol. 5, No. 4, pp. 93-108, 2008

[8] Jamuna V., Reddy S. R., "Modeling and Speed Control of Induction Motor Drives Using Neural Network", Annals of Diarea de Jos University of Galati, III, Vol. 33, No. 1, pp. 40-49, 2010

[9] Hasse K., "Zur Dynamik Drehzahlgeregelter Antriebe mit stromrichtergespeisten Asynchron-kurschlusslaüfer Maschinen", Techn. Hochsch: Darmstadt, Dissertation, 1969, pp. 74-78

[10] Blaschke F., "The Principle of Field Orientation as Applied to the New Transvektor Closed–Loop Control System for Rotating-Field Machines", Siemens Rev. 39 (5): 1972, pp. 217-220

[11] Meziane S., Toufouti R., Benalla H., "Nonlinear Control of Induction Machines Using an Extended Kalman Filter", Acta Polytechnica Hungarica, Vol. 5, No. 4, pp. 41-58, 2008

[12] Žilková J., Timko J., Berko J., "Speed Sensor-less Control of an Induction Motor Drive Using Extended Kalman Filter", In: Acta Technica ČSAV 50, No. 4, Prague, 2005, pp. 279-289

[13] Timko J., Žilková J., Balara D., "Artificial Neural Networks Application in Electrical Drives", (in Slovak), Calypso s.r.o., Košice, 2002, p. 239, ISBN 80-85723-27-1

[14] Timko J., Žilková J., Girovský P., "Shaft Sensor-less Vector Control of an Induction Motor", In: Acta Technica CSAV, Vol. 52, No. 1 (2007), pp. 81-91, ISSN 0001-7043

[15] Timko J., Žilková J., Girovský P., "Modeling and Control of Electrical Drives Using Neural Networks", (in Slovak), C-Press, Košice, 2009, p. 202, ISBN 978-80-8086-124-7

[16] Žilková J., Timko J., Girovský P., "Nonlinear System Control Using Neural Networks", Acta Polytechnica Hungarica, Vol. 3, No. 4, pp. 85-94, 2006

Fines Content and Cyclic Preloading Effect on Liquefaction Potential of Silty Sand: A Laboratory Study

Ahmed Arab, Mostefa Belkhatir

Laboratory of Materials sciences and Environment
Civil Engineering Department
Hassiba Bebouali University of Chlef
BP 151 Route de Sendjes
02000 Chlef
Algeria
ah_arab@yahoo.fr; abelkhatir@yahoo.com

Abstract: This paper presents a laboratory study of the influence of low plastic fines and preloading on the cyclic behaviour of silty sand. The study is based on undrained triaxial cyclic tests which were carried out for fines content ranging from 0 to 40%. The paper is composed of three parts. The first one presents the characteristics of soils used in this study; the second provides an analysis of the effect of low plastic fines on the cyclic behaviour of the sand-silt mixtures. The third part presents the effect of the preloading on the soil liquefaction. The test results indicate that the liquefaction potential of the mixtures decreases with increasing the fines content until Fc=20%, after which the potential of liquefaction increases moderately with the fines content Fc=40%. The over-consolidation and the cyclic drained preloading of low stress amplitude improved the liquefaction resistance of the sand-silt mixtures.

Keywords: fines; sand; undrained; potential; liquefaction; over-consolidation; cyclic preloading

1 Introduction

The city of Chlef (Algeria) was touched by many earthquakes last century. One of the most violent earthquakes occurred on October 10, 1980 with a magnitude of 7.3 degrees on the Richter scale, causing several thousand casualties among the population (more than 2500 dead) and much damage to buildings. The phenomenon of liquefaction was observed in several places and especially on the banks of the Wadi Chellif (river), as shown in Figs. 1, 2 and 3. The soil deposits in the area of Chlef are composed of a mixture of sand and fine grained soils.

The former investigations and researches were initially carried out to study the liquefaction of clean sands. Since the nineties special interest has been given to the liquefaction of the sands mixed with fine grained soils in order to study the effect of plastic and non-plastic fines on the behaviour of those mixtures and their liquefaction. These studies did not lead to a consensus on the influence of fines on the behaviour of those soils. Indeed, certain studies reported that an increase in the amount of silt leads to an increase in the resistance to liquefaction [1, 5, 6]; others showed that this increase reduces the resistance to liquefaction [2, 8, 16, 22, 25, 26, 28, 29,]. Other studies showed that the increase in the fraction of fines initially leads to a decrease of the resistance to liquefaction until a certain limiting silt content, then the resistance increases [3, 13, 17, 19, 23]. Finally, some of the more recent studies [14, 22, 24, 26] showed that the resistance to liquefaction of silty sands is more closely related to the soil skeleton void ratio than to its silt content (fines).

The over-consolidated soils (preloaded) are often found in nature especially with the thawing of the glaciers, scouring of the grounds, phenomenon of erosion, fluctuation of ground water, etc.

Several researchers studied and showed that the over-consolidation ratio has a significant effect on the liquefaction resistance of soils [9, 20, 21]. They observed that liquefaction resistance increases with the over-consolidation ratio. This effect becomes very important with the increase in the percentage of fines. By carrying out cyclic tests on the sand of Hostun, [3] found that with an over-consolidation ratio of 7, the liquefaction is obtained at the end of 17 cycles, whereas with a normally consolidated sample, 6 cycles are needed to reach liquefaction phenomenon. [7] studied the influence of stress history on the liquefaction resistance of sands. They showed that this resistance increases when the sample is initially subjected to a small cycles of loading followed by a drainage; other authors showed that a pre-shearing with great amplitudes can lead to a reduction of the cyclic resistance. [10, 18] interpreted this behaviour using the concept of 'phase transformation line/characteristic state', while specifying that any cyclic loading followed by a drainage in the contracting zone leads to the densest state of the material without modifying its structure, resulting in an increase in the cyclic shear strength. In contrast, if the preloading comprises a way in the dilating field, there is an untangling of grains, leading to a new structure with a lower resistance.

Ishihara and Okada [11] studied the influence of preliminary strong distortions on the liquefaction resistance of sands. The tests were carried out with the traditional triaxial apparatus according to two procedures. In the first procedure, the initial cycle is stopped when the deviator stress is cancelled; in the second, the initial cycle is completed by a loading in extension or compression to cancel the residual strain. They showed that there is insignificant influence of the preliminary strain on the cyclic shear strength when the preloading ends with a compression phase. In contrast, when the initial loading ends with an extension phase, the behaviour of sand shows a contractance and the cyclic shear strength is greatly reduced. [4]

realized a series of tests on samples subjected initially to preloading for two levels of axial strain. They found that a large preloading in compression (ε_1= +5%) or in extension (ε_1= -5%) induces a significant reduction of the liquefaction resistance. [27] found that a sample without initial preshearing presented a low liquefaction potential whereas the samples subjected to preshearing of low amplitude had a very great potential of liquefaction.

In this paper, we present a laboratory study dealing with the behaviour of a sand-silt mixture to study the cyclic undrained shear response using different fines contents ranging from 0 to 40% for the purpose of reinforcing the soil that is prone to liquefaction. These tests were carried out to better understand the effect of low plastic fines on the cyclic behaviour of the mixtures (potential of liquefaction) as well as the influence of the preloading. The paper is composed of three parts. In the first part we present the materials. The second part provides an analysis of the test results and discusses the influence of fines on the behaviour of the sand-silt mixtures. The third part presents the effect of preloading on the liquefaction potential of the sand-silt mixtures.

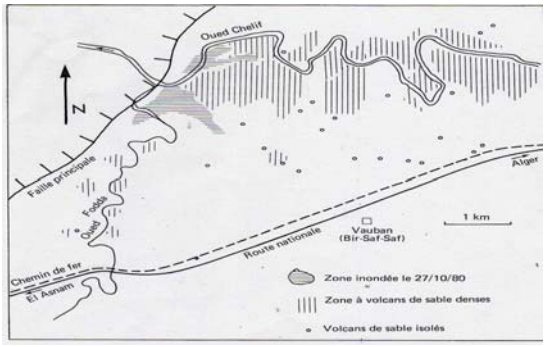


Figure 1
Location of liquefied soil (El Asnam, 1980)

Figure 2
Sand craters of liquified zones (El Asnam, 1980)

Figure 3
Sliding of the Chlef River banks (El Asnam, 1980)

2 Materials Tested

The laboratory tests were carried out on reconstituted uniform sand samples of Rass (Algeria) mixed with the silt of Sidi-M'hamed having a plasticity index of 2.33 (classified as a low plastic silt).

The sand of Rass comes from the accesses of the bed of Oued/Rass River (confluence of Oued/Chlef and Oued/Rass). It is a medium sand with an average diameter of D_{50} = 0.39 mm. Fig. 4 shows the grain size distribution curve of Rass sand and the silt used (named with the symbol SM). Tables 1, 2 and 3 give the summary on index properties of the sand, silt and the chemical analysis.

The dimensions of the samples were 70 mm in diameter and 70 mm in height in order to avoid the appearance of the instability (sliding surfaces) and buckling. After the specimen was formed, the specimen cap was placed and sealed with O-rings, and a partial vacuum of 20 kPa is applied to the specimen to reduce disturbances. Saturation was performed by purging the dry specimen with carbon dioxide for approximately 30 min. De-aired water was introduced into the specimen from the bottom drain line. Water was allowed to flow through the specimen until an amount equal to the void volume of the specimen was collected in a beaker through the specimen's upper drain line. A minimum Skempton [25] test specimen were isotropically consolidated at a mean effective pressure of 100 kPa subjected to undrained monotonic and cyclic triaxial loading with a constant strain rate of 0.167% per minute.

The experimental program includes alternate cyclic undrained tests on the mixture sand-silt with a relative density of Dr = 65% and a silt content ranging from 0 to 40% (ratio of the mass of fines on the mass of the sample) under an initial effective confining pressure σ'_c = 100 kPa (Cell pressure = 500 kPa and Back pressure = 400 kPa). The amplitude levels of these cycles (q_m) are equal to 30, 50 and 70 kPa. The loading level (CSR) is defined by:-

$$CSR = \frac{q_m}{2\sigma'_c} \tag{1}$$

q_m and σ'_c are the cyclic loading amplitude and the initial mean effective stress, respectively.

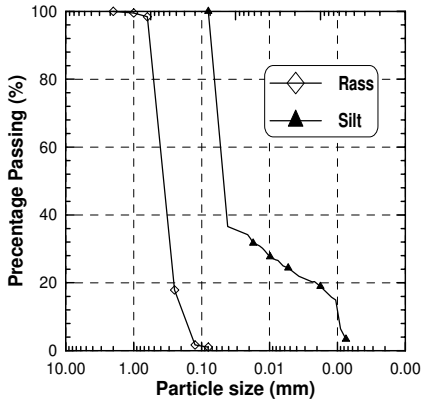


Figure 4
Grain size distribution curve of tested materials

Table 1
Properties of the Rass sand

Properties	Rass sand
γ_s (g/cm^3)	2.664
e_{max}	0.770
e_{min}	0.490
Cu	2.42
D_{10}	0.227
D_{50}	0.39
Shape of particles	Round

Table 2
Properties of the tested silt (SM)

Properties	S/M'hamed Silt
γ_s (g/cm^3)	2.58
e_{max}	1.17
e_{min}	0.76
Liquid limit (ω_L)	25.6
Plastic limit (ω_p)	22.3
Index of plasticity (I_p)	2.33

Table 3
Chemical analysis

Sand-Rass
- Fire loss 6.23
- Total Silica.....78.20
- Alumina (A_l2O_3).....2.03
- Oxide of iron ($Fe2O_3$).....5.58
- lime (CaO).....8.13
- Magnésia (MgO).....Not dosed
- Potash (K2O).....Traces
- Oxide of sodium (Na2O).....Traces
- Sulphates SO_4.....0.24
- Chlorides CL-Solubles in water..…...0.14
- Carbonates $CaCO_3$.....13.94
- Insolubles.....0.93
- Organic materials.....Conform

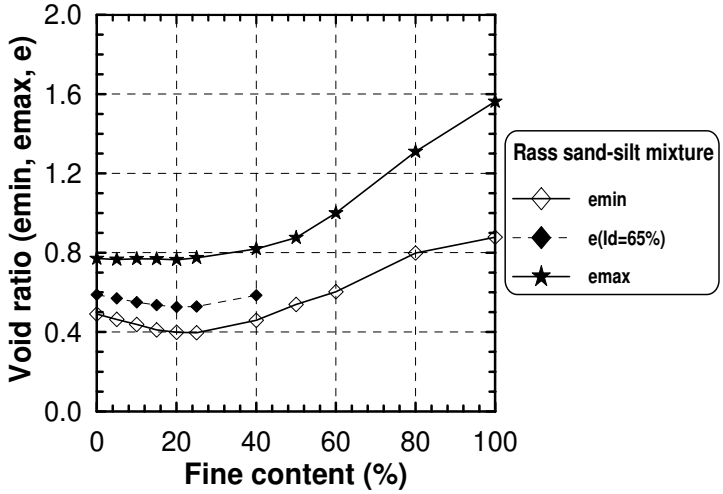


Figure 5
Minimal and maximal void ratios versus fines content

Table 4
Physical characteristics of the sand-silt mixture

Silt Content (%)	0	5	10	15	20	25	40	50	60	80	100
Gs	2.664	2.668	2.673	2.768	2.682	2.687	2.701	2.711	2.720	2.739	2.758
e_{max}	0.770	0.766	0.769	0.767	0.764	0.773	0.710	0.742	0.786	0.947	1.317
e_{min}	0.490	0.463	0.437	0.410	0.398	0.396	0.386	0.393	0.455	0.594	0.776

Fig. 5 and Table 4 show the variation of maximal and minimal void ratios versus the fines content (Fc=0%… 100%). The maximal and minimal void ratios of the sand and mixture were determined according to ASTM standards (D 4253) and Modified Proctor Compaction (D 1557).

These tests were carried out by using a triaxial apparatus of mark GDS (Minidyn2Hz) with samples of diameter and a height of 70 mm; the confinement and the back pressure were applied through GDS devices. In order to obtain a uniform density throughout the specimen, the under-compaction method of specimen preparation was used, as suggested by Ladd [15]. The under-compaction method consists of placing each layer at a density slightly greater than the density of the layer below it in order to account for a decrease in volume and increase in density that occurs in the lower layers when the new layer is placed. Relative density was varied by 1% per layer in general. The experimental equipment included a vacuum pump, a demineralised water tank, a pressure gauge of depression, and a mould used for the preparation of the sample. The sample was saturated by flushing with CO_2 and de-aired water.

The samples were isotropically consolidated to reach the value of the effective confining pressure preceding the cyclic loading.

3 Presentation and Discussion of Results

3.1 Effect of Loading Level

Figs. 6 and 7 show the evolution of the pore water pressure and axial strain versus the time of a series of alternate cyclic tests realized on clean sand samples. The amplitudes of these cycles (q_m) are equal to 30, 50 and 70 kPa, respectively; this gives a cyclic stress ratio CSR = 0,15, 0,25 and 0,35, (CSR = q_m / 2σ'c). Fig. 6 illustrates clearly the increase in the rate of the water pore pressure with the increasing amplitude of cyclic loading. This increase is very important when q_m increases from 30 to 50 kPa.

For the test with CSR = 0.35, the sample develops shear stresses after 8 cycles, corresponding to 2% of double amplitude, and liquefaction was reached during the 9^{th} cycle. For the samples with CSR = 0.25 and 0.15, they develop shear stresses after 9 cycles corresponding to an axial strain of 2% of double amplitude; liquefaction was reached during the 12^{th} cycle and 173 cycles for 2% of double amplitude, while liquefaction was observed after 176 cycles, respectively. It is clear that an increase in the amplitude of cyclic loading (q_m) accelerates the liquefaction process.

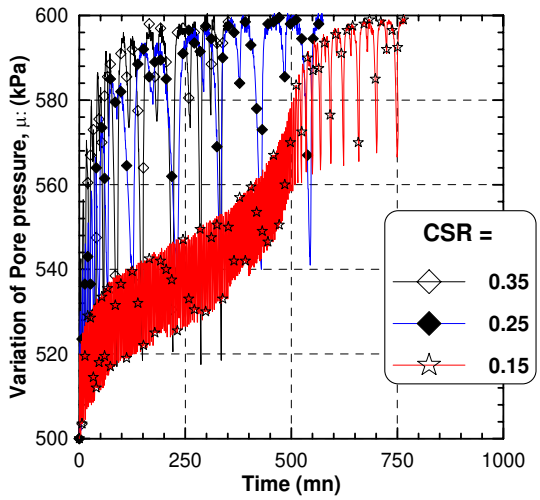


Figure 6
Evolution of the pore water pressure versus time

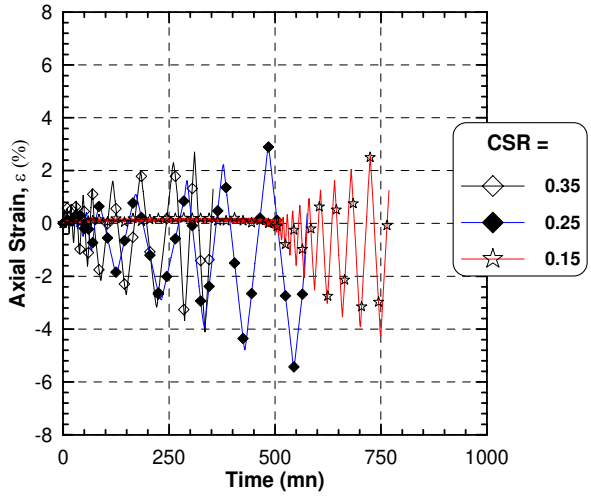
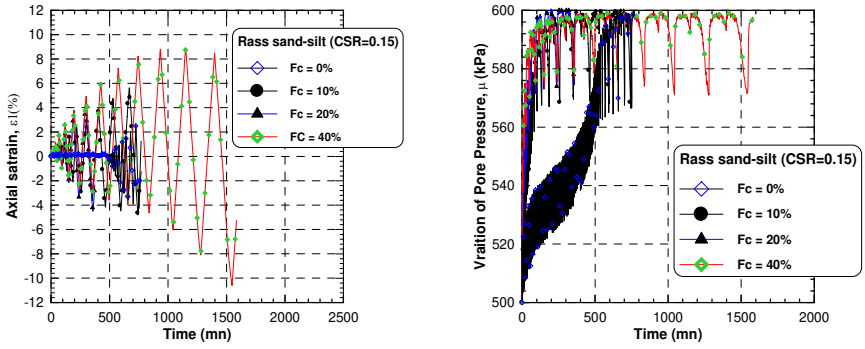


Figure 7
Evolution of the axial strain versus time

3.2 Effect of Fines on the Liquefaction Potential

Figs. 8a and 8b show the evolution of the axial strain and pore water pressure versus time for samples sheared with an amplitude $q_m = 30$ kPa (CSR = 0.15).

In Fig. 8a we notice that the sample with a silt content Fc=10% develops shear stresses after 15 cycles, increasing gradually to reach a double amplitude strain of 5% during the 22nd cycle; liquefaction was reached during the 23rd cycle. The sample with a silt content Fc=20% develops shear stresses after 7 cycles, reaching a double amplitude strain of 4% during the 12th cycle showing instantaneous liquefaction, whereas the sample with a silt content Fc=40% reached an axial strain of 4% during the 7th cycle, and 8% during the 13th cycle and the liquefaction of the sample was reached during the 14th cycle. To recall, the liquefaction of the clean sand (Fc=0%) occurred at 176th cycle; however, with the sample with a silt content Fc=10% the liquefaction was reached after 23 cycles. With the sample with a silt content Fc = 20%, the liquefaction was reached after 12 cycles; whereas with the sample with Fc=40%, it was reached after 14 cycles (Figure 9). In Fig. 8b, we notice that the sample with FC = 20% generates the pore water pressure quickly, whereas the samples with Fc=10% and 40% required much more time to generate the pore water pressure. This is due to the effect of fines causing an increase in the sample contractancy until a certain limiting value beyond it, they cause an increase in the dilatancy of the mixture sand-silt.



a) (b)

Figure 8
Influence of fines on the cyclic behaviour of the sand-silt mixtures

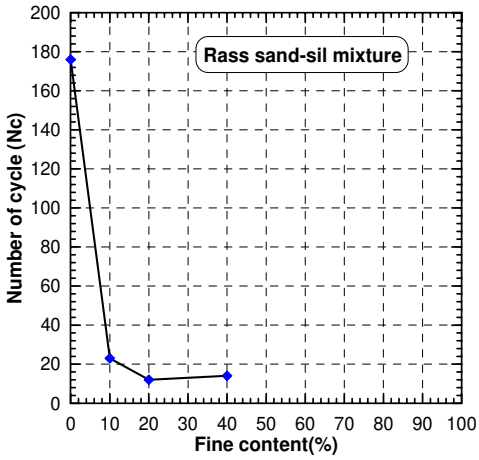
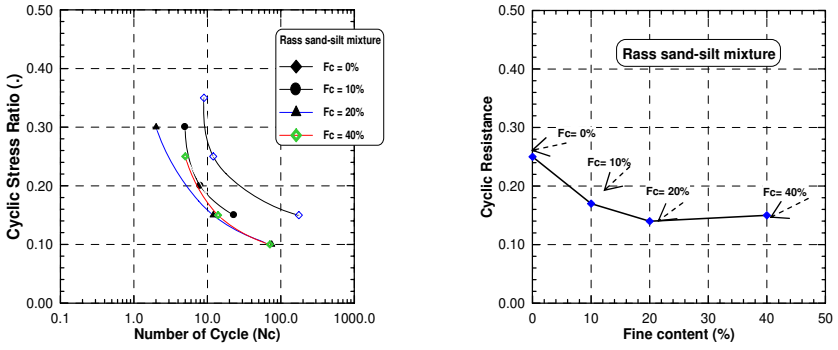


Figure 9
Evolution of the number of cycles versus fines content

Figs. 10a and 10b show the variation of the cyclic stress ratio (CSR=qm/2σ'_c) and cyclic liquefaction resistance (CLR) versus the number of cycles (Nc). Resistance to liquefaction (RLC) is defined by the cyclic stress ratio giving liquefaction for 15 cycles (Ishihara 1993). We notice in Fig. 10a that the resistance of liquefaction of the mixture Rass sand-SM silt decreases with an increase in the amount of fines until the fines content Fc = 20%; then we note a small increase in the liquefaction potential with the fines content Fc = 40%. This increase is due to the active role of fines beyond 20% which take part in the resistance to liquefaction. Fig. 10b shows cyclic liquefaction resistance versus the fines content. We note that the cyclic liquefaction resistance decreases with an increase in the fines content up to 20% having RLC = 0.14, then it re-increases slightly to reach the value of RLC = 0.15 for the fines content Fc = 40%.



(a) (b)

Figure 10
Effect of fines on the liquefaction potential of the Rass sand –SM silt mixtures

3.3 Effect of Preloading on the Resistance of Liquefaction

3.3.1 Effect of the Over-Consolidation

A series of undrained cyclic tests was carried out on Rass sand–SM silt mixture over-consolidated with an OCR = 5 (Figs. 11, 12 and 13). The tests were carried out with a initial relative density Id = 0.65 for three loading amplitudes (q_m = 30, 40 and 60 kPa) in order to determine the influence of the over-consolidation on the liquefaction potential of the mixtures. As can be seen from Figs. 11, 12 and 13, the three loadings levels lead to the liquefaction of the soils. For the highest loading (q_m = 60 kPa), liquefaction was observed after 8 cycles, whereas for the same loading, the normally consolidated soil underwent a liquefaction after 5 cycles.

For the loading with an amplitude q_m = 40 kPa, the over-consolidated soil (OCR=5) was liquefied after 26 cycles whereas the normally consolidated soil (OCR=1) was liquefied after 8 cycles (Fig. 11).

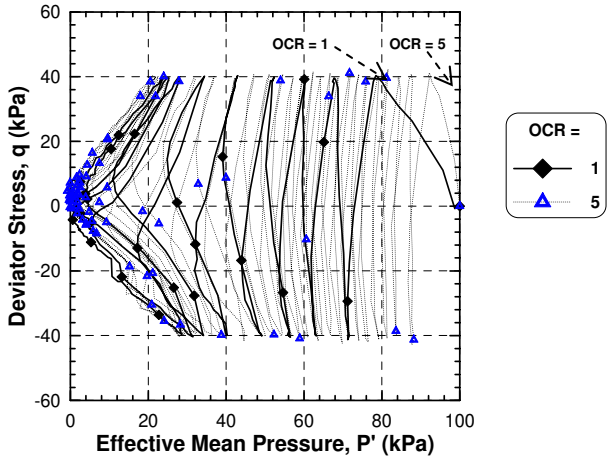
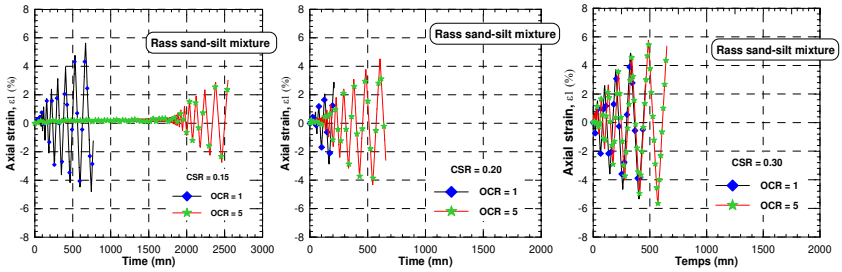


Figure 11

Undrained cyclic test on Rass sand-silt mixture (Fc = 10%, qm = 40 kPa, Dr = 65%, σ'_c= 100 kPa)

The influence of the lowest amplitude is even stronger. Indeed, for a loading q_m = 30 kPa, the normally consolidated soil was liquefied after 25 cycles compared to the 236 cycles necessary for the liquefaction of the over-consolidated soil. Figs. 12 and 13 show clearly that the over-consolidation of the soil attenuates the rate of increase in the axial strain and pore pressure, causing a delay in liquefaction. The normally consolidated sample develops double amplitude axial strain (2.5%) after 21 cycles, for which the full liquefaction was reached after 25 cycles; the over-consolidated sample develops double amplitude axial strain (2.5%) after 234 cycles and reached the full liquefaction after 236 cycles. It clearly shown in Fig. 13 that the over-consolidation delays the generation of pore pressure.



q_m = 30 kPa q_m = 40 kPa q_m = 60 kPa

Figure 12

Influence of the over-consolidation on the liquefaction potential (axial strain versus time)

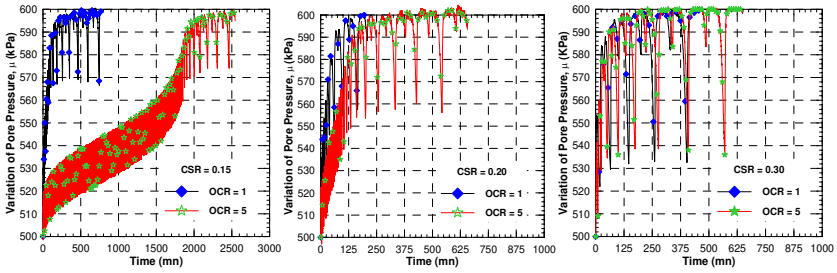


Figure 13

Influence of the over-consolidation on the liquefaction potential (pore pressure versus time)

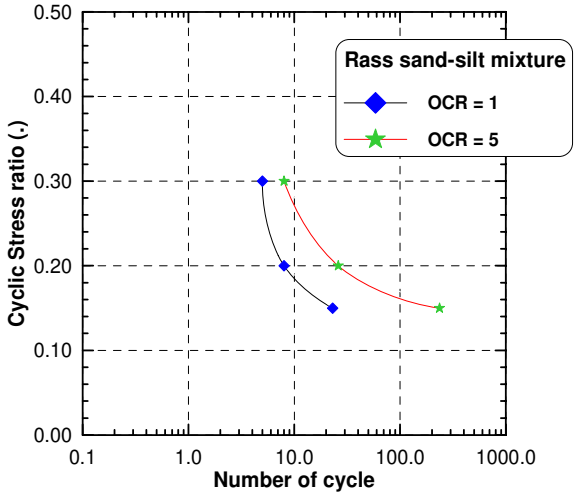


Figure 14

Influence of the over-consolidation on the liquefaction potential of Rass sand-sil mixture (FC = 10%)

Fig. 14 shows the influence of the over-consolidation on the resistance of liquefaction. It confirms clearly the results presented above, namely that the over-consolidation of soil increases its resistance to liquefaction. This is due to the fact that the over-consolidation amplifies the dilating character of the soils, inducing the attenuation in the rate of the water pressure under undrained loading condition.

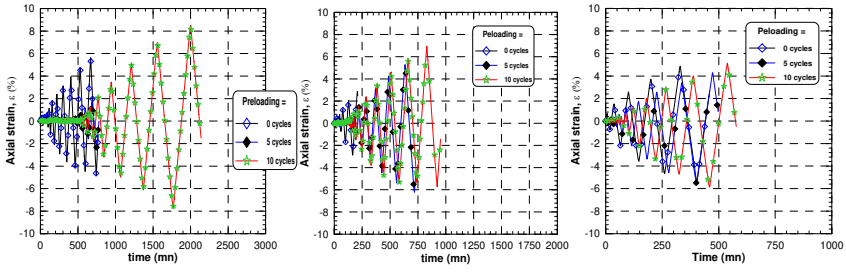
3.3.2 Influence of Cyclic Preloading

A series of tests were carried out on the mixture Rass sand– RS silt (FC = 10%) on samples subjected to a cyclic drained loading in order to study the influence of a cyclic preloading on the potential of liquefaction.

Two test series were carried out. In the first, the samples were initially subjected to 5 cycles, while in the second they underwent 10 cycles of loading (q_m = 30, 40 and 60 kPa). Each series comprises several tests carried out for three amplitude levels (q_m = 30, 40 and 60 kPa).

We note that the whole number of the selected loadings leads to liquefaction. The effect of the preloading is to increase the liquefaction resistance. As an example:

- for the loading q_m = 60 kPa, liquefaction was observed after 11 cycles for the sample subjected to 5 cycles and after 16 cycles for the sample subjected to10 cycles; it should be noted that the soil that did not undergo a cyclic loading was liquefied after 5 cycles.
- for the loading q_m = 40 kPa, liquefaction was observed after 31 cycles for the sample having undergone 5 cycles and after 37 cycles for the sample having undergone 10 cycles; it should be noted that the soil that did not undergo a cyclic loading was liquefied after 8 cycles.
- for the loading q_m = 30 kPa, liquefaction was observed after 80 cycles for the sample subjected to 5 cycles and after 95 cycles for the sample preloaded with 10 cycles; it should be noted that the soil that did not undergo cyclic loading was liquefied after 23 cycles.



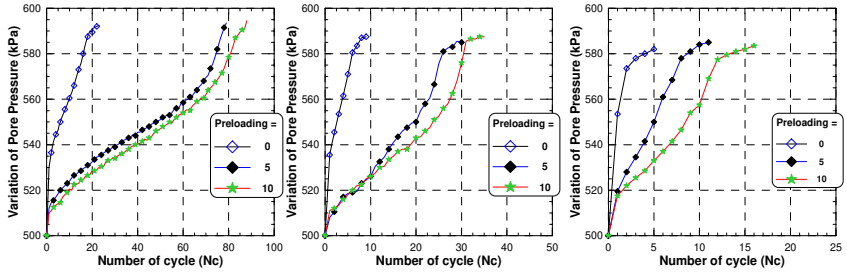
q_m = 30 kPa q_m = 40 kPa q_m = 60 kPa

Figure 15

Influence of the cyclic preloading on the liquefaction potential (axial strain versus time)

Fig. 15 illustrates the influence of the preloading level on the evolution of the axial strain. It clearly shows that an increase in Np (number of preloading) from 5 to 10 delays the development of the soil deformation. The sample preloaded with Np = 5 and 10 reached an axial strain with double amplitude of 4% after 78 cycles, 92 respectively, with a loading q_m= 30 kPa. The same observation was then done for the samples with a loading of q_m= 40 kPa and 60 kPa. Fig. 16 shows the evolution of pore pressure versus cycle number. The three figures show clearly that the cyclic preloading delays in considerable manner the rate of pore pressure. The pore pressure values are determined on the top of deviator.

These results are summarized in Fig. 17, which confirms that the cyclic loading improves the resistance to liquefaction of the soils. We note that the effect of the first 5 cycles is more significant than that of the last 5 cycles. This result can be explained by the fact that the cyclic loading improves the soil density and consequently increases its dilatancy.



q_m = 30 kPa q_m = 40 kPa q_m = 60 kPa

Figure 16
Influence of the cyclic preloading on the potential of liquefaction

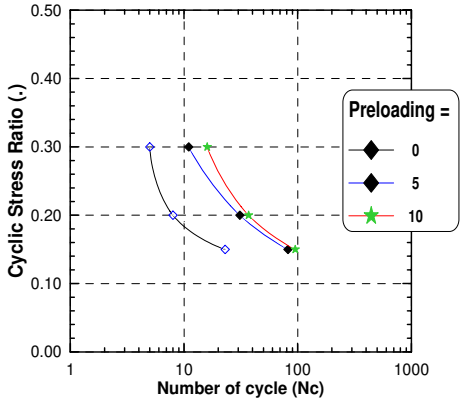


Figure 17
Cyclic preloading effect on the liquefaction potential of the mixture Rass sand - SM silt (FC = 10%, Id = 0,65)

Fig. 18 illustrates the evolution of the cyclic resistance to liquefaction versus the number of preloading. We observe a clear increase in the cyclic liquefaction resistance (previously defined) versus the level of preloading. The cyclic resistance increases in a linear manner with increasing the number on preloading. Drained cyclic preloading improves the resistance of the soil compared to the over-consolidation. The results give a CLR (cyclic liquefaction resistance) equal to 0.3 and 0.25 for the mixtures subjected to a preloading of Np=10 and 5; for the mixture subjected to an over-consolidation ratio of OCR=5, the CRL is equal to 0.23. However, the mixture not that was subjected to any preloading the CLR is 0.15.

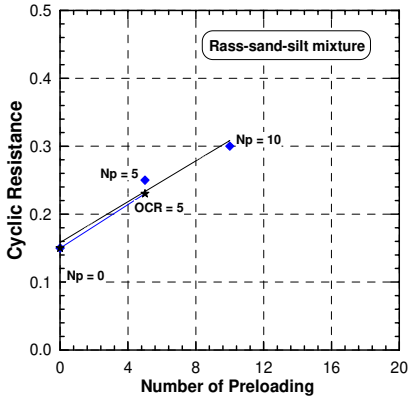


Figure18

Evolution of the liquefaction resistance (RLC) versus the preloading level

Conclusions

In this paper we have presented the results of a laboratory study of the influence of low plastic fines and preloading on the cyclic behaviour of silty sand. This study comprises cyclic undrained triaxial tests which were carried out with an index of density $I_D = 0.65$ for fines content varying between 0 and 40%.

The test results show that an increase rate in water pressure increases with the increase of the cyclic loading amplitude, and it becomes appreciable when the amplitude of loading varies from 30 to 50 kPa. The fines content (Fc) until a certain threshold (Fc=20% in our case) increases the sample contractancy; beyond this limit, it increases the dilatancy of the soil. The liquefaction potential of the mixture of Rass sand and SM silt decreases with an increase of the fines content until 20%, then we note a small increase in the liquefaction potential with the fines content of 40%. This increase is due to the active role of fines beyond 20% and which take part in the liquefaction resistance. The cyclic liquefaction resistance decreases from 0.26 to 0.14 with the increase of the fines content until the value of 20% having, then it re-increases slightly to reach the value of CRL = 0.15 for the fines content of 40%.

The over-consolidation of the soil increases its liquefaction resistance; due to the fact that it amplifies the dilating character of the soil, attenuating increase rate in the water pressure under undrained loading.

The drained cyclic loading improves the resistance to liquefaction of the soils; indeed, the effect of the first 5 cycles is more significant than that of the last 5 cycles. This result can be explained by the fact that the cyclic loading densifies the soil state and consequently increases its dilatancy. Drained cyclic preloading improves the resistance of the soil comparing to the over-consolidation.

References

[1] Amini F. & Qi G. Z. (2000) Liquefaction Testing of Stratified Silty Sands. Journal of Geotechnical Engineering Division, Proc. ASCE, Vol. 126 (3), pp. 208-217

[2] Arab, I. Shahrour, S. Hamoudi, L. Lancelot (2008) Influence of Fines Fraction on the Behaviour of a Silty Sand. Revue Française de Géotechnique, N° 122, 1er trimestre 2008, 37-43

[3] Bouferra Rachid (2000) Etude en laboratoire de la liquéfaction des sols. Thèse de doctorat, Ecole Universitaire des Ingénieurs de Lille USTLille, 2000, p. 110

[4] Bouferra R., & Shahrour I., (2004) Influence of Fines on the Resistance to Liquefaction of a Clayey Sand. Ground Improvement 8, No. 1, 1-5

[5] Chang N. Y., Yeh S. T. & Kaufman L. P. (1982) Liquefaction Potential of Clean and Silty Sands. Proc., 3rd Int. Earthquake Microzonation Conf., Vol. 2, 1017-1032

[6] Dezfulian H. (1982) Effects of silt content on dynamic properties of sandy soils. Proc., 8th World Con. on earthquake Engrg., 63-70

[7] Finn W. D., Bransby PL., Pickering DJ. (1970) Effect of Strain History on Liquefaction of Sands. Journal of soils Mech. Foundation Div., ASCE; 96(SM6), 1917-34

[8] Finn W. D., Ledbetter R. H. & Wu G. (1994) Liquefaction on Silty Soils: Design and Anlysis. Ground Failures under Seismic Conditions, Geotech. Spec. Publication, N°44, ASCE, New York, 51-76

[9] Ishihara K. & Takatsu H (1979) Effects of Oversurconsolidation and K_0 Conditions the Liquefaction Characteristics of Sands. Soils and Foundations, Tokyo, Japon, 59-68

[10] Ishihara K. & Okada S. (1978) Effects of Stress History on Cyclic Behaviour of Sands. Soils Found, 18(4), 31-45

[11] Ishihara K. & Okada Y. (1982) Effects of Large Preshearing on Cyclic Behaviour of Sand. Soils Mechanics and Foundations Engineering, Vol. 22, No. 3, 109-123

[12] Ishihara K. (1993) Liquefaction and Flow Failure during Earthquakes. The 33rd Rankine lecture, Geotechnique, 43(3), 351-415

[13] Koester J. P. (1994) The Influence of Fines Type and Content on Cyclic Strength. Geotechnical Special Publication N°44, S. Prakash and P. Dakoulas, eds., ASCE, New York, 17-33

[14] Kuerbis R., Negussey D ; & Vaid V. P. (1988) Effect on Gradation and Fines Content on the Undrained Response of Sand. Proceedings Hydraulic Fill Structures, Fort Collins, USA, 330-345

[15] Ladd R. S. (1978) Specimen Preparation Using Undercompaction. Geotechnical testing Journal, 1(1), 16-23

[16] Lade P. V. & Yamamuro J. A. (1997) Effects of Nonplastic Fines on Static Liquefaction Sands. Canad. Geotech. Journal, Ottawa 34, 918-928

[17] Law K. T. & Ling Y. H. (1992). Liquefaction of Granular Soils with Noncohesive Fines. Proc., 10th World Conf. on Earthquake Engrg., 1491-1496

[18] Luong M. P. (1980) Phénomène cyclique dans les sols pulvérulents. Revue Française de géotechnique, N°10, 39-53

[19] Polito Carmine Paul (1999) The Effects of Non-Plastiques and Plastiques Fines on the Liquefaction of Sandy Soils. Ph.D. dissertation, Faculty of Virginia Polytechnic Institute and State University, U.S.A.

[20] Seed H. B., Idriss I. M. & Lee K. L. (1975) Dynamics Analysis of the Slide in the Lower San Fermondo Dam During the Earthquake of February 1971. Journal Geotechnical Engineering, division ASCE, Vol. 101, GT 9, 889-911

[21] Seed H. B & Peacock W. H. (1971) Test Procedures for Measuring Soil Liquefaction Characteristics. Journal of the Soils Mechanics and Foundation Division. ASCE, Vol. 97 (8), 1099-1119

[22] Shen C. K., Vrymoed J. L. & Uyeno CK. (1977) The Effects of Fines on Liquefaction of Sands. Proc. 9th Int. Conf. on Soil Mech. and Found. Engineering, Vol. 2, 381-385

[23] Singh S. (1996) Liquefaction Characteristics of Silts" Geotechnical and Geological Engineering, 14, 1-19

[24] Skempton Skempton, A. W. (1954) The Pore Pressure Coefficients A and B, Geotechnique, Vol. IV, pp.143-147

[25] Troncoso J. H. & Verdugo R. (1985) Silt Content and Dynamic Behaviour of Tailing Sands. Proc., 12th Int. Conf. on Soil Mech. and Found. Engrg., 1311-1314

[26] Vaid V. P. (1994) Liquefaction of Silty Soils. Ground Failures under Seismic Conditions, Geotechnical Special Publication, N°44, ASCE, New York, 1-16

[27] Wichtmann T., Niemunis A., Triantafyllidis Th. & Poblete M. (2005) Correlation of Cyclic Preloading with the Liquefaction Resistance. Soil Dynamics and Earthquake Engineering, Vol. 25, 923-932

[28] Yamamuro J. A. & Lade P. V. (1997) Static Liquefaction of Very Loose Sands. Canad. Geotech. Journal, Ottawa 34, 905-917

[29] Zlatovic S. & Ishihara K. (1997) Normalised Behaviour of Very Loose Nonplastic Soil/ Effects of Fabric. Soils and Foundations, Tokyo, 37(4), 47-56

Comparative Analysis of Parallel Gene Transfer Operators in the Bacterial Evolutionary Algorithm

Miklós F. Hatwágner

Department of Information Technology
Jedlik Ányos Faculty of Engineering
Széchenyi István University
Egyetem tér 1, H-9026 Győr, Hungary
e-mail: miklos.hatwagner@sze.hu

András Horváth

Department of Physics and Chemistry
Jedlik Ányos Faculty of Engineering
Széchenyi István University
Egyetem tér 1, H-9026 Győr, Hungary
e-mail: horvatha@sze.hu

Abstract: The Bacterial Evolutionary Algorithm (BEA) is an evolutionary method, originally meant to optimize the parameters of fuzzy systems. The authors have already proposed three modified versions of the original algorithm in a previous paper to make it usable in engineering applications with time-consuming object functions as well. Section 1 summarizes the earlier results. It presents the operators of the original BEA and the suggested parallel version. In Section 2, the optimal parameter settings and the analytical estimation of wall clock time in parallel computations are investigated. In Section 3, the paper deals with genetic diversity in different BEA versions. The effect of the modified gene transfer operators on genetic diversity is measured. The conclusion is that the proposed methods have quite good efficiency in all cases, and we can reach the ideal case if we have full control over the parameters.

Keywords: optimization; Bacterial Evolutionary Algorithm; genetic diversity; parallel computing; parallel efficiency

1 Introduction

The Bacterial Evolutionary Algorithm (BEA) [13] [12] is a relatively new member of the populous family of evolutionary algorithms [2]. It is a descendant of the Genetic Algorithm (GA) [6] and the Pseudo-Bacterial Genetic Algorithm (PBGA) [14]. The BEA was proposed by Norberto Eiji Nawa and Takeshi Furuhashi in the late '90s.

The BEA inherited many properties of the GA: it is also a global search algorithm, which is useful if a near optimal, approximate solution of a problem is acceptable. The algorithm is able to solve complex optimization problems even if they have non-linear, non-continuous, multimodal, high-dimensional properties. In contrast to gradient based methods, the BEA does not demand the use or the existence of the derivatives of the objective functions. Furthermore, the operators of the BEA achieve some functions, e.g. elitism, that can be implemented in the canonical GA only with additional code. This nature of the BEA helps to keep the program more compact and reliable as well.

The BEA and the GA are of course heuristic type optimization methods, thus there is no guarantee of finding the location of the global extreme value. Despite this, these algorithms perform well in real-life optimization problems, and theoretically the probability of finding the global optima can be made arbitrarily large (see [15]).

The BEA was originally developed to optimize fuzzy systems' parameters, but it could be a proper tool to solve complex design and engineering optimization problems related to computational fluid dynamics (CFD) or finite element models (FEM). However, such models need huge computational power, because every object function evaluation in the optimization process contains a full CFD or FEM calculation, which can take 0.1 to 5 hours of CPU-time. In a typical industrial application, the number of design variables is 10 to 30 and the whole optimization needs thousands of object function evaluations [10]. Thus the necessary CPU-time is in the order of 1 week to some months, and therefore parallelization is necessary. Sometimes the problem itself can be parallelized, but it is more adequate if the optimization process is executed in a parallel way. Unfortunately, the BEA is inherently sequential so this method in its original version is practically inapplicable in this area.

1.1 Shortcomings of the Bacterial Evolutionary Algorithm

[13] contains an exhaustive review of the BEA, and thus we will give only a short introduction here.

Similarly to the GA, the BEA also uses a record of possible solutions. These solutions are often called bacteria as well. The bacteria together form the population.

There are two main operators of the BEA: bacterial mutation and gene transfer. The repeated utilization of these operators results in a series of generations. When some kind of termination condition is fulfilled, the best bacterium of the last population is accepted as the result of the optimization.

Bacterial mutation (Fig. 1) optimizes the bacteria individually. That is why all the bacteria can be mutated at the same time. The mutation functions in the following way. Every bacterium has K clones. Initially the clones are copies of the original bacterium. In each step of the mutation, exactly one gene at a specified position is modified randomly in every clone. If a better gene value (allele) has been found, it is copied into the other clones. At the end of mutation, if the objective value of the best clone is better than the value of the original bacterium, the bacterium is replaced with this clone.

Consequently, the objective function has to be evaluated K times in one step, and such a step is repeated g (the number of genes) times during the operation. As was already shown in [9], the mutation operator evaluates the objective function $E_m = PKg$ times (P is the population size). Because several genes cannot be evaluated in parallel, the theoretical maximum speedup of the evaluation of the bacteria is $S_m = E_m/g = PK$, and it can be achieved with $C = PK$ processors. (It is assumed that the evaluation time of all the bacteria is the same and it is independent of the alleles.)

In a typical calculation, $P \approx 30\text{-}100$, $K \approx 20\text{-}50$, thus $C_{max} \approx 600\text{-}5000$. In most cases this number is much bigger than the number of processors in today's systems, and thus bacterial mutation is suitable to run on most of the multiprocessor systems without modifications.

The second operator of the BEA is the so-called gene transfer (Fig. 2). It operates with the ordered list of bacteria. The bacteria with better objective values get into the superior half, the others into the inferior half. The operator repeats T times the following: it chooses one bacterium from the superior half and one from the inferior half. After that it selects one portion of the genes of the superior bacterium and copies it into the inferior bacterium. This modification of the inferior bacterium involves the re-evaluation of its objective function, and the re-sorting of the bacteria. Depending on the objective value of the modified bacterium it may get into the superior half.

Since the modified bacterium can belong to any of the two halves, it is obvious that the consecutive gene transfers are not independent. Because of this behaviour parallel gene transfers cannot be realized, and therefore modification of the gene transfer operator is needed for parallelization.

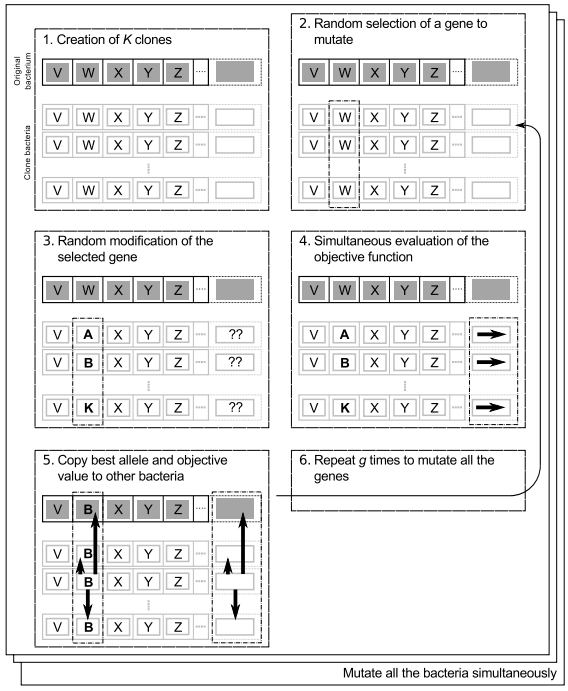


Figure 1
Schematic view of the bacterial mutation operator

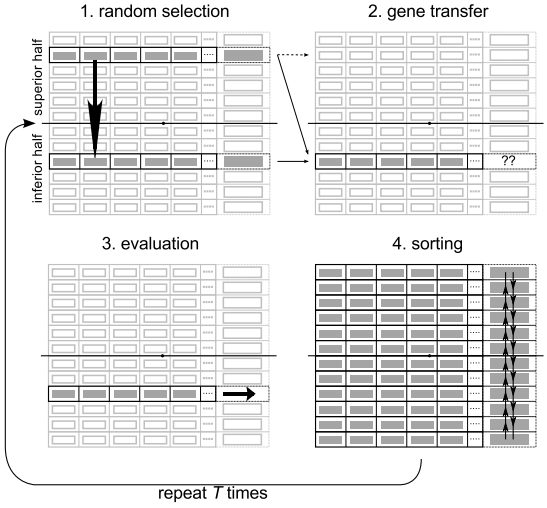


Figure 2
Schematic view of the gene transfer operator

1.2 Overview of the Suggested Modified Gene Transfer Operators

Three modified gene transfer operators were suggested in [9]. All of them are the derivatives of the original version with slight modifications.

"Original gene transfer with auxiliary population" (BEA Aux., Fig. 3) also keeps a record of superior and inferior bacteria based on the objective values. Even the selection of the superior and inferior bacterium is the same as before. This version of gene transfer keeps the inferior bacterium untouched; the modified bacterium goes into an "auxiliary population". The operator first fills the whole auxiliary population with A modified bacteria, and only then starts to evaluate the objective values of them simultaneously. After the evaluation the best P of $P+A$ bacteria form the population, and the other, worse bacteria are dropped. This procedure has to be repeated until the desired number of total gene transfers (T) is reached.

The second suggested gene transfer was called "gene transfer inspired by Microbial Genetic Algorithm" (pMGA, Fig. 4). The Microbial Genetic Algorithm (MGA) [8] is a simplistic GA. The MGA gave the idea of a new gene transfer because its selection and crossover can be regarded as a gene transfer. pMGA creates random and disjoint pairs of bacteria. The better bacterium (the so called "winner") of such a pair transfers some portion of its genetic material to the worse bacterium (loser). Because the pairs are disjoint, the gene transfer and the evaluation of the modified bacteria are independent from other pairs. This property allows parallel execution. Unfortunately, the size of the population limits the number of parallel gene transfers to $P/2$. If more gene transfers are required, the operation has to be repeated.

The MGA inspired other researchers as well to modify and use some of its simple genetic operators in the Bacterial Memetic Algorithm [7] [11] [14]. The MGA inspired gene transfer operation performs well on several important problems, e.g. the travelling salesman problem [4]. It was pointed out that it is easy to implement the parallel version of this gene transfer operator.

The third suggested version of gene transfer, "gene transfer inspired by MGA with auxiliary population," (see [9]) is a mixture of the previous two, in order to eliminate the $P/2$ barrier of the pMGA. The pMGA Aux. uses an auxiliary population (like the BEA Aux.) and places the modified bacteria into it, instead of the instant overwriting of the loser bacteria.

Note that the usage of the modified gene transfers suggested above influences the optimization process. For example, in the case of the original gene transfer, the inferior bacterium is always overwritten, no matter how good or bad it is; but with an auxiliary population it can survive if the auxiliary population contains mostly worse individuals. This is somehow similar to elitism, and thus it increases the average fitness of the next population but keeps the genetic diversity lower.

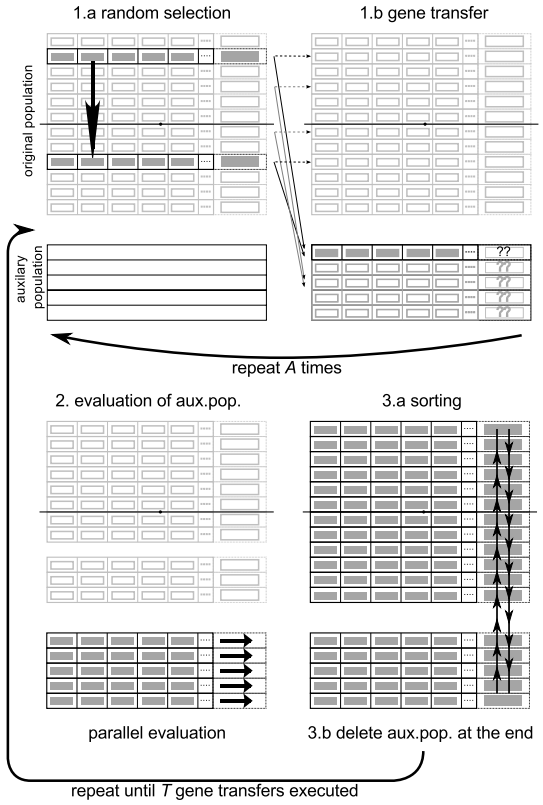


Figure 3
Schematic view of the original gene transfer with auxiliary population (BEA Aux.)

Another side effect of the modified operators can also be realized. The original gene transfer evaluates the objective function T times. This means that the modified allele has at most $T-1$ chances to infiltrate into other bacteria during the same gene transfer. But in the case of using an auxiliary population, the number of chances to infiltrate drops to at most $(T/A)-1$. The situation is similar in the case of the pMGA: a better allele can be inherited at most $(T/(P/2))-1$ times.

At this point, some important questions arise, e.g.: Do modified gene transfers sacrifice genetic diversity on the altar of parallel execution? What are the scaling properties of the different gene transfer versions? These questions can be answered easily after exhaustive empirical tests.

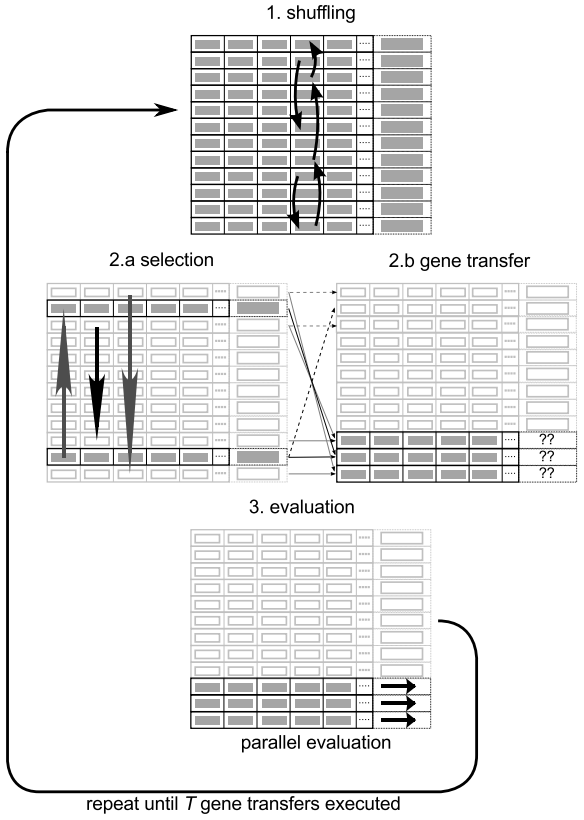


Figure 4
Schematic view of the gene transfer inspired by MGA (pMGA)

1.3 Preceding Results

The first investigation of parallel versions of the BEA and the MGA can be found in [9]. Here a custom optimization program with a master-slave model was used. The slave computers calculated the value of the objective function, while all other tasks fell upon the master computer, including the execution of the genetic operators.

Five test functions (De Jong's 1st and 3rd, Step, Rastrigin, Keane [3] [11] [16]) were used. These functions are well known in literature, and their qualitative properties and global extremes are also known. They are of different types, e.g. Step is non-continuous, Rastrigin is smooth but has many local minima, etc., and therefore they represent different kind of problem types from real life. The big difference between these functions and practical problems is the execution time: in an engineering application, where parallel execution is important, an objective

function evaluation takes many seconds, and therefore communication time is negligible; but in a modern hardware test function evaluation takes only a small fraction of a second. In order to simulate real life problems, a small artificial delay (approx. 0.005 sec.) was built into the test functions. The overhead of communication became negligible with this trick.

Based on the test calculations it can be concluded that the three modified gene transfer operators are applicable in real life problems. On the contrary, the acceleration of the optimization using the original gene transfer is the consequence of bacterial mutation only. It is not recommended to use the original version if more than one CPU is available, but in some cases the modified operators proved to be faster even with one CPU.

All the modified versions have good scaling properties. In [9] the authors used at most 16 processors. In this range the pMGA was slightly the fastest one, but the difference was very small. In this paper the range of investigation is extended to a higher number of CPUs. The ideal setting of number of gene transfers and the genetic diversity in different methods is also examined.

2 Analysis of the Suggested Operators

2.1 Right Settings of the Optimization Program

2.1.1 Maximizing CPU Utilization in Gene Transfer

In [9] the authors drew their conclusions using the results of their custom optimization program. The settings of this program were carefully chosen before the start of the executions. These settings were optimal in the sense that the program produced the same result with the least evaluation of the objective function. However, in a system containing a lot of CPUs the number of object function evaluations may be not proportional to wall clock time if we have a lot of idle CPUs due to bad parameters. In this section a simple model for CPU utilization is presented.

Let's assume that the evaluation time of an object function:

- is constant and one unit long,
- takes much more CPU time than bacterial (genetic) operators and master-slave communication.

Using these assumptions we can divide the calculation process into "computational rounds": if we have C CPUs, the master can send at most C object

function evaluations to them simultaneously, then it collects the results and sends another object functions again. CPU utilization is ideal if in every computational round exactly C object function evaluation is to be sent to the slaves.

Let N denote the maximum number of new bacteria that can be evaluated simultaneously during gene transfer. Without an auxiliary population, it is the half of the population size: $N_{woa} = \lfloor P/2 \rfloor$; with an auxiliary population, it is the size of the auxiliary population: $N_{wa} = A$. ($\lfloor . \rfloor$ is the rounding down function often referred to as "floor".)

If N is not a multiple of C, there will be idle CPUs during the evaluation of these individuals. This means that $\lceil N/C \rceil$ computational rounds are needed to evaluate N individuals. ($\lceil . \rceil$ is the rounding up function often referred to as "ceil".)

Another case is when idle CPUs appear if the total number of transfers (T) is not the multiple of N. In this case the $\lceil N/C \rceil$ computational rounds mentioned above must be executed $\lfloor T/N \rfloor$ times, but $T - \lfloor T/N \rfloor N$ evaluations remain, which implies $\lceil (T - \lfloor T/N \rfloor N)/C \rceil$ extra computational rounds.

Thus the utilisation of the slave CPUs during gene transfer can be specified with the following formula:

$$U_{C,T} = \frac{T}{\left\lfloor \frac{T}{N} \right\rfloor \left\lceil \frac{N}{C} \right\rceil C + \left\lceil \frac{T - \lfloor T/N \rfloor N}{C} \right\rceil C} \tag{1}$$

For example, in [9] in the case of the optimization of the Rastrigin function the authors used $P = 40$, $T = 400$, $A = 20$, $C = 16$. In this case the CPU utilisation is only $U_{C,T} = 62.5\%$, whether an auxiliary population was used or not.

It is easy to achieve 100% utilisation, if we know the number of processors in advance: let N be a multiple of C and T a multiple of N, thus we get $U_{C,T} = 1.0$.

2.1.2 The Optimal Number of Gene Transfers

Changing the gene transfer algorithm may change the optimal number of using this operator. A series of test calculations was performed to study this effect. For the sake of compactness, only the results of the optimization of the Rastrigin function are reviewed in this section. The main parameters of the test were: $K = 1$, $C = 64$, $P = 128$, $A = 64$. ($U_{C,T} = 100\%$) Wall clock time of reaching 0.01 object function value was measured and averaged over 20 independent calculations. (1 to 4% relative standard deviation was observed.)

For the sake of compactness, only the results of the optimization of the Rastrigin function are reviewed in Table 1. The phenomenon is the same in the case of other test functions.

Table 1
The effect of various *T/P* ratios on optimization time

T/P	BEA	BEA Aux.	pMGA	pMGA Aux.
4	*130.982*	13.588	15.699	13.255
6	146.793	11.736	13.619	11.786
8	169.898	*10.723*	12.634	11.371
12	190.434	10.688	12.117	*10.829*
16	242.417	11.177	*11.837*	11.156
24	322.155	11.970	12.899	12.956
32	429.840	13.300	14.649	13.696

Table 1 shows that for original BEA a small *T/P* ratio is optimal. The reason is simple: gene transfers in the BEA scale poorly for 64 CPUs, and for small number of transfers, the mutation operator dominates. (However, the wall clock time value is much higher than the one in parallel versions.)

For parallel versions the *T/P* ratio has an optimal range. One can conclude that the best *T/P* ratio for the parallel gene transfer operators is between 8 and 16 and there is only a small difference within this range. The tests showed similar results for other test problems and CPU numbers.

2.2 Test Methodology

Considering the observations mentioned above, one can choose good parameters for *T*, *A*, based on the *C* number of CPUs. There is however another problem: to measure the efficiency of different methods a lot of independent calculations must be performed for all the test problems. It takes a lot of time if we use real life object functions with many seconds of CPU-time consumption.

One possible solution is to use easily formulated test functions with very small calculation time, but apply a small amount of artificial delay, which makes the object function evaluation longer than the communication time. This method was used successfully in [9] for at most 16 CPUs, but it is not a good method for a much higher number of processors. Test calculations with 64 CPUs and 0.05s artificial delay showed more than 100% extra time originating from the communication bottlenecks. (Note that the communication between master and slaves consists of smaller than 1kB data blocks, but at the end of a computational round, when all the slaves want to send the data back to the master, a significant bottleneck arises.)

Increasing the artificial delay may help and it could bring the test calculations closer to real life, but results in very long test calculation time. To overcome these difficulties, another approach is used in this paper:

- No artificial delay is used.

- Load balancing is realised in calculations.
- All object function evaluations are logged with their sequential number, and value.

Load balancing is a key part of the measurements. Instead of sending the next job to the first idle slave, the master sends the jobs to the slaves in a predefined order to ensure as similar CPU loads as possible. Otherwise if there are several slaves (e.g. 64 or more) in the system and the evaluation time of the objective function is short compared to the communication time between master and slave, the first slaves would be fully loaded while the rest of the slaves remain idle. In this way the artificial delay included in the objective functions is not needed anymore.

This kind of test calculation gives enough information to reconstruct how many computational rounds would be needed if the optimization was executed in a load-balanced way on a *C*-processor machine. Assuming nearly identical time for object function evaluations, the number of computational rounds is proportional to wall-clock time. In the next subsection deduction of the number of computational rounds is presented.

2.3 The Number of Computation Rounds

The flow of a bacterial-type optimization begins with a random population generation and evaluation of the individuals. This means that P object function calls happen in the 0^{th} generation. After this initialization new generations are produced by $E_m = PKg$ mutations and $E_T = T$ gene transfers. (See Sec 1.1 for notations.) Thus the E_G number of objective function evaluation needed to create the next generation can be expressed as:

$$E_G = E_m + E_T = PKg + T \tag{2}$$

In a test calculation we measure how many objective function evaluations are required to reach a specific target objective value. (Naturally, an average number of independent calculations is used.) Let us denote this number of evaluations with M. If this number is known, one can calculate how many generations and how many computational rounds are needed for the optimization, and thus get a good approximation of wall-clock time.

The number of fully completed generations (except the 0^{th} generation, which needs P evaluations) can be expressed as:

$$G_f = \left\lfloor \frac{M - P}{E_G} \right\rfloor \tag{3}$$

For the last (possibly non-finished) generation E_l objective function evaluations remain.

$$E_l = (M - P) - G_f E_G \tag{4}$$

In the last generation the number of mutations and gene transfers may be less than E_m and E_T. If we perform the mutation first, the number of objective function evaluations used by mutation in the last generation will be:

$$E_{l,m} = \min(E_m, E_l) \tag{5}$$

and the number of objective function evaluations executed by gene transfers in the last generation is:

$$E_{l,T} = \max\left(0, E_l - E_{l,m}\right) \tag{6}$$

Now it is easy to express the needed number of computational rounds.

All the methods need R_0 computational rounds to evaluate the 0th generation:

$$R_0 = \lceil P / C \rceil \tag{7}$$

All the methods examined here use PK independent mutations, therefore

$$E_{m,p} = \min(C, PK) \tag{8}$$

evaluations can be made in parallel (in one computational round).

The original BEA needs $R_{f,BEA}$ computational rounds for every fully evaluated generations (remember that gene transfers are sequential operations in the BEA.):

$$R_{f,BEA} = \lceil PK / E_{m,p} \rceil g + E_T \tag{9}$$

Thus the number of computational rounds required by optimization using the BEA can be expressed as:

$$R_{BEA} = R_0 + G_f R_{f,BEA} + \lceil E_{l,m} / E_{m,p} \rceil + E_{l,T} \tag{10}$$

Using similar considerations the number of computational rounds of parallel versions can be expressed also.

$$\begin{aligned} &R_{parallel} = R_0 + \\ &G_f \left(\left\lceil \frac{PK}{E_{m,p}} \right\rceil g + \left\lfloor \frac{T}{N} \right\rfloor \left\lceil \frac{N}{C} \right\rceil + \left\lceil \frac{T - \lfloor T/N \rfloor N}{C} \right\rceil \right) + \\ &\left(\left\lceil \frac{E_{l,m}}{E_{m,p}} \right\rceil + \left\lfloor \frac{E_{l,T}}{N} \right\rfloor \left\lceil \frac{N}{C} \right\rceil + \left\lceil \frac{E_{l,T} - \lfloor E_{l,T}/N \rfloor N}{C} \right\rceil \right) \end{aligned} \tag{11}$$

$R_{pMGA\ Aux}$ is the number of computational rounds needed by optimization using the pMGA Aux. The value of it is the same as the number of computational rounds used up by the BEA Aux, $R_{BEA\ Aux}$.

$$R_{pMGAAux} = R_{BEAAux} = R_0 + G_f\left(\left\lceil \frac{PK}{C} \right\rceil g + \left\lfloor \frac{T}{A} \right\rfloor \left\lceil \frac{A}{C} \right\rceil + \left\lceil \frac{T - \left\lfloor \frac{T}{A} \right\rfloor A}{C} \right\rceil\right) + \left(\left\lceil \frac{E_{l,m}}{R_m} \right\rceil + \left\lfloor \frac{E_{l,T}}{A} \right\rfloor \left\lceil \frac{A}{C} \right\rceil + \left\lceil \frac{E_{l,T} - \left\lfloor \frac{E_{l,T}}{A} \right\rfloor A}{C} \right\rceil\right) \quad (12)$$

Here we used the same notation as in Sec 2.1.1, namely N is the maximum number of new bacteria that can be evaluated simultaneously during gene transfer. For the BEA Aux. and the pMGA Aux. methods $N = N_{wa} = A$, for pMGA $N = N_{woa} = \lfloor P/2 \rfloor$.

2.4 Test Results

Some test optimizations with different settings were performed to check the correctness of the above formulas, but there were no differences between the calculated and the measured number of computational rounds.

Six standard problems were used for testing the behaviour of modified gene transfers: Rastrigin, Keane, Step, Ackley, DeJong's 1st and DeJong's 3rd functions. (See [3], [11], [16].) Some of these are unimodal (eg. De Jong's 1st), others are multimodal (eg. De Jong's 3rd). Some of them are continuous (eg. Rastrigin) while others are not (eg. Step). This means the results are valid for a wide range of problems.

Even though the test functions have very different properties, and thus they represents a wide range of problems, the authors plan to execute more sophisticated tests with a much wider and more easily parameterizable set of test functions in the future. These test problem sets could be generated with the appropriate functions, see e.g. [1] [5].

Table 2 shows the measured values of M for the Rastrigin function with the target objective value of 0.01, and the calculated number of computational rounds as well. The main settings of the optimization program were the following: g = 20, P = 128, A = 64. According to the considerations, the numerical experiment must be performed only for one specific C value, and the others can be calculated from this. (Because the authors have been able to use a 64-core machine, C = 64 was used in the calculations.)

Table 2
Computational rounds needed with different number of CPUs

C	1 (=M)	2	8	64	256
R_{BEA}	128048	74520	34374	22665	21829
R_{BEAAux}	122732	61366	15342	1918	1117
R_{pMGA}	141582	70791	17698	2213	1291
$R_{pMGAAux}$	124608	62304	15576	1947	1134

The parallel efficiency of the calculations is also presented in Table 3. ($E_p=(R_1/R_X)/C$, where R_x is the number of computational rounds in x CPU-case.)

Table 3
Parallel efficiency of optimization with different number of CPUs (Rastrigin fn.)

C	1	2	8	64	256
$E_{p,BEA}$	1.000	0.859	0.466	0.088	0.023
$E_{p,BEAAux}$	1.000	1.000	1.000	1.000	0.429
$E_{p,pMGA}$	1.000	1.000	1.000	1.000	0.428
$E_{p,pMGAAux}$	1.000	1.000	1.000	1.000	0.429

Tables 2 and 3 show that the parallel versions scale ideally until full utilization is achieved. In the last column $N > A$, and therefore this is not true, and all the methods will slow down, but the original BEA shows bad performance for a much smaller number of CPUs. These results are in good correspondence with the ones in [9].

The R values for the $C = 1$ case shows the difference of the methods in 1 CPU case. It is not obvious that parallel versions are comparable with the original BEA in this case. Table 2 shows that the pMGA is worse than the BEA, but all the methods with auxiliary populations are better than the original bacterial algorithm even on 1 core. Due to the good scaling properties, even the pMGA beats the BEA in all $C > 1$ cases.

Testing with other functions show completely similar structure, and therefore only a small, significant part of the results are presented here.

Table 4 shows the ratio of the computational rounds needed by the modified gene transfers and the original version. Using only one slave CPU, the pMGA usually needs slightly more computational rounds (ie. wall-clock time) than the original gene transfer. In every other case, except for the Keane-function, all of the modified gene transfers perform better even in the $C = 1$ case, but the difference is only 1-2% in this case. The two versions using auxiliary populations are the best. These gene transfer methods provide practically the same performance.

Table 4
Ratios of computational rounds in case of 1 CPU (C=1)

	Rastrigin	Keane	Step	Ackley	DeJong's 1st	DeJong's 3rd
$\frac{R_{BEAAux.}}{R_{BEA}}$	0.958	1.018	0.832	0.935	0.947	0.949
$\frac{R_{pMGA}}{R_{BEA}}$	1.106	1.169	1.045	1.126	0.997	1.090
$\frac{R_{pMGAAux.}}{R_{BEA}}$	0.973	1.023	0.791	0.966	0.928	0.876

Table 5 shows the same ratios as Table 4 for the C = 64 case. Because of the good scaling properties of BEA Aux., pMGA and pMGA Aux. methods, all the values are lower than 1, which means that they are significantly better than the original BEA. The two methods with auxiliary populations are approximately efficient in the same degree, and the pMGA is slightly worse than these two methods.

It is clear that for a known C value, one can choose the other parameters for ideal scaling. But in practice sometimes the number of CPUs is not a fixed, predefined number. For example, some of the CPUs in the cluster are allocated for other jobs. It is important to examine the scaling properties of these methods for a "random" number of CPUs also. The (12) and (13) formulas can be used for such calculations.

Table 6 shows the parallel efficiency values in case of non-optimal values of C. (Other parameters are the same as in the above optimization of Rastrigin function.)

Table 5
Ratios of computational rounds in case of 64 CPUs

	Rastrigin	Keane	Step	Ackley	DeJong's 1st	DeJong's 3rd
$\frac{R_{BEAAux.}}{R_{BEA}}$	0.085	0.083	0.141	0.056	0.062	0.146
$\frac{R_{pMGA}}{R_{BEA}}$	0.097	0.095	0.177	0.067	0.065	0.168
$\frac{R_{pMGAAux.}}{R_{BEA}}$	0.086	0.083	0.134	0.058	0.061	0.135

Table 6
Parallel efficiency of optimization with non-optimal number of CPUs

C	15	30	45	60	75
$E_{p,BEA}$	0.299	0.170	0.121	0.091	0.075
$E_{p,BEAAux}$	0.932	0.828	0.900	0.678	0.855
$E_{p,pMGA}$	0.931	0.826	0.898	0.674	0.853
$E_{p,pMGAAux}$	0.932	0.828	0.899	0.676	0.855

The original BEA scales poorly, the other ones scale in a very similar way, and the efficiency of parallel versions is never lower than 0.67 in these examples. One can construct a parameter set when the parallel versions scale significantly worse, but calculations showed that for plausible cases the efficiency is always above 0.5.

Note that the formula used to calculate the number of computational rounds (11) is the same in the last three cases in our parameter settings, but the corresponding values of M are different in each line. That is why the number of computational rounds and efficiency values are different.

Figure 5 also shows the same effect for a wider range of CPU numbers. The parallel methods give the same curves according to the paper, and therefore only one of them is plotted. It is clear that if we can control the parameters, ideal efficiency can be achieved, but even with an unexpected number of CPUs the efficiency is mostly above 0.8.

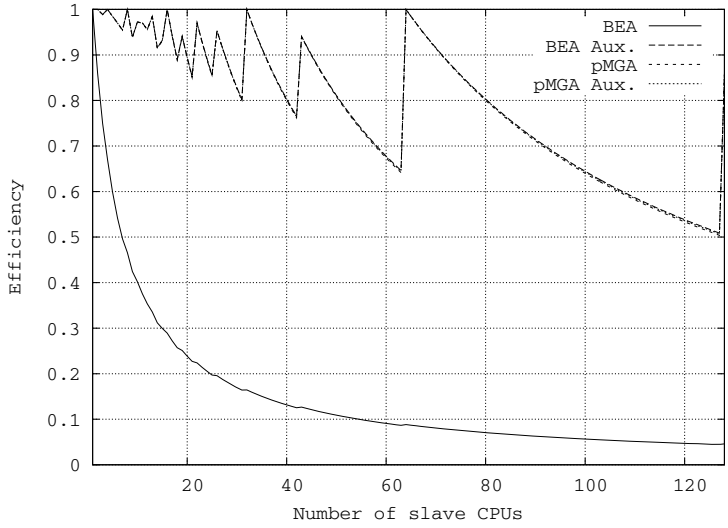


Figure 5
Efficiency of the algorithm as the function of the number of CPUs

3 The Effect of the Modified Gene Transfers on Genetic Diversity

3.1 Measuring Genetic Diversity

The modified gene transfer operators (BEA Aux., pMGA, pMGA Aux.) significantly differ from the original variant in one aspect: the new bacteria created by gene transfer are able to share their genetic information with other members of the population to a smaller degree, i.e., not more than $\lceil T/N \rceil$ times. This can decrease the genetic diversity, which can result in increasing the runtime of the program, and therefore it is important to measure the genetic diversity and prevent it from being too small.

The difference between two bacteria were defined with the following formula:

$$d_{ij} = \sqrt{\frac{\sum_{k=1}^{m}\left(\frac{x_{ik} - x_{jk}}{X_{k,\max} - X_{k,\min}}\right)^2}{g}} \tag{13}$$

This "distance" was originally proposed by Goldberg to realise niching with [6]. Here x_{ik} is the k^{th} chromosome of the i^{th} bacterium, $X_{k,max}$ and $X_{k,min}$ are the possible maximum and minimum values of a chromosome. Using this formula, the genetic diversity of a population can be determined in the following way:

$$D = \frac{1}{P-1}\sum_{i=1}^{P} d_{i,best} \tag{14}$$

The expressive meaning of this "genetic diversity" is straightforward: the value is between 0 and 1 and shows the average relative difference between chromosome values.

Figure 6 shows the change of genetic diversity during the optimization of the Rastrigin function. The chart shows the average of 15 repeated measurements. The main settings were the following: $C = 64$, $P = 64$, $K = 1$, $g = 20$, $T = 512$, $A = 64$.

It can be seen that all the methods show very similar genetic diversity functions. It is good news in the sense that the parallel methods are not worse than the original BEA. But all the methods show extremely low diversity at the end of the calculations, when the population consists of almost identical copies of an individual. When this happens, only the random search, due to the mutation operator, produces new values, which has very slow convergence.

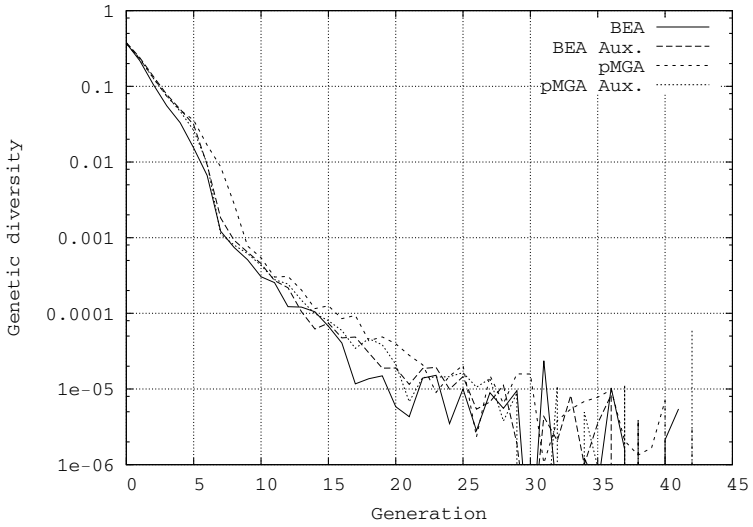


Figure 6
Average genetic diversity during optimization

The same phenomenon was discovered in the case of other test functions as well. This means that all the bacterial type optimizations produce slow convergence near the optimum because of the low genetic diversity. This conclusion is not surprising: the methods examined above have no mechanism to protect them against the reproduction of identical or very similar individuals.

Conclusions

The authors examined the effect of the recommendations and some other phenomenon as well in [9].

It was concluded that the ratio of gene transfers and population size heavily affects the optimization time. The ratio should be between 8 and 16 in case of using the parallel gene transfer operators. It is a rule of thumb if exhaustive tuning of the settings of the optimization cannot be realized.

Formulas have been given to estimate the change of wall clock time needed by optimization programs if the same optimization is executed with a different number of CPUs. It has been shown that the proposed parallel methods scale quite well even when the number of CPUs is not know in advance, while the original BEA's parallel efficiency is extremely low.

Lastly, it was pointed out that the parallel gene transfer operators do not worsen the genetic diversity in a considerable measure.

Based on this study the authors can recommend the parallel bacterial type methods with the optimal parameter setting described in this paper.

Acknowledgement

The authors' research is supported by the National Development Agency and the European Union within the frame of the project TAMOP 4.2.2-08/1-2008-0021 at the Széchenyi István University entitled "Simulation and Optimization - basic research in numerical mathematics".

References

[1] Addis, B., Locatelli, M.: A New Class of Test Functions for Global Optimization, Journal of Global Optimization, Vol. 38(3), 2007, pp. 479-501

[2] Bäck, T., Fogel, D. B., Michalewicz, Z.: Handbook of Evolutionary Computation, IOP Publishing and Oxford University Press, Abingdon, 1997

[3] De Jong, K. A.: An Analysis of the Behavior of a Class of Genetic Adaptive Systems, dissertation, University of Michigan, 1975

[4] Farkas, M., Földesi, P., Botzheim, J., Kóczy, L. T.: A Comparative Analysis of Different Infection Strategies of Bacterial Memetic Algorithms, in Proc. of 14th International Conference on Intelligent Engineering Systems (INES 2010), Las Palmas of Gran Canaria, Spain, May 5-7, 2010

[5] Gaviano, M., Kvasov, D. E., Lera, D., Sergeyev, Y. D.: Algorithm 829: Software for Generation of Classes of Test Functions with Known Local and Global Minima for Global Optimization, ACM Transactions on Mathematical Software, Vol. 29, No. 4, December 2003, pp. 469-480

[6] Goldberg, D. E.: Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Company, Inc., USA, 1989

[7] Grosan, C., Abraham, A.: Hybrid Evolutionary Algorithms: Methodologies, Architectures, and Reviews, Studies in Computational Intelligence, Vol. 75/2007, 2007, pp. 1-17

[8] Harvey, I.: The Microbial Genetic Algorithm, in Proceedings of the Tenth European Conference on Artificial Life, editor G. Kampis et al, Springer LNCS., Heidelberg, 2009

[9] Hatwágner, M., Horváth, A.: Parallel Gene Transfer Operations for the Bacterial Evolutionary Algorithm, Acta Technica Jaurinensis, Vol. 4, 2011, pp. 89-113

[10] Horváth, A., Horváth Z.: Optimal Shape Design of Diesel Intake Ports with Evolutionary Algorithm, Proceedings of 5th European Conference on Numerical Mathematics and Advanced Applications (ENUMATH 2003), edited by Feistauer, M. et al., Springer Verlag, 2004, pp. 459-470

[11] Mühlenbein, H., Schomisch, D., Born, J.: The Parallel Genetic Algorithm as Function Optimizer, Parallel Computing, Vol. 17, 1991, pp. 619-632

[12] Nawa, N. E., Furuhashi, T.: A Study on the Effect of Transfer of Genes for the Bacterial Evolutionary Algorithm, Second Intenational Conference on Knowledge-based Intelligent Electronic System, editors Jain, L. C., Jain, R. K., Adelaide, Australia, 21-23 April 1998, pp. 585-590

[13] Nawa, N. E., Furuhashi, T.: Fuzzy System Parameters Discovery by Bacterial Evolutionary Algorithm, IEEE Transactions on Fuzzy Systems, Vol. 7, No. 5, 1999, pp. 608-616

[14] Nawa, N. E., Hashiyama, T., Furuhashi, T., Uchikawa, Y.: A Study on Fuzzy Rules Discovery Using Pseudo-Bacterial Genetic Algorithm with Adaptive Operator, Proceedings of IEEE Int. Conf. on Evolutionary Computation, ICEC'97, 1997

[15] Pintér, J. D.: Global Optimization in Action, Kluwer Academic Publishers, Dordrecht, Netherlands, 1996

[16] Törn, A., Zilinskas, A.: Global Optimization, Lecture Notes in Computer Science, No. 350, Springer-Verlag, Berlin, 1989

Robust Multiobjective Optimization of Cutting Parameters in Face Milling

Şeref Aykut[1], Aykut Kentli[2], Servet Gülmez[3], Osman Yazıcıoğlu[4]

[1] Bitlis Eren University, Department of Mechanical Engineering, Faculty of Engineering Architecture, 13000 Bitlis, Turkey; E-mail: saykut@beu.edu.tr

[2] University of Marmara, Department of Mechanical Engineering, Faculty of Engineering, Göztepe Kampüsü 81040 Kadıköy – İstanbul, Turkey
E-mail: akentli@marmara.edu.tr

[3] University of Kocaeli, Gölcük Higher School of Vocational Education, Mechanical Department, Kocaeli, Turkey; E-mail: servet.gulmez@kocaeli.edu.tr

[4] Istanbul Commerce University, Department of Engineering and Design, 34840 Istanbul, Turkey; E-mail: oyazicioglu@iticu.edu.tr

Abstract: In this paper, a new multiobjective optimization approach is proposed for the selection of the optimal values for cutting conditions in the face milling of cobalt-based alloys. This approach aims to handle the possible manufacturing errors in the design stage. These errors are taken into consideration as a change in design parameter, and the design most robust to change is selected as the optimum design. Experiments on a cobalt-based superalloy were performed to investigate the effect of cutting speed, feed rate and cutting depth on the cutting forces under dry conditions. Material removal rate values were also obtained. Minimizing cutting forces and maximizing the material removal were considered as objectives. It is believed that the used method provides a robust way of looking at the optimum parameter selection problems.

Keywords: face milling; robust optimization; cobalt-based superalloy; sensitivity; multiobjective optimization; optimum cutting parameters

1 Introduction

Cobalt-based superalloys are used extensively in applications that require good wear, corrosion and heat resistance [1, 2]. Such features make them preferable in the nuclear and aerospace industries [3-5]. Among the cobalt-based superalloys, the most common ones are stellite alloys, especially the well-known stellite 6. The use of this alloy in industry has been increasing recently. Application areas

include pulp and paper processing, oil and gas processing, pharmaceuticals, chemical processing and medical applications. It is also employed in applications where corrosion resistance is an important factor.

As the use of cobalt-based stellite alloys has extended into various industrial sectors, the need for improving corrosion resistance of stellite alloys has increased as well. It has been observed that processing changes most probably affect the corrosion performance due to its effect on the microstructure of stellite alloy [6].

Cobalt-based superalloys are primarily based on carbides in Co matrix form. Their strength at grain boundaries, distribution, size and shape of carbides depend on processing conditions. Solid solution strength of Co-base alloys is normally provided by tantalum, tungsten, molybdenum, chromium and columbium [7-9]. Today, these alloys exist in a variety of more than 20 commercially available products, being used extensively in high temperature applications requiring superior wear, corrosion and heat resistance [10-11].

There are two main problems in machining cobalt-based superalloys. The first one is short tool life due to the working hardening and attrition properties of the superalloys. The second is the severe hardening of the surface of machined work pieces due to heat generation and plastic deformation. In order to achieve adequate tool life and the surface integrity of the machined surface, it is crucial to select reasonable machining conditions and parameters [4].

It is difficult to machine superalloys. The machinability of superalloys has not been improved enough, although there are new improvements in cutting tools. Machinability can be improved by minimizing tool-chip connection area, providing a sharp cutting edge and minimizing cutting depth. Machinability can be also improved further by providing minimum heat extraction, which results in a slow cutting speed and feed rate [4].

Metal machining not only requires knowledge of related areas of science and technology, but also plays an important role in manufacturing [12]. Because of its significance and complication, much attention has been paid to the cutting process, and many approaches have been attempted to get a better understanding of metal cutting principles. So far these methods have been mainly confined to either theoretical or experimental works. It is well-known that experimental studies are reliable and practical, but they are usually time-, labor- and material-consuming. Regarding theoretical analyses, there is experience in establishing and handling mathematical models, but much less experience and even avoidance of in experimental studies. The optimization method used in this paper utilizes few experimental results; therefore, it avoids lengthy operations. In addition, it uses a simple mathematical model of the cutting forces. Thus, it avoids complicated mathematical models. The combination of both approaches achieves a robust and reliable estimation.

There are several studies on surface milling [13-15]. These studies show that cutting forces also increase when feed rate and cutting depth increase. As cutting speed is a parameter directly affecting tool life, cutting forces are not directly related to the cutting speed. Since tool life is longer in asymmetric milling than in symmetric milling [4, 14], asymmetric face milling was preferred in this study. Additionally, inclined cutting theory is used in which cutting the tool grasps the work piece well and chip is removed as soon as possible.

This paper mainly focuses on finding the optimum parameters considering the cutting forces and material removal rate for milling of cobalt-based alloys. The cutting tests were carried out under dry conditions using PVD coated inserts. The machining parameters are optimized by using a new approach based on robustness. The practical cutting parameters can be different from what the manufacturers predict due to the uncertainties in material properties and the variations of the parameters in manufacturing. The used method takes care of these uncertainties by giving small deviations to parameters. From this perspective, this study is unique as an application to machining of cobalt-based alloys. Furthermore, suitability of the method is also analyzed by finding the optimum parameters.

The commonly used quantitative methods consider a single objective, such as minimization of cost or maximization of profit, for the optimization of the machining operations. For the process of the single objective optimization, several different techniques were proposed such as differential calculus [16], geometric and stochastic programming [17], regression analysis [18, 19], linear programming [20], genetic algorithm [21], and computer simulation [22]. In addition, there are also other local search methods, such as tabu search, ant colony optimization, pattern search, scatter search and fuzzy possibilistic programming [23].

In this study, optimization of the machining operation is considered as a multiobjective optimization problem. The new approach, considering robustness that does not require gradient calculations, useful with discrete variables, has shown its effectiveness and usability [24].

2 Optimization Methodology

Generally, uncertainty can be classified into two types: reducible and irreducible [25, 26, 27, 28]. Reducible uncertainty, often referred to as epistemic uncertainty, is used to represent incomplete information about an event such as a simulation or model of an engineering problem. In contrast, irreducible uncertainty, often referred to as aleatory uncertainty [25, 26], arises due to the inherent uncertainty associated with an engineering system under consideration. Irreducible uncertainty

refers to the uncertainty or a part of uncertainty that cannot be reduced at any expense due to its inherent nature such as the likelihood of the fractional components in raw crude oil. Thus, it is treated as irreducible. Research streams have been extensively developed to understand and deal with uncertainty in design problems along two inter-related, but different directions: robust optimization [29] and sensitivity analysis [30]. Li [31] proposed an integrated approach that incorporates two existing approaches into one optimization procedure: a robust optimization approach used to design around irreducible uncertainty [32] and a global sensitivity analysis to deal with reducible uncertainty [33, 34]. Despite the fact that this study employs the same approach, it focuses on implementation problems having a discrete solution set and on the investigation of the effect of certain change in parameters, instead of performing sensitivity analysis. The implementation of the proposed approach to find the optimum cutting conditions in machining a superalloy does exist in literature, although it has already implemented in two different problems [24, 35].

This approach has two main steps: obtaining Pareto optimum points and selecting the robust optimum point. In this study, the design space is formed by the obtained data from the experiments. The Pareto points are obtained by using weighting function methodology and the optimum point is selected among them according to the new approach considering robustness. A comprehensive survey on robust optimization can be found in [36]. According to this analogy, this study can be described as a tolerance design treating uncertainty at deterministic parameters. This approach has proved that it is quite useful in dealing with discrete variables defined on a population of cutting condition values obtained from experiments.

2.1 Obtaining Pareto Optimum Set

Over the past few decades, multiobjective optimization has been acknowledged as an advanced design technique in optimization. The reason is that the most real-world problems are multidisciplinary and complex, since it is common to have more than one important objective in each problem. To accommodate many conflicting design goals, one needs to formulate the optimization problem with multiple objectives.

A multiobjective optimization problem can be formulated as follow:

Min $[f_1(x), f_2(x),..., f_n(x)]$

subject to

$$g_j(x) \geq 0 \quad j = 1,2,...,m \tag{1}$$

$$h_j(x) = 0 \quad j = 1,2,...,p<n$$

where x is a n-dimensional design variable vector, $f_i(x)$ is the objective function, $g_j(x)$ and $h_j(x)$ are inequality and equality constraints.

A variety of techniques and applications of multiobjective optimization have been developed over the past decade. The progress in the field of multiobjective optimization was summarized by Marler and Arora [37] and later by Chinchuluun and Pardalos [38]. It is inferred from these surveys that if one has decided that an optimal design is to be based on the consideration of several objectives, then the multiobjective theory (Pareto theory) provides the necessary framework. If the minimization or maximization is the objective for each criterion, then an optimal solution should be a member of the corresponding Pareto set. In addition, further improvements in one criterion require a clear tradeoff with at least one another criterion.

Radfors, et al [39] in their study has explored the role of Pareto optimization in computer-aided design. They used the weighting method, the noninferior set estimation (NISE) method, and the constraint method for generating the Pareto optimal. Marler and Arora [40] have investigated the fundamental significance of the weights in terms of preferences, the Pareto optimal set, and objective-function values. Kim and de Weck [41] presented an adaptive weighted sum (AWS) method for multiobjective optimization problems. In the first phase, the usual weighted sum method is performed to approximate the Pareto surface quickly, and a mesh of Pareto front patches is identified. Each Pareto front patch is then refined by imposing additional equality constraints that connect the pseudonadir point and the expected Pareto optimal solutions on a piecewise planar hypersurface in the m-dimensional objective space. In this study, the weighted sum method was used and a brief explanation of the method is given at the following paragraphs.

Pareto serves optimality as the basic multiobjective optimization concept in virtually all of the previous literature [42]. The Pareto optimal is stated in simple words as follows: A vector X* is Pareto optimal if there exists no feasible vector X which would decrease some objective function without causing a simultaneous increase in at least one objective function. This definition can be explained graphically. An arbitrary collection of feasible solutions for a two-objective minimization problem is shown in Figure 1. The area inside of the shape and its boundaries are feasible. The axes of the graph are the objectives: F' and Q'. It can be seen from the graph that the noninferior solutions are found in the portion of the boundary between points A and B. Thus, here arises the decision-making problem from which a partial or complete ordering of the set of nondominated objectives is accomplished by considering the preferences of the decision maker. Most of the multiobjective optimization techniques are based on how to elicit the preferences and determine the best compromise solution. From this perspective, the used approach differs from other techniques. This approach chooses the optimum point by considering the change in parameters and the effect of change to objectives.

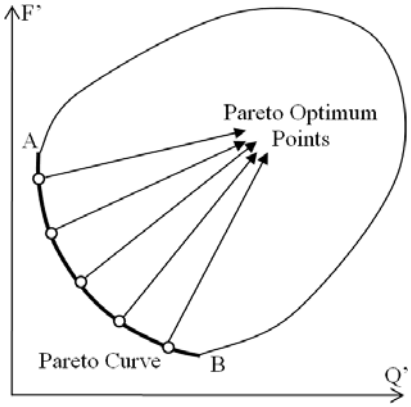


Figure 1
Graphical interpretation of Pareto optimum

The weighted sum method is based on the preference techniques of the weights' prior assessment for each objective function. It transforms the multiobjective function to a single criterion function through a parameterization of the relative weighting of the objectives. With the variation of the weights, the entire Pareto set can be generated. This means that we change the multiobjective optimization problem to a single optimization problem by creating one function of the form.

$$f(x) = \sum_{i=1}^{k} w_i f_i(x) \tag{2}$$

where $w_i \geq 0$ are the weighting coefficients representing the relative importance of the objective.

The best results are usually obtained if objective functions are normalized. In this case, the vector function is normalized to the following form

$$\tilde{f}(x) = \left[\tilde{f}_1(x), \tilde{f}_2(x), ..., \tilde{f}_k(x)\right]^T \tag{3}$$

where $$\tilde{f}_i(x) = \frac{f_i(x)}{f_i^o} \quad i=1,2,...,k \tag{4}$$

Here, f_i° is generally the maximum value of i^{th} objective function (A condition $f_i^\circ \neq 0$ is assumed).

In this study, the total force and cutting flow of material are considered as objectives. The total force value is the resultant force of the obtained forces in experiments, and the cutting flow value is obtained by using Equation 5.

$$Q = \frac{a_p \,.\, a_e \,.\, f}{1000} \ \ (m^3 / \min) \tag{5}$$

where a_p, a_e and f represents cutting depth, cutting width (constant) and feed, respectively. The cutting flow of material, Q, should be maximized, and total force F should be minimized to minimize the tool wear and used power. Thus, to maximize the composite weighted function, inverse of the force is taken as objective and the objective function (J) is set as

$$J = \frac{1}{F} + w.Q \tag{6}$$

where w is a weighting co-efficient varied to obtain Pareto optimum points. In order to bring the values in the same range, Q and (1/F) are normalized with their maximum values where the relation in Equation 7 is used to obtain Pareto optimum values;

$$J = F' + w.Q' = \frac{F_{\max}}{F} + w.\frac{Q}{Q_{\max}} \tag{7}$$

The design space is related with the allowed maximal dimension of the controlled variable vectors used during the machining operation. The design variables are the cutting speed (V_c), the feed (f) and the cutting depth (a_p). The design space is a typical discrete and non-convex domain.

2.2 Selecting the Robust Optimum Point

At the second step, according to the Pareto optimum points, the optimum point is selected based on changes in the objective function when small variations are permitted in design variables. In this study, equal contributions of each variable are considered. Based on positive/negative variations in design variables, and average changes in the objective function values are calculated at every Pareto optimum point. Figure 2 shows the change in parameter and objective values for two parameter case.

The optimum point is selected as the one having the minimum changes on

$$\Delta V_j = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[F'(z_i) - F'(z_o)\right]^2 + \left[Q'(z_i) - Q'(z_o)\right]^2\right\} \tag{8}$$

where n is the number of design variable change around every Pareto optimum point, $F'(z_o)$, $Q'(z_o)$ are the objective function values at the Pareto optimum point, $F'(z_i)$, $Q'(z_i)$ are the objective function values when a certain change is applied to a design parameter, and j is the index of the Pareto optimum point [24]. While calculating the change in objectives, the objective values that are not in the

feasible region are not taken into account. For example, the changes in objective values at point 1, 5 and 6 in Figure 2 are not considered because they are not in the feasible region.

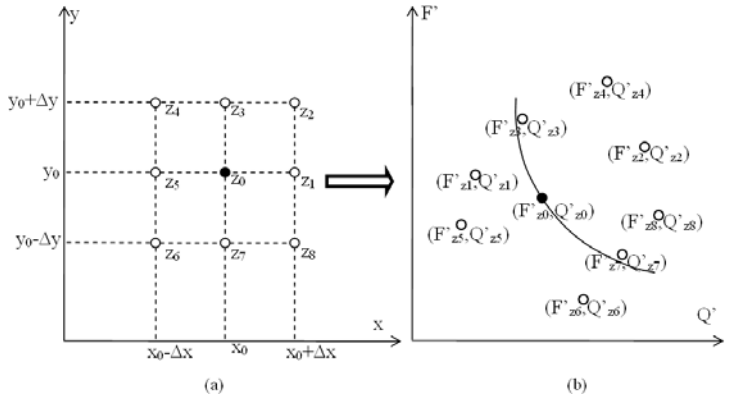


Figure 2
Change in design parameters (a) and objectives (b)

3 Experimental Setup

The experiments to investigate the cutting forces for asymmetric face milling were carried out on a CNC milling machine. The influence of the other machining conditions (feed rate, axial depth of cut and feed rate per tooth) on the cutting forces in dry cutting were also considered. A 9 kW Johnford WMC-850 series of CNC milling machine was used. The cutting forces were measured by using a Kistler 9265B series dynamometer.

Surface machining was done with the parameters selected by considering the recommended values of ISO for superalloys [43]. The experiments are given parameter values as shown in Table 1.

Table 1
Cutting conditions for face milling

Cutting speeds	V_s	m/min	30,35,40
Feed rates	f	mm/min	60,70,80,90,100
Depths of cut	a_p	mm	0.25, 0.5, 0.75
Widths of cut	a_e	mm	50
Feed rates per	F_z	mm/tooth	0.1
Coolant	--	--	Dry

To avoid thermal effects, lower cutting speeds were chosen. In addition, higher cutting speeds result in severe tool wear and the higher feeds cause a large deformation rate. Ranges for process parameters and the obtained results are shown in Table 2.

The stellite 6 workpiece used in the machining test is made from cast material. The chemical composition of the workpiece material is given in Table 3. The hardness of the workpiece is 44 HRC. The tool material ISO P30 (SECO grade H40, quality insert) was coated using PVD (Physical Vapor Deposition) [44].

4 Results and Discussion

The approach described above was applied to the experimental data given in Section 3. In this study, three parameters (the cutting speed, the feed and the cutting depth) were considered. The experiments in Table 3 were used in the calculation. The change in cutting depth was assumed as 0.25 mm, the change in feed was assumed as 10 mm/min. and the change in cutting speed was assumed as 5 m/min. The objective functions were evaluated under these assumptions. Then, the Pareto points were evaluated using weighting function. A sample calculation is given in Table 4 for the first data of Table 2 and w=0.1.

For every point, objective function values were calculated. Even though the weighting coefficient of the objective function is firstly changed from 10^{-6} to 10^{6}, it has been seen that it is enough to change from 10^{-1} to 10 to get the Pareto optimum points (Table 5).

Table 2
The experimental values obtained in machining

Data	a_p (mm)	f (mm/min)	V_c (m/min)	F_z(N)	F_y(N)	F_x(N)
1	0.25	60	30	140	50	40
2	0.25	70	30	150	70	70
3	0.25	80	30	170	90	90
4	0.25	90	30	210	130	140
5	0.25	100	30	300	140	200
6	0.25	60	35	220	125	180
7	0.25	70	35	240	140	175
8	0.25	80	35	360	175	240
9	0.25	90	35	380	150	250
10	0.25	100	35	400	180	265
11	0.25	60	40	250	160	140

12	0.25	70	40	280	165	150
13	0.25	80	40	300	170	170
14	0.25	90	40	320	180	175
15	0.25	100	40	360	200	180
16	0.50	60	30	240	120	160
17	0.50	70	30	280	120	170
18	0.50	80	30	310	150	175
19	0.50	90	30	340	280	180
20	0.50	100	30	380	300	200
21	0.50	60	35	150	200	160
22	0.50	70	35	190	210	200
23	0.50	80	35	200	200	200
24	0.50	90	35	250	210	250
25	0.50	100	35	350	220	300
26	0.50	60	40	280	130	150
27	0.50	70	40	300	210	200
28	0.50	80	40	320	200	210
29	0.50	90	40	330	210	230
30	0.50	100	40	500	310	300
31	0.75	60	30	325	160	300
32	0.75	70	30	350	170	320
33	0.75	80	30	365	185	325
34	0.75	90	30	400	225	335
35	0.75	100	30	450	250	360
36	0.75	60	35	250	180	180
37	0.75	70	35	280	200	220
38	0.75	80	35	300	220	250
39	0.75	90	35	310	250	250
40	0.75	100	35	410	310	380
41	0.75	60	40	330	300	310
42	0.75	70	40	380	320	325
43	0.75	80	40	425	280	355
44	0.75	90	40	500	330	360
45	0.75	100	40	530	375	375

Table 3
Composition of the experimental material Stellite 6

Element	C	Si	Mn	Cr	Ni	Mo	W	Ti	Fe	Ta	Co
Weight (%)	1.09	1.07	0.49	28.17	1.92	0.96	5.17	0.01	2.88	0.04	Balanced

Table 4
Calculation of an objective function value (F_{max}=153.948 and Q_{max}=3.75 for the case)

Data	a_p (mm)	f (mm/min)	V_c (m/min)	Q	Q/Q_{max}	F_z(N)	F_y(N)	F_x(N)	F_r	F_{max}/F_r	J
1	0.25	60	30	0.75	0.2	140	50	40	153.948	1	1.02

Table 5
Pareto optimum points

Data	a_p (mm)	f (mm/min)	V_c (m/min)	Weighting Coefficients
1	0.25	60	30	0.1, 0.5
2	0.75	100	30	1, 5, 10

Change is given to the design parameters of obtained Pareto optimum points and deviation in objectives is calculated (Table 6). Only feasible points are given in this table.

Table 6
Change in parameter and objective for Pareto optimum points

Change in Pareto Point 1	Obj. Func. F_{max}/F	Obj. Func. Q/Q_{max}	Deviation in F_{max}/F	Deviation in Q/Q_{max}
0.25;60; 30	1	0.2	0	0
0.25;60; 35	0.495769	0.2	0.504231	0
0.25;70; 30	0.85659	0.233333	0.14341	0.033333
0.25;70; 35	0.468829	0.233333	0.31171	0.033333
0.5; 60; 30	0.492776	0.4	0.507224	0.2
0.5; 60; 35	0.518664	0.4	0.481336	0.2
0.5; 70; 30	0.441295	0.466667	0.558705	0.266667
0.5; 70; 35	0.44404	0.466667	0.55596	0.266667
Change in Pareto Point 2	Obj. Func. F_{max}/F	Obj. Func. Q/Q_{max}	Deviation in F_{max}/F	Deviation in Q/Q_{max}
0.75;100;30	0.245073	1	0	0
0.75; 90; 30	0.27094	0.9	0.025867	0.1
0.75;100;35	0.240838	1	-0.00424	0
0.75; 90; 35	0.327401	0.9	0.082328	0.1
0.5; 100; 30	0.293888	0.666667	0.048815	0.333333
0.5; 90; 30	0.323546	0.6	0.078473	0.4
0.5; 100; 35	0.301396	0.666667	0.056323	0.333333
0.5; 90; 35	0.374371	0.6	0.129298	0.4

As a last step, the squares of total deviations and their mean were calculated by using Equation 8. It is seen that the first Pareto point has a total average deviation of 0.236451 and the second Pareto point has a total average deviation of 0.074765. the second Pareto point is the optimum point, having the minimum deviation. The optimum cutting condition found is at a_p=0.75 mm; f=100 mm/min; V_c=30 m/min.

Conclusions

Since cutting conditions regulate the machining process through the developed cutting forces, the optimization of machining parameters is important. Uncontrollable variations are unavoidable in machining due to the quality of manufacturing tools, measurement tools, operators' mistakes, imperfections during the manufacturing processes, etc. The method used was to evaluate average deviations from the Pareto optimum points because of uncontrollable variations. The selection of the optimum design point with the minimum deviation is the criterion to find the robust optimum point.

Although there are several methods in literature for the multiobjective optimization of machining processes, a new approach is used in this work. The main advantage of this approach is to get the robust optimum point. In addition, there is no need to calculate complex modeling formulations or simulations of the process, which requires a lot of time and hardware. Instead, simple statistical calculations are enough to get acceptable results. Moreover, this approach gives much more reliable solutions because experimental data were used, and these data were the exact values to represent the process.

The used method has proved that it is very useful when dealing with discrete variables defined on a population of cutting condition values obtained from experiments. It is believed that this method provides a robust way of looking at the optimum parameter selection problem. In addition, it can easily handle those cases where each of the design variables has different uncertainty ranges.

The results of the case study have shown the benefits of the new approach. The optimum cutting conditions are determined for the machining of Cobalt-based alloy stellite 6 material as a_p = 0.75 mm; f = 100 mm/min and V_c = 30 m/min.

When the results are compared with previous study which considered surface roughness, it is seen that feed values were same, but depth of cut and cutting speed get higher values since material removal rate and resultant force are considered in the meantime.

References

[1] Agarwal, S. C., Ocken, H. (1990) The Microstructure and Galling Wear of a Laser-melted Cobalt-base Hardfacing Alloy. Wear, Vol. 140, pp. 223-233

[2] Crook, P. (1993) Metals Handbook Vol. 2: Properties and Selection: Non-ferrous Alloys and Special-Purpose Materials, 10th edition. USA: ASM Int.

[3] Murray, J. D., McAlister, A. J. (1984) Bulletin in Alloy Phase Diagrams, Vol. 5, p. 90

[4] Aykut, Ş. (2005) The Investigation of Effects of Machinability on Chip Removal Parameters for Face Milling of Cobalt-Based Superalloy Steels. Thesis (PhD) Marmara University, Istanbul, Turkey

[5] Kuzucu, V., Ceylan, M., Celik, H., Aksoy, İ. (1997) Microstructure and Phase Analyses of Stellite Plus 6 wt.% Mo Alloy. Journal of Materials Processing Technology, Vol. 69, pp. 257-263

[6] Mohamed, K. E., Gad, M. M. A., Nassef, A. E., El-Sayed, A. W. (1999) Localized Behaviour of Powder Metallurgy Processed Cobalt-based Alloy Stellite 6 in Chloride Environments. Zeitschrift fuer Metallkunde, Vol. 90, pp. 195-201

[7] Balazinski, M., Songmene, V. (1995) Improvement of Tool Life through Variable Feed Milling of Inconel 600. Annals of CIRP, Vol. 44, No. 1, pp. 55-58

[8] Natural, N., Yamaha, Y. (1993) High Speed Machining of Inconel 718 with Ceramic Tools. Annals of CIRP, Vol. 42, No. 1, pp. 103-106

[9] Alauddin, M., El-Baradie, M. A. and Hashmi, M. S. J. (1996) End Milling Machinability of Inconel 718. Journal of Engineering Manufacturing, Vol. 210, pp. 11-23

[10] Field, M. (1968) Machining Aerospace Alloys. Iron and Steel Institute, Special Report 94

[11] Warburton, P. (1967) Problems of Machining Nickel-based Alloys. Iron and Steel Institute, Special Report 94

[12] Milton, C. (1984) Metal cutting Principles. Oxford: Oxford University Press

[13] Alauddin, M., Mazid, M. A., El Baradi, M. A., Hashmi, M. S. J. (1998) Cutting Forces in the End Milling of Inconel 718. Journal of Materials Processing Technology, Vol. 77, pp. 153-159

[14] Diniz, A. E., Filho, J. C. (1999) Influence of the Relative Positions of Tool and Workpiece on Tool Life, Tool Wear and Surface in the Face Milling Process. Wear, Vol. 232, pp. 67-75

[15] Shunmugam, S. V., Bhaskara, R. T., Narendran, T. (2000) Selection of Optimum Conditions in Multi-Pass Face Milling Using a Genetic Algorithm. International Journal of Machine & Tools Manufacture, Vol. 40, pp. 4014-4414

[16] Lavernhe, S., Tournier, C., Lartigue, C. (2008) Optimization of 5-axis High-Speed Machining Using a Surface-based Approach. Computer-aided Design, Vol. 40, pp. 1015-1023

[17] Ye, T., Xiong, C.-H. (2008) Geometric Parameter Optimization in Multi-Axis Machining. Computer-aided Design, Vol. 40, pp. 879-890

[18] Bağcı, E., Aykut, Ş. (2006) A Study of Taguchi Optimization Method for Identifying Optimum Surface Roughness in CNC Face Milling of Cobalt-based Alloy (Stellite 6) The International Journal of Advanced Manufacturing Technology, Vol. 29, pp. 940-947

[19] Cus, F., Balic, J. (2000) Selection of Cutting Conditions and Tool Flow in Flexible Manufacturing System. International Journal for Manufacturing Science and Technology, Vol. 2, pp. 101-106

[20] Tan, F. P., Creese, R. C. (1995) A Generalized Multi-Pass Machining Model for Machining Parameter Selection in Turning. International Journal of Production Research, Vol. 33, pp. 1467-1487

[21] Davim, J. P., Conceicçao Antonio, C. A. (2001) Optimisation of Cutting Conditions in Machining of Aluminium Matrix Composites Using a Numerical and Experimental Model. Journal of Materials Processing Technology, Vol. 112, pp. 78-82

[22] Milfelner, M., Cus, F. (2000) System for Simulation of Cutting Process. International Scientific Conference on the Occasion of the 50th Anniversary of Founding the Faculty of Mechanical Engineering, Ostrava, pp. 349-352

[23] Onwubolu, G. C., Kumalo, T. (2002) Multi-Pass Turning Optimisation Based on Genetic Algorithms. International Journal of Production Research, Vol. 39, No. 16, pp. 3727-3745

[24] Kentli, A., Kar, A. K. (2002) A Multiobjective Optimization Approach to Buckling Problem of Non-Prismatic Columns. 6th Biennial Conference on Engineering Systems Design and Analysis (ESDA 2002), Istanbul, Turkey

[25] Oberkampf, W. L., DeLand, S. M., Rutherford, B. M., Diegert, K. V., Alvin, K. F. (2002) Error and Uncertainty in Modeling and Simulation. Reliability Engineering & System Safety, Vol. 75, No. 3, pp. 333-357

[26] Oberkampf, W. L., Helton, J. C., Joslyn, C. A., Wojtkiewicz, S. F., Ferson, S. (2004) Challenge Problems: Uncertainty in System Response Given Uncertain Parameters. Reliability Engineering & System Safety, Vol. 85, No. 1-3, pp. 11-19

[27] O'Hagan, A., Oakley, J. E. (2004) Probability is Perfect, but We can't Elicit It Perfectly. Reliability Engineering & System Safety, Vol. 85, No. 1-3, pp. 239-248

[28] Guo, J., Du, X. (2007) Sensitivity Analysis with the Mixture of Epistemic and Aleatory Uncertainties. AIAA Journal, Vol. 45, No. 9, pp. 2337-2349

[29] Taguchi, G. (1978) Performance Analysis Design. International Journal of Production Research, Vol. 16, pp. 521-530

[30] Saltelli, A., Chan, K., Scott, E. M. (2000) Sensitivity analysis. NewYork, NY: JohnWiley & Sons

[31] Li, M., Azarm, S., Williams, N., Al Hashimi, S., Almansoori, A., Al Qasas, N. (2009) Integrated Multi-Objective Robust Optimization and Sensitivity Analysis with Irreducible and Reducible Interval Uncertainty. Engineering Optimization, Vol. 41, No. 10, pp. 889-908

[32] Li, M., Azarm, S., Boyars, A. (2006) A New Deterministic Approach Using Sensitivity Region Measures for Multi-Objective and Feasibility Robust Design Optimization. Journal of Mechanical Design, Vol. 128, No. 4, pp. 874-883

[33] Li, M. (2007) Robust Optimization and Sensitivity Analysis with Multi-Objective Genetic Algorithms: Single- and Multidisciplinary Applications. Thesis (PhD). University of Maryland, College Park, Maryland, USA

[34] Li, M., Williams, N., Azarm, S. (2009) Interval Uncertainty Reduction and Single-Disciplinary Sensitivity Analysis with Multi-Objective Optimization. Journal of Mechanical Design, Vol. 131, No. 3, pp. 1-11

[35] Işık, B., Kentli, A. (2009) Multicriteria Optimization of Cutting Parameters in Turning of UD-GFRP Materials Considering Sensitivity. The International Journal of Advanced Manufacturing Technology, Vol. 44, pp. 1144-1153

[36] Beyer, H., Sendhoff, B. (2007) Robust Optimization – a Comprehensive Survey. Computer Methods in Applied Mechanics and Engineering, Vol. 196, pp. 3190-3218

[37] Marler, R. T., Arora, J. S. (2004) Survey of Multi-Objective Optimization Methods for Engineering. Structural and Multidisciplinary Optimization, Vol. 26, pp. 369-395

[38] Chinchuluun, A., Pardalos, P. M. (2007) A Survey of Recent Developments in Multiobjective Optimization. Annals of Operations Research, Vol. 154, pp. 29-50

[39] Radford, A. D., Gero, J. S., Roseman, M. A., Balachandran, M., (1985) Pareto Optimization as a Computer-aided Design Tool. In Optimization in Computer-aided Design. (J. S. Gero Eds.) pp. 47-69, Amsterdam: North-Holland

[40] Marler, R. T., Arora, J. S. (2010) The Weighted Sum Method for Multi-Objective Optimization: New Insights. Structural and Multidisciplinary Optimization, Vol. 41, No. 6, pp. 853-862

[41] Kim, I. Y., de Weck, O. L. (2006) Adaptive Weighted Sum Method for Multiobjective Optimization: a New Method for Pareto Front Generation. Structural and Multidisciplinary Optimization, Vol. 31, pp. 105-116

[42] Lee, K. Y., El-Sharkawi, M. A. (2008) Modern Heuristic Optimization Techniques: Theory and Applications to Power Systems. IEEE/John Wiley Publishing

[43] ISO 8688-1 (1989) Tool Life Testing in Milling, Part I, Face Milling, 1st Edition. International Standards for Business, Government and Society

[44] SECO (2003) Catalogue and Technical Guide Milling. Sweden

After Information Security – Before a Paradigm Change (A Complex Enterprise Security Model)

Pál Michelberger Jr.

Óbuda University, Keleti Károly Faculty of Business and Management, Institute of Management and Organisation, Népszínház u. 8, H-1081 Budapest, Hungary
e-mail: michelberger.pal@kgk.uni-obuda.hu

Csaba Lábodi

QLCS Kft., Cholnoky J. u. 1/a, H-8200 Veszprém, Hungary
e-mail: csaba.labodi@gmail.hu

Abstract: Security management for business enterprises is currently undergoing major changes. Instead of the separate regulation of distinct areas (guarding infrastructure, work safety, security technology, information security, etc.) there is an emerging holistic approach based on new management methods and company culture. Partly as a result of existing traditions, the professional business background to the implementation of these changes is rather fragmented and may not yet exist at all. Our survey calls attention to a new opportunity. In our opinion the expected level of company security and business continuity may be reached departing from information security-related international standards and recommendations, by business risk analysis and compliance with a wide range of security expectations.

Keywords: information security; risk analysis; business continuity; process security; enterprise security model

Introduction

The spread of standardised information security management systems clearly attests to the rise in the demand of enterprise security [17]. Protecting information is nevertheless not a sufficient measure in itself. Maintaining reliability and satisfying business partners' demands in time, volume and quality requires much more. The tools of enterprise value generation as well as major and subordinate processes also have to be made secure. Enterprise security and disaster recovery plans and ideas, already known in connection with information security, may

serve as a good starting point for the creation of a holistic business enterprise security model. The present paper departs from the demands of the information security management system, discusses enterprise risk analysis and surveys various international security-related standards and recommendations. After summarising real-life experience, an attempt will be made to compile a complex enterprise security model that can deal with real-life threats.

Business enterprise security management is able to control the critical processes and tools of an enterprise to reach company targets derived from strategic plans. This means an uninterrupted series of planning, organisation, management and control and coordination activities that guarantee a desired and sustainable level of security for all internal and external parties of the enterprise. Instead of technical issues, the workings of the organisation system gain focus, allowing the measurement, continuous development and optimisation of security aims [2].

1 The Point of Departure: Information Security

Information is value for the enterprise, being a basis for managerial decisions and business success. It may be relevant with respect to products, services, technological know-how and available resources as well as business partners. If lacking or false, inaccurate or delivered into unauthorised hands, information may cause damage to the organisation. It must therefore be protected.

Information security is a far more complex issue than IT security. Today it is not enough to think in terms of firewalls, reliable hardware and well-defined identification systems. A conscious buildup of technological background is no longer sufficient.

The integrity, availability, and confidentiality [23] of information is primarily threatened by negligent handling or purposeful damage caused by internal employees (through company information management systems and the intranet) and strategic partners with access to company databases through the internet, extranet or Electronic Data Interchange (suppliers, retailers, cooperation partners and financial service providers).

Several other qualities, such as accuracy, accountability, non-repudiation, and dependability may also be linked to information security.

Generally, it is the handling of information and information carriers that is regulated in order to protect information property. This is independent from the form the information is presented in. Protection functions well if the information to be protected is defined, along with internal and external threats, the risks posed by these, and the regulations and means of protection [24].

The aim of information security is to ensure the continuity of business in a structured manner and mitigate damage caused by security events. Information security may be achieved by the application of protective measures, taking risks into consideration. These consist of regulations defining enterprise processes, the enterprise structure reflecting these processes, and the regulated operation of IT tools (hardware, software, telecommunications devices) appropriate to them [11].

In order to ensure the long-term operation of business organisations, the application of a system providing security is necessary (instead of means and devices that give a false sense of security).

Several interconnected dimensions or levels of security may be defined [9];

- Information technology infrastructure level (hardware, software and network protection)
- Information management level (data entry, modification, deletion, information gathering and data query)
- Conduct of affairs/Workflow level (process management, workflow)
- Organisational level (information security strategy, risk management)

The creation of n environment supportive of information security is also important, which means an accepted information security policy, clearly defined areas of responsibility, training, and the assurance of financial resources. In addition, this involves the full registration of IT devices and documentation, risk assessment (for IT devices and the challenges of the environment), and the handling of user authorisations (for access to documentation, networks, servers, workstations, application software, and the information itself). This means the assurance of business continuity and disaster recovery based on not only information security, but also on a process-centred vision [1].

The Business Continuity Plan (BCP) ensures the availability of business process backup IT resources at given times and functional levels as well as the minimalisation of damage caused by unexpected events. It is important for this document to include potential threats to the various processes, the likelihood of their occurrence, and the damage potentially resulting from the breakdown of the process. It is in the course of the so-called Business Impact Analysis (BIA) that procedures for the maintenance of operations are determined (Fig. 1).

The Disaster Recovery Plan (DRP) contains substitute solutions for the case of major damage and events resulting in the breakdown of information technology service. The aim is to facilitate the minimalisation of negative effects and the fast restoration of original circumstances at acceptable costs. This plan must also include supplementary measures and tools for the case of a limited availability or complete breakdown of resources that ensure the continuity of processes critical for the existence of the organisation. The Disaster Recovery Plan is thus usually linked to the Business Continuity Plan. Good BCPs and DRPs examine

organisational processes and consider the links and ties between these two. They contain practicable and risk-commensurate intervention orders, are known and accepted by the higher management of the organisation, and constantly undergo testing, maintenance and development.

The preparation of the business continuity and disaster recovery plans involves the surveying of all organisational processes and thus may take a long time (several months in certain cases) to introduce. An external adviser may have to be employed to process the internal interviews and systematise results. At the same time, internal experts in full knowledge of the organisation's workings are also needed to construct the complete system. This means considerable costs, both for introduction and maintenance. The education and continuous training of responsible leaders and subordinate workers must be a top priority.

In the course of the business impact analysis, the processes of the organisation are classified according to risk levels (low, medium or high priority). The Maximum Tolerable Downtime (MTD) is determined. The effects and probability of potential threats are also examined.

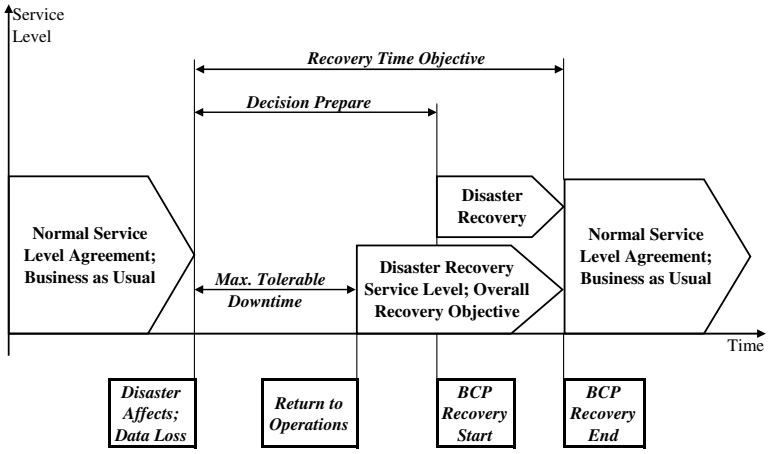


Figure 1
The Business Continuity Plan Model

2 Enterprise Risk Analysis

Risk is the potential occurrence of events or disturbances within the enterprise and in its environment (including its markets, etc.…) that endanger the fulfillment of customer demands or the security of any involved enterprise parties (stake- and stockholders).

The risk of security-related incidents may be expressed with a money/time unit quantity or, if it is not definable in this manner, with a "mark" showing the magnitude and tolerability of the risk [3]. Risk depends on the probability of harmful event occurrence as well as the resulting damage calculated in terms of finances (Fig. 2).

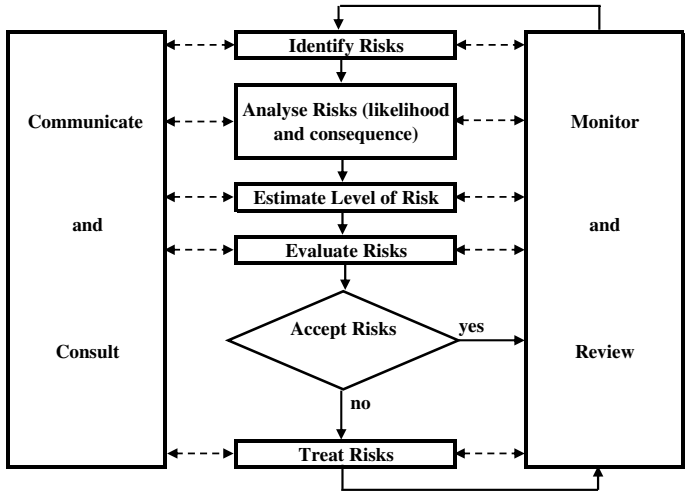


Figure 2
The process of risk management (based on Standards Australia, AS/NZS 4360 [32])

In the absence of accurately definable data related to potential risks, the umbrella term 'vulnerability' has been introduced instead of the traditional risk-centred approach. The vulnerability factors of supply chains may be classified into five groups [6], which may be complemented by two further risk sources [7]:

1) disturbances in the value-added process (manufacturing, purchasing, storage, delivery, scheduling)
2) control (non-existence or failure)
3) demand (lack of information, unpredictability, unexpected events)
4) supply (unreliability, lack of capacity, vis major)
5) environmental (economic and political events, accidents, natural disasters)

and

6) enterprise structure (if non-conformance with enterprise processes)
7) a supply or sale chain or a "network" composed of several individual companies (disturbances in communication or uncertainties in cooperation)

These seven points embrace virtually all security perspectives and thus may serve as a foundation for our security model with respect to enterprise functional risk analysis.

3 Standards and Recommendations for the Basis of Enterprise Security

There exist several internationally recognized and accepted documents in this area. We shall mention a few of these – the ones we deemed important based on our own professional perspective. Naturally, there are various other standards and recommendations that could be introduced and adapted to serve enterprise security.

Most regulations listed below are process-centred.

The joint application of standards and recommendations is no thought of evil. The structures of e.g. ISO 14001 and ISO 27001 standards are quite similar. An integrated environmental and information security management system may be created and run on these. Regulations based on the COSO enterprise risk management framework [16] may easily be introduced to a framework built on ITIL and/or CObIT [12].

3.1 ISO 14000

Enterprises may use an environment-centred standardised management system in support of their environment protection tasks in the course of their operations. Among other targets, this standard aims to lower the environmental impact resulting from enterprise operation, promote corporate image and make concerned parties interiorise environment-related behaviour patterns.

According to the standard, the environment-centred management system deals with those enterprise activities that have an impact on the environment, risk assessment, compliance with operation-related legal and other security requirements and the achievement of a still acceptable environmental impact level. The resources and capabilities for an environmentally conscious operation are defined along with forms of internal and external communication. Preparation guidelines for emergency situations, troubleshooting, controlling and prevention are also regulated. The standard does not include concrete requirements and control measures. Its implementation is primarily ethically driven with legal and economic factors gaining ground [18, 19].

3.2 BS OHSAS 18001

The main resource for enterprises is the well-trained, value-creating employee. Workplace health protection and security are effectively supported by a management system constructed according to the BS OHSAS 18001 standard [20]. Its primary aim is the definition and management of risk events that may negatively affect workers' performance or may cause an accident or health damage.

The management system handles the risks of work processes with a view toward the relevant legal environment as well as security requirements and targets characteristic for the given enterprise. It also regulates workplace health maintenance tasks. It is a valuable tool for the observation and evaluation of processes, and also for the definition of resources and capabilities necessary to maintain the management system. It prescribes the documentation and after-the-event investigation and evaluation of hazardous occurrences. Reactions to hazardous situations and corrective preventive activities are prioritised. The system standard does not prescribe concrete requirements or control methods but its application results in target-oriented and process-centred enterprise operation. By its application a further step may be taken towards the creation of a work environment that serves the better protection of human resources [21].

3.3 ISO/IEC 38500

An international standard that may also serve as a management framework, ISO/IEC 38500 of Australian origin is an aid for the intra-enterprise management of information and communication technologies [22]. Based on the document, a management cycle may be created (Fig. 3) which regulates the intratechnological support of business processes, evaluating and controlling them. Here managerial responsibility is also tackled along with information technology aspects of enterprise strategy, the acquisition of IT tools, their performance, compliance with business targets, and human behaviour.

The elements of the GRC model based on the standard are as follows

- Governance – enterprise targets, processes and the organisation running these processes, with special emphasis on IT supporting the targets;
- Risk Management – the identification of expected events and related risks, the definition of an acceptable security level for all enterprise processes and supporting IT tools (COSO ERM);
- Compliance – the enterprise must comply with internal prescriptions and laws, standards and contractual requirements.

The application of the model involves the creation of a comprehensive requirements list which is continuously adapted to changing circumstances. Management is aware of the risks as well as what expectations it has to fulfill in the given moment. This is a self-maintaining management circle which may lead to risk-based managerial decisions. It handles corporate strategy as well as financial processes and workflow, technology and employees.

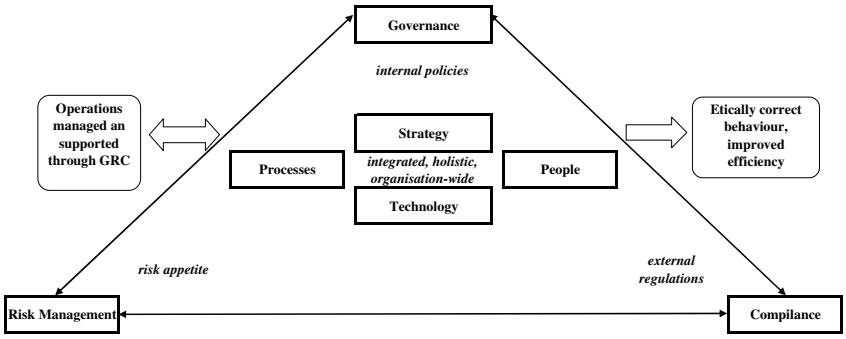


Figure 3
The GRC model [10]

3.4 ISO/IEC 27001

ISO/IEC 2700x is an information security management system or standard package of British origin providing guidance to information security activities [33]. Companies define security requirements and related measures on the basis of business objectives and organisational strategy. Information security (integrity, confidentiality, and availability) is treated with special emphasis. It is not linked to any sort of information technology. The standard (ISO/IEC 27001) [23] divides company operations and the related requirements into 11 protection areas and, within these, 39 targets and 133 protection measures. The information security management system, once it has been implemented and documented, may be accredited by an independent accreditation organisation (ISO/IEC 27002) [24]. In the standard package there appear a few supplementary sections, presented as individual standards (e.g.: ISO/IEC 27005 information security risk management standard with advice on selecting appropriate risk analysis and management tools and methods). Development never stops. There are plans for further standards (e.g.: implementation guide – ISO/IEC 27003; guidance on information security management for sector-to-sector communication – ISO/IEC 27010; information security in telecommunication – ISO/IEC 27013).

3.5 CObIT

The ISACF (Information Systems Audit and Control Foundation, IT Governance Institute, USA) has developed a recommendation entitled "CObIT" (Control Objectives for Information and related Technology) [14].

Practically, this material is a management tool which helps users to understand and handle risks and advantages connected to information and information technology. This internationally approved and developed "framework" was created primarily for business enterprises and is aimed at harmonising information technology services and the operational processes of the organisation as well as facilitating the measurability of the security and management features of information technology services.

CObIT is a collection of documents grouping best practices according to a set of criteria. In order to ensure the necessary information for the fulfillment of organisational (business) aims, information technology resources must be managed within a framework of connected procedures. With its use, we may bridge the gap between business risks, control requirements, and issues of technical nature. The system may be used by the higher management, the users, IT professionals, and the controllers of the information system at the same time. The real aim of COBIT is the achievement and maintenance of information technology security at a minimum risk and maximum profit…

Its structure is as follows:

- Executive Summary
- Framework

Control Objectives (34 processes, or process + management guidelines and maturity model + management guidelines, critical success factors, Key Goal Indicators, to define target levels of performance; and Key Performance Indicators, to measure whether an IT control process is meeting its objective) Supplements (summary overview, case studies, frequently asked questions)

The recommendation defines 34 "management" goals in connection with information technology processes, dividing them into four areas:

1 Plan and Organise,

2 Acquire and Implement,

3 Deliver and Support,

4 Monitor and Evaluate.

There are 215 specific and detailed control objectives throughout the 34 high-level IT processes.

3.6 ITIL

The ISO/IEC 20000-1, -2 standard [25, 26] was created on the basis of and in harmony with the British-developed ITIL (Information Technology Infrastructure Library), dealing with the operation issues of information systems (Fig. 3) [15]. The first part of the document is a set of formal requirements concerning acceptable information technology services, while the second part is a guide to service management and auditing according to the first part. Service management activities are connected to the currently popular PDCA model, which is applied in several standards.

In addition to the management system, the issues of planning and implementation of information technology systems, and the planning and creation of new services, there are five basic areas of complete service management:

- Service security (service level, service reporting, capacity, service continuity, availability, information security, budgeting and accounting for IT services)
- Management processes (configuration and change management)
- Release (documents, operational description distribution management, the documentation of approved modifications)
- Solution processes (incident and problem management)
- Relationship (customer service, business and supplier relationship management)

3.7 BS 25999

The British standard package concerning business continuity (BS 25999-1, -2) [27, 28] also facilitates the creation of a corporate process management system. It is applicable to all types of organisations. The assessment of potential threats and risk factors is the result of a complex impact analysis (Business Impact Analysis, BIA). The key products of the company and the steps in their manufacturing as well as service support processes-, and the maximum acceptable period of business breakdown and dependence on external business partners are examined. Based on the impact analysis the company creates a Business Continuity Plan which helps avoid problems even in case of unexpected events (natural disaster, shortage of raw materials, utility failure, labour force shortage, breakdown of technological equipment, IT problems, customer complaints, etc.). The firm retains its good reputation and is able to carry on value-added processes and maintain connections with business partners.

3.8 Supply Chain Continuity Models and Standards

According to the definition established by the US Supply Chain Council (SCOR model) [8], a supply chain comprises all activities connected to manufacturing and delivery, from the suppliers' suppliers to the final consumers. The five major processes determining the supply chain are

1 planning (supply/demand analysis and the determination of quality, quantity and scheduling factors for products or services),

2 sourcing (raw materials, spare parts and cooperational services),

3 making (the manufacturing of spare parts and assembly),

4 delivery (stockpiling, order management, distribution, and serving the final consumers),

5 returning (handling faulty or superfluous products and maintenance needs, customer service work).

Here we do not see the accumulation of discrete results reached by individual organisations within the supply chain but synergic effects are created in various domains of production due to the allocation of resources. At the same time this is also true for risks. The management of the supply chain means conscious collaboration on behalf of the companies. Its existence is accepted by the participants as a contributing factor to the improvement of their competitive position. The members of the chain are willing to sacrifice their individual, short-term advantages to facilitate the optimal operation of the whole chain. This in turn presupposes protective activities to ensure the safety of "supply" and joint risk management. Standard package ISO 28000 may be a helpful instrument of regulation here because it includes supply chain security management requirements [29, 30, 31].

For supply chains the effective operation of the whole network is more important than the optimal resource utilisation of its individual member enterprises. The widely used and highly practicable CPFR (Collaborative Planning, Forecasting and Replenishment) process model also pushes companies into this direction [13]. The basis of demand planning is the final customer demand. The application of the model results in a consensus-based forecast which will in turn determine the plans for distribution, manufacturing and purchasing, also broken down to individual members. Supply chain members try to use forecast data as accurate as possible. This works to improve supply security.

3.9 Enterprise Risk Management

The Enterprise Risk Management framework first compiled by COSO (Committee of Sponsoring Organisations of Treadway Comission) in 1992 is designed for the use by higher management and decision-makers. It focuses on the internal

processes of the enterprise, their management and control. With constant attention to business strategy, it may be applied to almost any type of risk but is mostly used in the finance area [16].

4 A New, Holistic and Process-centred Enterprise Security Model

The basis of successful enterprise operation is the conscious assumption and management of risks. Business organisations need a type of Enterprise Risk Management (ERM) which

- identifies and handles risk factors;
- encompasses the whole organisation and the surrounding environment;
- allows managers an overview of the entire risk profile;
- aids strategic and operative decision-making.

Thus the protection of enterprise (organisational) processes may be ensured and the security of processes may be reached.

Process security may be regarded as a state: if the prescribed input factors (resources needed for the completion of the process) are ensured, the organisational units involved in the process will produce the required output (product, service, or information) in the prescribed time, quantity and quality; in the case of a disturbance, the normal course of procedure is restored with minimal effort and in the acceptable minimum time [4].

The majority of standards and recommendations discussed so far are process-centred but mostly concern one individual function area within enterprises. The execution of all organisational processes requires resources, the most important of which is information provided in the appropriate time and place and to the authorised persons – information being a fundamental precondition for value-added processes. This is why the introduction of an information security management system may serve as the foundation of a security model for the whole organisation. By the regulation of workflow – or by its restoration in case of a disturbance – the security of the "virtual functioning" of the organisation may be created. A major role in enterprise functioning may be given to the job definition of persons applying a virtual enterprise model (e.g. ERP, EAM). What data may they register, modify, delete; what data queries and transactions may they initialise? Users occupy predetermined positions that incorporate security requirements [5]. The handling of this (role analysis, -design, -management and -maintenance) is professionally termed role life-cycle.

Simultaneous with the creation of information security – as a state – the security management of further "subdomains" may follow (human resources, the environment, production, internal logistics, supply and delivery chains, infrastructure, R&D). The concept of logical physical and organisational security defined there may then be extended to other areas. By logical protection we mean the assurance of data integrity, virus and computer intervention protection and classification methods; under physical protection come plant entry, uninterrupted energy sources, surveillance, and fire and flood protection; organisational or administrative protection means protection against internal fraud or misconduct, and purposeful or accidental damage. The integration of these may substantially lower enterprise security risks.

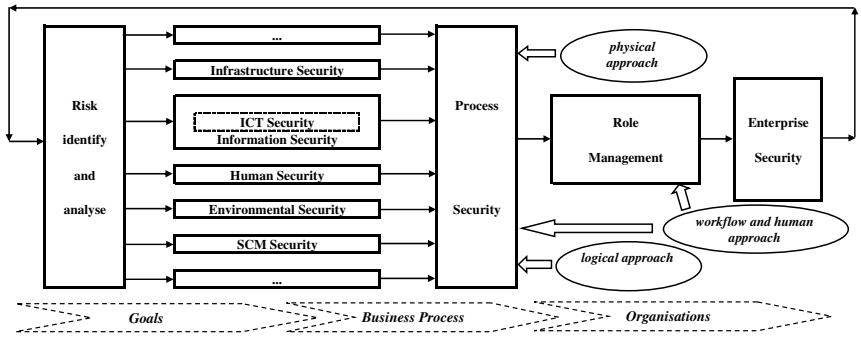


Figure 4
The attainment of enterprise security

The security of processes may in turn usher in total enterprise security (Fig. 4). A holistic security perspective takes into account and prioritises the organisation's strategic targets, the value-added processes and tools as well as supporting information technology. The management of the organisation must then work together to define and reach security targets, and should control and measure the execution of these steps [2].

Conclusions

Enterprise security is a status as well. This however cannot be regarded as static. Only continuous security activity requiring constant development and control based on risk analysis and management can be productive in the business organisations. The preliminary model introduced in this study has been made up in consideration of numerous standards and recommendations. The classical "goals – process – organisation" sequence or control loop can also be predominant here. In addition to and after the security demands of specialised enterprise areas a major role is given to processes and their representations in workflow. The latter demands the definition of jobs and responsibilities. It is we, the human beings, who are the main security risk of the organisations and the organisational security

can only be realized if the work of those participating in the processes is regulated and also if these persons are prepared to manage the unexpectedly happened risk events.

The subjects of business continuity and disaster recovery appear in standards and recommendations primarily dealing with information security. The narrowly interpreted information technology approach (BCP – Business Continuity Plan, DRP – Disaster Recovery Plan) can be extended to guarantee the conditions of any other business processes as well as to execute the tasks related to these processes in accordance with the regulations and finally in the case of malfunction normal operation can be restored.

References

[1] Aagedal, Jan Øyvind – den Braber, Folker – Dimitrakos, Theo – Gran, Bjørn Axel – Raptis, Dimitris – Stølen, Ketil: Model-based Risk Assessment to Improve Enterprise Security. Proceeding of the 6th International Enterprise Distributed Object Computing Conference (EDOC'02) September 17-20, 2002, pp. 51-64, ISBN 0-7695-1742-0, www.itsec.gov.cn (downloaded: 30 September 2011)

[2] Carelli, Richard A. – Allen, Julia H. – Stevens, James F. – Willke, Bradford J. – Wilson, William R.: Managing for Enterprise Security. Networked Systems Survivability Program, Carnegie Mellon University, 2004, p. 55 (CMU/SEI-2004-TN-046)

[3] Chapman, Robert J.: Simple Tools and Techniques of Enterprise Risk Management. John Wiley and Sons, 2006, p. 466, ISBN 13 978-0-470-01466-0

[4] Harrington, James H.: Business Process Improvement (The Breakthrough Strategy for Total Quality, Productivity and Competitiveness). McGraw-Hill, Inc. 1991, ISBN 0-07-026768-5

[5] Kern, Axel – Kuhlmann, Martin – Schaad, Andreas – Moffett, Jonathan: Observations on the Role Life-Cycle in the Context of Enterprise Security Management. SACMAT'02 Proceedings of the 7th ACM Symposium on Acces Control Models and Technologies, Monterey, CA, USA, June 3-4, 2002, pp. 43-51, ISBN 1-58113-496-7

[6] Christopher, Martin – Peck, Helen: Building the Resilient Supply Chain. International Journal of Logistics Management, Vol. 15, No. 2, 2004, pp. 1-13

[7] Smith, Gregory, E. – Watson, Kevin J. – Baker, Wade H. – Pokorski, Jay A.: A Critical Balance: Collaboration and Security in the IT-enabled Supply Chain. International Journal of Production Research. Vol. 45, No. 11, June 2007, pp. 2595-2613

[8] Supply Chain Council, Supply-Chain Operations Reference-model (SCOR). Overview. Version 10.0, 2010, http://supply-chain.org/f/Web-Scor-Overview.pdf (downloaded: 12 September 2011)

[9] Ji-Yeu Park – Rosslin John Robles - Chang-Hwa Hong – Sang-Soo Yeo – Tai-hoon Kim: IT Security Strategies for SME's. International Journal of Software Engineering and its Applications, Vol. 2, No. 3, July 2008, pp. 91-98

[10] Racz, Nicolas - Weippl, Edgar - Seufert, Andreas: A Frame of Reference for Research of Integrated Governance, Risk & Compliance (GRC). In: Bart De Decker, Ingrid Schaumüller-Bichl (Eds.), Communications and Multimedia Security, 11th IFIP TC 6/TC 11 International Conference, CMS 2010 Proceedings. Berlin: Springer, pp. 106-117

[11] Szádeczky, Tamás: Problems of Digital Sustainability. Acta Polytechnica Hungarica, Vol. 7, No. 3, 2010, pp. 123-136

[12] Wilder, Dan: The New Business Continuity Model. White paper, 2008, p. 58. www.talkingbusinesscontinuity.com/downloads/pdf/The-New-Business-Continuity-Model. (downloaded: 18 November 2011)

[13] Collaborative Planning, Forecasting and Replenishment (CPFR). Overview, 2004, Voluntary Interindustry Commerce Standards (VICS) www.vics.org (downloaded: 5 March 2012)

[14] CObIT version 4.1 Excerpt, Executive Summary, Framework, IT Governance Institute, USA, 2007 www.isaca.org/Knowledge-Center/cobit/Documents/COBIT4.pdf (downloaded: 5 March 2012)

[15] An Introductory Overview of ITIL V3. IT Service Management Forum, 2007 www.itsmfi.org (downloaded: 6 October 2011)

[16] Enterprise Risk Management - Integrated Framework Executive Summary. Committee of Sponsoring Organizations of the Treadway Commission. September, 2004 (www.coso.org/documents/COSO_ERM_ExecutiveSummary.pdf - downloaded: 6 October 2011)

[17] www.iso27001certificates.com (downloaded continuously)

[18] ISO 14001:2004 Environmental management systems – Requirements with guidance for use

[19] ISO 14004:2004 Environmental management systems – General guidelines on principles, systems and support techniques

[20] BS OHSAS 18001:2007 Occupational health and safety management systems. Requirements

[21] BS OHSAS 18002:2008 Occupational health and safety management systems. Guidelines for the implementation of OHSAS 18001:2007

[22] ISO/IEC 38500:2008 Corporate governance of information technology

[23] ISO/IEC 27001:2005 Information technology – Security techniques – Information security management systems – Requirements

[24] ISO/IEC 27002:2005 Information technology – Security techniques – Code of practice for information security management

[25] ISO/IEC 20000-1:2011 Information technology – Service management – Part 1: Service management system requirements

[26] ISO/IEC 20000-2:2012 Information technology – Service management – Part 2: Guidance on the application of service management systems

[27] BS 25999-1:2006 Business Continuity Management, Code of Practice

[28] BS 25999-2:2007 Business Continuity Management, Specification

[29] ISO 28000:2007 Specification for security management systems for the supply chain

[30] ISO 28001:2007 Security management systems for the supply chain - Best practices for implementing supply chain security, assessments and plans - Requirements and guidance

[31] ISO 28002:2011 Security management systems for the supply chain - Development of resilience in the supply chain - Requirements with guidance for use

[32] AS/NZS 4360:2004 Risk management

[33] www.iso27001security.com (downloaded continuously)

Verification of Communication Protocols Based on Formal Methods Integration

Slavomír Šimoňák

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice
Letná 9, 042 00 Košice, Slovakia
e-mail: slavomir.simonak@tuke.sk

Abstract: Communication protocols define the set of rules needed to exchange messages between communicating entities. Networked and distributed systems, built around communicating protocols, are widely used nowadays. Since such systems are often deployed in safety-critical applications, confidence in protocol correctness is highly required. We propose an approach based on formal method integration to support the modeling and analysis of communication protocols. Process algebra and Petri nets are used together to combine the best properties from both methods – the exceptional properties for system description offered by process algebra, and the powerful analytical properties of Petri nets. The ideas described within the paper are demonstrated by an example – the Trivial FTP (TFTP) protocol.

Keywords: protocol correctness; formal methods integration; Petri nets; process algebra

1 Introduction and Motivation

A protocol is a set of rules which must be followed in the course of some activity. Originally, the term was used in connection to human (more or less formal) activities. If the protocol is not followed, the activity will not be successful. Nowadays the term is increasingly used when the communication between computers, computer components or computer systems is considered [19]. Within the paper we will focus on this type of *communication protocols*. Communication protocols thus define the set of rules needed to exchange messages between two or more communicating entities [4, 5, 17]. Such protocols are elements of great importance when networked and distributed systems are considered [25]. Nowadays such systems are very common, and incorrect communication, or no communication at all, can cause complications, the severity of which can range from financial loss (internet banking, e-shopping) to issues like human health and lives (medical or transport systems).

2 Protocol Correctness

The mentioned risks are motivating factors for the development and use of protocols to ensure a correct information exchange between communicating entities. But what is the protocol *correctness,* and how can it be shown that the particular protocol is correct? We can look at the problem from different points of view. In the event that we want to check if the system complies with the requirements and performs the functions for which it is intended, we are talking about *validation*. A process used to determine if the system is consistent, adheres to standards, uses reliable techniques, performs the selected functions in the correct manner is referred to as *verification* [26, 10, 15]. Confidence in protocol correctness can be increased in different ways. *Testing* is one method and involves building a prototype and observing it, or observing the behavior of a real system. The main disadvantage of testing is that it can be used to show errors, but not to prove correctness. *Simulation* is a method based on the construction of an executable model of the system and its observation. Simulation requires the generation of test cases, which are difficult to design for complex systems. In the case of critical properties, simulation is not believed to provide sufficient confidence [2]. Methods such as simulation and testing are usually used with success to show the performance characteristics of the system considered. *Formal methods,* on the other hand, are mathematically-based techniques offering a framework in which systems can be specified, developed and analyzed in systematic manner. Generally, they are not suitable for assessing a system's performance, but there exist methods specifically designed for this purpose (e.g. Performance Evaluation Process Algebra – PEPA) [9]. Since performance analysis is not our goal here, we will focus our attention on employing formal methods in protocol correctness analysis further in this paper.

3 Related Works

Many attempts have been made to perform protocol analysis based on formal methods [7, 27, 2, 6, 13]. Process algebras, Petri nets and other methods have been employed for the specification and analysis of these systems. Process algebraic constructs, like basic operators for constructing finite processes (alternative and sequential composition), communication, encapsulation, abstraction and other operators, form a solid basis for specification and analysis of wide range of systems. In particular, they are suitable for the specification of communication protocols. The usual method for performing the system analysis in this case is the following: Firstly, the desired external behavior of the protocol is specified in the form of a process term (usually using basic operators and recursion). Next, the implementation of the protocol is specified in the form of a process term (usually

including basic operators, parallel operators and recursion). Then, internal actions are forced to communicate using the encapsulation operation and the internal communication actions are made invisible using an abstraction operator, so effectively only the input/output relation of implementation is visible. Finally, using the axiom system, we try to show the terms are equal. By this equality we prove the desired external behavior and the input/output relation of the implementation are (rooted branching) bisimilar [7]. The above mentioned process enables us to show that the system meets properties included in the term validation.

4 The Method

Instead of the verification process just described (or as an addition to it), we propose to use a method based on a combination of two formal methods: process algebra and Petri nets [22]. Our aim here is to use the best properties from both worlds – the exceptional properties for system description offered by process algebra, and the powerful analytical properties of Petri nets. We believe it is easier to create a description of a communication protocol using the constructs of process algebra. The main reason for our belief here is that de/composition and communication can be expressed more naturally using special algebraic constructs than using Petri nets. The larger and more complex the modeled system, the greater is the impact of this advantage. In the case of communication protocols, the system usually consists of communicating entities, medium and maybe other parts, which can be specified separately and put together by means of the parallel composition operator. On the other hand, we believe the analysis is usually better/easier done using Petri nets. Petri nets are a well-known formal method, mainly due to their valuable analytical properties and intuitive graphical representation [27, 3]. Two types of properties can be investigated using the Petri net models: properties which depend on the initial marking (behavioral properties) and those which are independent of it (structural properties). Problems connected with analysis of behavioral properties include reachability, boundedness, liveness, reversibility and home state, coverability and other problems. Structural properties, depending on the topological structures of Petri nets, on the other hand, hold for any initial marking or are concerned with the existence of certain firing sequences from some initial marking. Properties of this kind include structural boundedness, conservativeness, repetitiveness and consistency [14, 11]. Invariants of the system can be derived from the structure of the net, so the construction and analysis of the state space (which can be of great size) is not necessary here. Invariant-based analysis alone is a powerful tool for studying the structural properties of Petri nets.

Many different dialects of Petri nets are available today and differ by the properties such as *modeling power* and *decision power*. Modeling and decision power are in some respects antagonistic properties; by increasing modeling power, decision power usually decreases. In our case, ordinary Petri nets are considered, which represents a good balance between modeling and decision power.

The key element of the method proposed here is the automatic, semantic-preserving transformation of process-algebraic specification into the Petri net-based one. After the transformation is performed, the powerful analytical properties of Petri nets can be used. By the analysis we can disclose defects in the internal consistency and correctness of the specification, which can potentially be hidden within the specification.

5 The TFTP Protocol

The transformation method developed by the author, described in deeper detail, can be found in [20, 21]. Transformation is automated by tools like ACP2Petri and PATool [23], both developed at the author's home institution. ACP2Petri is the tool performing the transformation itself. It accepts a process algebraic, ACP-based [1] specification in the PAML language as the input and produces the corresponding Petri net in the standard PNML format, which is supported by various analytical tools, such as TINA, Netlab and PNtool2 [24, 16, 12].

The TFTP (Trivial File Transfer Protocol) [8, 18] was chosen as an example to demonstrate the ideas presented above. TFTP is a simple protocol to move files between machines. It is designed to be small and easy to implement, so it lacks most of the features of a regular FTP. The protocol only supports reading and writing files from/to a remote server. It cannot list directories and currently has no support for a user authentication. TFTP is a protocol with strict data transfer restrictions. When an error occurs, the current transfer is stopped and connection is terminated, so it is necessary to establish the connection and start the transfer again. Communication between the server and the client will be described at the level of packets exchange. According to the type of data within a packet, packets can be subdivided into types summarized in Table 1.

Table 1
TFTP packet types

Packet type	**Description**
RRQ	Read request
WRQ	Write request
D1	Data 1 (first packet of a file)
DL	Data L (last packet of a file)

DN	Data N (N-th packet of a file)
ACK0	Answer after the request for writing to the server was received
ACK1	Answer after receiving the first packet of the file
ACKN	Answer after receiving the N-th packet of the file
ERR	Error

All the packets of the communication serve one of the following purposes:

- To transfer the (parts of the) file, i.e. data packets (D1, DL, DN).
- To control the transfer, i.e. control packets (ACK0, ACK1, ACKN, ERR, RRQ, WRQ).

Basic TFTP functionality includes reading a file from the server and writing a file to the server, respectively. Let us describe those activities in deeper detail.

5.1 Reading a File from the Server

The operation is initiated by sending the RRQ packet by the client. The server can respond to this request in three ways:

- By sending ERR, if the file requested does not exist, or it is unable to read the file.
- By sending D1 – a positive answer to the request and a first packet of the file.
- By sending DL – a positive answer too, but also a signal, that this is the only packet of the file.

When a client receives the ERR packet, the reading of the file is terminated and it is able to send its new request. The client replies to D1 and DL (received as the first after RRQ packet) by sending the ACK1 packet. If server receives the ACK1 packet after sending DL, it terminates the connection and is ready for the next request. In the case that the server receives ACK1 after sending D1, it responds with the next part of the file in form of a DN or DL packet, where the latter of the two is used, when the last part of the file is to be sent. The client replies to the received DN by the ACKN packet, where N is the packet number, and to packet DL (when it is not only packet of the file), replies also with DN.

5.2 Writing a File to the Server

The writing operation is very similar to the reading one, with one significant exception: after receiving the writing request in form of the WRQ packet, the server replies by ACK0, which is a packet type reserved especially for this purpose only. The rest of communication runs analogically to the reading operation explained above.

The protocol operation from the client's point of view is depicted in Figure 1.

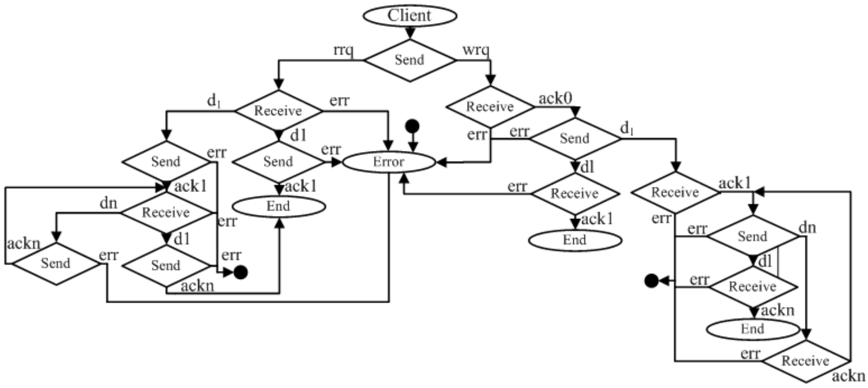


Figure 1
Protocol operation, client point of view

5.3 Formal Specification

Formal specification of the TFTP protocol is based on the analysis of its operation given above. The rules for naming the actions used within the specification are as follows. The first symbol of the action name is one of the three following: c (client), s (server) and m (medium); the second symbol is '_' (underscore); and the third one gives a type of communication (s-send, r-receive). The next symbols represent an abbreviation of the message (packet) type. Some packet types are used in both directions of communication, so the direction in these cases is expressed by the suffix w (write to server) or r (read from server). Packets of ERR type are distinguished similarly according to the direction by adding a suffix s (from the server) or c (from the client), respectively. The two operations supported by the TFTP protocol (read and write) are mutually independent, so it is possible to perform decomposition and specify (and analyze) each of them separately. The reading operation (TFTPR) specification only is given within the rest of this paper. In the case of interest in a whole TFTP specification, please refer to [8]. The following TFTPR specification is given in textual form [23] of process algebra ACP [1].

```
*Communication
gamma (c_srrq,m_srrq) = srrq         gamma (s_sdnr,m_sdnr) = sdnr
gamma (m_rrrq,s_rrrq) = rrrq         gamma (m_rdnr,c_rdnr) = rdnr
gamma (c_sack1r,m_sack1r) = sack1r   gamma (s_sdlr,m_sdlr) = sdlr
gamma (m_rack1r,s_rack1r) = rack1r   gamma (m_rdlr,c_rdlr) = rdlr
gamma (c_sacknr,m_sacknr) = sacknr   gamma (s_serrs,m_serrs) = serrs
gamma (m_racknr,s_racknr) = racknr   gamma (m_rerrs,c_rerrs) = rerrs
gamma (s_sd1r,m_sd1r) = sd1r         gamma (c_serrc,m_serrc) = serrc
gamma (m_rd1r,c_rd1r) = rd1r         gamma (m_rerrc,s_rerrc) = rerrc
```

We started with the definition of communication between actions, where the ACP-style binary communication function gamma gives the action that is the result of communication. Actions not essential from our point of view are hidden (encapsulated) further in order to concentrate on the protocol operation.

```
*Encapsulation
encset[H](c_srrq,m_srrq,m_rrrq,s_rrrq,c_sack1r,m_sack1r,m_rack1r,s_ra
ck1r,c_sacknr,m_sacknr,m_racknr,s_racknr,s_sd1r,m_sd1r,m_rd1r,c_rd1r,
s_sdnr,m_sdnr,m_rdnr,c_rdnr,s_sdlr,m_sdlr,m_rdlr,c_rdlr,s_serrs,m_ser
rs,m_rerrs,c_rerrs,c_serrc,m_serrc,m_rerrc,s_rerrc)
```

The specifications of client, server and the messages transferring medium are given. The recursive specifications used here reflect their repeated activity.

```
*Client
CRN=c_rdnr.(c_sacknr.CRN+c_serrc)+c_rdlr.(c_sacknr+c_serrc)+c_rerrs.C
CR1=c_rd1r.(c_sack1r.CRN+c_serrc)+c_rdlr.(c_sack1r+c_serrc)+c_rerrs.C
C = (c_srrq.CR1).C

*Server
SRN=s_sdnr.(s_racknr.SRN+s_rerrc)+s_sdlr.(s_racknr+s_rerrc)+s_serrs.S
SR1=s_sd1r.(s_rack1r.SRN+s_rerrc)+s_sdlr.(s_rack1r+s_rerrc)+s_serrs.S
S = (s_rrrq.SR1).S

RRQ = m_srrq.m_rrrq.RRQ
ACK1 = (m_sack1r.m_rack1r).ACK1
ACKN = (m_sacknr.m_racknr).ACKN
D1 = (m_sd1r.m_rd1r).D1
DN = (m_sdnr.m_rdnr).DN
DL = (m_sdlr.m_rdlr).DL
ERRS = (m_serrs.m_rerrs).ERRS
ERRC = (m_serrc.m_rerrc).ERRC

*Composition
TFTP = encaps[H](C||S||RRQ||ACK1||ACKN||D1||DL||DN||ERRS||ERRC)
```

At the end of the specification, all the components are put together by means of the parallel composition operation. The encapsulation set (H) is applied to the whole composition. The specification given above is translated into a machine readable, XML-based PAML format using the PATool [23]. The resulting algebraic specification is transformed subsequently into the corresponding Petri net (Figure 2) and stored in PNML format. The resulting Petri net generated by the ACP2Petri tool has a simple (matrix) layout and was edited manually to the form depicted in Figure 2 using the TINA Toolbox [24].

5.4 TFTPR Analysis

Petri net analysis can be used for the investigation of several properties of modeled systems, as was mentioned in Section 4 of the paper. The methods to analyze Petri net models may be subdivided into the following three groups: the coverability (reachability) tree method, the matrix-equation approach and the reduction/decomposition techniques [4, 11, 14].

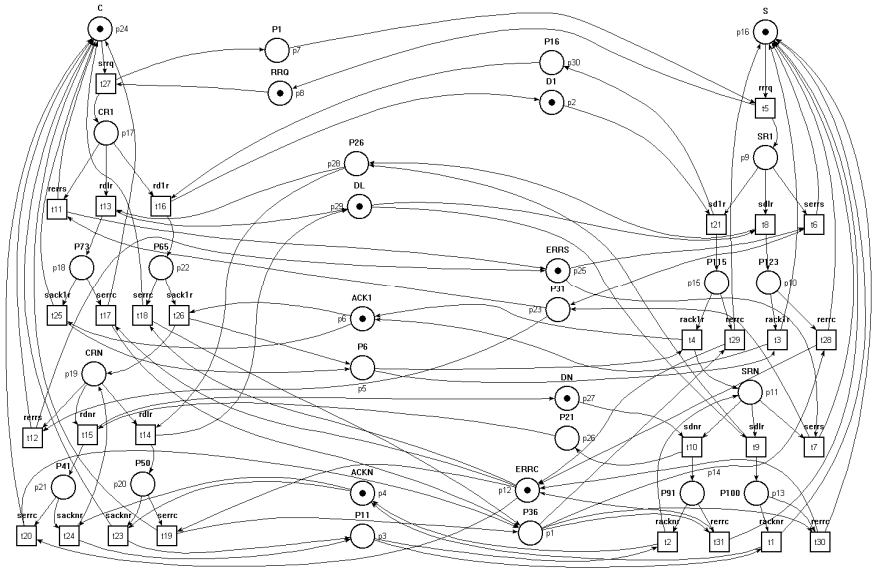


Figure 2
Petri net of TFTPR protocol

The execution of operations or occurrence of events within the modeled system is simulated by the firing of Petri net transitions. State changes of the system thus are reflected by the changes in distribution of tokens (marking) in places of Petri net. By analysis of these changes, one can study the dynamic behavior of the modeled system. For instance, when the sequence of messages (*srrq* – send read request, *rrrq* – receive read request, *sd1r* – send first packet of data read from the server, *rd1r* – receive first packet of data) is exchanged between a client and a server, starting from the initial state of our system (Figure 2), the client can respond by sending one of two messages – error (*serrc*) or acknowledgement (*sack1r*) respectively, depending on the successfulness of receiving the first packed of requested data file. Within the corresponding Petri net, the situation is modeled by two enabled transitions (*serrc* and *sack1r*), as depicted in Figure 3.

Invariants alone have numerous applications and form the basis for many necessary/sufficient conditions of Petri net model properties. There are a variety of tools available today which help with invariants calculation. In our case, the Netlab [16] tool was used and invariants of places and transitions calculated are summarized in Table 2 and Table 3, respectively.

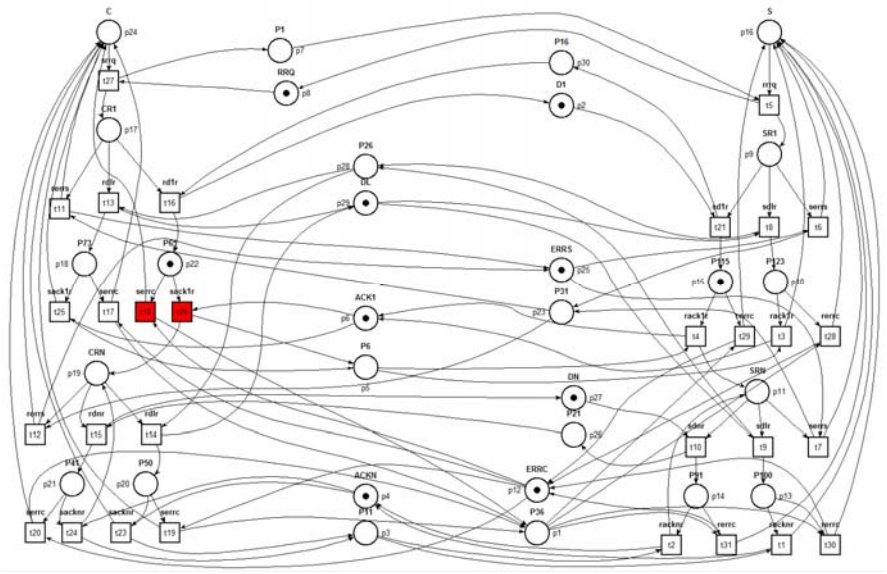


Figure 3
Studying the dynamic behavior using the Tina [24] stepper simulator

Table 2
Invariants of places

p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}	p_{18}	p_{19}	p_{20}	p_{21}	p_{22}	p_{23}	p_{24}	p_{25}	p_{26}	p_{27}	p_{28}	p_{29}	p_{30}
P36	*D1*	*P11*	*ACKN*	*P6*	*ACK1*	*P1*	*RRQ*	*SR1*	*P123*	*SRN*	*ERRC*	*P100*	*P91*	*P115*	*S*	*CR1*	*P73*	*CRN*	*P50*	*P41*	*P65*	*P31*	*C*	*ERRS*	*P21*	*DN*	*P26*	*DL*	*P16*
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Moreover, Netlab provides the summary of net analysis results in textual form based on invariants and a reachability graph. Some of them are listed below in an abbreviated form.

- `Dead transitions (RG): none.`
- `Total deadlock (RG): none.`
- `Reversibility (RG, condensed): The net is reversible.`
- `Necessary conditions for invariants: There exists a non-negative T-invariant. Therefore, the necessary condition for reversibility is satisfied, and the net may be reversible.`
- `Partial deadlocks exist in the following sinks (RG, condensed): none.`

- `Liveness (RG, condensed): The net is live.`
- `Necessary conditions for invariants: There exists a positive T-invariant. Therefore, the necessary condition for liveness is satisfied, and the net may be live.`
- `Boundedness (RG): The net is bounded.`
- `Sufficient conditions for invariants: There exists a positive P-invariant. Therefore, the sufficient condition for boundedness is satisfied, and the net is bounded.`

PNtool2, as an addition to the invariant analysis, provides also the reachability analysis based on results of the research performed at the author's home institution [11, 12].

Table 3

Invariants of transitions

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	t_{21}	t_{22}	t_{23}	t_{24}	t_{25}	t_{26}	t_{27}	t_{28}	t_{29}	t_{30}
racknr	*sdnr*	*rerrs*	*rerrs*	*rdlr*	*rdlr*	*rdnr*	*rd1r*	*serrc*	*serrc*	*serrc*	*racknr*	*serrc*	*sd1r*	*sacknr*	*sacknr*	*sack1r*	*sack1r*	*srrq*	*rerrc*	*rerrc*	*rack1r*	*rerrc*	*rerrc*	*rack1r*	*rrrq*	*serrs*	*serrs*	*sdlr*	*sdlr*
0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0

Conclusions

Within this work a method for the specification and verification of communication protocols is discussed. It is based on combining process algebra and Petri nets in order to simplify both the production of protocol specification (using process algebra) and protocol analysis (using Petri nets). The Trivial FTP protocol is taken as an example of practical employment of the method proposed. In this case, the resulting Petri net has the properties (boundedness, liveness, deadlock freeness) which a correct communication protocol should have.

Our future activities will include several items to be explored. Firstly, a lot of time was spent editing the generated Petri net model into the shape depicted in Figure 2. We have made some attempts in the field of generating the layout of Petri net models, but there is still a lot of space to improve. Another idea is to incorporate the notion of time into the process of transformation. This would lead to an update of the transformation method (and the ACP2Petri tool implementing it) with the support of corresponding timed process algebra and Petri net formalisms. And finally, we plan to specify some other protocols to recognize the potential strengths and limits of the method proposed above and compare it more deeply with other approaches.

Acknowledgement

This work is the result of the project implementation: Development of the Center of Information and Communication Technologies for Knowledge Systems (ITMS project code: 26220120030) supported by the Research & Development Operational Program funded by the ERDF.

References

[1] Baeten, J. C. M., Weijland, W. P.: Process Algebra, Cambridge University Press, 1990

[2] Barjaktarovic, M., Shiu-Kai, C., Jabbour, K.: Formal Specification and Verification of Communication Protocols Using Automated Tools, Proceedings of ICECCS'95, pp. 246-253, 1995

[3] Češka, M., Marek, V., Novosad, P., Vojnar, T.: Petri nets, BUT, 2009

[4] Diaz, M.: Petri Nets: Fundamental Models, Verification and Applications, John Wiley and Sons, 2009

[5] East, I.: Computer Architecture and Organization, Pitman Publishing, 1990

[6] Edwards, J.: Process Algebras for Protocol Validation and Analysis, Proceedings of PREP 2001, pp. 1-20, Keele, England, 2001

[7] Fokkink, W.: Introduction to Process Algebra, Springer-Verlag, 2007

[8] Fürdős, F.: Verification of Communication Protocols, diploma thesis, Technical university of Košice, 2009

[9] Hillston, J.: Process Algebras for Quantitative Analysis, Proceedings of the 20th Annual IEEE Symposium on Logic in Computer Science (LICS' 05), pp. 239-248, Chicago, 2005

[10] Holzmann, G. J.: Design and Validation of Computer Protocols, Prentice-Hall, 1991

[11] Hudák, Š.: Reachability Analysis of Systems Based on Petri Nets, Elfa, Košice, 1999

[12] Hudák, Š., Zaitsev, D. A., Korečko, Š., Šimoňák, S.: mfdte/pntool – a Tool for the Rigorous Design, Analysis and Development of Concurrent and Time-critical Systems, Acta Electrotechnica et Informatica, Vol. 7, No. 4, 2007

[13] Lanet, J. L.: Using the B Method to Model Protocols, AFADL98 (LISI/ENSMA), pp. 79-90

[14] Murata, T.: Petri-Nets: Properties, Analysis and Applications, Proceedings of the IEEE, 77(4), 1989

[15] Oberkampf, W. L., Trucano, T. G., Hirsch, C.: Verification, Validation, and Predictive Capability in Computational Engineering and Physics,

Foundations for Verification and Validation in the 21st Century Workshop, Hopkins University, Maryland, 2002

[16] Petri net tool Netlab, available at: http://www.irt.rwth-aachen.de/en/fuer-studierende/downloads/petri-net-tool-netlab

[17] Sharp, R.: Principles of Protocol Design, Springer-Verlag, 2008

[18] Sollins, K.: The TFTP Protocol, 1992, available at: http://tools.ietf.org/html/rfc1350

[19] Szádeczky, T.: Problems of Digital Sustainability, Acta Polytechnica Hungarica, Vol. 7, No. 3, 2010

[20] Šimoňák, S.: Formal Methods Integration Based on Petri nets and Process algebra Transformations, PhD thesis, Technical University of Košice, 2003

[21] Šimoňák, S., Hudák, Š., Korečko, Š.: ACP2Petri: a Tool for FDT Integration Support, Proceedings of EMES'05, pp. 122-127, 2005

[22] Šimoňák, S., Hudák, Š., Korečko, Š.: Protocol Specification and Verification Using Process Algebra and Petri Nets, Proceedings of CSSim 2009, pp. 110-114

[23] Šimoňák, S., Peťko, I.: PATool – A Tool for Design and Analysis of Discrete Systems Using Process Algebras with FDT Integration Support, Acta Electrotechnica et Informatica, Vol. 10, No. 1, 2010, pp. 59-67

[24] TINA (Time Petri Net Analyzer) home, available at: http://homepages.laas.fr/bernard/tina/description.php

[25] Tomášek, M.: Language for a Distributed System of Mobile Agents, Acta Polytechnica Hungarica, Vol. 8, No. 2, 2011

[26] Verification and Validation, available at: http://en.wikipedia.org/wiki/Verification_and_validation

[27] Zaitsev, D. A., Zaitsev, I. D.: Verification of Ethernet Protocols via Parametric Composition of Petri Net, 12th IFAC Symposium on Information Control Problems in Manufacturing, pp. 122-127, 2006

Product Configurator Self-Adapting to Different Levels of Customer Knowledge

Igor Fürstner, Zoran Anišić

Subotica Tech – College of Applied Studies
Marka Oreškovića 16, Subotica, Serbia
ifurst@vts.su.ac.rs; azoran@vts.su.ac.rs

Márta Takács

Óbuda University
Bécsi út 96/B, H-1034 Budapest, Hungary
takacs.marta@nik.uni-obuda.hu

Abstract: When selling a customized product with the support of a product configurator, there is a risk that customers will abort the configuration process if the configuration dialogue does not suit the customer well. To reduce this risk, a product configurator that self-adapts to different levels of customer knowledge instead of vice versa is needed. In this paper, a self-adapting approach for product configurators is proposed, one which relies on a fuzzy logic-based algorithm. The approach is implemented for the configuration of the thermal insulation of buildings. The product configurator is tested by users with different capabilities and a comparison of results with professionally performed calculations is performed. It is shown that the proposed approach allows inexperienced customers, too, to make appropriate decisions about thermal insulation. This is an advancement that can considerably expand the scope of the application of product configurators.

Keywords: mass customization; customer profiling; decision support systems; thermal insulation; fuzzy logic

1 Introduction

Nowadays, customers expect to get exactly what they would like. Therefore products should be customized, i.e. mass customized, which results in a drastic increase in the product variety offered by enterprises [1], [2]. This changes the role of the customer from being the consumer of a product to being an active partner in a process of adding value. Active customer participation is crucial for

the successful incorporation of customer needs into the product, which directly influences the final product offering [3], [4], [5].

Recent developments in product configuration systems support the process of customized product development and production [6], [7], [7], [9]. Product configurators involve the customer into the configuration process. This raises several issues that need to be addressed: one of these is that despite the fact that nowadays customers are knowledgeable in general, they are still far from being experts that can really co-create a product or a service [10]. Customers usually only want the product alternatives that exactly meet their requirements; if too much of a choice is offered or an offer is too complex, customers can feel frustrated or confused, and therefore incapable of making proper decisions. This can lead to the abortion of the configuration process [1], [11]. This overload of information is caused by the limited information processing capacity of humans, the lack of customer knowledge about the product, and customer ignorance about his or her real individual needs [12].

In order to reduce the probability of the customer aborting the configuration process, some authors suggest the use of different types of configuration dialogues that are adapted to different types of customers [1]. Unfortunately, these authors do not provide any indication whether the assignment of the dialogue type to a customer is chosen by the customer or defined by the system.

The present paper contributes to fill this gap by developing a configuration approach where the configurator self-adapts to the customer's expertise in a given area and provides a successful product configuration, while maintaining the necessary level of detail and results accuracy. The approach is tested on a real product configurator for the outer thermal insulation of individual residential buildings (not a block of flats) which is an area of importance involving private building owners without expertise, installers with great experience but limited theoretical knowledge, and experts with a great amount of experience and knowledge.

2 Theoretical Background

The theoretical basis of the present work, on the one hand, is the design of configuration dialogues and on the other hand includes the principles of thermal insulation. Even though the aim is to contribute to the design of adapted configuration dialogues, one needs to be familiar with the theoretical bases of thermal insulation to appreciate the proposed application. This blend of different fields is a common characteristic of applied research on product configuration. For the sake of clarity, the two fields involved will be discussed separately.

2.1 Product Configuration and Choice Complexity Experienced by Customer

The identification and implementation of customer requirements are significant issues for successful product development [13] as well as for product configuration [1]. To be able to select or filter objects for an individual, information is required about the individual. Such information is also necessary to be able to decide which options to present and how to present them. For example, performance-oriented language can be used for non-expert customers while components-oriented language can be used for highly interested and experienced customers [1], [14]. The problem of adapting the configuration process to different customers can be addressed with the identification of different customer profiles that group individual customers based on some crucial characteristics such as the knowledge about the product. Based on the defined customer profiles, different descriptions of the same product can and should be provided.

Customer profiles consist of general, specific and contextual information about customers. General information usually deals with basic and demographic attributes, information about general interests, information about relationships to other customers, information about purchase history and usage/interaction behaviour and ratings of products, product components and certain attributes [15], [16]. Specific information refers to the specific requirements of the customer [17], while contextual information about customers is information such as the time of the day, the date, etc. [18].

In order to define an exact customer profile for each individual customer, a proper customer profile definition model has to be developed. Today IT technology makes it possible to collect information from customers implicitly or explicitly. Customer profiles can be obtained from the customers' purchase history [19], [20]. They can also be created by specifying information explicitly at the beginning of the configuration process. Also, customer behaviour during the configuration process is used to profile the customer [21], [22]. Besides this, information on behaviour of other customers that have similarities to the given customer is also used to define the customer [23]. Even though many customer profile models have been suggested in the literature, there is no specific discussion about customer profile definition in the field of outer thermal insulation of buildings.

2.2 The Configuration of Outer Thermal Insulation

In recent years, we have witnessed the fact that energy resource prices have risen considerably. In addition, environmental issues have become more relevant than ever before [24]. The Directive 2002/91/EC and 2010/31/EU of the European Parliament and Council on the energy performance of buildings, which can be

universally taken into account as recommendations beyond the borders of the European Union, state among other things that since buildings account for 40% of total energy consumption in the Union, and since the sector is expanding, energy consumption is growing. Therefore, reduction of energy consumption in the construction sector constitutes important measures needed to reduce energy dependency and greenhouse gas emissions.

Therefore, the outer thermal insulation of buildings is becoming more and more important. The most favourable insulation must be calculated based on a particular lifetime for the building [25], [26]. If the structure has been correctly designed, there will be nothing to affect the insulation when it is in place. Insulation makes good economic sense as it reduces energy consumption in buildings. Insulation as a single investment pays for itself many times over during the life cycle of a building [27], [28]. A high insulation standard for floors, walls, roof and windows does not only mean lower energy consumption. It will also reduce the power need and makes the heating period shorter [29]. It improves the conservation of existing free energy and creates conditions for simpler heating systems [30].

Research in the field of outer thermal insulation defines several rules that have to be considered when one wants to make the necessary calculations regarding thermal, ventilation, solar and other losses or gains. Those rules require knowledge about different parameters, such as the overall position of the building, the building's characteristics (structure, measures, materials, etc.), and the conditions regarding the surroundings (weather data, etc.) in order to be able to calculate the required level of thermal insulation [31]. On a professional level, i.e. if one wants to have highly accurate results, these calculations include all required information in detail and will not be discussed in this paper, because several professional software packages exist on the market that deal with the problem, such as *Bausoft Winwatt, Resfen, CASAnova,* etc.

Nowadays, a number of different product configurators are used to configure various types of products, yet in the field of outer thermal insulation of buildings, there is not any product configurator that can satisfy the requirements of different users. Existing literature does not discuss any calculation methods that can provide acceptable results if just a fraction of required information is available, which is the case of customers who are not professionals in the field of thermal insulation.

3 Objective and Method

3.1 Objective

The paper aims at contributing to fill a gap in literature by introducing the necessary elements with which a successful product configurator in the field of thermal insulation of individual residential buildings can be created and implemented. Thus, a product configuration approach is proposed that self-adapts to the individual capabilities of the given customer, where these may range from having no knowledge about thermal insulation at all to being professionals in the related field.

3.2 Method

In order to propose and test the self-adapting configuration approach for thermal insulation, the authors have integrated their knowledge gained from literature on customer profile definition and thermal insulation with an analysis of various product configurators on the web, and with the specific experience of a company that produces thermal insulation materials. The authors' years of personal field experience on the topic have been another important element in making feasible decisions.

The development process has involved several cycles of refinement in the algorithm and the entire product configurator. At each cycle, first the algorithm for customer profiling was discussed. Subsequently, the quantity and the required level of details regarding input parameters, constraints and calculations to obtain the desired insulation results were specified. Each cycle was then concluded with a number of trials by asking some volunteers to act as test persons. Feedback was always required from the test subjects until finally no significant new insights were obtained.

Once the profiling algorithm and the configuration procedure were established, a test was performed in order to evaluate the results. The test was performed by two groups of potential users. For one group, it was assumed that they had some technical knowledge in the related field, while for the other groups this was not the case. The first group was comprised of 27 university students of mechanical engineering who either had to work with insulation issues at their home, during summer work or due to specific examinations at the university. The second group was made up of 26 university students who up to that point had not had any kind of experience in this field, but were expected to attain some in the future. The results of the configurations were compared to exact calculations performed by the authors.

4 Customer Profile Definition

The need for various customer profiles is based on the experience of using a previous version of the developed configurator for outer thermal insulation of buildings [32]. The configurator was meant to be used by customers with technical knowledge ranging from the almost non-existent to professional but without the differentiation in the configuration dialogue. The results of its use showed that most of the problems arose because some of the previous non-professional customers had found the product configurator too complex to use and did not finish the configuration process. On the other hand, some of the professional customers found that the configurator lacked the possibility of defining exact and precise input parameters. Other problems included the lack of variance in the degree of result accuracy and the duration of the configuration process. This need led to the introduction of a product configuration procedure shown in Fig. 1.

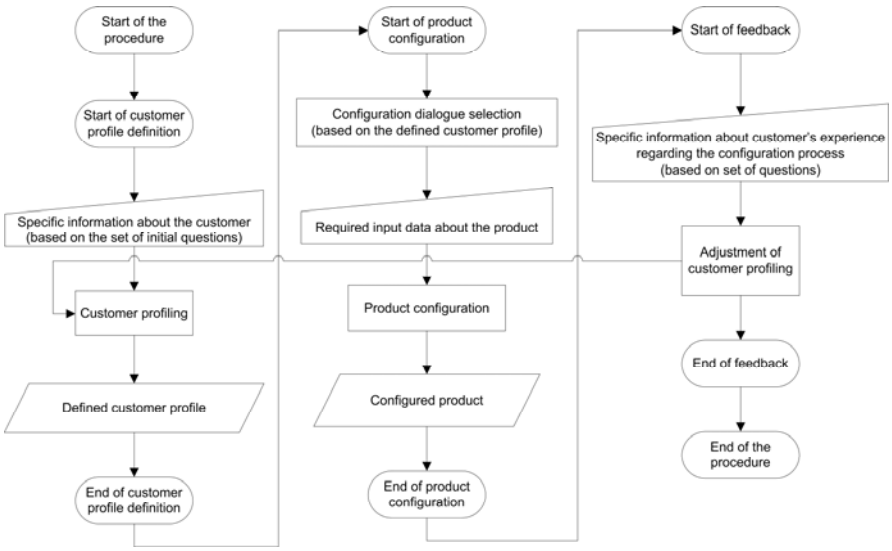


Figure 1
The product configuration procedure

4.1 Overview of Customer Profile Definition

As an answer to observed problems, three types of customer profiles have been introduced: *beginner*, *intermediate* and *professional*. The "*beginner customer*" is a customer without proper technical knowledge about thermal insulation, or maybe a customer with no need for highly accurate results, or a customer with a need of a fast enough result, etc. The "*intermediate customer*" is a customer with average technical knowledge about thermal insulation, but it can also be a customer

without proper technical knowledge about thermal insulation but with more time for completing the configuration process or with a need for more accurate result, etc. The "*professional customer*" is a customer with great knowledge about the problem of thermal insulation; it may also be a customer with general technical knowledge about thermal insulation but with more time for completing the configuration process or with a need for more accurate result, etc.

In order to define the appropriate customer profile, an algorithm for collecting and using specific information about customers is introduced. The customer is given a set of initial questions to be answered at the beginning of the configuration process. The possible answers are defined within a range of very low to very high. Along with the answers, the order of answering the questions is also taken into account, because customers usually, based on their belief, sooner answer questions that are of higher importance to them than questions that are not. There is also an option to leave unanswered a question the customer feels is unimportant [33].

To analyze the answers generated by each customer and to use them to form a customer profile, a number of different approaches may be used. Nevertheless, the linguistic nature of the questions and answers refers to the use of a non-crisp logic. This led to the use of fuzzy-based technique to make a decision about the appropriate customer profile [34], [35], [36]. The technique used is Mamdani's fuzzy inference method, which is one of the most commonly seen fuzzy methodologies. Also, after the configuration process, the customer's feedback to his/her satisfaction with the configured profile is analyzed and the algorithm for customer profile definition in the future is adapted according to the feedback.

All the questions asked have an associated linguistic variable for storing the answers that can have values ranging from 0 to 1. For the same answers, the $\mu(x)$ membership functions of linguistic variables may change according to the order in which they had been answered. If the answer to the question is the first one (each question is defined by a linguistic variable with different values), the membership functions taper, which is formulated as: $\mu^{1st}(x) = [\mu(x)]^{y_i}$, where $y_i \geq 1$. It results in a more unique response. If the answer to the question is the last one, the membership functions expand, i.e. the equations are changed in the same manner, but: $y_i \leq 1$. It results in a more vague response.

The fuzzy output from the system, i.e. the decision is made in a manner that for i initial questions, each of which can have y_i values, $y_1 * y_2 * ... * y_i$ if-then rules can be defined. The rules are designed to produce j different outputs o with defined membership functions. After the evaluation of if-then rules, an aggregated output is generated. The aggregated output is then defuzzified by using the centroid calculation method.

Changes in input membership functions influence the customer profile definition. For the same answers, but for a different answering order, the configured customer profile can be different.

After the configuration task is finished, the customer is asked to answer a new set of questions, where the values of the answers can range from -0.5 to 0.5. The answers to the questions are the feedback about how well the configurator has recognized the customer's needs and limitations. Initially, all the answers are set to the value of 0, which means that the customer is satisfied with the configuration process. Based on the answers to questions, the values for input linguistic variables are modified to new values as: $new\_value = previous\_value + \frac{feedback}{2}$, where: $0 \le new\_value \le 1$. This is the input for a new fuzzy output from the system, i.e. a new decision. This new output o_{new} takes into consideration whether a customer is satisfied with a configured customer profile. Based on the difference between an original and a new output, the membership functions for o_{i+1}, where o_{i+1} is the output for customer profile configuration in the future, are shifted left or right to better articulate the future customers' preferences. The amount of shifting sa is calculated as: $sa = \frac{o - o_{new}}{10}$. The division by 10 is used to assure that the shift is not too big.

4.2 Initial Customer Profile Definition

The initial questions asked before the start of the configuration process are:

- What is your estimate of your knowledge about thermal insulation?
- What are your needs considering the accuracy of the configuration results?
- How much time do you have for completing the configuration process?

The answers can range from "I have no knowledge about thermal insulation at all" (Where the value of the answer is 0) to "I am a professional in the field of thermal insulation" (Where the value of the answer is 1) for the first question; from "I need as accurate result as possible" (Where the value of the answer is 0) to "I just want a rough estimate" (Where the value of the answer is 1) for the second question; and from "I have enough time for completing the configuration process" (Where the value of the answer is 0) to "I have limited time for completing the configuration process" (Where the value of the answer is 1) for the third question. Initially, all the answers are set to the value of 0.5. The answers are used as input data for customer profile configuration.

Based on asked questions and answers, three linguistic variables are defined:

- *Knowledge about thermal insulation k*, whose values are: very poor, poor, average, good and very good;
- *Accuracy of the configuration results a*, whose values are: high, average, low;
- *Time for the configuration process t*, whose values are: enough, average, not enough.

The membership functions for the variables are chosen based on testing and experience and are presented in equations (1)-(3).

$$\mu_{k=very\_poor}(x)=\begin{cases}1, & 0\le x\le 0.05\\ \dfrac{0.5-x}{0.5-0.05}, & 0.05<x\le 0.5\\ 0, & 0.5<x\le 1\end{cases} \qquad \mu_{k=poor}(x)=\begin{cases}\dfrac{x}{0.3}, & 0\le x\le 0.3\\ \dfrac{0.6-x}{0.6-0.3}, & 0.3<x\le 0.6\\ 0, & 0.6<x\le 1\end{cases}$$

$$\mu_{k=average}(x)=\begin{cases}0, & 0\le x\le 0.2\\ \dfrac{x-0.2}{0.5-0.2}, & 0.2<x\le 0.5\\ \dfrac{0.8-x}{0.8-0.5}, & 0.5<x\le 0.8\\ 0 & 0.8<x\le 1\end{cases} \qquad \mu_{k=good}(x)=\begin{cases}0, & 0\le x\le 0.4\\ \dfrac{x-0.4}{0.7-0.4}, & 0.4<x\le 0.7\\ \dfrac{1-x}{1-0.7}, & 0.7<x\le 1\end{cases} \qquad (1)$$

$$\mu_{k=very\_good}(x)=\begin{cases}0, & 0\le x\le 0.5\\ \dfrac{x-0.5}{0.95-0.5}, & 0.5<x\le 0.95\\ 1, & 0.95<x\le 1\end{cases}$$

$$\mu_{a=high}(x)=\begin{cases}1, & 0\le x\le 0.1\\ \dfrac{0.75-x}{0.75-0.1}, & 0.1<x\le 0.75\\ 0, & 0.75<x\le 1\end{cases} \qquad \mu_{a=average}(x)=\begin{cases}0, & 0\le x\le 0.1\\ \dfrac{x-0.1}{0.5-0.1}, & 0.1<x\le 0.5\\ \dfrac{0.9-x}{0.9-0.5}, & 0.5<x\le 0.9\\ 0 & 0.9<x\le 1\end{cases} \qquad (2)$$

$$\mu_{a=poor}(x)=\begin{cases}0, & 0\le x\le 0.25\\ \dfrac{x-0.25}{0.25-0.9}, & 0.25<x\le 0.9\\ 1, & 0.9<x\le 1\end{cases}$$

$$\mu_{t=enough}(x)=\begin{cases}1, & 0\le x\le 0.1\\ \dfrac{0.75-x}{0.75-0.1}, & 0.1<x\le 0.75\\ 0, & 0.75<x\le 1\end{cases} \qquad \mu_{t=not\_enough}(x)=\begin{cases}0, & 0\le x\le 0.25\\ \dfrac{x-0.25}{0.25-0.9}, & 0.25<x\le 0.9\\ 1, & 0.9<x\le 1\end{cases} \qquad (3)$$

$$\mu_{t=average}(x)=\begin{cases}\dfrac{x-0.1}{0.5-0.1} & 0\le x\le 0.5\\ \dfrac{0.9-x}{0.9-0.5}, & 0.5<x\le 1\end{cases}$$

If the answer to the question is the first one, the used membership functions change in a manner described by equations (4)-(6). If the answer to the question is the last one, the membership functions change in a manner described by equations (7)-(9).

$$\mu_{k=very\_poor}^{1st}(x) = \left[\mu_{k=very\_poor}(x)\right]^2 \qquad \mu_{k=good}^{1st}(x) = \left[\mu_{k=good}(x)\right]^2$$
$$\mu_{k=poor}^{1st}(x) = \left[\mu_{k=poor}(x)\right]^2 \qquad \mu_{k=very\_good}^{1st}(x) = \left[\mu_{k=very\_good}\right]^2 \tag{4}$$
$$\mu_{k=average}^{1st}(x) = \left[\mu_{k=average}(x)\right]^2$$

$$\mu_{a=high}^{1st}(x) = \left[\mu_{a=high}(x)\right]^2 \qquad \mu_{a=poor}^{1st}(x) = \left[\mu_{a=poor}(x)\right]^2 \tag{5}$$
$$\mu_{a=average}^{1st}(x) = \left[\mu_{a=average}(x)\right]^2$$

$$\mu_{t=enough}^{1st}(x) = \left[\mu_{t=enough}(x)\right]^2 \qquad \mu_{t=not\_enough}^{1st}(x) = \left[\mu_{t=not\_enough}(x)\right]^2 \tag{6}$$
$$\mu_{t=average}^{1st}(x) = \left[\mu_{t=average}(x)\right]^2$$

$$\mu_{k=very\_poor}^{1st}(x) = \left[\mu_{k=very\_poor}(x)\right]^{0.9} \qquad \mu_{k=good}^{1st}(x) = \left[\mu_{k=good}(x)\right]^{0.75}$$
$$\mu_{k=poor}^{1st}(x) = \left[\mu_{k=poor}(x)\right]^{0.75} \qquad \mu_{k=very\_good}^{1st}(x) = \left[\mu_{k=very\_good}\right]^{0.9} \tag{7}$$
$$\mu_{k=average}^{1st}(x) = \left[\mu_{k=average}(x)\right]^{0.25}$$

$$\mu_{a=high}^{last}(x) = \left[\mu_{a=high}(x)\right]^{0.25} \qquad \mu_{a=poor}^{last}(x) = \left[\mu_{a=poor}(x)\right]^{0.25} \tag{8}$$
$$\mu_{a=average}^{last}(x) = \left[\mu_{a=average}(x)\right]^{0.75}$$

$$\mu_{t=enough}^{last}(x) = \left[\mu_{t=enough}(x)\right]^{0.25} \qquad \mu_{t=not\_enough}^{last}(x) = \left[\mu_{t=not\_enough}(x)\right]^{0.25} \tag{9}$$
$$\mu_{t=average}^{last}(x) = \left[\mu_{t=average}(x)\right]^{0.75}$$

The fuzzy output from the system is designed to produce three different outputs *o*: *beginner*, *intermediate* and *professional*. Its membership functions are defined in equation (10). The decision is made using 45 *if-then* rules that take into consideration all of the possible answers. The rules are defined as the examples given in equation (11). After the evaluation of *if-then* rules, an aggregated output is generated that maximizes the values for the membership functions to the maximum values obtained from the rules. The final output, i.e. the customer profile is determined after defuzzification.

$$\mu_{o=beginer}(x) = \begin{cases} 1, & 0 \le x \le \alpha_0 \\ \dfrac{\beta_0 - x}{\beta_0 - \alpha_0}, & \alpha_0 < x \le \beta_0 \\ 0, & \beta_0 < x \le 1 \end{cases} \qquad \mu_{o=professional}(x) = \begin{cases} 0, & 0 \le x \le \phi_0 \\ \dfrac{x - \phi_0}{\phi_0 - \varphi_0}, & \phi_0 < x \le \varphi_0 \\ 1, & \varphi_0 < x \le 1 \end{cases}$$

$$\mu_{o=\text{int}ermediate}(x) = \begin{cases} 0, & 0 \le x \le \chi_0 \\ \dfrac{x - \chi_0}{\delta_0 - \chi_0}, & \chi_0 < x \le \delta_0 \\ \dfrac{\varepsilon_0 - x}{\varepsilon_0 - \delta_0}, & \delta_0 < x \le \varepsilon_0 \\ 0 & \varepsilon_0 < x \le 1 \end{cases} \tag{10}$$

$$\alpha_0 = 0.2, \beta_0 = 0.5$$
$$\chi_0 = 0.3, \delta_0 = 0.5$$
$$\varepsilon_0 = 0.7, \phi_0 = 0.5$$
$$\varphi_0 = 0.8$$

$$
\begin{aligned}
&if\ (\mu_{k=very\_poor} \neq 0 \wedge \mu_{a=high} \neq 0 \wedge \mu_{t=enough} \neq 0) \Rightarrow \\
&1.\ \mu_{o=\text{int}ermediate} = \min(\mu_{k=very\_poor}, \mu_{a=high}, \mu_{t=enough}) \\
&... \\
&if\ (\mu_{k=average} \neq 0 \wedge \mu_{a=poor} \neq 0 \wedge \mu_{t=not\_enough} \neq 0) \Rightarrow \\
&27.\ \mu_{o=beginner} = \min(\mu_{k=average}, \mu_{a=poor}, \mu_{t=not\_enough}) \\
&...
\end{aligned}
\tag{11}
$$

4.3 Feedback

After the configuration task is finished, the customer is asked to answer a set of three new questions:

- Are you satisfied with the complexity of the configurator? *c*;
- Is the result satisfactory? *s*;
- Are you satisfied with the time spent for the configuration process? *i*.

The answers can range from "The configurator is too complex" (where the value of the answer is -0.5) to "The configurator is too easy " (where the value of the answer is 0.5) for the first question; from "The results should be more detailed and precise" (where the value of the answer is -0.5) to "The results are too detailed" (where the value of the answer is 0.5) for the second question; and from "I could have spent more time for the configuration process" (where the value of the answer is -0.5) to "The configuration process was too long" (where the value of the answer is 0.5) for the third question. Initially, all the answers are set to the value of 0, which means that the customer is satisfied with the configuration process.

Based on the answers to questions, the input values for k,a,t are modified to k_{new},a_{new},t_{new}, as was described in *4.1 Overview of customer profile definition*. This is the input for a new fuzzy output from the system, i.e. a new decision. This new output o_{new} takes into consideration whether a customer is satisfied with a configured customer profile. Based on the difference between an original and a new output, the membership functions for o_{i+1} are shifted left or right to better articulate the customers' preferences in the future. The shifted membership functions for o with corrections are presented in equation (12).

$$
\begin{aligned}
\mu_{o=dummy}^{\,i+1}(x) &= \begin{cases} 1, & 0 \le x \le \alpha_{i+1} = (\alpha_i + sa) \\ \dfrac{\beta_{i+1} - x}{\beta_{i+1} - \alpha_{i+1}}, & \alpha_{i+1} = (\alpha_i + sa) < x \le \beta_{i+1} = (\beta_i + sa) \\ 0, & \beta_{i+1} = (\beta_i + sa) < x \le 1 \end{cases} \\
&\quad if \alpha_{i+1} < 0.05 then \alpha_{i+1} = 0.05 \\
&\quad if \alpha_{i+1} > 0.35 then \alpha_{i+1} = 0.35 \\
&\quad if \beta_{i+1} < 0.35 then \beta_{i+1} = 0.35 \\
&\quad if \beta_{i+1} > 0.65 then \beta_{i+1} = 0.65 \\
\mu_{o=\text{int}ermediate}^{\,i+1}(x) &= \begin{cases} 0, & 0 \le x \le \chi_{i+1} = (\chi_i + sa) \\ \dfrac{x - \chi_{i+1}}{\delta_{i+1} - \chi_{i+1}}, & \chi_{i+1} = (\chi_i + sa) < x \le \delta_{i+1} = (\delta_i + sa) \\ \dfrac{\varepsilon_{i+1} - x}{\varepsilon_{i+1} - \delta_{i+1}}, & \delta_{i+1} = (\delta_i + sa) < x \le \varepsilon_{i+1} = (\varepsilon_i + sa) \\ 0 & \varepsilon_{i+1} = (\varepsilon_i + sa) < x \le 1 \end{cases} \\
&\quad if \chi_{i+1} < 0.15 then \chi_{i+1} = 0.15 \\
&\quad if \chi_{i+1} > 0.45 then \chi_{i+1} = 0.45 \\
&\quad if \delta_{i+1} < 0.35 then \delta_{i+1} = 0.35 \\
&\quad if \delta_{i+1} > 0.65 then \delta_{i+1} = 0.65 \\
&\quad if \varepsilon_{i+1} < 0.55 then \varepsilon_{i+1} = 0.55 \\
&\quad if \varepsilon_{i+1} > 0.85 then \varepsilon_{i+1} = 0.85 \\
\mu_{o=professional}^{\,i+1}(x) &= \begin{cases} 0, & 0 \le x \le \phi_{i+1} = (\phi_i + sa) \\ \dfrac{x - \phi_{i+1}}{\phi_{i+1} - \varphi_{i+1}}, & \phi_{i+1} = (\phi_i + sa) < x \le \varphi_{i+1} = (\varphi_i + sa) \\ 1, & \varphi_{i+1} = (\varphi_i + sa) < x \le 1 \end{cases} \\
&\quad if \phi_{i+1} < 0.35 then \phi_{i+1} = 0.35 \\
&\quad if \phi_{i+1} > 0.65 then \phi_{i+1} = 0.65 \\
&\quad if \varphi_{i+1} < 0.65 then \varphi_{i+1} = 0.65 \\
&\quad if \varphi_{i+1} > 0.95 then \varphi_{i+1} = 0.95
\end{aligned}
\tag{12}
$$

5 Configuration of Thermal Insulation

Different customer profiles ask for different levels and complexity of input parameters and constraints. Due to extensive amount of data and required calculations, this chapter presents just a brief overview of required input parameters, calculations and results that are shown in Fig. 2.

6 Experiment

The developed configurator was tested to obtain results regarding the accuracy of the calculated heat losses, configured thermal insulation and the necessary time for the configuration process. To be able to assess the quality of the configured results, exact calculations using professional software were performed by the authors.

The configuration process was carried out by two groups of potential customers, with differently assumed capabilities regarding their knowledge on thermal insulation. All participants of the experiment tested the configurator individually. In the initial phase of the configuration process, for each participant a customer profile was defined based on input questions. The customer sample and the result of customer profiling is presented in Table 1. Differences in assumed and profiled customers resulted mostly because some of the participants assumed to have no related knowledge at all asked for results of higher accuracy and/or declared themselves to have more time available. At the same time, some of the participants assumed to have proper capabilities asked for short configuration process or had no need for accurate results.

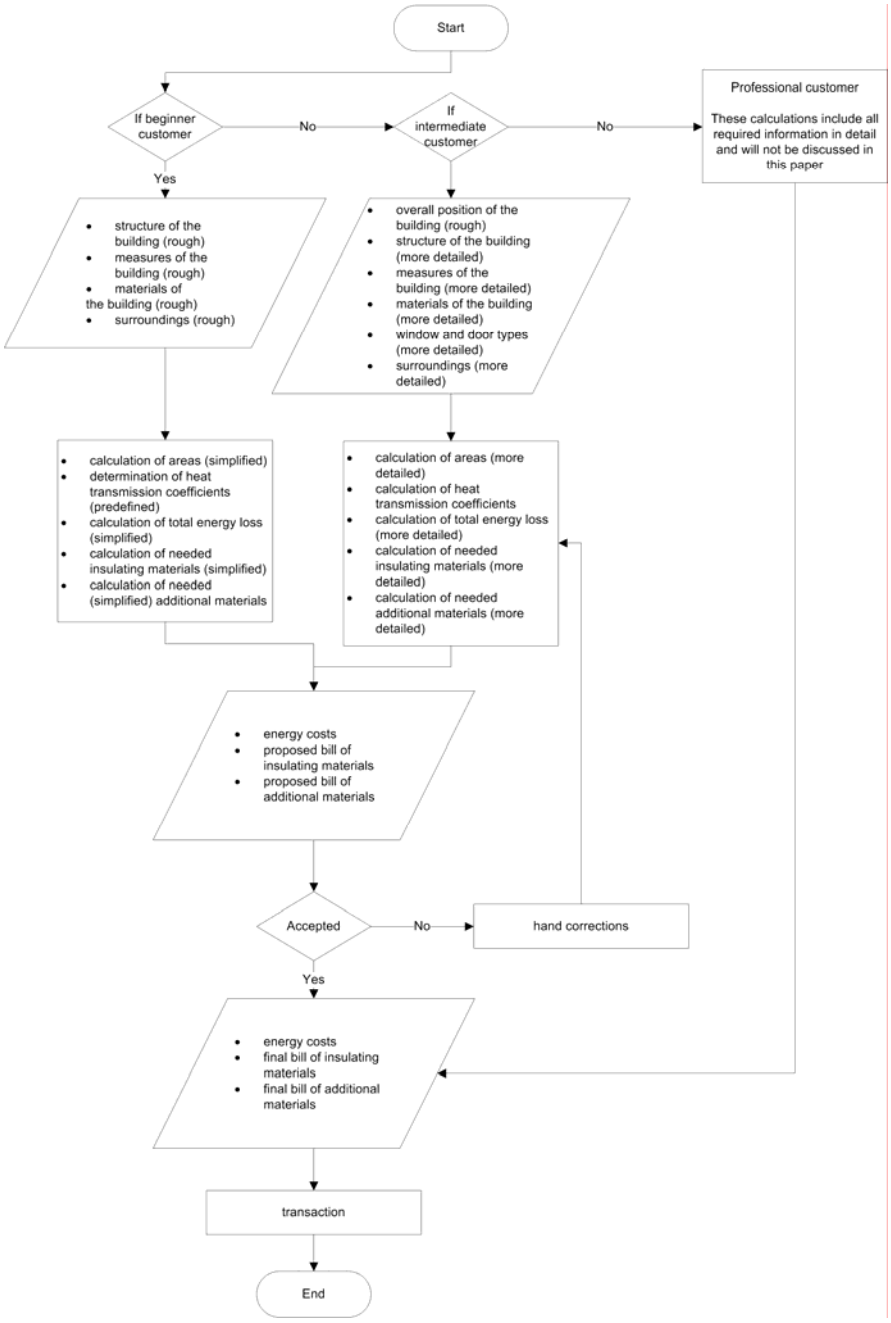


Figure 2
Simplified calculation algorithm

Table 1
The customer sample and the result of customer profiling

Participant characteristics (defined before customer profiling)		Participant profiles (resulting after customer profiling)	
		Profiled as *beginner customer*	Profiled as *Intermediate customer*
Assumed to have some technical knowledge	27	5	22
Assumed to have no technical knowledge	26	19	7

Based on the defined customer profile, each participant configured the outer thermal insulation for five different individual residential buildings that were common in design in the region of Central and East Europe and available for measurement. An example of input data regarding two types of buildings in the case of beginner and intermediate customer is presented in Appendix A. The necessary data used in the case of the professional customer is obtained from construction documentation of the buildings and is not presented in the paper.

Heat loss is calculated for following input temperatures [31]:

- Outer average air temperature - $268[K]$;
- Average ground temperature - $285[K]$;
- Average air temperature of the cellar - $285[K]$;
- The building's inner average air temperature - $293[K]$;
- Unused average air temperature of the loft - $268[K]$.

Calculated average heat losses without insulation and with the proposed insulation, as well as average relative deviations of calculated heat losses without insulation and with the proposed insulation from the exact calculations for different customer profiles, are shown in Table 2.

The result of the configuration process is the bill of necessary insulating and additional materials for the configured insulation. An example of one calculated bill of insulating and additional materials for the second type of analyzed buildings is shown in Appendix B.

7 Discussion of Results

The developed product configurator is to be used by a wide range of customers, those with average or no technical knowledge at all, as well as those who are professionals in the related field. Based on the customer profile selected by the configurator by interacting with the customer, the product configurator provides different results and satisfies different objectives regarding the complexity of the dialogue, the time required by the configuration process and the accuracy of the obtained results i.e. the resulting configurations.

Table 2

Average calculated heat losses [kWh/month] and relative deviations to detailed calculations

Building	Without insulation			With proposed insulation		
	Type of customer			Type of customer		
	Beginner	Intermediate	Professional	Beginner	Intermediate	Professional
No. 1 (Detached building, "square" shaped ground plan, no cellar, one floor, brick wall, etc.)	11380	12165	12560	5340	3684	3964
	-9.39%	*-3.14%*	*0%*	*34.71%*	*-7.06%*	*0%*
No. 2 (Detached building, "T" shaped ground plan, cellar, two floors with loft, brick wall with plaster, etc.)	17456	21165	20560	8465	5830	6156
	-15.10%	*2.94%*	*0%*	*37.51%*	*-5.30%*	*0%*
No. 3 (In contact with another building, "L" shaped ground plan, cellar, one floor with loft, brick wall, etc.)	6087	5179	5347	1236	1125	1034
	13.84%	*-3.14%*	*0%*	*19.54%*	*8.80%*	*0%*
No. 4 (In contact with another building, "square" shaped ground plan, no cellar, one floor, brick wall with plaster, etc.)	7891	7773	7650	1956	2236	2123
	3.15%	*1.61%*	*0%*	*-7.87%*	*5.32%*	*0%*
No. 5 (In contact with another building, "C" shaped ground plan, no cellar, one floor, brick wall, etc.)	8035	7435	7095	5203	4620	4205
	13.25%	*4.79%*	*0%*	*23.73%*	*9.87%*	*0%*

The results for the "*beginner customer*" show that absolute average deviation from the exact calculations of heat losses ranges from 3.15% to 15.10% for calculations without thermal insulation and from 7.87% to 37.51% for calculations with thermal insulation (Table 2). The deviation from the detailed calculations

regarding the calculated amounts of the required insulating and additional materials is up to 100% for some materials (Appendix B). The necessary time for the configuration process is about 3-5 minutes. The required number of input data is 11. The number and type of input data indicates that customers without proper technical knowledge can input the data successfully and in a relatively short time. This fulfills one of the objectives of the research in terms of the usability for non-professional customers. However, input data limits the possibilities of obtaining highly accurate results because, compared with the detailed calculations, the calculation algorithm needs to be considerably simplified (Fig. 2). The simplification of the calculation procedure leads to results, both of the calculations of heat losses and the required materials, which cannot be acceptable in a technical sense, though these results will lend useful insight into the needs, purpose and required level of thermal insulation. Keeping in mind the definition of the beginner customer, the objective of providing acceptable accuracy is also fulfilled.

The results for the "*intermediate customer"* show that the absolute average deviation from the exact calculations of heat losses ranges on the one hand from 1.61% to 4.79% for calculations without thermal insulation and on the other hand from 5.30% to 9.87% for calculations with thermal insulation (Table 2). The deviation of the calculated amounts of the necessary insulating and additional materials is up to 6% (Appendix B). It must be pointed out that on this level all of the required insulating and additional insulating materials are included in the result. The time required for the configuration process amounts to approximately 6-12 minutes. At this level the required number of input data varies due to the differences that appear in the structure of the analyzed buildings. The number of input data range from proximately 60 to more than a 100 if the building consists of several floors or has a more complex ground plan (Appendix A). This indicates that the configuration task is extensive; however, the type of data and the fact that many data are inherited based on previous input will ensure that the configuration task will take low cognitive effort and a relatively short time. The input data enable the definition of a more complex calculation algorithm (Fig. 2) that provides results acceptable in a technical sense. It must be mentioned that when purchasing materials in practice, it is usually recommended to buy quantities that are 5-10% larger due to waste occurring in the process of applying the insulation. Given the definition of the "*intermediate customer*", objectives regarding complexity and cognitive effort are fulfilled.

Based on input data, the obtained model of the building for the beginner and intermediate customer differs to greater or lesser extent from the actual ground plan shape, structure and dimensions of an original building. These differences lead to calculation results that may have higher or lower results compared to detailed calculations (Table 2). Therefore deviations of results compared to detailed calculations may in some cases be negative, and in other cases positive values. For the same building, the configuration of the beginner customer can

yield a result that is for instance lower than that of detailed calculation, while at the same time the case of the intermediate customer can have higher calculated values. The reason for this occurrence lies in the fact that the calculation algorithms in these two cases are different.

The results for the "*professional customer"* are of high accuracy, yet this also means that the calculation algorithm is rather complex and requires an extensive amount of data both specific and typical for the area of investigations. Therefore, the necessary time for the whole process is several times longer than on the previously described levels and it may take up to several hours if the analyzed building has a more complex structure. While the results for this level will be more precise, they also require expert technical knowledge on thermal insulation from the customer. This, on the other hand, may also mean that this type of calculation is no longer considered a configuration in the classical sense of the term (which takes place in the sales process), but a full-scale calculation process (which takes the form of a professional service that should be paid for).

Conclusions

The aim of this research has been to develop and test a product configuration approach that can self-adapt to customers who have different levels of knowledge related to the configured products as well as their own real individual needs. In order to be able to develop a product configurator for such a wide range of customers and to avoid the abortion of the configuration process, and thus of the final economic transaction, three customer profiles have been identified: *beginner*, *intermediate* and *professional*. In order to map the customer onto one of these profiles, an algorithm has been proposed that uses specific information about customers and their behaviour during the profile definition process. For each customer profile a different configuration dialogue has been defined so that the amount and complexity of input information decreases in lower customer profiles. To manage the differences of input data, an associated calculation algorithm has been developed for each of the identified customer profiles.

To evaluate the performance of the proposed configuration approach, a product configurator in a highly specific technical field of thermal insulation has been developed. Thermal insulating materials are mass products in themselves, but when installed on a specific building the final configured product is unique for each individual solution. This calls for greater customer involvement in the configuration process as the input data are unknown in the beginning due to the specific characteristics of each building. The results of testing the proposed configuration approach showed that the gained results were usable in practice regardless of the simplifications of the input data and calculations on the lower-level customer profiles.

This leads to the conclusion that if this approach is usable for customization of a product that requires high customer involvement and is unique for each solution, it could also be successfully used in the customization of products with more stable

product structures, attributes and constraints. These products include capital goods such as buildings, machinery or tools. Further, they include consumer goods, both simple, such as shoes or clothing, and complex, such as cars, computers and furniture, or even services like insurance and tourism offers. If the possible structure of the product is known, regardless of its complexity, there is no need to have simplifications in the calculation algorithm, i.e. in the configuration algorithm. Following this logic, the input on lower customer profiles will give a more generalized description of customer requirements, while on higher level, customer requirements will be defined in greater detail. This will eventually lead to product solutions that will in each case be exact and correct but meet the exact requirements of customers to a greater or lesser extent in correspondence to the level of input and customer knowledge. Therefore the contribution of this research lies in the introduction of a product configuration approach that self-adapts the configuration process to the capabilities and needs of the customers, who can range from non-experts to experts in the related field of investigation.

Certain issues arise from the fact that the presented configurator is to be used by customers both with or without specific technical knowledge and from the fact that the configured product is connected to a specific technical field. One of the problematic points is that a simple configuration algorithm with a relatively small number of input data given by a non-expert customer may provide a somewhat incorrect result. However this same non-expert customer will benefit from having performed this configuration in any case. Should the customer then decide to purchase insulation material without expert help, they will still be better off and are more likely to make a "good purchase". In terms of directions for possible future research, this points towards the investigation of the limits of simplifying the configuration process. The other question at hand comes from the trend whereby nowadays customers buy complex products like computers and cars using product configurators even though they are far from being experts in the fields. When does a proposed configuration process become too complex and time demanding? How long before the customer becomes frustrated and decides to abort the process? Future research should thus focus on investigating what is the maximum possible engagement of the customer when using a product configurator and whether there are products that are not configurable at all by using product configurators. Understanding the limits of simplification and complexity will likely turn this specified configurator into a generally applicable, useful tool for mass customization.

Acknowledgements

This research is supported by *"Masterplast Yu"* Ltd.

Appendix A: Example of Input Data

Type of information	Type of building							
	Building No. 1			Building No. 2				
	Type of customer			Type of customer				
	Beginner	Intermediate		Beginner	Intermediate			
General								
Area where the building is situated		Normal			Normal			
Location of the building		Protected			Protected			
Type of building		Detached			Detached			
Cellar	No	No		Yes	Yes			
Percentage of cellar underground [%]	0	0		40	40			
Insulation of the cellar's walls		No			No			
Insulation of the floor which is in contact with the ground		No			No			
Number of floors in the building	1	1		2	2			
Socle covered with plaster		No			No			
Type of roof	1	1		6	6			
Insulation type of loft	2	2		2	2			
Loading of attic/loft, i.e. roof		No			No			
Estimated area of the ground-plan of the building [m2]	100			130				
High or low ceiling	Low			Low				
Estimated total number of rooms	6			12				
Estimated type of the building regarding the prevailing materials used	Full brick			Full brick				
Previous insulation	No			No				
Estimated quality of the windows and doors	Average			Average				
Specific								
Level		First floor	Roof		Cellar	First floor	Second floor	Roof
Shape of the ground plan		1	1		4	4	4	4
Linear dimensions of the ground plan (circumferential dimensions) [m]								
d1		11	11		11	11	11	11
d2		9	9		13	13	13	13
d3		-	-		5	5	5	5
d4		-	-		4	4	4	4
d5		-	-		2	2	2	2
Height i.e. heights of the floor [m]								
h1		3	3		2.6	3	3	5
h2		-	-		-	-	-	3
h3		-	-		-	-	-	1
Percentage of free surface area of exterior walls of the building's floor [%]								
a1		100	100		100	100	100	100
a2		100	100		100	100	100	100
a3		100	100		100	100	100	100
a4		100	100		100	100	100	100
a5		-	-		100	100	100	100
a6		-	-		100	100	100	100
a7		-	-		100	100	100	100
a8		-	-		100	100	100	100
Correction of exterior surface of floor walls [m2]		0	0		0	0	0	0
Initial profile [m]		40	0		48	0	0	0
Edge protectors [m]		8	0		12	0	0	0
Number of wall layers		1	1		2	2	2	2
Dominating type and thickness of the floor's wall								
Layer1 - type		Full brick	Full brick		Full brick	Full brick	Full brick	Full brick
Layer1 - thicness [m]		0.35	0.35		0.35	0.35	0.35	0.35
Layer2 - type		-	-		Plaster	Plaster	Plaster	Plaster
Layer2 - thicness [m]		-	-		0.02	0.02	0.02	0.02
Layer3 - type		-	-		-	-	-	-
Layer3 - thicness [m]		-	-		-	-	-	-
Layer4 - type		-	-		-	-	-	-
Layer4 - thicness [m]		-	-		-	-	-	-
Layer5 - type		-	-		-	-	-	-
Layer5 - thicness [m]		-	-		-	-	-	-
Dominating type and thickness of the floor's (level's) floor								
Layer1 - type		Armed concrete	Armed concrete		Armed concrete	Armed concrete	Armed concrete	Armed concrete
Layer1 - thicness [m]		0.2	0.2		0.2	0.2	0.2	0.2
Layer2 - type		-	-		-	-	-	-
Layer2 - thicness [m]		-	-		-	-	-	-
Layer3 - type		-	-		-	-	-	-
Layer3 - thicness [m]		-	-		-	-	-	-
Layer4 - type		-	-		-	-	-	-
Layer4 - thicness [m]		-	-		-	-	-	-
Layer5 - type		-	-		-	-	-	-
Layer5 - thicness [m]		-	-		-	-	-	-
Correction of the surface area of the floor's (level's) floor [m2]		0	0		0	0	0	0
Dominating type, total surface area, and number of windows of the floor		Wood double layer	-		Wood double layer	PVC	PVC	PVC
Number of windows		4	0		6	8	8	2
Total surface area of all windows [m2]		6	0		4	16	16	4
Profiles for window joint [m]		0	0		0	0	0	0
PVC profile for balconies with glass net [m]		0	0		0	0	0	0
Dominating type, total surface area, and number of doors of the floor		Wood	-		PVC	Wood	-	-
Number of doors		2	0		2	1	0	0
Total surface area of all doors [m2]		4	0		6	2	0	0

Appendix B: Example of Calculated Bill of Materials

Type of customer											
Beginer					Intermediate				Professional		
Insulating materials	Material code	Amount	Packaging	Relative deviation to detailed calculation	Material code	Amount	Packaging	Relative deviation to detailed calculation	Material code	Amount	Packaging
Styrodur	-	0 [m2]	0	100.00%	0510-25006000	125 [m2]	36	1.68%	0510-25006000	127.13 [m2]	37
Isomaster	0501-08008000	230 [m2]	79	7.79%	0501-08006000	248.77 [m2]	100	0.27%	0501-08006000	249.44 [m2]	100
Styrodur	0510-30008000	13.5 [m2]	6	2.17%	0510-30008000	13.15 [m2]	6	4.71%	0510-30008000	13.8 [m2]	6
Styrodur	0510-30008000	46 [m2]	19	6.05%	0510-30008000	46.08 [m2]	20	5.88%	0510-30008000	48.96 [m2]	20
Isomaster	-	0 [m2]	0	100.00%	0501-08006000	18.6 [m2]	5	2.77%	0501-08006000	19.13 [m2]	5
Isomaster	0301-07010000	43.33 [m2]	18	-0.74%	0301-07010000	43.11 [m2]	18	-0.23%	0301-07010000	43.01 [m2]	18
Isover rio twin	0508-rt126000	101.07 [m2]	7	-12.80%	0508-rt126000	85.6 [m2]	6	4.46%	0508-rt126000	89.6 [m2]	6
Additional materials	Material code	Amount	Packaging	Relative deviation to detailed calculation	Material code	Amount	Packaging	Relative deviation to detailed calculation	Material code	Amount	Packaging
Masterfix	0103-10001005	49.92 [l]	11	33.63%	0103-10001005	72.62 [l]	15	3.46%	0103-10001005	75.22 [l]	16
Masterfix	0103-01101025	2026.5 [kg]	83	36.85%	0103-01101025	3087 [kg]	124	3.81%	0103-01101025	3209.15 [kg]	129
Masternet	0101-117wh055	318.45 [m2]	6	12.62%	0101-117wh055	347.6 [m2]	7	4.63%	0101-117wh055	364.46 [m2]	7
Thermomaster	0110-0z080000	46 [m]	23	50.00%	0110-0z080000	92 [m]	46	0.00%	0110-0z080000	92 [m]	46
Ejot	0111-03000000	23 [pcs]	1	50.00%	0111-03000000	46 [pcs]	1	0.00%	0111-03000000	46 [pcs]	1
Ejot	0112-03000000	46 [pcs]	1	50.00%	0112-03000000	92 [pcs]	1	0.00%	0112-03000000	92 [pcs]	1
Ejot	0112-05000000	46 [pcs]	1	50.00%	0112-05000000	92 [pcs]	1	0.00%	0112-05000000	92 [pcs]	1
Ejot	0112-10000000	46 [pcs]	1	50.00%	0112-10000000	92 [pcs]	1	0.00%	0112-10000000	92 [pcs]	1
Thermomaster	0118-00130250	920 [pcs]	4	7.79%	0118-00130250	995.08 [pcs]	4	0.27%	0118-00130250	997.76 [pcs]	4
Plug dowel	0717-64002000	230 [pcs]	2	50.00%	0717-64002000	460 [pcs]	3	0.00%	0717-64002000	460 [pcs]	3
Thermomaster	0105-10100000	8 [m]	4	66.67%	0105-10100000	24 [m]	10	0.00%	0105-10100000	24 [m]	10

References

[1] C. Forza, F. Salvador: Product Information Management for Mass Customization, Palgrave Macmillan, London, UK, 2007

[2] S. Stankovski, M. Lazarević, G. Ostojić, I. Ćosić, R. Purić: RFID Technology in Product/Part Tracking during the Whole Life Cycle, Assembly Automation, Assembly Automation 29 (4) (2009) pp. 364-370

[3] X. Du, J. Jiao, M. M. Tseng: Understanding Customer Satisfaction in Product Customization, International Journal of Advanced Manufacturing Technology 31 (3/4) (2006) pp. 396-406

[4] Q. Zhang, M. M. Tseng: Modelling and Integration of Customer Flexibility in the Order Commitment Process for High Mix Low Volume Production, International Journal of Production Research 47 (22) (2009) pp. 6397-6416

[5] M. M. Tseng, R. J. Jiao, C. Wang: Design for Mass Personalization, CIRP Annals – Manufacturing Technology 59 (1) (2010) pp. 175-178

[6] D. Yang, R. Miao, W. Hongwei, Y. Zhou: Product Configuration Knowledge Modeling Using Ontology Web Language, Expert Systems with Applications 36 (3) (2009) pp. 4399-4411

[7] G. Zülch, H. I. Koruca, M. Börkircher: Simulation-supported Change Process for Product Customization – A Case Study in a Garment Company, Computers in Industry 62 (2011) pp. 568-577

[8] A. Trentin, E. Perin, C. Forza: Overcoming the Customization-Responsiveness Squeeze by Using Product Configurators: Beyond Anecdotal Evidence, Computers in Industry 62 (2011) pp. 260-268

[9] Z. Chen, I. Wang: Personalized Product Configuration Rules with Dual Formulations: a Method to Proactively Leverage Mass Confusion, Expert Systems with Applications 37 (1) (2010) pp. 383-392

[10] J. R. Galbraith: Designing the Customer-Centric Organization, Jossey-Bass, San Francisco, CA, 2005

[11] F. Salvador, C. Forza: Principles for Efficient and Effective Sales Configuration Design, International Journal of Mass Customisation 2 (1/2) (2007) pp. 114-127

[12] T. Blecker, N. Abdelkafi: Mass Customization: State of The Art and Challenges, in: T. Blecker, G. Friedrich (Eds.), Mass Customization: Challenges and Solutions, Springer, New York, NY, 2006, pp. 1-25

[13] P. Engelbrektsson, M. Soderman: The Use and Perception of Methods and Product Representations in Product Development: a Survey of Swedish Industry, Journal of Engineering Design 15 (2) (2004) pp. 141-154

[14] X. Luo, Y. Tu, J. Tang, C. K. Kwong: Optimizing Customer's Selection for Configurable Product in B2C e-commerce Application, Computers in Industry 59 (2008) pp. 767-776

[15] T. Leckner, M. Lacher: Simplifying Configuration through Customer-oriented Product Models, in: Proceedings of the International Conference on Engineering Design ICED 03, Stockholm, Sweden (2003)

[16] L. B. Romdhane, N. Fadhel, B. Ayeb: An Efficient Approach for Building Customer Profiles from Business Data, Expert Systems with Applications 37 (2010) pp. 1573-1585

[17] G. Hong, D. Xue, Y. Tu: Rapid Identification of the Optimal Product Configuration and its Parameters Based on Customer-Centric Product Modeling for One-of-a-Kind Production, Computers in Industry 61 (3) (2010) pp. 270-279

[18] M. Koch, K. Moeslein: User Representation in Ecommerce and Collaboration Applications, in: Proceedings of the 16th Bled eCommerce Conference eTransformation, Bled, Slovenia (2003) pp. 649-661

[19] Y. B. Cho, Y. H. Cho, S. H. Kim: Mining Changes in Customer Buying Behavior for Collaborative Recommendations, Expert Systems with Application 28 (2) (2005) pp. 359-369

[20] W. Fan, M. D. Gordon, P. Pathak: Effective Profiling of Consumer Information Retrieval Needs: A Unified Framework and Empirical Comparison, Decision Support Systems 40 (20) (2005) pp. 213-233

[21] S. S. Weng, M. J. Liu: Feature-based Recommendation for One-to-One Marketing, Expert Systems with Applications 26 (4) (2004) pp. 493-508

[22] X. Zhang, J. Edwards, J. Harding: Personalised Online Sales Using Web Usage Data Mining, Computers in Industry 58 (2007) pp. 772-782

[23] Y. Park, K. Chang: Individual and Group Behavior-based Customer Profile Model for Personalized Product Recommendation, Expert Systems with Application 36 (2009) pp. 1932-1939

[24] D. Lalic, K. Popovski, V. Gecevska, P.S. Vasilevska, T. Zdravko: Analysis of the Opportunities and Challenges for Renewable Energy Market in the Western Balkan Countries, Renewable and Sustainable Energy Reviews 15 (2011) pp. 3187-3195

[25] I. Sartori, A. G. Hestnes: Energy Use in the Life Cycle of Conventional and Low-Energy Buildings: A Review Article, Energy and Buildings 39 (2007) pp. 249-257

[26] N. Huberman, D. Pearlmutter: A Life-Cycle Energy Analysis of Building Materials in the Negev Desert, Energy and Buildings 40 (2008) pp. 837-848

[27] T. Ramesh, R. Prakash, K. K. Shukla: Life Cycle Energy Analysis of Buildings: An Overview, Energy and Buildings 42 (2010) pp. 1592-1600

[28] J. Morrissey, R. E. Horne: Life Cycle Cost Implications of Energy Efficiency Measures in New Residential Buildings, Energy and Buildings 43 (2011) pp. 915-924

[29] N. Daouas: A Study on Optimum Insulation Thickness in Walls and Energy Savings in Tunisian Buildings Based on Analytical Calculation of Cooling and Heating Transmission Loads, Applied Energy 88 (2011) pp. 156-164

[30] M. S. Alhomoud: Performance Characteristics and Practical Applications of Common Building Thermal Insulation Materials, Building and Environment 40 (2005) pp. 353-366

[31] E. R. Schramek: Taschenbuch für Heizung und Klimatechnik, R. Oldenbourg Verlag, München, Germany, 1995

[32] I. Fuerstner, Z. Anisic: Intelligent Product Configurator – the New Approach in Thermo Insulation of Buildings, Annals of the Faculty of Engineering Hunedoara 7 (2) (2009) pp. 165-170

[33] C. Chen: Human-centered Product Design and Development, Advanced Engineering Informatics 23 (2) (2009) pp. 140-141

[34] H. J. Zimmermann: Fuzzy Set Theory – and its Applications, Kluwer-Nijhoff Publishing, Boston, MA, 1997

[35] M. Takács: Multilevel Fuzzy Approach to the Risk and Disaster Management, Acta Polytechnica Hungarica 7 (4) (2010) pp. 91-102

[36] B. Zemeková, J. Talašová: Fuzzy Sets in HR Management, Acta Polytechnica Hungarica 8 (3) (2011) pp. 113-124

Characterization of Peak-Rate-limited Bandwidth-Efficient Discriminatory Processor Sharing

Pál L. Pályi[1], Attila Kőrösi[1], Balázs Székely[*2], József Bíró[*1], and Sándor Rácz[3]

[1] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, H-1529 P.O. box 91, e-mail: palyi@tmit.bme.hu

[2] Institute of Mathematics, Budapest University of Technology and Economics

[3] Traffic Analysis and Network Performance Laboratory, Ericsson Research, Ericsson Hungary

Abstract: In this paper we characterize the state space of the discriminatory processor sharing service discipline with peak-rate limitations of the flows. We analyze a bandwidth-efficient rate sharing model, in which the unused capacity of the server by peak-rate limited flows is re-distributed among the non-limited flows. An efficient algorithmic approach is presented to determine which classes are subject to peak-rate limitations and based on this the bandwidth shares of flows of classes in a given state of this system.

Keywords: Discriminatory Processor Sharing; bandwidth-efficient; Peak rate

1 Introduction

Modern mobile telecommunications networks and high speed data packet services need the elaboration of new resource sharing, congestion avoidance and dimensioning methods, in order to ensure the appropriate Quality of Service (QoS) and Service Differentiation. For link dimensioning [1] [2] purposes bandwidth sharing models are needed, especially for the elastic-type (compressible) traffic flows. Such models describing the flow-level performance of elastic flows have been widely studied in the literature [3, 4, 5, 6]. In this paper, processor sharing-like models are considered, in which the service capacity (the

* The research of Székely and Bíró was partially supported by NKTH-OTKA Foundation research grant #77778 and #77802, respectively.

bandwidth) of the server (the link) is shared among the jobs (flows) according to some sharing principles.

Probably the very first and simplest (egalitarian) processor sharing model was presented by Kleinrock in [7], mainly motivated by the modeling of time-shared computer systems. In [8] a single-server processor-sharing system with several classes was analyzed. Classes are distinguished based on weights, and jobs in the classes have no limitations on their possible service rates (there are no peak-rate limitations), except the server capacity itself. The scheduling strategy considered divides the total capacity in unequal fractions among the different flows according to the corresponding weights; hence such models are called as discriminatory processor sharing (DPS). The paper provides solutions for the conditional expected response time (conditional average waiting time) of a class−*k* job with a given service time requirement (with a given size of the class−*k* flow) as well as for the unconditional response times.

In [9] the authors use the results from [8] and prove that – assuming the system is stable – for each class the expected unconditional response time is finite and that the expected conditional response time has an asymptote.

In [10], for multi-class egalitarian processor sharing queues, the authors show that the marginal queue length distribution for each class is equal to the queue length distribution of an equivalent single class processor sharing model with a random number of permanent customers. Similarly, the mean sojourn time (conditioned on the initial service requirement) for each class can be obtained by conditioning on the number of permanent customers.

Peak-rate limitations have been introduced and analyzed in a single class processor sharing system first in [11] (called M/G/R Processor Sharing Model) and several improved versions have been studied and proposed for dimensioning IP access networks, e.g. in [12] and [13]. In [3] multi-rate (peak rate limited) loss models for elastic traffic are evaluated. A structural characterization of reversibility is developed and used to build a non-egalitarian processor-sharing queueing discipline that admits a product-form solution. However, in this model a special type of Discriminatory Processor Sharing is considered, where corresponding peak rates and weights are in proportion to each other.

1.1 Discriminatory Processor Sharing

Discriminatory Processor Sharing (DPS) [14] is an important generalization of the (multi-class) egalitarian processor sharing discipline. In DPS, to each traffic class a weight is assigned; the weight of class-*i* is denoted by g_i. The bandwidth shares of flows are proportional to these weights. Flows from the same class always get the same bandwidth share. More formally, two requirements can be identified on the capacity shares in DPS: requirement-A: $c_i/c_j=g_i/g_j$ and requirement-B:

$\sum_{i=1}^{K} N_i c_i = C$ (c_i's are the bandwidth shares, and C is the server capacity), which are uniquely fulfilled[1] by

$$c_i = \frac{g_i C}{\sum_{j=1}^{K} g_j N_j}, \quad i \in \{1, \ldots, K\},$$

where N_i is the number of class$-i$ users in the system. This bandwidth share is also the solution of the following optimization problem:

$$\max_{\underline{x}} \sum_{i=1}^{K} N_i g_i \log x_i \quad \text{s.t.} \sum_{i=1}^{K} N_i x_i = C. \tag{1}$$

In this paper, we characterize the state space and the bandwidth sharing scheme of the peak-rate limited DPS with bandwidth-efficient rate sharing. The peak rate limitation means that each traffic class has its own maximal rate that is denoted by b_i for class-i. If there is enough capacity then the flows receive their peak bandwidths. When there is not enough capacity for all ongoing flows to get their peak rates, that is, $\sum_{i=1}^{K} N_i b_i > C$, then some flows or all flows will be "compressed" in the sense of their reduced service rates. This is the elastic "regime" of the model. On bandwidth-efficient rate sharing we mean that requirement-B should be fulfilled in the elastic regime of the model; that is, all bandwidth left by the uncompressed flows is to be redistributed among the compressed flows. This type of rate sharing is also referred to as Pareto-efficient in the literature [14].

The paper is organized as follows. In the next section we show that the bandwidth redistribution leads to a simpler and well interpretable bandwidth share calculation in the case of compressed flows. In Section 3 we present that there is a strict order of compression and we give a method for determining the set of compressed classes and the bandwidth shares.

[1] Work-conserving property, i.e., either all flows get all the bandwidth they required or the system is serving on its full capacity.

2 Bandwidth Share Calculations in Peak-Rate Limited DPS

The non bandwidth-efficient processor sharing has been widely studied in the literature, but it does not prove to be a realistic model for real systems. In a non bandwidth-efficient case the total available capacity may not be used because residual capacity left by peak-rate limited flows is not (fully) redistributed among non peak-rate limited flows. In the models analyzed in [14, 15] there is no redistribution at all of unused capacity, hence bandwidth share of flow-*i* can be simply calculated in the following way:

$$c_i = \min\left(\frac{g_i C}{\sum_{j=1}^{K} g_j N_j}, b_i\right). \tag{2}$$

In [14] only the non bandwidth-efficient case is discussed and the bandwidth-efficient case is considered to be harder to analyze.

According to bandwidth-efficient rate sharing, unused capacity is redistributed among flows in proportion to their weights, so the calculation of the bandwidth shares and determining the set of compressed classes are somewhat more complicated. For a while, let us assume that the set of compressed ($\mathfrak{Z}: \{\forall i, c_i < b_i\}$) and uncompressed ($\mathfrak{A}$) classes are known in a given state $\underline{N} = (N_i,\ i \in \{1,\ldots,K\})$. In this case, $c_i = b_i,\ i \in \mathfrak{A}$. Since these flows cannot utilize their bandwidth shares, they leave

$$\sum_{i \in \mathfrak{A}} \left(\frac{g_i N_i}{\sum_{j=1}^{K} g_j N_j} C - N_i b_i \right)$$

Capacity, which is re-distributed among compressed flows. If $j \in \mathfrak{Z}$, the original bandwidth share is increased due to the redistribution. The redistribution should be proportional to the weights g_i, in order to keep a similar requirement to requirement-A. Between two compressed classes, $c_i / c_j = g_i / g_j,\ i, j \in \mathfrak{Z},$

and between a compressed and an uncompressed class, $c_i > c_k \frac{g_i}{g_k}, \forall i \in \mathfrak{Z}, \forall k \in \mathfrak{A}$.[2] This results

$$c_i = \frac{g_i}{\sum_{j \in \mathfrak{A} \bigcup \mathfrak{Z}} g_j N_j} C + \frac{g_i}{\sum_{k \in \mathfrak{Z}} g_k N_k} \sum_{l \in \mathfrak{A}} \left(\frac{g_l N_l}{\sum_{j=1}^{K} g_j N_j} C - N_l b_l \right), \; i \in \mathfrak{Z}. \tag{3}$$

Due to our assumption, constraint $c_i \leq b_i$ is fulfilled for $i \in \mathfrak{Z}$. This formula shows that identifying the service rate of classes and the set of compressed classes are more complicated in the bandwidth-efficient approach.

For an illustration of the differences between the bandwidth-efficient and the non bandwidth-efficient approaches see the Appendix. In the following, we consider the bandwidth-efficient method.

For implementing a calculation of bandwidth-efficient rate shares based on (3), we first show a simpler form of that equation, and then using this simpler form, we present a method for determining $\mathfrak{Z}$ and $\mathfrak{A}$.

Lemma 1 *The service rate of the compressed classes' users formulated in* (3) *can be re-written as*

$$c_i = \frac{g_i}{\sum_{j \in \mathfrak{Z}} g_j N_j} \left(C - \sum_{k \in \mathfrak{A}} N_k b_k \right), \; i \in \mathfrak{Z}. \tag{4}$$

The proof of this lemma is based on taking the right-hand side of (3) over a common denominator and performing a simplification, which eventually results in the right-hand side of (4). The immediate consequence is that c_i, $i \in \mathfrak{Z}$ can be considered as the bandwidth allocation of a reduced Discriminatory Processor Sharing system with capacity $(C - \sum_{k \in \mathfrak{A}} N_k b_k)$ and traffic classes $\mathfrak{Z}$ in state $\underline{N}$.

Note that $\mathfrak{Z}$ is unique for a given $\underline{N}$. We can distinguish between two cases. If $\sum_{i=1}^{K} N_i b_i \leq C$, $\mathfrak{Z}$ is empty so solution $c_i=b_i$ evidently fulfills constraint $\sum_{i=1}^{K} N_i c_i \leq C$. In the second case, $\sum_{i=1}^{K} N_i b_i > C$ and hence

[2] This requirement is needed to ensure that $\mathfrak{Z}$ is unique for given $\underline{N}$.

$\sum_{i=1}^{K} N_i c_i = C$. From this, it follows that there exists a class i^* for which $c_{i^*}<b_{i^*}$, that is, there is at least one compressed class (i^*). If class j is also compressed, $\frac{g_{i^*}}{g_j}=\frac{c_{i^*}}{c_j}$ holds by definition. Thus, $c_j = \min\left(b_j, \frac{g_j}{g_{i^*}} c_{i^*}\right)$, which means that $c_j, \forall j$ can be calculated from the bandwidth share c_{i^*} of one compressed class i^*. As a consequence,

$$\sum_{j=1}^{K} N_j \cdot \min\left(b_j, \frac{g_j}{g_{i^*}} c_{i^*}\right) = C. \tag{5}$$

The left-hand side of (5) is monotonously increasing function of c_{i^*}, so while $\sum_{i=1}^{K} N_i b_i > C$, there is one solution for c_{i^*}. Therefore, there is one solution for each c_j.

Preliminary numerical calculations lead us to conjecture that the bandwidth allocation from (4) is a global solution of the following optimization problem, which differs from (1) in the constraint $x_i \in [0, b_i]$:

$$\max_{\underline{x}} \sum_{i=1}^{K} N_i g_i \log x_i \quad \text{s.t.} \quad \sum_{i=1}^{K} N_i x_i \le C \quad \& \quad \forall i \; x_i \in [0, b_i].$$

In the following section we present an algorithmic approach to determine the set of compressed classes $\mathfrak{Z}$.

3 Determining the Compression of Classes

In this section, we propose a method for determining the set of compressed classes $\mathfrak{Z}$ and also the bandwidth shares of flows from each class.

Let C denote the considered capacity. We distinguish among three disjunct cases considering the compression of classes; namely all classes are uncompressed, all classes are compressed, and there are compressed classes but at least one class is uncompressed.

Let $\mathfrak{S}$ be a subset of $\{1,\ldots,K\}$. In the following three steps below, $\mathfrak{S}$ will be adjusted. Initially, let $\mathfrak{S}=\{1,\ldots,K\}$.

Step-1: Check whether all classes are uncompressed. If it is true, then C is enough for peak rates of all flows, i.e.

$$\sum_{i\in\mathfrak{S}} N_i b_i \le C\,. \tag{6}$$

In this case every flow gets its peak rate: $c_i = b_i,\ \forall i \in \mathfrak{S}$ and all classes are contained by $\mathfrak{A}$. No further steps are needed.

Step-2: Check whether all classes are compressed. If it is true, the bandwidth share of all flows are less than their peak rates, and there is no unused capacity to be redistributed, i.e.,

$$\frac{g_i}{\sum_{j\in\mathfrak{S}} N_j g_j} C < b_i, \forall i \in \mathfrak{S}\,. \tag{7}$$

The bandwidth share of flows are

$$c_i = \frac{g_i}{\sum_{j\in\mathfrak{S}} N_j g_j} C,\ \forall i \in \mathfrak{S},$$

so every class gets bandwidth share in proportion to their weights and all classes are contained by $\mathfrak{Z}$. No further steps are needed.

Step-3: In this case there are compressed classes, but at least one class is uncompressed, because none of the conditions in Step-1 (6) and Step-2 (7) is fulfilled. To determine which class is surely uncompressed we use the following equivalence

$$\frac{g_i}{b_i} < \frac{\sum_{j\in\mathfrak{S}} N_j g_j}{C}, \forall i \in \mathfrak{S} \Leftrightarrow \max_{i\in\mathfrak{S}} \frac{g_i}{b_i} < \frac{\sum_{j\in\mathfrak{S}} N_j g_j}{C} \tag{8}$$

The left-hand side of the relation simply comes from (7) by rearrangement. The equivalence above also means that if (7) is not fulfilled, then the right-hand side of (8) is also not fulfilled. Consequently, this surely uncompressed class is i',

$$i' = \arg\max_{i\in\mathfrak{S}} \frac{g_i}{b_i},\ \text{with bandwidth share}\ \ c_{i'} = b_{i'}.$$

For the remaining classes we should evaluate a reduced system where the effect of this class is considered according to Lemma-1; that is, the considered capacity is reduced by $N_{i'}b_{i'}$ ($C \leftarrow C - N_{i'}b_{i'}$) and only the remaining classes are considered ($\mathfrak{S} \leftarrow \mathfrak{S} \setminus \{i'\}$). For the reduced system we continue with Step-2.

The above described method can be summarized as the following algorithm (Algorithm 1):

1. $\mathfrak{Z} = \{1, 2, ..., K\}$

2. while $\max_{i \in \mathfrak{Z}} \left\{ \frac{g_i}{b_i} \right\} \geq \frac{\sum_{j \in \mathfrak{Z}} N_j g_j}{C}$ and $\mathfrak{Z} \neq \varnothing$ do

$$i' = \arg\max_{i \in \mathfrak{Z}} \left\{ \frac{g_i}{b_i} \right\}$$

$$\mathfrak{Z} \leftarrow \mathfrak{Z} \setminus \{i'\}$$

$$C \leftarrow C - N_{i'} b_{i'}$$

3. for $i = 1$ to K do

$$\text{if } i \in \mathfrak{Z} \text{ then } c_i = \frac{g_i}{\sum_{i \in \mathfrak{Z}} N_i g_i} C$$

$$\text{else } c_i = b_i.$$

Algorithm 1 Determining the set of compressed classes

An important consequence of the above described method is that the following order of classes:

$$\frac{g_1}{b_1} \leq \frac{g_2}{b_2} \leq \ldots \frac{g_K}{b_K},$$

based on the ratios g_i/b_i, is directly related to the compressed and uncompressed classes in such a way that:

if a class with higher g/b value is compressed, then all classes with lower g/b are compressed, and if a class with lower g/b value is uncompressed, then all classes with higher g/b are uncompressed. Also, observe that the compression order depends on neither the server capacity nor the number of users. In addition to this, the above described method also provides bandwidth shares of flows in each classes, as formulated in (4).

Conclusion

This paper is concerned with discriminatory processor sharing model with peak-rate limitations and a bandwidth-efficient rate sharing. We have shown that at a given state space (as a snapshot) the peak-rate limited DPS system includes a reduced-capacity DPS system over the compressed classes. Based on this, we have given a method to determine the set of compressed and uncompressed classes of flows and their bandwidth shares. It has also been shown that there is a strict order

of classes in which they become compressed, and this order coincides with the order of ratios of the corresponding class weights and class peak rates (g_i/b_i).

The significance of the result presented in this paper lies in the fact that these are inevitably the starting points for both the analysis and simulation of Discriminatory Processor Sharing constrained by peak-rate limitations and unused capacity redistribution.

Acknowledgment

The fruitful discussions with Szilveszter Nádas from Ericsson are highly appreciated.

References

[1] R. Jain, "Congestion Control and Traffic Management in Atm Networks: Recent Advances and a Survey," in *COMPUTER NETWORKS AND ISDN SYSTEMS*, 1995, pp. 1723-1738

[2] S. Nádas, S. Rácz, and P. Pályi, *Handbook of HSDPA/HSUPA Technology*. CRC Press, Taylor & Francis Group, 2010, ch. HSPA Transport Network Layer Congestion Control, pp. 297-330

[3] V. Koukoulidis, "A Characterization of Reversible Markov Processes with Applications to Shared-Resource Environments," Ph.D. dissertation, Concordia University Montreal, Canada, 1993

[4] S. Rácz, B. P. Gerö, and G. Fodor, "Flow Level Performance Analysis of a Multi-Service System Supporting Elastic and Adaptive Services," *Performance-Evaluation*, Vol. 49, pp. 451-469, 2002

[5] P. A. V. J. Lassila, P., "Dimensioning of Data Networks: a Flow-Level Perspective," *European Transactions on Telecommunications*, Vol. 20, No. 6, pp. 549-563, 2008

[6] J. Roberts, "A Survey of Statistical Bandwidth Sharing," *Computer Networks*, Vol. 45, No. 3, 2004

[7] L. Kleinrock, "Time-shared Systems: A Theoretical Treatment," *J. of ACM*, Vol. 14, No. 2, pp. 242-261, 1967

[8] G. Fayolle, I. Mitrani, and R. Iasnogorodski, "Sharing a Processor among Many Job Classes," *J. ACM*, Vol. 27, No. 3, pp. 519-532, 1980

[9] K. Avrachenkov, U. Ayesta, P. Brown, and R. Núńez-Queija, "Discriminatory Processor Sharing Revisited," in *In: Proc. IEEE Infocom 2005, Miami FL*, 2005, pp. 784-795

[10] S. Cheung, H. van den Berg, and R. Boucherie, "Decomposing the Queue Length Distribution of Processor-Sharing Models into Queue Lengths of Permanent Customer Queues," *Performance Evaluation*, Vol. 62, No. 1-4, pp. 100-116, 2005

[11] K. Lindberger, "Balancing Quality of Service, Pricing and Utilisation in Multiservice Networks with Stream and Elastic Traffic," in *ITC-16: International Teletraffic Congress*, 1999, pp. 1127-1136

[12] A. Riedl, T. Bauschert, and J. Frings, "A Framework for Multi-Service IP Network Planning," in *International Telecommunication Network Strategy and Planning Symposium (Networks)*, 2002, pp. 183-190

[13] Z. Fan, "Dimensioning Bandwidth for Elastic Traffic," *NETWORKING 2002: Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications*, pp. 826-837, 2006

[14] U. Ayesta and M. Mandjes, "Bandwidth-Sharing Networks under a Diffusion Scaling," *Annals Operation Research*, Vol. 170, No. 1, pp. 41-58, 2009

[15] A. Lakshmikantha, R. Srikant, and C. Beck, "Differential Equation Models of Flow-Size-based Priorities in Internet Routers," *International Journal of Systems, Control and Communications*, Vol. 2, No. 1, pp. 170-196, 2010

Appendix

Comparison of bandwidth-efficient and non bandwidth-efficient limited approaches

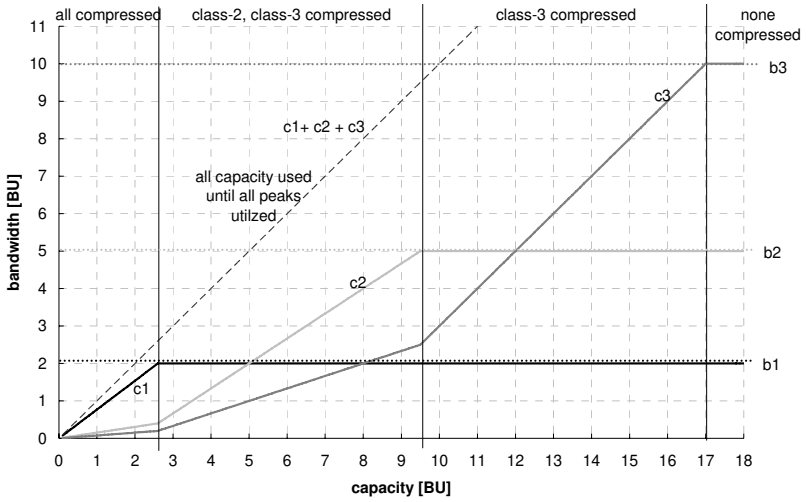


Figure 1
Peak-rate limited bandwidth-efficient DPS

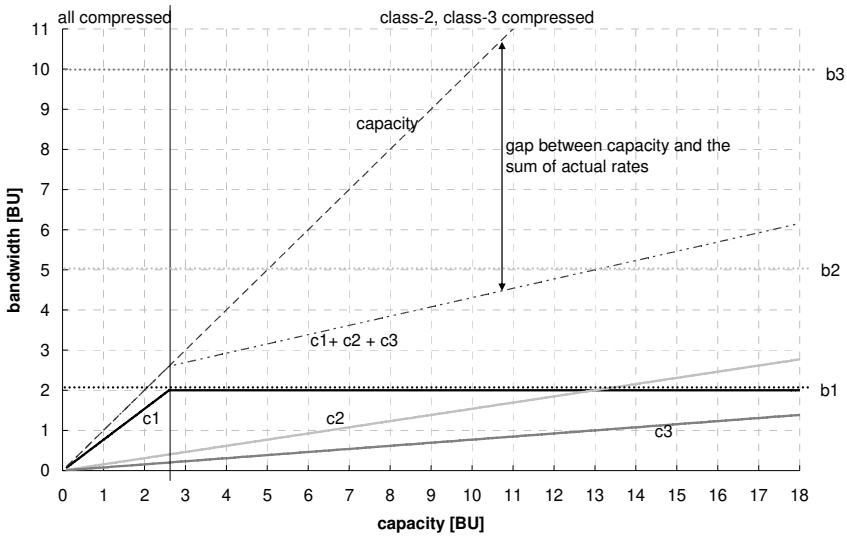


Figure 2
Peak-rate limited *non* bandwidth-efficient DPS

Table 1
Parameter settings

	Class-1	Class-2	Class-3
Peak rate (b_i)	2	5	10
Weight (g_i)	10	2	1
Number of users (N_i)	1	1	1

Figure 1 and **Figure 2** illustrate how the bandwidth-efficient and non bandwidth-efficient approaches are different from each other. Both figures show the same scenario, the only difference is the bandwidth-efficiency. See class parameters in **Table 1**. On the horizontal axis the capacity is shown. It is increased from 0 to 18 Bandwidth Units. On the vertical axis the actual service rate of the given class (c_i) is plotted. The peak rate of each class (b_i) is also shown.

The main difference between the two approaches is that in the case of the non-bandwidth-efficient approach, the total capacity is fully utilized only if all classes are compressed (**Figure 2**, where capacity is less than 2.7), i.e., not the peak rates are the limiting factors in the service rates. Otherwise, in this case the capacity is not utilized because residual capacity left by the peak-rate limited class 1 is not fully redistributed among non peak-rate limited class 2 and 3. In the case of the bandwidth-efficient approach (**Figure 1**), the available capacity is always fully utilized, except when the service rates of all flows are limited by their peak rates.

In **Figure 1**, four regions can be distinguished. If capacity is less than 2.7, all classes are compressed. This is the only region, where the bandwidth-efficient and the non bandwidth-efficient approaches give the same service rates, because there is no unused capacity left from peak-rate limited classes. If the capacity is not less than 2.7 but less than 9.5, then class-1 is no longer compressed, i.e., it gets its peak rate (b_1). Class-2 and class-3 are still compressed in proportion of their weights. If the capacity is not less than 9.5 but less than 17 then class-2 receives its peak rate also and is no longer compressed. In this region only class-3 is compressed, but gets all the capacity left from both peak-rate limited classes. If the capacity is not less than 17 then all classes receives their peak rates. In this region, because all classes are limited by their peak rates, further increase of the capacity does not increase the sum of the services rates of the three classes.

In **Figure 2**, only two regions can be distinguished. If capacity is less than 2.7, all classes are compressed. This is the only region, where the bandwidth-efficient and the non bandwidth-efficient approaches give the same service rates. Therefore, this region of **Figure 2** is the same as that of **Figure 1**. If the capacity is not less than 2.7 then class-1 already gets its peak rate (b_1), class-2 and class-3 are still compressed and their service rate is calculated according to the same formula as in region 1. Since capacity left by the peak rate limited class-1 is not redistributed among class-2 and class-3, there is a gap between the capacity and the sum of the actual service rates. The higher the capacity, the larger this gap gets.

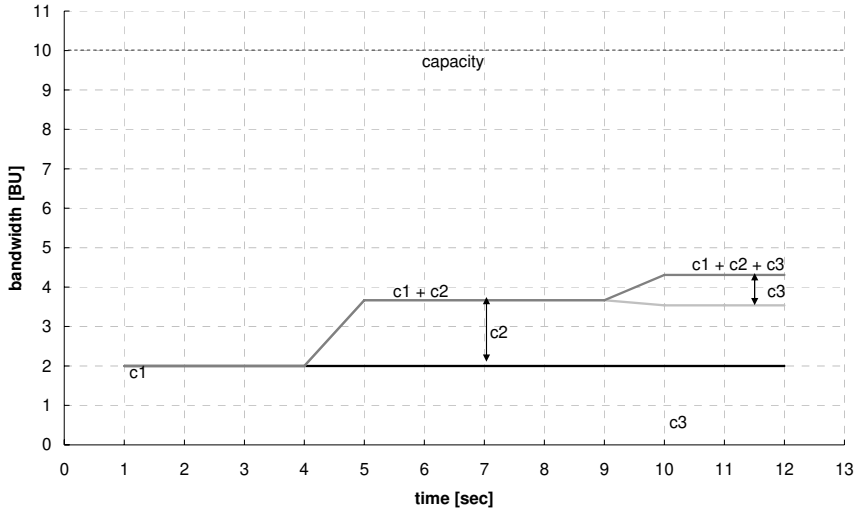


Figure 3
Peak-rate limited *non* bandwidth-efficient DPS, capacity=10 BU

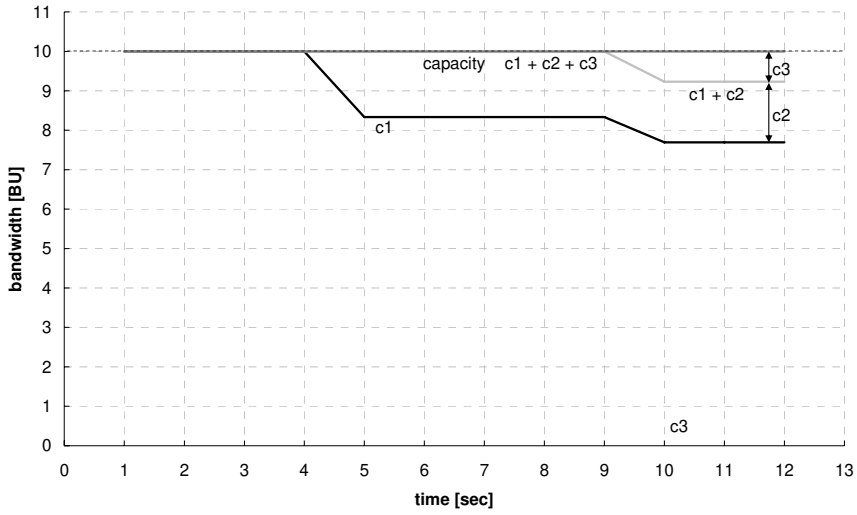


Figure 4
DPS with no peak rates, capacity=10 BU

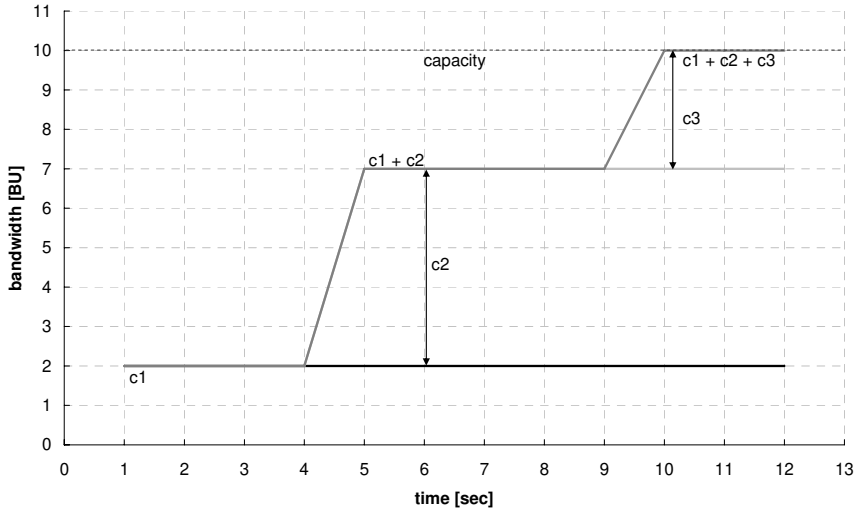


Figure 5
Peak-rate limited bandwidth-efficient DPS, capacity=10 BU

Figure 3 and **Figure 5** give an other illustration of the difference between the bandwidth-efficient and the non-bandwidth efficient approaches. In both figures the same scenario is depicted, apart from the bandwidth-efficiency. See class

parameters in **Table 1**. Capacity is now fixed at 10 BU. **Figure 4** also shows the same scenario but no peak rates are used. In these three figures, time is plotted on the horizontal axis. On the vertical axis the actual service rate of the given class (c_i) is shown in a cummalative way. It means that instead of plotting c_1, c_2, c_3 individually, c_1, the sum of c_1 and c_2, and the sum of c_1, c_2, and c_3 is plotted. At 0 sec no flows are in the system. At 1 sec a flow from class-1 arrives, at 5 sec a flow from class-2 arrives, finally at 10 sec, a flow from class-3 arrives.

Figure 3 shows the peak-rate limited non bandwidth-efficient approach. When class-1 arrives at 1 sec, it gets its peak rate. When class-2 arrives at 5 sec, its service rate is calculated according to (2), therefore it cannot utilize its peak rate. The same applies to class-3 when it arrives at 10 sec. The total capacity can still not be utilized.

Figure 4 depicts the same scenario with *no* peak rate limitations. When class-1 arrives at 1 sec, it can use the total capacity. When class-2 arrives at 5 sec the two classes share the capacity in proportion of their weights. When class-3 arrives at 10 sec the capacity is shared among three flows in proportion of their weights. The total capacity is always utilized since no peak rates are limiting the flows service rates.

Figure 5 shows the peak-rate limited bandwidth-efficient approach. When class 1 arrives at 1 sec, it gets its peak rate. When class-2 arrives at 5 sec it also gets its peak rate since the sum of the peak rates (7 BU) is still less than the capacity (10 BU). When class-3 arrives at 10 sec, it gets compressed since it has the smallest weight and otherwise the capacity would be exceed. The total capacity is only utilzed after 10 sec because until this time peak rate is limiting both flows.

A Simple Method to Forecast Travel Demand in Urban Public Transport

Balázs Horváth

Széchenyi István University
Department of Transport
Egyetem tér 1, H-9026 Győr, Hungary
balazs.horvath@sze.hu

Abstract: The key to the planning of public transport systems is the accurate prediction of the traffic load, or the correct execution of the planning stage assignment. This requires not only a well-functioning assignment method, but also reliable passenger data. Reliable passenger data means a time-dependent origin-destination matrix.
To solve the problem of the lack of time-dependent passenger data, we have developed a forecasting method. It consists of three stages.
In the first stage, we collect full scope cross-section data. This can be done either with personnel or with an automatic counting system. If personnel are used it costs a lot, and there is the chance for many possible errors. However, the results in most cases are good enough. Automatic counting system can be either a counter machine or even a simple "Check in" E-ticketing system.
In the second stage, we link boarding and alighting. As result, we get the origin-destination matrix for each run. This method is based on the likelihood of alighting at a given stop.
In the third stage, we combine origin-destination matrices of the runs through transfers. At this stage we assume that the probability of a transfer between two runs in a given stop is proportional to the travel possibilities in this relation.
To view the entire method in practice we proved it in a Hungarian city (Dunaújváros). The results were reliable, so they could be used in the planning process.

Keywords: public transport; transport planning; demand; OD matrix

1 Introduction

One of the possible solutions to handling the anomalies in city transport is the preference of public transport. Conversely, it is important to plan and operate a high level public transport service. The bottleneck of the planning of such systems is the knowledge of user demand. Without this knowledge, even the smallest change in the system is only a guess work, and the effect is unpredictable.

The cognition's methods of travel demand have been known for a long time [1], but their use has its limitations. Through our research we have built up a model which is able to generate a time-dependent O-D matrix for public transport with the use of the present transport system's characteristics.

2 Difficulties of Travel Demand's Determination

There are several methods to discover travel demand. Such can be the use of questionnaires or the application of a "check in – check out" e-ticketing system.

The application of an e-ticketing system can collect very detailed and accurate time dependent data day by day. But the establishment of such a system is very expensive and for a small bus operator unrealistic.

Another possible way is to organise questioners, but this costs a lot of financial resources and needs a large number of employees. Furthermore, the accuracy and reliability of the data are not always perfect, because reliable data need a large sample several times a day.

As an example, we can see a Hungarian city with 30,000 trips /day. This means about 7,000 trips in the morning peak period. The city can be divided into 25 zones, which means 625 possible trip relations. Some of these possibilities are not realistic or are used by only a few passengers; therefore, the number of real transport connections are about 100. If it is needed to take a sample from the morning peak, we have a basis of 7,000 persons. We can discover only the relations with at least 70 passengers, which means the incidence ratio is P=0,01. With the use of the common reliability of 95% and 10% relative error, the sample size is as follows (Eq. (1)).

$$n = \frac{t^2 \cdot (1-P)}{h^2 \cdot P} = \frac{1{,}96^2 \cdot (1-0{,}01)}{0{,}1^2 \cdot 0{,}01} = 38031{,}8 \tag{1}$$

where t reliability level of 95% (t=1,96)

P incidence ratio

h relative error

This value should be corrected because of the finite number of the basis (Eq. (2)).

$$n_0 = \frac{n}{1+\frac{n}{N}} = \frac{38031{,}8}{1+\frac{38031{,}8}{7000}} = 5911{,}8 \tag{2}$$

where N number of the elements in the basis

This means an 85% sample to have accurate and reliable O-D data for the morning peak. Under the same conditions for the whole day (the basis is 30,000 persons) it is necessary to ask 16,771 persons, which means 56% of the users should be questioned.

Literature recommends for such a network with home interviews a sample of 25% [2], [3]. This means 50% of the users under a modal-split of 50% should be asked. This means that every 4th household needs to be questioned.

It is easy to understand that this task (a sample of 50%) is hard or even impossible to perform. Therefore it is clear that there is a need for other methods to produce a time dependent O-D matrix for public transport.

3 Calculation of the O-D Matrix from a Cross-Section Counting

We have been dealing with the evaluation and planning of public transport systems for a long time. We are in almost cases we facing the problem: How to produce a reliable O-D matrix?. Considering that in all of the analyses there were cross-section countings, it was obvious to start our method with these countings. On the basis of this revelation we built up a method to calculate the O-D matrix for public transport.

The method has two major parts:

- Calculation of the O-D matrix for runs (services)
- Forecasting of the O-D matrix for the whole network considering transfers

3.1 Calculation of the O-D Matrix for Runs

After the performance of the full-scope cross-section counting we know the number of boarding and alighting passengers for each run in each stop (Fig. 1).

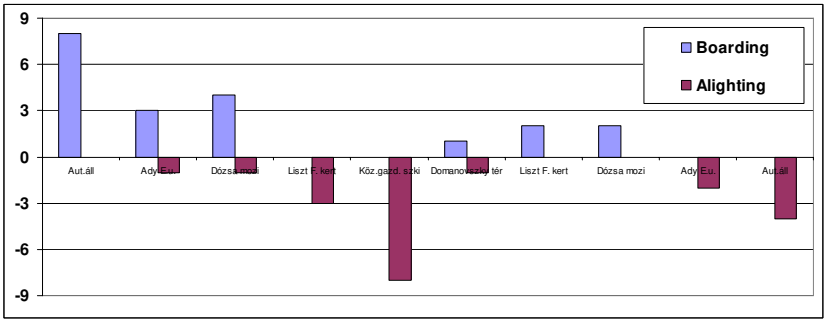


Figure 1
Number of boarding and alighting through the route of a line in one run

Assuming that the destinations (alighting) of the passengers boarding at a given stop are commensurate with the ratio of the alighting passengers of the remaining stops, we can forecast with likelihood the possible alighting stop of a passenger boarded at a given stop. The probability that a passenger boarded in stop i will travel to stop j can be calculated (Eq. (3)).

$$P_{i,j} = \frac{out_j}{\sum_{k=i+1\rightarrow n} out_k} \tag{3}$$

where out_j number of passengers alighting at stop j

n number of stops on the line route

With the help of this probability, the estimated number of trip makers between stop i and j on a given run of a given line is as follows (Eq. (4)).

$$f_{i,j} = in_i \cdot P_{i,j} \tag{4}$$

where in_i number of passengers boarding in stop i

The method can be refined if we take into account the average trip length (l_t) on that given run. The average trip length can be calculated if we know the number of boarding and alighting and the stop distance (distance between stops - l_s) (Eq. (5)).

$$l_{t,avg} = \frac{\sum_{k=1\rightarrow n-1} l_{s,k} \cdot p_k}{\sum_{r=1\rightarrow n} in_r} \tag{5}$$

where $l_{s,k}$ length of the k^{th} link (distance between stop k and k+1)

n number of stops on the line route

p_k number of passengers on the k^{th} link

In this case the probability of someone's travelling from stop i to stop j is as follows (Eq. (6)):

$$P_{i,j} = \frac{\frac{1}{\left|l_{t,avg} - l_{i,j}\right| + 1} \cdot out_j}{\sum_{k=i+1\rightarrow n} \frac{1}{\left|l_{t,avg} - l_{i,k}\right| + 1} \cdot out_k} \tag{6}$$

where $\left|l_{t,avg} - l_{i,j}\right|$ absolute value of the difference between average trip length and length of the analysed trip

Before we start the calculation of the probabilities to estimate the O-D matrix of the run, we can isolate some of the trips. These are the so-called definite trips. Either before or after these trips the vehicle is empty. This means that the vehicle is truly empty, or that a full passenger change has happened (after alighting the vehicle is empty, but there is also boarding). At this point, the line route of the run can be divided into sub-runs or sub-line routes.

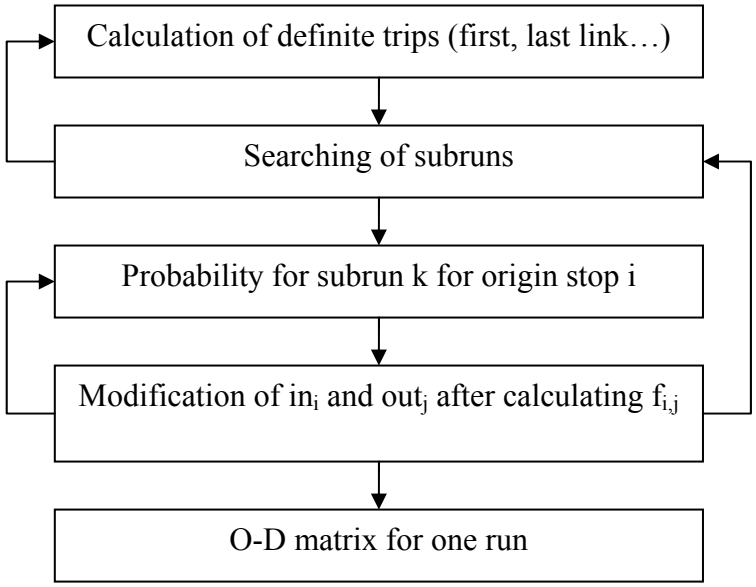


Figure 2
Process of the O-D matrix calculation for one run

Our task is to search the possible trips inside these sub-runs. This method gives much more accurate results than the calculation of the original one without any prudence. A special case is the first and the last link because before or after them the vehicle is empty, so the alighting after the first link, and the boarding before the last one are definite trips. The process of the model is shown in Figure 2.

3.2 Transfers between Runs

The O-D matrix of the runs should be corrected because a trip from i to j with a transfer in k will be two trips after the first step. These two trips are: i-k and k-j. The connection between these two trips is the transfer. In this second step of the method we must search this connection. To determine these trips with transfers, first we need to calculate the transfer ratio for stops and/or trips.

There are two different kinds of transfer ratio: the transfer ratio for alighting passengers and the transfer ratio for boarding passengers.

The transfer ratio for alighting passengers means the share between passengers who have reached their destination at a stop and the passengers who have alighted for transfer. Clearly it is the ratio of passengers who need to transfer and sum of alighted passengers.

The transfer ratio for boarding passengers is the other way around, and this means the share of passengers who are boarding because of a transfer among all of the boarding passengers.

If a trip from i to j needs to transfer at stop k, then the passengers of relation i,j has to be chosen from the passengers of the trips i-k and k-j. If we know the share of passengers who are transferring and those reaching destination from i to k, then we can give an estimation of the number of passengers travelling from i to j. Similarly to this, there is a ratio for passengers travelling from k to j. With this ratio, we know the share of trip starters and transfer makers. On the basis of this, there is another estimation of passenger numbers of i-j. We have to choose the right one from these two estimations. In general if a trip needs n transfers there are 2*n possible passenger numbers.

The process of this estimation can be done as follows: first, all of the possible routes between i and j must be calculated. Afterwards, trips with transfers will be broken up into sub-trips (arms). For all of the arms the number of passengers will be calculated. For all of the transfer points the transfer ratio will be calculated. On the basis of these, all of the trips will have several possible passenger numbers. Finally, out of these possible numbers the right one will be chosen.

The calculation of this step starts with the calculation of the routes. Through this calculation it is important to write down all the stops which are reachable from a given origin point with a given run. If a stop can be reached with several runs the transfer possibilities will be noted, and new possible destinations will be recorded. In the practice in works as follows: first line's first run starts from first stop. From this stop the second stop of this run can be reached. If there is a new line's new run, the process starts again, if not, then the third stop can be reached etc.

This will result in the list of all transport connections. Some of them are wrong or unrealistic connections. They have to be deleted either through the generation or afterwards. A connection is unrealistic if:

- it starts earlier and finishes together or later than another connection
- in a T time period the number of transfers of this connection is n+a
- in a T time period it is k times or at least with t minutes longer than the shortest path
- the trip time is longer than T_{max}

where n minimum number of transfers in a given relation

According to the literature [4] the parameters should be as follows:

- T=10 minutes
- a=2
- k=1,5
- t=10 minutes
- T_{max} in Hungarian cities can be 50 minutes

In the previous step the number of passengers was calculated for the direct trips. In this step the task is to calculate the number of passengers for the trips with a transfer.

The goal is to calculate the number of passengers in the relation i-j while the route leads through i-k_1-k_2-j with a transfer in k_1 and k_2. On the basis of the first step, the number of passengers for i-k_1, k_1-k_2 and k_2-j is known. The transfer ratios (calculation will be shown later) are known for both transfer points for alighting and boarding. It means for n transfers 2*n transfer ratios. It is impossible to use all of them separately because relation i-j has only one number of passengers. There is a method to solve this problem:

- Trips with one transfer: the average of the two numbers of passengers must be used
- Trips with more than one transfer: the highest and lowest number of passengers will be deleted. The searched number is the average of the rest

The number of passengers calculated in this way is higher than one of the arms (e.g.: i-k_1, k_1-k_2 or k_2-j); then the smallest number of the arms must be used.

3.3 Calculation of Transfer Ratios

The key to the transfer correction (shown previously) is the accurate calculation of the transfer ratio. There are two ways to calculate these:

- Questionnaires
- Estimation

With the use of questionnaires the basic problem returns. In this case, the situation is much better because a smaller sample can give good results due to the limited number of possible answers at few transfer points.

It is enough to ask passengers at selected transfer points about their behaviour at that stop: Are they starting here or transferring? If transferring, which line have they arrived from, and which line will they travel with further? Theoretically a small sample can give accurate results.

The other possibility to estimate the transfer ratio is based on the principle that the operated public transport system describes more or less the travel demand. Accepting this idea, it is then possible to build up a basis matrix with reciprocates of the journey times. This matrix must be corrected with the known boarding and alighting numbers. On the basis of this matrix, it is possible to determine the transfer ratio [5].

Following this, the transfer ratio for relation i-j in transfer point k_1 can be calculated:

$$tr_{ijk_1out} = \frac{f_{ij}^n}{\sum_l f_{il}^n \cdot tp_{k_1ij} + f_{ik_1}^n} \tag{7}$$

where f^n_{ij} member of the basis matrix after correction (iteration in n step)

tp_{k1ij} coefficient for transfer: 1 if trip from i to j has to transfer in k_1; 0 if not

Afterwards, the first possible passenger number for relation i-j on run r can be:

$$f_{ij,r}^1 = f_{ik_1,r} \cdot tr_{ijk_1out} \tag{8}$$

It is possible to calculate all the potential numbers for relation i-j. With the application of the method written in the previous section, the number of passengers can be calculated for relation i-j.

3.4 Utility of the O-D Matrix

After the calculation of all the possible relations, in the result is a “travel diary”. This diary will contains all the passenger numbers of all connections. These connections are true only in the existing public transport system, and therefore the diary is useless for planning in this format.

To use it in the planning, it must be aggregated in time (time periods for e.g. peak) and space (zones).

If there is a detailed plan for the future it is possible to use it as 10-15 minute matrices. Another way it can be useful to have it in the format of hour matrices.

4 Matrix Estimation in the Practice

4.1 Raw Data for Estimation

In 2008 there was a full scope cross-section counting in the Hungarian city of Dunaújváros [6]. Through this counting, the whole public transport system was analysed. The resulting numbers of the counting were processed with the demonstrated estimation method.

To make the plans and improvements in the public transport system, we built up the model of the transport system in the software Visum. The model had two pieces of input data:

- Supply (transport service)
  - Transport network (routes, stops)
  - Public transport service (lines, timetable)
- Demand (travel demand)

While the supply side was clear, the demand side was only partly known.

4.2 Application of the Estimation Method

We estimated the O-D matrix for each run with the help of the above described method. To accelerate the process, we divided the area into 24 zones and aggregate the data not after but before the process. In this way we searched zone-zone connections instead of stop-stop ones.

In this city there are few transfers. The transfer correction was needed in only a limited number of cases. Therefore the errors caused by this correction (if) were not significant.

As a result of the process, there was a matrix with 8819 rows and 24 columns. All of the rows symbolised passenger movement at stops on the basis of runs. After aggregation of these 8819 rows, there were 19 O-D matrices each for one hour through the day.

It was possible to divide them into smaller parts but it was enough for the planning.

These 19 O-D matrices was implemented into the transport model created by the software Visum (Figure 3).

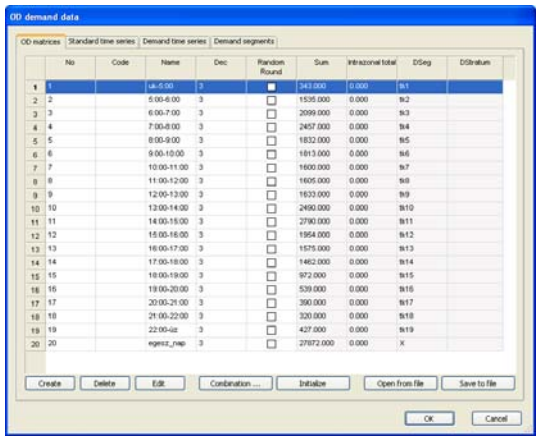


Figure 3
Travel demand in the transport model

After implementation of this demand into the model, the next step was to calibrate the assignment method. To calibrate it, some assignment was done on the present network.

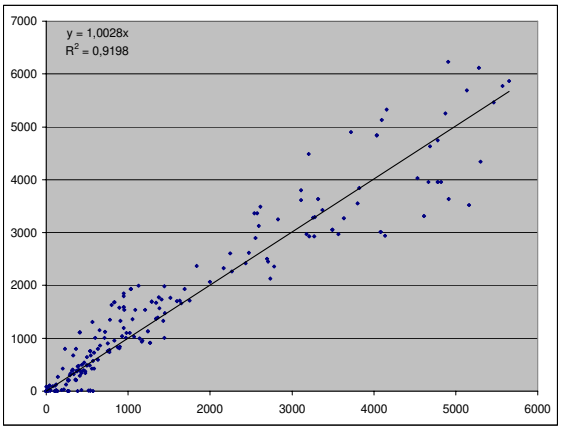


Figure 4
Calculated values in the function of measured values

In the calibration we used the timetable-based assignment method since the timetable was known for the present system.

The result of the assignment was checked on the basis of the links of the network. Theoretically the measured (counted) and calculated values should be equal. This means a graph of y=x, where y is the calculated and x is the measured (observed) value. In this study we get a graph of y=1.0028x which is very good. The indicator of the accuracy, the correlation coefficient was R^2=0.9198 (Figure 4). After these the estimation can be declared as good enough for planning purposes.

Figure 5
Measured and calculated values on the link-bar graph

The values can be compared also on graphical way, with the help of a link-bar graph.

Further modification of the assignment model it was possible to reach a correlation coefficient of 0.9991 although the graph was worse in this case (y=0.9245x). It is important to note that both situations give good results for network planning purposes.

Conclusions

The bottleneck of the planning of public transport systems is the knowledge of user demand. This knowledge is usually missing or only partly known. To solve this problem we built up a method to estimate the travel demand in time and space with high reliability. It means the estimation of a time-dependent O-D matrix for public transport systems.

The method was checked in practice. It proved that the method is good enough to use it in the normal day-to-day work for planning public transport systems.

References

[1] B. Horváth, G. Horváth, "Methodolgy of Public Transport Planning", Proceedings of European Society of Transport Institutes (ESTI) Young researcher's conference '99, Paris, France October 2-3, ESTI, 1999

[2] G. Fülöp, B. Hirkó, J. Mátyus, I. Prileszky, L. Szabó, *Közlekedési üzemtan II.*, SZIF-Universitas, Győr, hungary, 1999

[3] E. Nagy, D. Szabó, *Városi közlekedési kézikönyv*, Műszaki Könyvkiadó, Budapest, Hungary, 1984

[4] PTV AG, *Visum 10.0 User Manual*, PTV Planung Transport Verkehr AG, Karlsruhe, 2007

[5] B. Horváth, I. Farkas, R. Horváth, Á. Winkler, *Városi közforgalmú közlekedési szolgáltatás javításának lehetőségei Zalaegerszegre adaptált modell segítségével*, Széchenyi István University, Győr, Hungary, 2008

[6] I. Prileszky, G. Fülöp, B. Horváth, G. Horváth, I. Farkas, Á. Winkler, *Dunaújváros helyi tömegközlekedési szolgáltatásának fejlesztése komplex hatékonysági kritériumok alapján*, Universitas-Győr Nonprofit Kft, Győr, Hungary, 2009

Indirect Rotor Field-oriented Control (IRFOC) of a Dual Star Induction Machine (DSIM) Using a Fuzzy Controller

Radhwane Sadouni, Abdelkader Meroufel

Intelligent Control and Electrical Power Systems Laboratory (ICEPS)
Djillali Liabes University, BP 89 Sidi Bel-Abbes, Algeria
redouanesadouni@gmail.com; ameroufel@yahoo.fr

Abstract: We present in this paper, a comparative study between a PI regulator and fuzzy regulator for a control speed of a Dual Star Induction Machine (DSIM) supplied with a two PWM voltage source inverter (VSI) and decoupled by field-oriented control (FOC).The simulation results illustrate the robustness and efficiency of the fuzzy regulator to the parametric variations.

Keywords: dual star induction machine (DSIM); field-oriented control (FOC); fuzzy logic

1 Introduction

In industrial applications in which high reliability is demanded, a multi-phase induction machine instead of traditional three-phase induction machine is used. The advantages of multi-phase drive systems over conventional three-phase drives are: the total rating of system is multiplied, the torque pulsations will be smoothed, the rotor harmonic losses as well as the harmonics content of the DC link current will be reduced. And the loss of one machine phase does not prevent the machine working, thus improving the system reliability [1].

A common type of multiphase machine is the dual star induction machine (DSIM), also known as the six phase induction machine. These machines have been used in many applications (pumps, fans, compressors, rolling mills, cement mills, mine hoists …[2]) due to their advantages in power segmentation, reliability, and minimized torque pulsations. Such segmented structures are very attractive for high-power applications since they allow the use of lower rating power electronic devices at a switching frequency higher than the one usually used in three-phase AC machine drives [3].

The main difficulty in the asynchronous machine control resides in the fact that complex coupling exists between the field and the torque. The space vector control assures decoupling between these variables, and the torque is made similar to that of a DC machine [4].

In Field-oriented Control (FOC), three types of orientation exist: rotor field orientation, stator field orientation and rotating field orientation. In this paper, the rotor field-oriented control is applied to the DSIM using PI and fuzzy regulators.

2 Machine Model

A schematic of the stator and rotor windings for a machine dual three phase is given in Fig. 1. The six stator phases are divided into two wyes-connected three-phase sets labeled A_{s1}, B_{s1}, C_{s1} and A_{s2}, B_{s2}, C_{s2} whose magnetic axes are displaced by an angle $\alpha=30°$. The windings of each three-phase set are uniformly distributed and have axes that are displaced $120°$ apart. The three-phase rotor windings A_r, B_r, C_r are also sinusoidally distributed and have axes that are displaced apart by $120°$ [5].

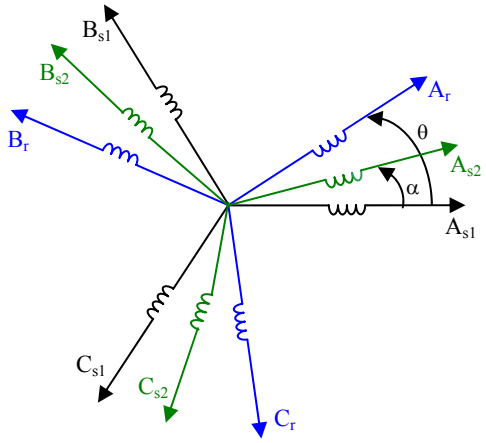


Figure 1
Windings of the dual star induction machine

The following assumptions are made: [4], [6]:

- Motor windings are sinusoidally distributed;
- The two stars have the same parameters;
- The magnetic saturation, the mutual leakage inductances and the core losses are negligible;
- Flux path is linear.

The voltage equations of the dual star induction machine are as follow [7] [8]:

$$\begin{aligned}
[V_{s1}] &= \begin{bmatrix} V_{sa1} \\ V_{sb1} \\ V_{sc1} \end{bmatrix} = [R_{s1}][I_{s1}] + \frac{d}{dt}[\Phi_{s1}] \\
[V_{s2}] &= \begin{bmatrix} V_{sa2} \\ V_{sb2} \\ V_{sc2} \end{bmatrix} = [R_{s2}][I_{s2}] + \frac{d}{dt}\left[\Phi_{s2}\right] \\
[0] &= \begin{bmatrix} V_{ra} \\ V_{rb} \\ V_{rc} \end{bmatrix} = [R_{r}][I_{r}] + \frac{d}{dt}[\Phi_{r}]
\end{aligned} \tag{1}$$

Where:

$R_{sa1} = R_{sb1} = R_{sc1} = R_{s1}$: Stator resistance 1.

$R_{sa2} = R_{sb2} = R_{sc2} = R_{s2}$: Stator resistance 2.

$R_{ra} = R_{rb} = R_{rc} = R_{r}$: Rotor resistance.

$$[R_{s1}] = \begin{bmatrix} R_{s1} & 0 & 0 \\ 0 & R_{s1} & 0 \\ 0 & 0 & R_{s1} \end{bmatrix};\ [R_{s2}] = \begin{bmatrix} R_{s2} & 0 & 0 \\ 0 & R_{s2} & 0 \\ 0 & 0 & R_{s2} \end{bmatrix};\ [R_{r}] = \begin{bmatrix} R_{r} & 0 & 0 \\ 0 & R_{r} & 0 \\ 0 & 0 & R_{r} \end{bmatrix} \tag{2}$$

$$[I_{s1}] = \begin{bmatrix} I_{sa1} \\ I_{sb1} \\ I_{sc1} \end{bmatrix};\ [I_{s2}] = \begin{bmatrix} I_{sa2} \\ I_{sb2} \\ I_{sc2} \end{bmatrix};\ [I_{r}] = \begin{bmatrix} I_{ra} \\ I_{rb} \\ I_{rc} \end{bmatrix} \tag{3}$$

$$[\Phi_{s1}] = \begin{bmatrix} \Phi_{sa1} \\ \Phi_{sb1} \\ \Phi_{sc1} \end{bmatrix};\ [\Phi_{s2}] = \begin{bmatrix} \Phi_{sa2} \\ \Phi_{sb2} \\ \Phi_{sc2} \end{bmatrix};\ [\Phi_{r}] = \begin{bmatrix} \Phi_{ra} \\ \Phi_{rb} \\ \Phi_{rc} \end{bmatrix} \tag{4}$$

The expressions for stator and rotor flux are [7]:

$$\begin{bmatrix} [\Phi_{s1}] \\ [\Phi_{s2}] \\ [\Phi_{r}] \end{bmatrix} = \begin{bmatrix} [L_{s1s1}] & [L_{s1s2}] & [L_{s1r}] \\ [L_{s2s1}] & [L_{s2s2}] & [L_{s2r}] \\ [L_{rs1}] & [L_{rs2}] & [L_{rr}] \end{bmatrix} . \begin{bmatrix} [I_{s1}] \\ [I_{s2}] \\ [I_{r}] \end{bmatrix} \tag{5}$$

Where:

$[L_{s1s1}]$: Inductance matrix of the star 1.

$[L_{s2s2}]$: Inductance matrix of the star 2.

$[L_{rr}]$: Inductance matrix of the rotor.

$[L_{s1s2}]$: Mutual inductance matrix between star 1 and star 2.

$[L_{s2s1}]$: Mutual inductance matrix between star 2 and star 1.

$[L_{s1r}]$: Mutual inductance matrix between star 1 and rotor.

$[L_{s2r}]$: Mutual inductance matrix between star 2 and rotor.

$[L_{rs1}]$: Mutual inductance matrix between rotor and star 1.

$[L_{rs2}]$: Mutual inductance matrix between rotor and star 2.

The expression of the electromagnetic torque is then as follows [7] [9] [10]:

$$T_{em} = \left(\frac{p}{2}\right).\left([I_{s1}]\frac{d}{d\theta}[L_{s1r}][I_r] + [I_{s2}]\frac{d}{d\theta}[L_{s2r}][I_r]\right) \quad (6)$$

The Park model of the dual star induction machine in the references frame at the rotating field (*d, q*) is defined by the following equations system (7) [11].

Figure 2 represents the model of the DSIM in the Park frame.

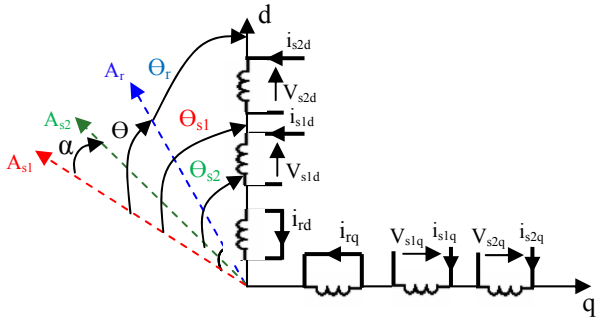


Figure 2
Representation of DSIM in the Park frame

$$\begin{aligned}
V_{s1d} &= R_{s1} I_{s1d} + \frac{d}{dt}\Phi_{s1d} - \omega_s \Phi_{s1q} \\
V_{s1q} &= R_{s1} I_{s1q} + \frac{d}{dt}\Phi_{s1q} + \omega_s \Phi_{s1d} \\
V_{s2d} &= R_{s2} I_{s2d} + \frac{d}{dt}\Phi_{s2d} - \omega_s \Phi_{s2q} \\
V_{s2q} &= R_{s2} I_{s2q} + \frac{d}{dt}\Phi_{s2q} + \omega_s \Phi_{s2d} \\
0 &= R_r I_{rd} + \frac{d\Phi_{rd}}{dt} - \omega_{sr} \Phi_{rq} \\
0 &= R_r I_{rq} + \frac{d\Phi_{rq}}{dt} + \omega_{sr} \Phi_{rd}
\end{aligned} \quad (7)$$

Where:

$$\Phi_{s1d} = L_{s1} I_{s1d} + L_m (I_{s1d} + I_{s2d} + I_{rd})$$

$$\Phi_{s1q} = L_{s1} I_{s1q} + L_m (I_{s1q} + I_{s2q} + I_{rq})$$

$$\Phi_{s2d} = L_{s2} I_{s2d} + L_m (I_{s1d} + I_{s2d} + I_{rd}) \quad (8)$$

$$\Phi_{s2q} = L_{s2} I_{s2q} + L_m (I_{s1q} + I_{s2q} + I_{rq})$$

$$\Phi_{rd} = L_r I_{rd} + L_m (I_{s1d} + I_{s2d} + I_{rd})$$

$$\Phi_{rq} = L_r I_{rq} + L_m (I_{s1q} + I_{s2q} + I_{rq})$$

L_m: Cyclic mutual inductance between stator 1, stator 2 and rotor.

The mechanical equation is given by [8]:

$$J \frac{d\Omega}{dt} = T_{em} - T_r - F_r \Omega \quad (9)$$

With :

$$T_{em} = p \frac{L_m}{L_r + L_m} \left[\Phi_{rd} (I_{s1q} + I_{s2q}) - \Phi_{rq} (I_{s1d} + I_{s2d}) \right] \quad (10)$$

3 Voltage Source Inverter Modelling

The voltage source inverter (VSI) is a static converter constituted by switching cells generally with transistors or thyristors GTO for high powers (Figure 3). The operating principle can be expressed by imposing on the machine the voltages with variable amplitude and frequency starting from a standard network 220/380 V – 50 Hz [12].Voltages at load neutral point can be given by the following expression [13]:

$$\begin{bmatrix} V_{an} \\ V_{an} \\ V_{an} \end{bmatrix} = \frac{E}{3} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} K_{11} \\ K_{12} \\ K_{13} \end{bmatrix} \quad (11)$$

This modelling for the two converters that feed the DSIM.

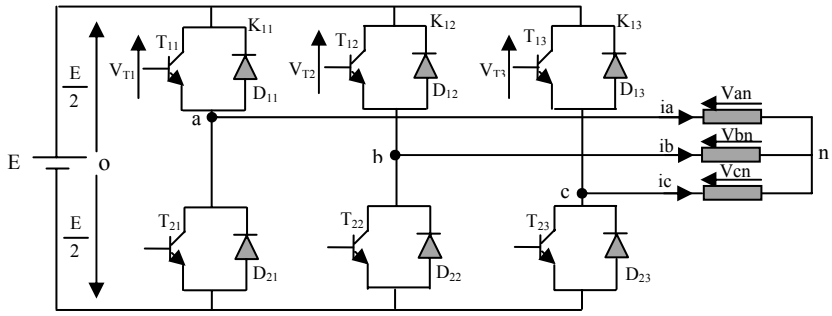


Figure 3
Voltage Source Inverter scheme

4 Field-oriented Control

The objective of space vector control is to assimilate the operating mode of the asynchronous machine at the one of a DC machine with separated excitation, by decoupling the torque and the flux control. The IRFOC consists in making $\Phi_{qr}=0$ while the rotor direct flux Φ_{dr} converges to the reference Φ_r^* [4] [14].

By applying this principle ($\Phi_{qr}=0$ and $\Phi_{dr}=\Phi_r^*$) to equations (7) (8) and (10), the finals expressions of the electromagnetic torque and slip speed are:

$$T_{em} = p\frac{L_m}{L_m + L_r}\Phi_r^*\,(I_{s1q}^* + I_{s2q}^*) \tag{12}$$

$$w_{sr}^* = \frac{R_r L_m}{(L_m + L_r)\Phi_r^*}(I_{s1q}^* + I_{s2q}^*) \tag{13}$$

The stators voltage equations are:

$$\begin{aligned}
V_{s1d}^* &= R_{s1}I_{s1d} + L_{s1}\frac{d}{dt}I_{s1d} - \omega_s^*(L_{s1}I_{s1q} + T_r\,\Phi_r^*\,w_{sr}^*)\\
V_{s1q}^* &= R_{s1}I_{s1q} + L_{s1}\frac{d}{dt}I_{s1q} + \omega_s^*(L_{s1}I_{s1d} + \Phi_r^*)\\
V_{s2d}^* &= R_{s2}I_{s2d} + L_{s2}\frac{d}{dt}I_{s2d} - \omega_s^*(L_{s2}I_{s2q} + T_r\,\Phi_r^*\,w_{sr}^*)\\
V_{s2q}^* &= R_{s2}I_{s2q} + L_{s2}\frac{d}{dt}I_{s2q} + \omega_s^*(L_{s2}I_{s2d} + \Phi_r^*)
\end{aligned} \tag{14}$$

The torque expression shows that the reference fluxes and stator currents in quadrate are not perfectly independent. Thus, it is necessary to decouple the torque and flux control of this machine by introducing new variables:

$$
\begin{aligned}
V_{s1d} &= R_{s1} I_{s1d} + L_{s1} \frac{d}{dt} I_{s1d} \\
V_{s1q} &= R_{s1} I_{s1q} + L_{s1} \frac{d}{dt} I_{s1q} \\
V_{s2d} &= R_{s2} I_{s2d} + L_{s2} \frac{d}{dt} I_{s2d} \\
V_{s2q} &= R_{s2} I_{s2q} + L_{s2} \frac{d}{dt} I_{s2q}
\end{aligned}
\tag{15}
$$

The equation system (15) shows that the stator voltages *(V_{s1d}, V_{s1q}, V_{s2d}, V_{s2q})* are directly related to the stator currents (I_{s1d}, I_{s1q}, I_{s2d}, I_{s2q}). To compensate the error introduced at decoupling time, the voltage references (V_{s1d}^{*}, V_{s2d}^{*}, V_{s1q}^{*}, V_{s2q}^{*}) at constant flux are given by:

$$
\begin{aligned}
V_{s1d}^{*} &= V_{s1d} - V_{s1dc} \\
V_{s1q}^{*} &= V_{s1q} + V_{s1qc} \\
V_{s2d}^{*} &= V_{s2d} - V_{s2dc} \\
V_{s2q}^{*} &= V_{s2q} + V_{s2qc}
\end{aligned}
\tag{16}
$$

With:

$$
\begin{aligned}
V_{s1dc} &= \omega_s^{*} (L_{s1} I_{s1q} + T_r \Phi_r^{*} w_{sr}^{*}) \\
V_{s1qc} &= \omega_s^{*} (L_{s1} I_{s1d} + \Phi_r^{*}) \\
V_{s2dc} &= \omega_s^{*} (L_{s2} I_{s2q} + T_r \Phi_r^{*} w_{sr}^{*}) \\
V_{s2qc} &= \omega_s^{*} (L_{s2} I_{s2d} + \Phi r^{*})
\end{aligned}
\tag{17}
$$

For a perfect decoupling, we add stator current regulation loops (I_{s1d}, I_{s1q}, I_{s2d}, I_{s2q}) and we obtain at their output stator voltages (V_{s1d}, V_{s1q}, V_{s2d}, V_{s2q}).The decoupling bloc scheme in voltage (Field-oriented control FOC) is given in Figure 4.

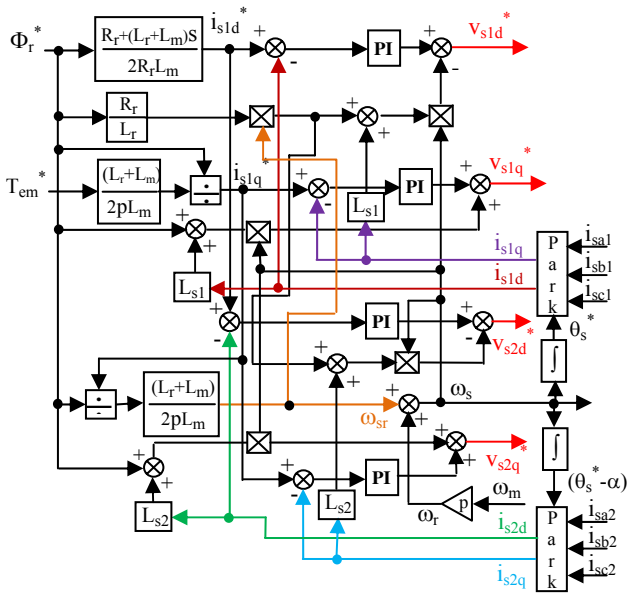


Figure 4
Decoupling bloc in voltage

5 Indirect Method Speed Regulation

The principle of this method consists in not using rotor flux magnitude but simply its position calculated with reference sizes. This method eliminates the need to use a field sensor, but only the one of the rotor speed [15].

The speed regulation scheme by IFOC of the DSIM is given in Figure 5.

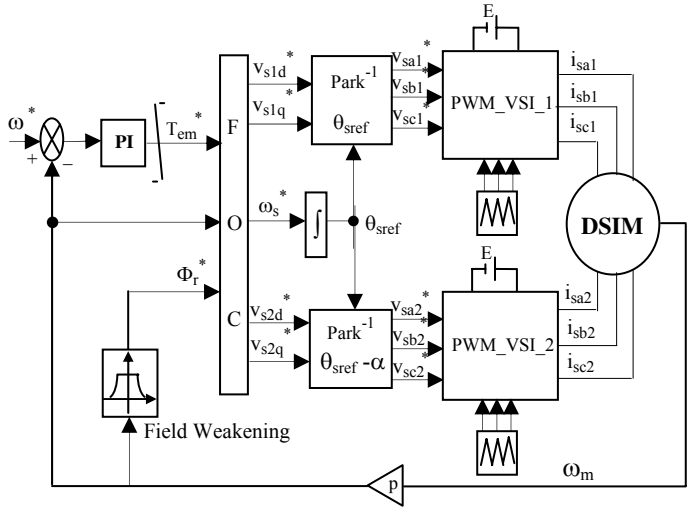


Figure 5
Indirect method speed regulation

5.1 Robustness Tests

The robustness of the indirect method speed regulation of the DSIM is visualized for two tests: the first is the varying of rotor resistance R_r ($R_r = 2R_{rn}$ at $t = 1$ s); the second increasing of inertia J ($J = 2J_n$ at $t = 1$ s).

5.2 Results and Discussion

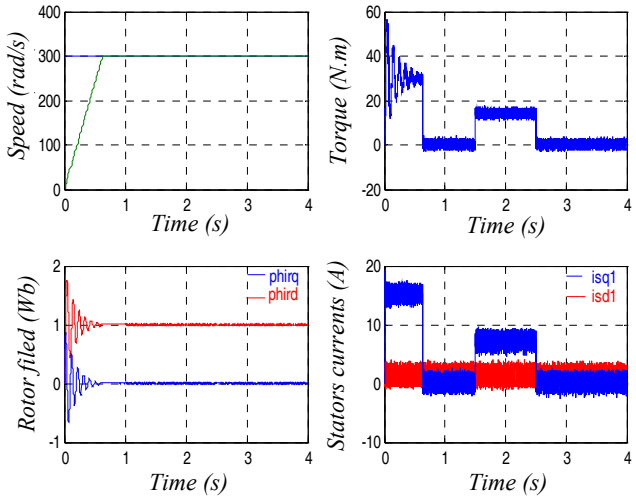


Figure 6

Indirect method speed regulation with load torque T_r = 14 N.m between [1.5 2.5] s

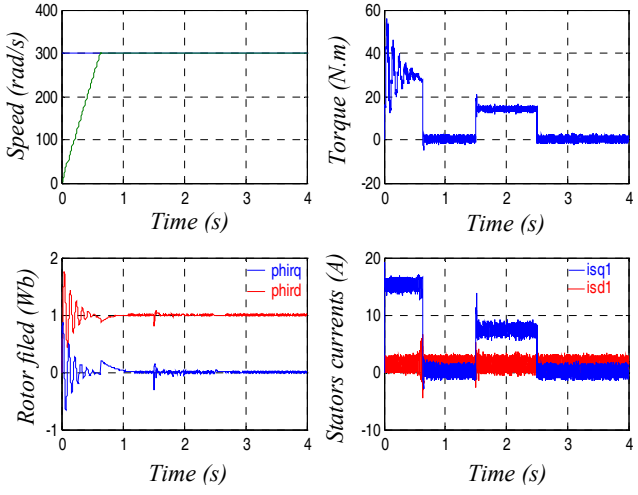


Figure 7

DSIM Comportment with rotor resistance variation (R = 2R$_n$ at t = 1 s)

The speed reaches its reference value (*300 rad/s*) after (*0.78 s*) with an overtaking of (*0.32%)* of the reference speed (Figure 6). The perturbation reject is achieved at *(0.1 s)*. The electromagnetic torque compensates the load torque and reaches at starting *(60 N.m)*.

Simulation results show the regulation sensibility with PI for rotor resistance variation. We note that the decoupling is affected. The inertia variation increases the inversion time of rotating direction (Figures 7 and 8).

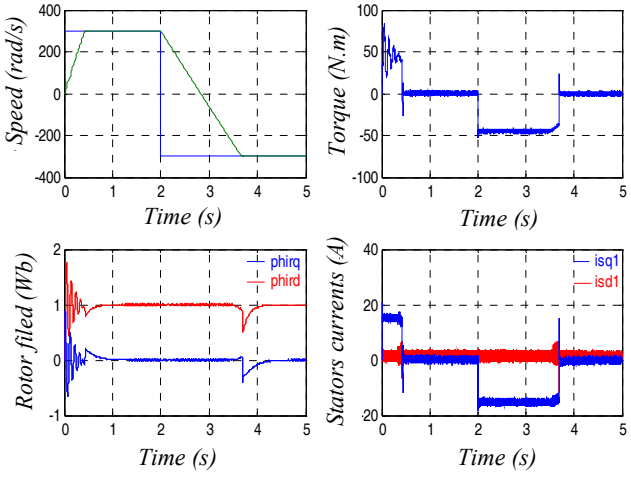


Figure 8
DSIM Comportment with inertia variation ($J=2J_n$ at t=1 s)

6 Fuzzy Logic Principle

The fuzzy logic control (FLC) has been an active research topic in automation and control theory since Mamdani proposed in 1974 based on the fuzzy sets theory of Zadeh (1965) to deal with the system control problems that are not to model [16].

The structure of a complete fuzzy control system is composed of the following blocs: Fuzzification, Knowledge base, Inference engine, Defuzzification. Figure 9 shows the structure of a fuzzy controller [16].

The Fuzzification module converts the crisp values of the control inputs into fuzzy values. A fuzzy variable has values which are defined by linguistic variables (fuzzy sets or subsets) such as: low, medium, high, big, slow . . . where each is defined by a gradually varying membership function. In fuzzy set terminology, all the possible values that a variable can assume are named the universe of discourse, and the fuzzy sets (characterized by membership function) cover the whole universe of discourse. The membership functions can be triangular, trapezoidal … [16].

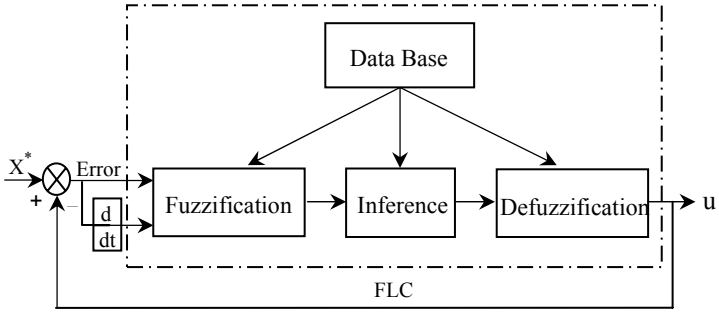


Figure 9
Fuzzy controller structure

The number of linguistic value (small negative, middle negative, positive...), represented by the membership functions can vary (for example three, five or seven). An example of Fuzzyfication is illustrated in (Figure 10) for a single variable of x with triangular membership function; the corresponding linguistic values are characterized by the symbols likewise:

NL: Negative Large.

NS: Negative Small.

ZE: Zero Equal.

PS: Positive Small.

PL: Positive Large.

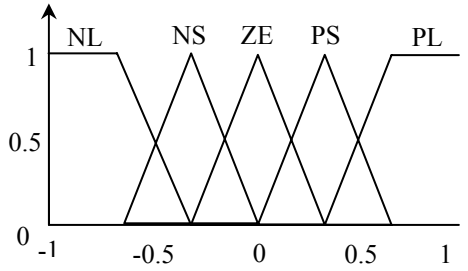


Figure 10
Fuzzyfication with five membership functions

A fuzzy control essentially embeds the intuition and experience of a human operator, and sometimes those of a designer and researcher. The data base and the rules form the knowledge base which is used to obtain the inference relation. The data base contains a description of input and output variables using fuzzy sets. The rule base is essentially the control strategy of the system. It is usually obtained from expert knowledge or heuristics; it contains a collection of fuzzy conditional

statements expressed as a set of *If-Then* rules [16]. An example of a rule type: if x1 is positive large, x2 is zero equal, then, u is positive small, where: x1 and x2 represent two input variables of the regulator likewise: the gap of variable to regulate and its variation, and u represent the control variable (output).

Table 1 presents a two linguistic variables of input; the speed error « *e* » and its variation « *de* » and the output variable « *du* ».

Table 1
Rules base for speed control

du		*e*				
		NL	*NS*	*ZE*	*PS*	*PL*
de	*NL*	*NL*	*NL*	*NS*	*NS*	*ZE*
	NS	*NL*	*NS*	*NS*	*ZE*	*PS*
	ZE	*NS*	*NS*	*ZE*	*PS*	*PS*
	PS	*NS*	*ZE*	*PS*	*PS*	*PL*
	PL	*ZE*	*PS*	*PS*	*PL*	*PL*

The mathematical procedure of converting fuzzy values into crisp values is known as 'Defuzzification'. A number of Defuzzification methods have been suggested. The choice of Defuzzification methods usually depends on the application and the available processing power. This operation can be performed by several methods of which center of gravity (or centroid) and height methods are common [16].

7 Speed Control by Fuzzy Regulator

The principle of the fuzzy speed control is the same one as that given in Figure 5, but we have changed the classical PI speed controller with a fuzzy logic controller (FLC); the other current regulators remain of classical type. The principle scheme of the speed regulation by fuzzy logic is given in Figure 11.

7.1 Results and Discussion

The speed reaches its reference value after (*0.43 s*) without overtaking. The electromagnetic torque compensates the load torque and presents at starting a value equal to *(80 N.m)* (Figure 12).

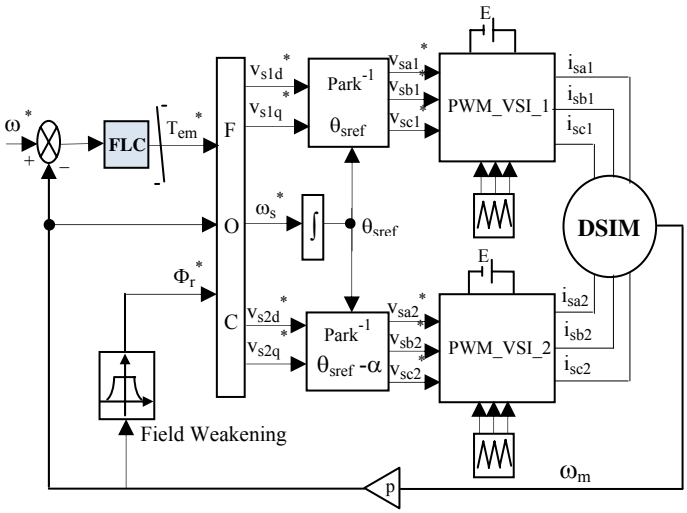


Figure 11

Indirect method fuzzy speed regulation

The simulation results show the insensitivity of fuzzy control to machine parameters variation (Increasing of R_r and J of *100%* of theirs nominal value) (Figure 13, and Figure 14). The inversion time of the speed is without overtaking, with a negative torque equal to *(45 N.m)* (Figure 14). Direct rotor field *(Φ_{dr})* follows the reference value (*1 Wb)* and the quadrature component (*Φ_{qr})* is null.

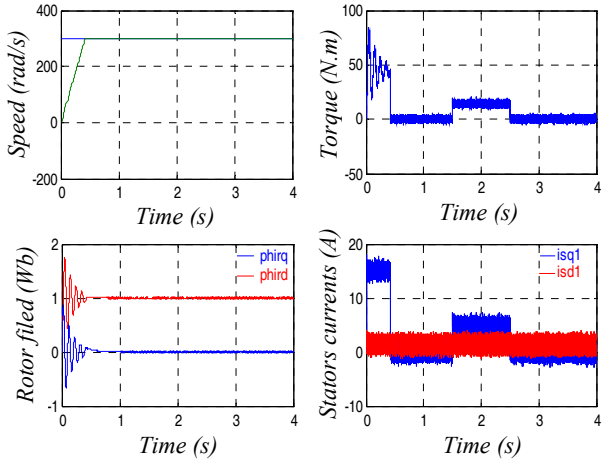


Figure 12

Speed regulation using fuzzy regulator, with applying resistant torque (T_r =14 N.m) between [1.5 2.5] s

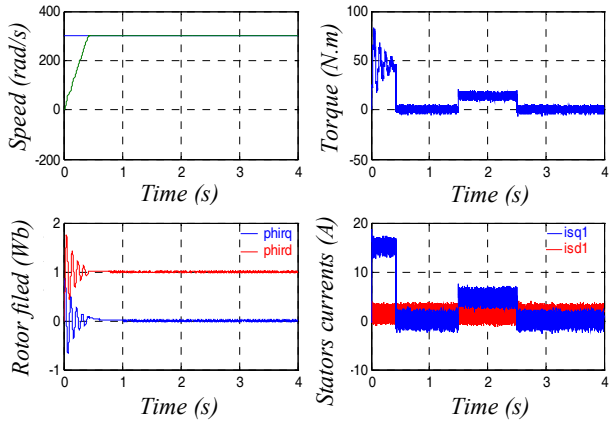


Figure 13
DSIM Comportment with rotor resistance variation (R = 2 R_n at t = 1 s)

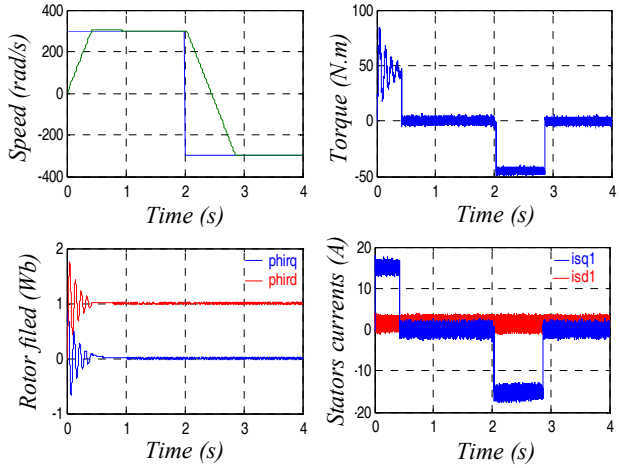


Figure 14
DSIM Comportment with inertia variation (J=2J_n at t=1 s)

Conclusions

In this paper we are presented a Field-oriented Control (FOC) of a Dual Star Induction Machine (DSIM). Two types of regulator are tested for the machine speed regulation: a PI regulator and a fuzzy regulator. The simulation results show the sensitivity of PI regulators to the parameters variation of the DSIM. The fuzzy regulator has very good dynamic performances compared with the conventional PI regulator: (a small response time, overtaking negligible, a small speed inversion time). Additionally, the robustness tests show that the fuzzy regulator is

insensitive to parameters variation (rotor resistance and inertia); this returns to the fact that the fuzzy regulator synthesis is realized without taking into account the machine model.

References

[1] R. Kianinezhad, B. Nahid, F. Betin, and G. A. Capolino: A Novel Direct Torque Control (DTC) Method for Dual Three Phase Induction Motors, IEEE, 2006

[2] Y. Zhao, T. A. Lipo: Space Vector PWM Control of Dual Three Phase Induction Machine Using Vector Space Decomposition, IEEE Trans. Ind. Appl., Vol. 31, No. 5, pp. 1100-1109, September/October 1995

[3] D. Hadiouche, H. Razik, and A. Rezzoug: On the Modeling and Design of Dual Stator Windings to Minimize Circulating Harmonic Currents for VSI Fed AC Machines, IEEE Transactions On Industry Applications, Vol. 40, No. 2, pp. 506-515, March /April 2004

[4] E. Merabet, R. Abdessemed, H. Amimeur and F. Hamoudi: Field-oriented Control of a Dual Star Induction Machine Using Fuzzy Regulators, CIP, Sétif, Algérie, 2007

[5] G. K. Singh, K. Nam and S. K. Lim: A Simple Indirect Field-oriented Control Scheme for Multiphase INDUCTION machine, IEEE Trans. Ind. Elect., Vol. 52, No. 4, pp. 1177-1184, August 2005

[6] R. Bojoi, M. Lazzari, F. Profumo and A. Tenconi: Digital Field-oriented Control for Dual Three-Phase Induction Motor Drives, IEEE Transactions on Industry Applications, Vol. 39, No. 3, May/June 2003

[7] E. M. Berkouk, S. Arezki: Modélisation et Commande d'une Machine Asynchrone Double Etoile (MASDE) Alimentée par Deux Onduleurs à Cinq Niveaux à Structure NPC, Conférence national sur le génie électrique, CNGE, Tiaret, Algérie 2004

[8] Bachir Ghalem, Bendiabdellah Azeddine: Six-Phase Matrix Converter Fed Double Star Induction Motor, Acta Polytechnica Hungarica, Vol. 7, No. 3, 2010

[9] A. Igoudjil; Y. Boudjema: Etude du changeur de fréquence à cinq niveaux à cellules imbriquées. Application à la conduite de la machine Asynchrone à Double Etoile, Mémoire d'ingéniorat de l'USTHB d'Alger, Algérie, Juin 2006

[10] D. Hadiouche: Contribution à l'étude de la machine asynchrone double étoile modélisation, alimentation et structure, Thèse de doctorat, Université Henri Poincaré, Nancy-1, Décembre 2001

[11] Z. Chen, AC. Williamson: Simulation Study of a Double Three Phase Electric Machine, International conference on Electric Machine ICEM'98, 1998, pp. 215-220

[12] S. Bazi: Contribution à la Commande Robuste d'une Machine Asynchrone par la Technique PSO, Mémoire de Magister de l'Université de Batna, Algérie, mai 2009

[13] G. Seguier, Electronique de Puissance, Editions Dunod 7ème édition. Paris, France, 1999

[14] R. N. Andriamalala, H. Razik and F. M. Sargos: Indirect-Rotor-Field-oriented-Control of a Double-Star Induction Machine Using the RST Controller, IEEE, 2008

[15] R. Sadouni: Commande par Mode Glissant Flou d'une Machine Asynchrone à Double Etoile, Mémoire de Magister, UDL de Sidi Bel Abbes, Algérie, Décembre 2010

[16] A. Aissaoui, M. Abid, H. Abid, A. Tahour and A. Zeblah: A Fuzzy Logic Controller for Synchronous Machine, Journal of Electrical Engineering, Vol. 58, No. 5, 2007, 285-290

Performance Evaluation Metrics for Software Fault Prediction Studies

Cagatay Catal

Istanbul Kultur University, Department of Computer Engineering, Atakoy Campus, 34156, Istanbul, Turkey, c.catal@iku.edu.tr

Abstract: Experimental studies confirmed that only a small portion of software modules cause faults in software systems. Therefore, the majority of software modules are represented with non-faulty labels and the rest are marked with faulty labels during the modeling phase. These kinds of datasets are called imbalanced, and different performance metrics exist to evaluate the performance of proposed fault prediction techniques. In this study, we investigate 85 fault prediction papers based on their performance evaluation metrics and categorize these metrics into two main groups. Evaluation methods such as cross validation and stratified sampling are not in the scope of this paper, and therefore only evaluation metrics are examined. This study shows that researchers have used different evaluation parameters for software fault prediction until now and more studies on performance evaluation metrics for imbalanced datasets should be conducted.

Keywords: performance evaluation; software fault prediction; machine learning

1 Introduction

Performance evaluation of machine learning-based systems is performed experimentally rather than analytically [33]. In order to evaluate analytically, a formal specification model for the problem and the system itself would be needed. This is quite difficult and inherently non-formalisable for machine learners, which are nonlinear and time-varying [40, 33]. The experimental evaluation of a model based on machine learning is performed according to several performance metrics, such as probability of detection (PD), probability of false alarm (PF), balance, or area under the ROC (Receiver Operating Characteristics) curve. As there are numerous performance metrics that can be used for evaluation, it is extremely difficult to compare current research results with previous works unless the previous experiment was performed by a researcher under the same conditions. Finding a common performance metric can simplify this comparison, but a general consensus is not yet reached. Experimental studies have shown that only a small portion of software modules cause faults in software systems. Therefore, the

majority of software modules are represented with non-faulty labels and the rest are marked with faulty labels during the modeling phase. These kinds of datasets are called imbalanced / unbalanced / skewed, and different performance metrics exist to evaluate the performance of fault prediction techniques that are built on these imbalanced datasets. The majority of these metrics are calculated by using a confusion matrix, which will be explained in later sections. Furthermore, ROC curves are very popular for performance evaluation. The ROC curve plots the probability of a false alarm (PF) on the x-axis and the probability of detection (PD) on the y-axis. The ROC curve was first used in signal detection theory to evaluate how well a receiver distinguishes a signal from noise, and it is still used in medical diagnostic tests [45].

In this study, we investigate 85 software fault prediction papers based on their performance evaluation metrics. In this paper, these metrics are briefly outlined and the current trend is reflected. We included papers in our review if the paper describes research on software fault prediction and software quality prediction. We excluded position papers that do not include experimental results. The inclusion of papers was based on the degree of similarity of the study with our fault prediction research topic. The exclusion did not take into account the publication year of the paper or methods used. We categorized metrics into two main groups: the first group of metrics are used to evaluate the performance of the prediction system, which classifies the module into faulty or non-faulty class; the second group of metrics are used to evaluate the performance of the system, which predicts the number of faults in each module of the next release of a system. Therefore, researchers can choose a metric from one of these groups according to their research objectives. The first group of metrics are calculated by using a confusion matrix. These metrics were identified through our literature review and this set may not be a complete review of all the metrics. However, we hope that this paper will cover the major metrics applied frequently in software fault prediction studies. This paper is organized as follows: Section 2 describes the software fault prediction research area. Section 3 explains the performance metrics. Section 4 presents the conclusions and suggestions.

2 Software Fault Prediction

Software fault prediction is one of the quality assurance activities in Software Quality Engineering such as formal verification, fault tolerance, inspection, and testing. Software metrics [30, 32] and fault data (faulty or non-faulty information) belonging to a previous software version are used to build the prediction model. The fault prediction process usually includes two consecutive steps: training and prediction. In the training phase, a prediction model is built with previous software metrics (class or method-level metrics) and fault data belonging to each

software module. After this phase, this model is used to predict the fault-proneness labels of modules that locate in a new software version. Figure 1 shows this fault prediction process. Recent advances in software fault prediction allow building defect predictors with a mean probability of detection of 71 percent and mean false alarm rates of 25 percent [29]. These rates are at an acceptable level and this quality assurance activity is expected to quickly achieve widespread applicability in the software industry.

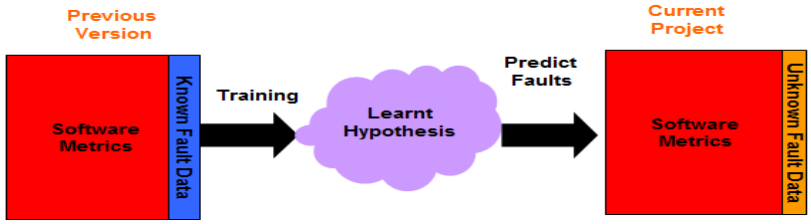


Figure 1
The software fault prediction process [34]

Until now, software engineering researchers have used Case-based Reasoning, Neural Networks, Genetic Programming, Fuzzy Logic, Decision Trees, Naive Bayes, Dempster-Shafer Networks, Artificial Immune Systems, and several statistical methods to build a robust software fault prediction model. Some researchers have applied different software metrics to build a better prediction model, but recent papers [29] have shown that the prediction technique is much more important than the chosen metric set. The use of public datasets for software fault prediction studies is a critical issue. However, our recent systematic review study has shown that only 30% of software fault prediction papers have used public datasets [5].

3 Performance Evaluation Metrics

According to the experimental studies, a majority of software modules do not cause faults in software systems, and faulty modules are up to 20% of all the modules. If we divide modules into two different types, faulty and non-faulty, the majority of modules will belong to the non-faulty class and the rest will be members of the faulty class. Therefore, datasets used in software fault prediction studies are imbalanced. Accuracy parameter cannot be used for the performance evaluation of imbalanced datasets. For example, a trivial algorithm, which marks every module as non-faulty, can have 90% accuracy if the percentage of faulty modules is 10%. Therefore, researchers use different metrics for the validation of software fault prediction models. In this section, the metrics identified during our literature review will be briefly outlined.

3.1 Metrics for Evaluation of Classifiers

Model validation for machine learning algorithms should ensure that data were transformed to the model properly and the model represents the system with an acceptable accuracy. There are several validation techniques for model validation, and the best known one is N-fold cross-validation technique. This technique divides the dataset into N number of parts, and each of them consists of an equal number of samples from the original dataset. For each part, training is performed with (N-1) number of parts and the test is done with that part. Hall and Holmes [17] suggested repeating this test M times to randomize the order each time [29]. Order effect is a critical issue for performance evaluation because certain orderings can improve / degrade performance considerably [13, 29]. In Table 1, a confusion matrix is calculated after N*M cross-validation.

Table 1
Confusion Matrix

	NO (Prediction)	YES (Prediction)
NO (Actual)	True Negative (TN) A	False Positive (FP) B
YES (Actual)	False Negative (FN) C	True Positive (TP) D

Columns represent the prediction results and rows show the actual class labels. Faulty modules are represented with the label YES, and non-faulty modules are represented with the label NO. Therefore, diagonal elements (TN, TP) in Table 1 show the true predictions and the other elements (FN, FP) reflect the false predictions. For example, if a module is predicted as faulty (YES) even though it is a non-faulty (NO) module, this test result is added to the B cell in the table. Therefore, number B is incremented by 1. After M*N tests, A, B, C, and D values are calculated. In the next subsections, these values (A, B, C, D) will be used to compute the performance evaluation metrics.

3.1.1 PD, PF, Balance

The equations used to calculate probability of detection (PD), probability of false alarm (PF), and accuracy metrics are shown in Formulas 1, 2, and 3 respectively. The other term used for PD metric is recall.

$$\text{PD} = \text{recall} = \frac{D}{C + D} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{PF} = \frac{B}{A + B} = \frac{FP}{FP + TN} \quad (2)$$

$$\text{Accuracy} = \frac{A + D}{(A + B + C + D)} \tag{3}$$

Balance metrics is the Euclidean distance between (0, 1) and (PF, PD) points. PD, accuracy, and balance parameters should be maximized and PF metrics should be minimized for fault predictors. Menzies et al. [29] reported that the best fault predictors provide 71% of PD and 25% of PF values. They used PD, PF, and balance parameters as the performance evaluation metrics in this study. Turhan and Bener [38] showed that the independence assumption in the Naive Bayes algorithm is not detrimental with principal component analysis (PCA) pre-processing, and they used PD, PF, and balance parameters in their study.

3.1.2 G-mean1, G-mean2, F-measure

Some researchers use G-mean1, G-mean2, and F-measure metrics for the evaluation of prediction systems, which are built on imbalanced datasets. Formulas 6, 7, and 8 show how to calculate these measures, respectively. Formula 4 is used for precision parameter and True Negative Rate (TNR) is calculated by using Formula 5. The formula for recall is given in Formula 1.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{True Negative Rate (TNR)} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{G-mean1} = \sqrt{\text{Precision} * recall} \tag{6}$$

$$\text{G-mean2} = \sqrt{recall * TNR} \tag{7}$$

$$\text{F-measure} = \frac{2\ (\text{recall} * \text{Precision})}{\text{recall} + \text{Precision}} \tag{8}$$

Ma et al. [26] used G-mean1, G-mean2, and G-mean3 to benchmark several machine learning algorithms for software fault prediction. They sorted algorithms according to their performance results for each metric and marked the top three algorithms for each metric. They identified the algorithm that provides G-mean1, G-mean2, and F-measure values in the top three. According to this study, Balanced Random Forests is the best algorithm for software fault prediction problems. Furthermore, they reported that boosting, rule set, and single tree classifiers do not provide acceptable results even though these algorithms have been used in literature. Koru and Liu [23] evaluated the performance of classifiers according to the F-measure value. Arisholm et al. [1] built 112 fault prediction models and compared them according to precision, recall, accuracy, Type-I error, Type-II error, and AUC parameters. The following sections will explain AUC, Type-I, and Type-II errors.

3.1.3 AUC

Receiver Operating Characteristics (ROC) curves can be used to evaluate the performance of software fault prediction models. In signal detection theory, a ROC curve is a plot of the sensitivity vs. (1-specificity) and it can also be represented by plotting the probability of false alarm on the X-axis and the probability of detection on the Y-axis. This curve must pass through the points (0, 0) and (1, 1) [29]. The important regions of ROC curve are depicted in Figure 2. The ideal position on ROC curve is (0, 1) and no prediction error exists at this point. A line from (0, 0) to (1, 1) provides no information and therefore the area under ROC curve value (AUC) must be higher than 0.5. If a negative curve occurs, this means that the performance of this classifier is not acceptable. A preferred curve is shown in Figure 2. The cost-adverse region has low false alarm rates and is suitable if the validation & verification budget is limited. In the risk-adverse region, even though the probability of detection is high, the probability of false alarm is also high, and, therefore, cost is higher. For mission critical systems, a risk-adverse region is chosen and for business applications, a cost-adverse region is more suitable.

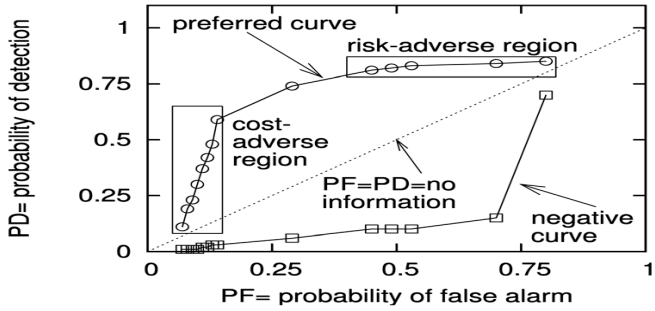


Figure 2
Regions of ROC curve [29]

The area under the ROC curve (AUC) is a widely used performance metric for imbalanced datasets. Ling et al. [25] proposed the usage of an AUC parameter to evaluate the classifiers and showed that AUC is much more appropriate than accuracy for balanced and imbalanced datasets. Van Hulse et al. [39] applied an AUC metric to evaluate the performance of 11 learning algorithms on 35 datasets. In addition to this metric, they also utilized Kolmogorov-Smirnov (K/S) statistics [18], geometric mean, F-measure, accuracy, and true positive rate (TPR) parameters. They stated that AUC and K/S parameters measure the capability of the classifier and showed AUC values of algorithms in tables. Li et al. [24], and Chawla and Karakoulas [7] used an AUC parameter for unbalanced datasets. For a competition in "11th Pacific-Asia Conference on Knowledge Discovery and Data Mining" (PAKDD2007), performance evaluations for an imbalanced dataset were

performed according to AUC values, and the model that provides 70.01% of AUC value was selected as the best algorithm. Catal and Diri [6] examined nine classifiers and compared their performance according to the AUC value. Mende and Koschke [28] used an AUC parameter to compare classifiers on thirteen datasets.

3.1.4 Sensitivity, Specificity, J Coefficient

El-Emam et al. [11] proposed the usage of the J parameter to measure the accuracy of binary classifiers in software engineering. The J coefficient was first used in medical research [41]; it is calculated by using sensitivity and specificity parameters. El-Emam et al. [12] used the J coefficient for performance evaluation of algorithms. Sensitivity, specificity, and the J parameter are calculated by using Formulas 9, 10, and 11 respectively.

$$\text{Sensitivity} = \frac{D}{C + D} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{Specificity} = \frac{A}{A + B} = \frac{TN}{TN + FP} \tag{10}$$

$$\text{J} = \text{sensitivity} + \text{specificity} - 1 \tag{11}$$

Sensitivity measures the ratio of actual faulty modules which are correctly identified and specificity measures the ratio of non-faulty modules which are correctly identified.

3.1.5 Type-I error, Type-II error, Overall Misclassification Rate

Some researchers used Type-I error and Type-II error parameters to evaluate the performance of fault prediction models [42, 15, 35, 1, 2]. The overall misclassification rate parameter takes care of these two error parameters. Formulas 12, 13, and 14 are used to calculate the Type-I error, Type-II error, and overall misclassification rate respectively. If a non-faulty module is predicted as a faulty module, a Type-I error occurs, and if a faulty module is predicted as a non-faulty module, a Type-II error occurs. A Type-II error is more significant than a Type-I error because faulty modules cannot be detected in that case.

$$\text{Type-I error} = \frac{B}{A + B + C + D} = \frac{FP}{TN + FP + FN + TP} \tag{12}$$

$$\text{Type-II error} = \frac{C}{A + B + C + D} = \frac{FN}{TN + FP + FN + TP} \tag{13}$$

$$\text{Overall misclassification rate} = \frac{C + B}{A + B + C + D} \tag{14}$$

3.1.6 Correctness, Completeness

Correctness and completeness parameters were used for the evaluation of fault prediction models [4, 44, 9, 16, 27]. Formulas 15 and 16 show how to calculate correctness and completeness measures.

$$\text{Correctness} = \frac{D}{B + D} = \frac{TP}{FP + TP} \tag{15}$$

$$\text{Completeness} = \frac{D}{C + D} = \frac{TP}{FN + TP} \tag{16}$$

3.1.7 FPR, FNR, Error

The false positive rate (FPR), the false negative rate (FNR), and error parameters are used for performance evaluation [41, 43].

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

$$FNR = \frac{FN}{FN + TP} \tag{18}$$

$$Error = \frac{FN + FP}{TP + FP + FN + TN} \tag{19}$$

These three performance indicators should be minimized, but there is a trade-off between the FPR and FNR values. The FNR value is much more crucial than the FPR value because it quantifies the detection capability of the model on fault-prone modules and high FNR values indicate that a large amount of fault-prone modules cannot be captured by the model before the testing phase. Therefore, users will probably encounter these problems in the field and the nondetected faulty modules can cause serious faults or even failures. On the other hand, a model having high FPR value will simply increase the testing duration and test efforts.

3.1.8 Cost Curve

Jiang et al. [19] recommended adopting cost curves for the fault prediction performance evaluation. This is the first study to propose cost curves for performance evaluation of fault predictors and it is not yet widely used. However, it is not easy to determine the misclassification cost ratio and the selection of this parameter can make the model debatable. Drummond and Holte [10] proposed cost curves to visualize classifier performance and the cost of misclassification was included in this technique. Cost curve plots the probability cost function on the x-axis and the normalized expected misclassification cost on the y-axis. Details of this approach will not be given here due to length considerations, and

readers may apply to papers by Jiang et al. [19] or Drummond and Holte [10] to learn the details of this approach. However, this approach is not widely used in the software fault prediction research area.

3.2 Metrics for the Evaluation of Predictors

Some researchers predict the number of faults in each module of the next release of a system, and the modules' classification is performed according to the number of faults. Modules are sorted in descending order with respect to the number of faults, and the modules which should be tested rigorously are identified according to the available test resources.

3.2.1 Average Absolute Error, Average Relative Error

Average absolute error and average relative error parameters have been used as performance evaluation metrics by numerous researchers for software quality prediction studies [22, 20, 21, 14]. Formulas 20 and 21 show how to calculate average absolute error (AAE) and average relative error (ARE) parameters, respectively. The actual number of faults is represented by y_i, while y_j represents the predicted number of faults, and n shows the number of modules in the dataset.

$$\text{AAE} = \frac{1}{n}\sum_{i=1}^{n} \left| y_i - y_j \right| \tag{20}$$

$$\text{ARE} = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{y_i - y_j}{y_i + 1} \right| \tag{21}$$

3.2.2 R^2

R^2 measures the power of correlation between predicted and actual number of faults [37]. Another term for this parameter is the *coefficient of multiple determination,* and this parameter is widely used in studies that predict the number of faults. Many researchers have applied this parameter in their studies [9, 36, 8, 37, 31, 3]. This metric's value should be near to 1 if the model is to be acceptable, and Formula 22 is used to calculate this parameter. The actual number of faults is represented by y_i, $\hat{y}_i$ represents the predicted number of faults, and $\overline{y}$ shows the average of fault numbers.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{\mathbf{y}})^2} \tag{22}$$

Conclusions

The use of different evaluation parameters prevents the software engineering community from easily comparing research results with previous works. In this study, we investigated 85 fault prediction papers based on their performance evaluation metrics and categorized these metrics into two main groups. The first group of metrics are used for prediction systems that classify modules into a faulty or non-faulty module and the second group of metrics are applied to systems that predict the number of faults in each module of the next release of a system. This study showed that researchers have used numerous evaluation parameters for software fault prediction up to now, and the selection of common evaluation parameters is still a critical issue in the context of software engineering experiments. From the first group, the most common metric for software fault prediction research is the area under ROC curve (AUC). The AUC value is only one metric and it is not a part of the metric set. Therefore, it is easy to compare several machine learning algorithms by using this parameter. In addition to AUC, PD, and PF, balance metrics are also widely used. In this study, we suggest using the AUC value to evaluate the performance of fault prediction models. From the second group of metrics, R^2 and AAE / ARE can be used to ensure the performance of the system that predicts the number for faults. We suggest the following changes in software fault prediction research:

- *Conduct more studies on performance evaluation metrics for software fault prediction.* Researchers are still working on finding a new performance evaluation metric for fault prediction [19], but we need more research in this area because this software engineering problem is inherently different than the other imbalanced dataset problems. For example, it is not easy to determine the misclassification cost ratio (Jiang et al., 2008) and therefore, using cost curves for evaluation is still not an easy task.
- *Apply a widely used performance evaluation metric.* Researchers would like to be able to easily compare their current results with previous works. If the performance metric of previous studies is totally different than the widely used metrics, that makes the comparison difficult.

References

[1] E. Arisholm, L. C. Briand, and E. B. Johannessen, A Systematic and Comprehensive Investigation of Methods to Build and Evaluate Fault Prediction Models, Journal of Systems and Software 83 (1) (2010) 2-17

[2] Y. Bingbing, Y. Qian, X. Shengyong, and G. Ping, Software Quality Prediction Using Affinity Propagation Algorithm, Proc. IJCNN 2008, 2008, pp. 1891-1896

[3] D. Binkley, H. Feild, D. Lawrie, M. Pighin, Software Fault Prediction Using Language Processing, Proc. Testing: Academic and industrial Conference Practice and Research Techniques - MUTATION, TAICPART-MUTATION 2007, Washington, DC, 2007, pp. 99-110

[4] L. C. Briand, V. Basili, C. Hetmanski, Developing Interpretable Models with Optimized Set Reduction for Identifying High Risk Software Components, IEEE Transactions on Software Engineering 19 (11) (1993) 1028-1044

[5] C. Catal, B. Diri, A Systematic Review of Software Fault Prediction Studies, Expert Systems with Applications 36 (4) (2009a) 7346-7354

[6] C. Catal, B. Diri, Investigating the Effect of Dataset Size, Metrics Sets, and Feature Selection Techniques on Software Fault Prediction Problem, Inf. Sci. 179 (8) (2009b) 1040-1058

[7] N. V. Chawla, G. J. Karakoulas, Learning from Labeled and Unlabeled Data. An Empirical Study across Techniques and Domains, Journal of Artificial Intelligence Research, 23 (2005) 331-366

[8] G. Denaro, Estimating Software Fault-Proneness for Tuning Testing Activities, Proc. 22nd Int'l Conf. on Soft. Eng., Limerick, Ireland, 2000, pp. 704-706

[9] G. Denaro, M. Pezzè, S. Morasca, Towards Industrially Relevant Fault-Proneness Models, International Journal of Software Engineering and Knowledge Engineering 13 (4) (2003) 395-417

[10] C. Drummond, R. C. Holte, Cost Curves: An Improved Method for Visualizing Classifier Performance, Machine Learning 65 (1) (2006) 95-130

[11] K. El-Emam, S. Benlarbi, N. Goel, Comparing Case-based Reasoning Classifiers for Predicting High Risk Software Components, Technical Report, National Research Council of Canada, NRC/ERB-1058, Canada, 1999

[12] K. El-Emam, W. Melo, J. C. Machado, The Prediction of Faulty Classes Using Object-oriented Design Metrics, Journal of Systems and Software 56 (1) (2001) 63-75

[13] D. Fisher, L. Xu, N. Zard, Ordering Effects in Clustering, Proc. Ninth Int'l Conf. Machine Learning, 1992

[14] K. Gao, T. M. Khoshgoftaar, A Comprehensive Empirical Study of Count Models for Software Fault Prediction, IEEE Transactions for Reliability 56 (2) (2007) 223-236

[15] P. Guo, M. R. Lyu, Software Quality Prediction Using Mixture Models with EM Algorithm, Proc. 1st Asia-Pacific Conference on Quality Software, Hong Kong, 2000, pp. 69-80

[16] T. Gyimothy, R. Ferenc, I. Siket, Empirical Validation of Object-oriented Metrics on Open Source Software for Fault Prediction, IEEE Transactions on Software Engineering 31 (10) (2005) 897-910

[17] M. Hall, G. Holmes, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, IEEE Trans. Knowledge and Data Eng. 15 (6) (2003) 1437-1447

[18] D. J. Hand, Good Practice in Retail Credit Scorecard Assessment, Journal of the Operational Research Society 56 (2005) 1109-1117

[19] Y. Jiang, B. Cukic, T. Menzies, Cost Curve Evaluation of Fault Prediction Models, Proc. 19th International Symposium on Software Reliability Engineering, IEEE Computer Society, Washington, DC, 2008, pp. 197-206

[20] T. M. Khoshgoftaar, N. Seliya, Tree-based Software Quality Estimation Models for Fault Prediction. 8th IEEE Symposium on Software Metrics. Ottawa, Canada, 2002, pp. 203-215

[21] T. M. Khoshgoftaar, E. Geleyn, K. Gao, An Empirical Study of the Impact of Count Models Predictions on Module-Order Models, Proc. 8th Int'l Symp. on Software Metrics, Ottawa, Canada, 2002, pp. 161-172

[22] T. M. Khoshgoftaar, N. Seliya, N. Sundaresh, An Empirical Study of Predicting Software Faults with Case-based Reasoning, Software Quality Journal 14 (2) (2006) 85-111

[23] A. G. Koru, H. Liu, An Investigation of the Effect of Module Size on Defect Prediction Using Static Measures, Proc. Workshop on Predictor Models in Software Engineering, St. Louis, Missouri, 2005, pp. 1-5

[24] X. Li, L. Wang, E. Sung, AdaBoost with SVM-based Component Classifiers, Eng. Appl. Artif. Intelligence 21 (5) (2008) 785-795

[25] C. X. Ling, J. Huang, H. Zhang, AUC: A Better Measure than Accuracy in Comparing Learning Algorithms, Canadian Conference on Artificial Intelligence, Halifax, Canada, 2003, pp. 329-341

[26] Y. Ma, L. Guo, B. Cukic, A Statistical Framework for the Prediction of Fault-Proneness, Advances in Machine Learning Application in Software Engineering, Idea Group Inc., 2006, pp. 237-265

[27] A. Mahaweerawat, P. Sophasathit, C. Lursinsap. Software Fault Prediction Using Fuzzy Clustering and Radial Basis Function Network, Proc. International Conference on Intelligent Technologies, Vietnam, 2002, pp. 304-313

[28] T. Mende, R. Koschke, Revisiting the Evaluation of Defect Prediction Models, Proc. 5th international Conference on Predictor Models in Software Engineering, Vancouver, Canada, 2009, pp. 1-10

[29] T. Menzies, J. Greenwald, A. Frank, Data Mining Static Code Attributes to Learn Defect Predictors, IEEE Transactions on Software Engineering 32 (1) (2007) 2-13

[30] S. Misra, Evaluation Criteria for Object-oriented Metrics, Acta Polytechnica Hungarica 8(5) (2011) 109-136

[31] A. P. Nikora, J. C. Munson, Building High-Quality Software Fault Predictors, Software Practice and Experience 36 (9) (2006) 949-969

[32] O. T. Pusatli, S. Misra, Software Measurement Activities in Small and Medium Enterprises: An Empirical Assessment, Acta Polytechnica Hungarica 8(5) (2011) 21-42

[33] F. Sebastiani, Machine Learning in Automated Text Categorization, ACM Comput. Surv. 34 (1) (2002) 1-47

[34] N. Seliya, Software Quality Analysis with Limited Prior Knowledge of Faults. Graduate Seminar, Wayne State University, Department of Computer Science, 2006

[35] N. Seliya, T. M. Khoshgoftaar, Software Quality Estimation with Limited Fault Data: A Semi-supervised Learning Perspective, Software Quality Journal 15 (3) (2007) 327-344

[36] M. M. Thwin, T. Quah, Application of Neural Networks for Software Quality Prediction Using Object-oriented Metrics, Proc. 19th International Conference on Software Maintenance, Amsterdam, The Netherlands, 2003, pp. 113-122

[37] P. Tomaszewski, L. Lundberg, H. Grahn, The Accuracy of Early Fault Prediction in Modified Code, Proc. 5th Conference on Software Engineering Research and Practice in Sweden, Västerås, Sweden, 2005, pp. 57-63

[38] B. Turhan, A. Bener, Analysis of Naive Bayes' Assumptions on Software Fault Data: An Empirical Study, Data Knowl. Eng. 68 (2) (2009) 278-290

[39] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, Experimental Perspectives on Learning from Imbalanced Data, 24th International Conference on Machine Learning, Corvalis, Oregon, 2007, pp. 935-942

[40] H. Wang, Y. Chen, and Y. Dai, A Soft Real-Time Web News Classification System with Double Control Loops, Proc. WAIM 2005, 2005, pp. 81-90

[41] W. Youden, Index for Rating Diagnostic Tests, Cancer 3 (1) (1950) 32-35

[42] X. Yuan, T. M. Khoshgoftaar, E. B. Allen, K. Ganesan, An Application of Fuzzy Clustering to Software Quality Prediction, Proc. 3rd IEEE Symposium on Application-Specific Systems and Software Engineering Technology, Richardson, Texas, 2000, p. 85

[43] S. Zhong, T. M. Khoshgoftaar, N. Seliya. Unsupervised Learning for Expert-based Software Quality Estimation, Eighth IEEE International Symposium on High Assurance Systems Engineering, 2004, pp. 149-155

[44] Y. Zhou, H. Leung, Empirical Analysis of Object-oriented Design Metrics for Predicting High and Low Severity Faults, IEEE Transactions on Software Engineering 32 (10) (2006) 771-789

[45] M. J. Zolghadri, E. G. Mansoori, Weighting Fuzzy Classification Rules Using Receiver Operating Characteristics (ROC) Analysis, Inf. Sci. 177 (11) (2007) 2296-2307

Route Choice Estimation Based on Cellular Signaling Data

Tamás Tettamanti[1], Hunor Demeter[2], István Varga[1]

[1] Department of Control and Transport Automation, Budapest University of Technology and Economics
Bertalan L. u. 2, H-1111 Budapest, Hungary
E-mail: tettamanti@mail.bme.hu; ivarga@mail.bme.hu

[2] Nokia Siemens Networks, CTO Research SWS M2M Group
Köztelek u. 6, H-1092 Budapest, Hungary
E-mail: hunor.demeter@nsn.com

Abstract: The rapid growth of mobile phone communications has induced novel and emerging technologies in the past decades. Signaling data of cellular phones can be used as valuable information for state-of-the-art traffic applications especially in urban areas. The paper focuses on the applicability for estimating the traces of traveling mobiles in the transportation system. By observing anonym mobile phones, typical route choices can be determined and thus other traffic characteristics can be obtained. These information may serve as efficacious basis for transportation forecasting and planning, traffic control measures or even for real-time route guidance.

Keywords: cellular network; Voronoi tessellation; handover; location update; mobile phone; estimation; route choice; traffic assignment

1 Introduction

If a mobile phone (terminal) is moving in its cellular network different types of signaling events are generated due to the principles of the radio frequency based telecommunication. The location data of terminals can be exploited for various goals. One of the potential applications is represented by road traffic related use. A large number of patents and papers relevant to this matter has been published in recent decades. One of the earliest paper investigating mobile location technique is presented by [1]. Without attempting to provide a detailed survey, we refer to some interesting research articles which are related to traffic applications based on radio signaling data. The studies are mostly focusing on two research directions. On the one hand, measurement and estimation technologies of the most important traffic parameters are investigated: speed, traffic flow, travel time, origin-destination (OD) traffic flow. On the other hand, the publications concern

methodologies for Intelligent Transportation Systems (ITS): navigation service, route guidance, incident detection, road monitoring, road use charging, traffic information service. [2] and [3] suggest methods of tracing a mobile by evaluating subsequent signal strength measurements to different base stations of GSM system by applying proper estimation methods. [4] discusses several map matching techniques (semi-deterministic algorithms, probabilistic, and fuzzy-logic-based algorithms) to reveal the trajectory of a vehicle from the point of view of the positioning sensors. [5] investigates the travel time estimation problem on freeways by using cell phones as probes. The potential of OD matrix estimation from mobile phone data is demonstrated by [6] with pilot study results. [7] proposes a method for real-time estimation of traffic flow and density on motorways. [8] provides a review of measurement concepts of traffic variables by using GSM network. [9] and [10] present advanced methodologies for cellular data based ITS applications.

These papers discuss the whole range of potential traffic applications. The problem of suitable route choice estimation without propagation modeling and 3D environment models, however, is not investigated. In urban road traffic network the route choice behavior of vehicles is a real challenge. Even if a series of proper signaling data can be obtained from the mobile network concerning a given traveler the exact match with the road links is not straightforward. The paper proposes a practical approach to estimate route choices of travelers by associating mobile phone data with traffic assignment.

The layout of the paper is as follows. First, some relevant preliminaries of the cellular communication system are provided. After the introductory sections the route choice estimation problem of traveling mobile is introduced. The filtering of signaling events and traffic assignment technique is described preparing the advised solution. Path estimation is proposed by deviation calculation concerning the results of appropriate route assignment. An application example with simulation study is provided on the basis of real-world test measurements. Finally, summary concludes the paper.

2 Preliminaries

The section aims to introduce the basic elements used in the paper. Handover (HO) and location area update (LAU) signaling events are briefly presented. Furthermore, the Voronoi tessellation based modeling for mobile network is also described.

2.1 Cellular Signaling Events in Mobile Network

A GSM telecommunication network consists of radio cells. A cell represents geographic coverage area of a base transceiver station. The main task of a base

station is to realize wireless communication between the terminal and the network. A set of base stations (tens or even hundreds of them) belongs to a location area (LA). A traveling mobile phone generates several types of signaling events depending on its behavior. The cell change and LA change are the most frequent and thus the most significant occurrences generating HO and LAU events. A HO is reported when a phone call in progress is redirected from the current cell and its used channel to a new cell and channel. In inactive (idle) mode a LAU is generated when a terminal enters a new location area on location area crossing. The terminal must also perform LAU periodically even if it is in the same area.

The main potential of the signaling events is the opportunity to use them without any additional infrastructure. As the terminals automatically report HO/LAU events to the communication system, the cell phone operator may exploit these data (as server-side processing). The possible applications of the HO/LAU reports have been widely investigated, e.g. in the papers referred in Section 1. To exploit the mobility events, first of all, the monitoring and collecting of signaling data must be realized. Several monitoring methods exist, which can be classified as active, passive, client-side, server-side, or third party solutions. A detailed survey on monitoring possibilities is provided by [11].

Nevertheless, it has to be noted that HO events are generated only if the mobile phone is active, i.e. a call is in progress. In idle mode, only LAU reports are generated, which are also valuable but infrequent data. Location area crossings are rather far from each other (even tens of kilometers). Thus, in urban area several phone users (subscribers) may be not observable. A possible solution can be the artificial generation of HO/LAU events by the operator at given locations.

Naturally, the legal aspects must also be taken into consideration and made clear when mobility data of phone users are aggregated. Technically, the anonymity of subscribers can be ensured for privacy protection.

2.2 Modeling Cellular Phone Network

The use of mobile signaling data requires a suitable model of the cellular network. This is a big challenge as the cells do not form circles. Furthermore, they overlap each other to a greater or lesser extent especially in urban areas. Nevertheless, a practical approach is the application of the Voronoi tessellation method [12], [13]. If the transceiver antenna coordinates are available the geographic coverage area of the cells can be estimated. The Voronoi partitioning represents a special decomposition (convex polygons) of a given plane with n generator points. Each generator is associated with a corresponding Voronoi cell. Practically, a Voronoi polygon is the set of all points where each point is closer to its generator than to any other generator in the given plane. Obviously, the locations of transceiver antennas may serve as the generating points for the tessellation. Figure 1 shows the basic concept of the modeling with Voronoi cells. The circles depict the coverage area of the base stations.

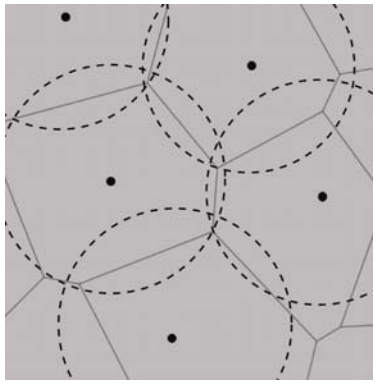


Figure 1
Mobile cells modeled by Voronoi polygons

The area of influence of each base station can be easily delimited without the knowledge of the physical aspects of the communication (e.g. radio wave propagation, interference, etc.). The locations of antennas are only taken into consideration. Therefore, Voronoi tessellation may provide an efficient method to trace the spatial locomotion of a call, e.g. by observing mobile signaling events.

Nevertheless, the drawbacks of this approach must also be mentioned. The Voronoi based modeling applies two assumptions. It is hypothesized that the power of all transceiver antennas are identical in the given plane. Furthermore, the effects of spatial objects (such as buildings, vehicles, etc.) are neglected. A possible solution to improve the accuracy of the tessellation was proposed by [12] and [14]. If the power threshold assigned for each base station is known, a multiplicatively weighted Voronoi diagram can be created. Thus, the different transmitted signal strengths of the antennas can be taken into account.

3 Route Choice Estimation

The route choice estimation problem is the main aim of this paper. If the Voronoi diagram of the mobile network is known and mobile signaling events are captured, a potential solution is provided by applying the route assignment (an efficient tool of traffic engineering practice).

3.1 The Problem of the Accurate Determination of Mobile Traces

HO report contains the codes of the previous and the new cell as well. In case of LA change LAU event gives information on previous LA and the new cell. In Section 2.2 the Voronoi tessellation was proposed as an efficient modeling

approach to identify cells in the mobile network. If HO/LAU signaling data are available and it is possible to aggregate them, the routes of the subscribers can be estimated. The estimation, however, is not straightforward especially in dense road traffic network, i.e. in urban areas. Figure 2 shows an example in Budapest by assuming a given path between two locations in District 9 and 11.

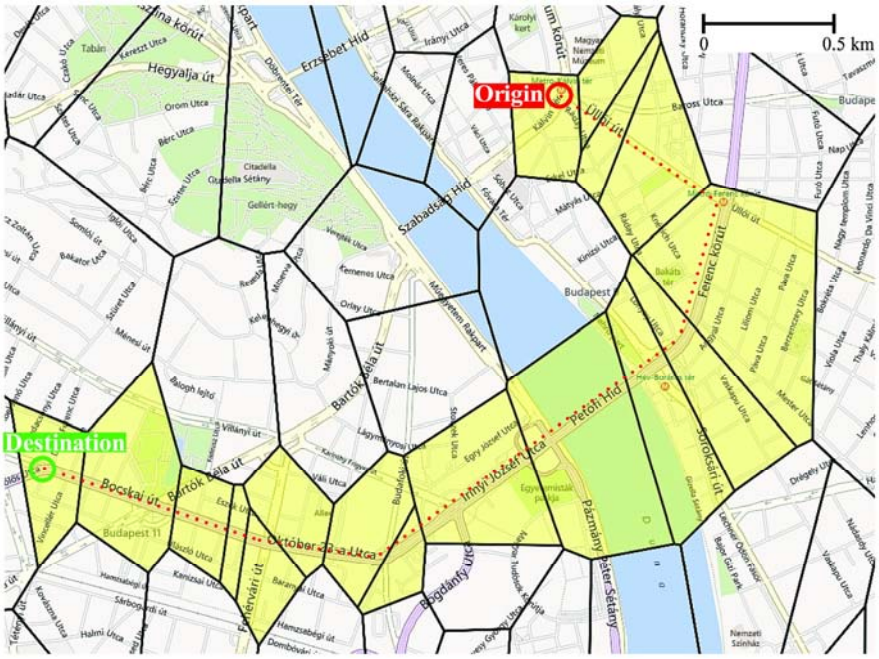


Figure 2

The ideal sequence of Voronoi polygons generated by a traveling terminal

The base station data (used here and in the following parts) were gathered from the OpenCellID project's database (www.opencellid.org). The Voronoi tessellation (black lines) was computed concerning one of the Hungarian mobile operator. The tessellation was carried out easily by Quantum GIS, software of another open source project (www.qgis.org). The red dotted line shows the trip of the subscriber. Figure 2 represents that one cell may cover several road links. Therefore, if only HO/LAU reports are available the trace of the subscribers cannot be unambiguously determined.

3.2 Filtering HO Events

Beyond the problem introduced in the previous section other difficulties must be considered as well. Figure 2 demonstrates an ideal case where a clear sequence of Voronoi cells can be observed in time and space. Nevertheless, this is not the case in practice for several reasons, e.g. interference effect, varying signal strength,

load of the network. Due to the intra-cell effect the cells may show incoherent order. Intra-cell HO is a special signaling event when only the used channel changes during the handover and the cell remains the same. Another potential problem may be caused by short-period cell changes which occur the ping-pong HO effects. In case of ping-pong HO the terminal is "bouncing" quickly between two base stations back and forth. A similar effect is the rapid cell change representing short-period cell transitions among multiple cells. This may be possibly caused by reason of faulty coverage of base stations.

As consequence, it is worth to filter incoherent HO events mentioned above. Filtered handovers result in better sequence of cells which is more convenient for route choice estimation as well. Figure 3 displays the chain of Voronoi cells after filtering for the OD pair denoted by Fig. 2. The data was gathered by the Mobile Quality Analyzer (MQA) software developed by Nokia Siemens Networks. MQA logged all HO events during the trip. Therefore, Fig. 3 shows a realistic version of the ideal case displayed by Fig. 2.

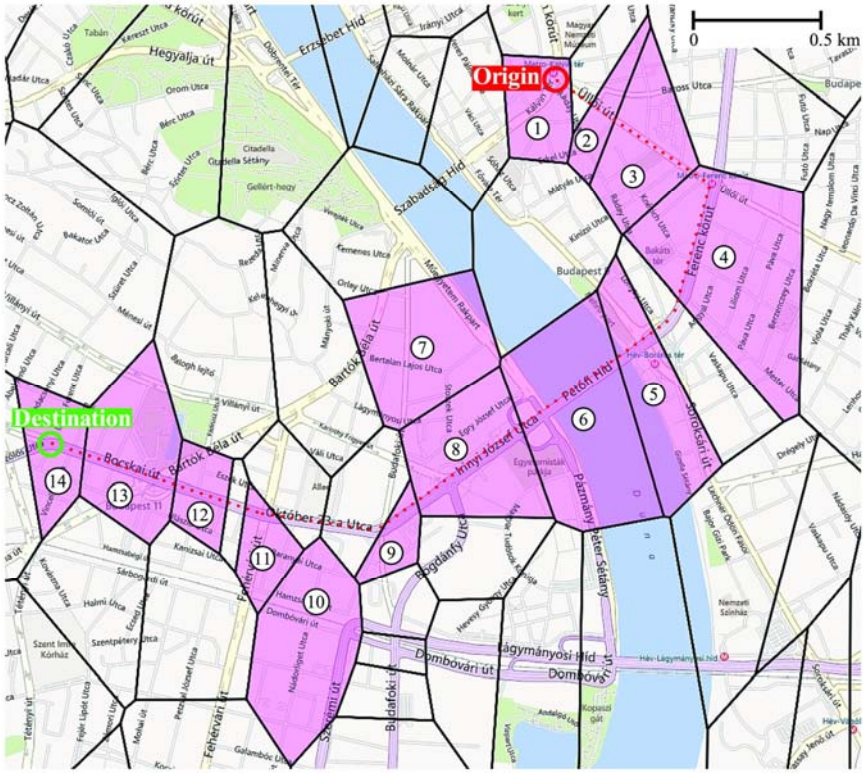


Figure 3
The chain of filtered Voronoi polygons generated by a moving terminal

Some remarks concerning the measurement results must be done. The generated sequence of the mobile phone is semi-deterministic. For several reasons the handovers may be produced diversely on the same path. Typically, moving obstacles (e.g. bus in the other lane), weather conditions, or load of the network may affect the wave propagation resulting in different signaling events. Nevertheless, the main characteristic of the sequence is able to reflect the trip.

3.3 Traffic Assignment

A potential solution to the problem introduced in Section 3.1 is the application of traffic assignment methods to assist the mobile signaling data. The classic four-step road traffic forecasting consists of trip generation, trip distribution, modal split, and route assignment [15], [16]. Trip generation aims to define the total number of trips generated by the set of N origins (zones or nodes) and attracted by the set of M destinations within the study area. The trip distribution step intends to predict the number of trips concerning each OD pair, i.e. construct the trip (OD) matrix. The estimation of modal split gives the percentage of travelers using a particular type of transportation (public transport, private cars, etc.). The last element of the traffic forecasting process is the distribution of traffic among all related origins and destinations, i.e. selection of potential paths in the available transportation network. Several assignment algorithms have been elaborated since the 1950s. The basic classification of traffic assignments may be done by the ability of considering time-varying parameters. Static or time-independent assignment deals with traffic demand assumed to be essentially constant. On the other hand, dynamic assignment methodologies consider time-dependent parameters (e.g. dynamic OD matrix, congestion effects).

The basic input parameter for all assignment methods is the OD matrix which can be derived by household surveys or estimated by using traffic counts and historical OD data. Trip matrix may be determined in a static as well as a dynamic fashion (e.g. [17]) corresponding to the available information. Furthermore, several additional input parameters (static and dynamic) can be involved into the process depending on the applied assignment.

3.4 Route Choice Estimation Assisted by Traffic Assignment

The investigated problem in this section can be summarized as follows.

- A cellular phone network is given, represented by Voronoi diagram.
- Mobile signaling events (HO/LAU) can be aggregated and evaluated concerning the network.
- Method is sought for determining the most probable path for each moving subscriber.

By knowing the signaling sequence of a terminal, the corresponding Voronoi cells can be determined, e.g. as displayed in Fig. 3. Naturally, at least two signaling events are needed to create a chain of cells. The first and last HO/LAU reports define the origin and destination cells of the trip forming an OD pair. Thus, the final problem is to determine the exact path of the given OD pair. Nevertheless, due to the scales of the Voronoi cells a plethora of possible routes could be found. The cell size may be several thousands of square meters. Therefore, a traffic assignment must be carried out concerning the given OD pair.

The first step of assignment is to define the trip matrix which is generally available as result of a proper trip distribution. In assignment procedures the origins and destinations are defined as zones or nodes (e.g. road intersections). In our case it is assumed that such OD information is available. Moreover, the Voronoi polygons can be identified as OD zones. Hence, the corresponding zones are defined as the first and last cells generated by the subscriber. Then, traffic assignment can provide the distribution of traffic, i.e. the number of trips on the assigned paths. The next step is to find the most probable route among the assigned ones. The cells are indexed by $i = 1,2,\ldots m$ and the assigned routes by $j = 1,2,\ldots n$. A simple and practical solution is proposed by measuring the squared deviations between each cell and the assigned paths. The measurement means the calculation of the shortest Euclidean distance ($d_{i,j}$) from the centroid of the cell (i) to the given path (j) such that the drawn line is perpendicular to the tangent of the path. A simple example is shown by Fig. 4.

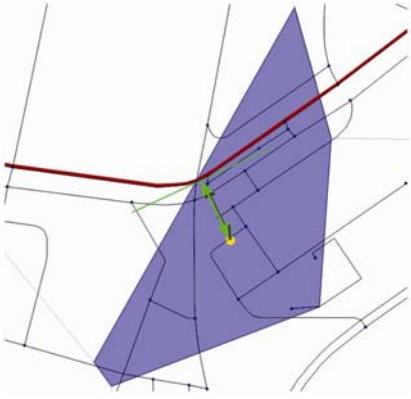


Figure 4
Shortest distance between the centroid and the assigned route

The centroid (denoted by yellow circle) is determined as the center of gravity of the Voronoi cell (blue polygon), which usually does not coincide with the location of the base station. The shortest distance ($d_{i,j}$) to the assigned route (red line) is represented by the green arrow.

For each potential path j all square deviations can be calculated and summarized as follows:

$$D_j = \sum_{i=1}^{m} d_{i,j}^2 \,. \tag{1}$$

Thus, to characterize the digressions compared to the assigned routes the squared deviations are used. Finally, the lowest scalar represents the most likely route choice of the subscriber, i.e. $\min(D_j),\ \ j = 1,2,\ldots n$. This means that the chosen path among the n assigned routes best fits the Voronoi cells.

The proposed method can be applicable in offline or online fashion with appropriate sample interval. Practically, any traffic assignment method can be used to assist the estimation problem. If real-time functioning is realized the use of time-varying OD matrices (if available) can be advantageous. The dynamic traffic assignment may result in more accurate estimates on route choices.

4 Application Example

To validate the elaborated estimation method an application example is presented based on real-world data and test measurements.

4.1 Simulation Environment

The OD pair shown by Fig. 3 is investigated. By using real-world network data and MQA measurements the route choice estimation procedure is realized as detailed in Section 3.4. VISUM software [18] was applied as simulation environment. VISUM is a macroscopic traffic simulator commonly used in practice for diverse problems related to road traffic analysis, forecasting, and planning. The simulator allows GIS-based data management. Hence, the GIS files of the Voronoi diagram of the test field were simply imported into the simulator. VISUM provides several assignment methods. For the test case the incremental assignment was applied. The incremental approach takes two following basic assumptions: each trip-maker chooses a path which minimizes the travel time, the travel time on a link depends on the link flow. Incremental assignment divides the traffic volumes into fractions to be assigned in steps. In each step, a part of the total demand is assigned based on the all-or-nothing assignment, i.e. assuming constant link travel times.

4.2 Simulation Results

Figures 5-8 represent the results of the incremental assignment calculated in VISUM between the origin and destination zones of the test area. As already

mentioned in the previous section the incremental method assigns a given fraction of the demand in each step. In our case the entire traffic demand of the OD pair was proportionally distributed into four parts. Thus, four distinct paths were assigned by VISUM.

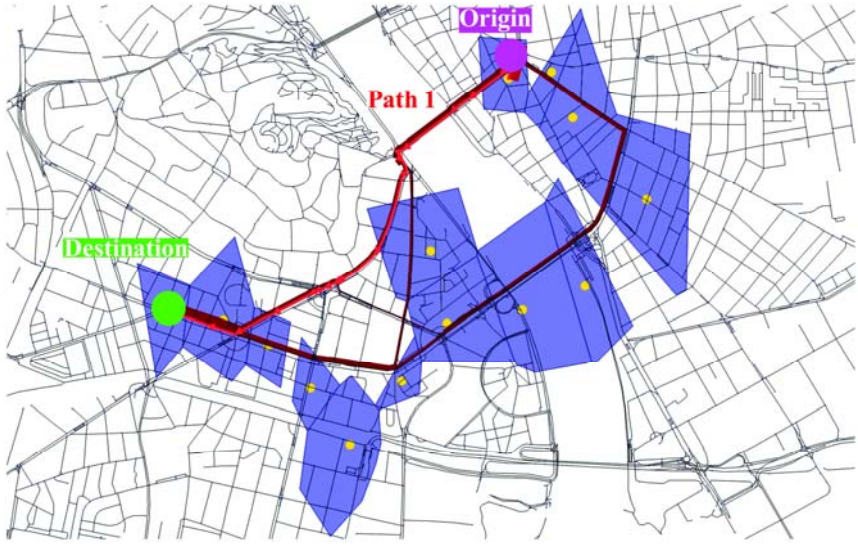


Figure 5
Path 1 of the test OD pair

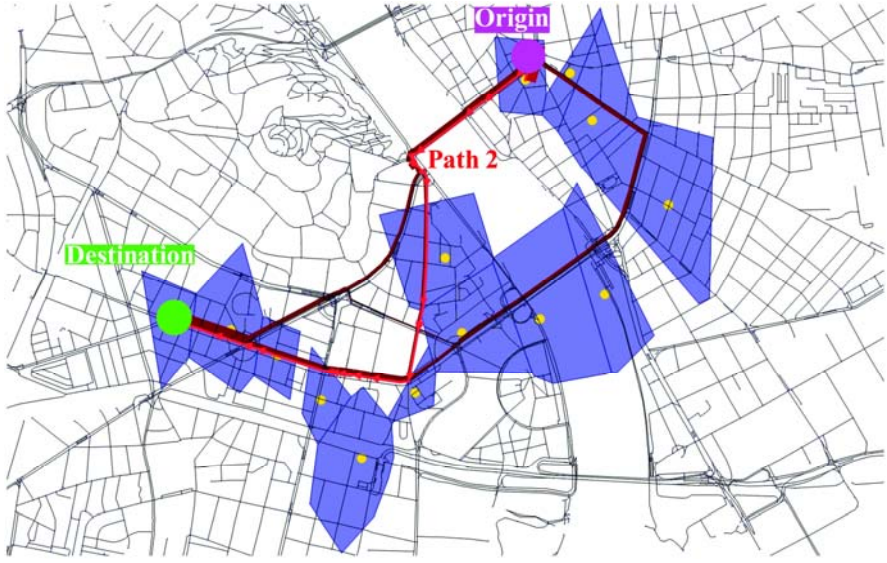


Figure 6
Path 2 of the test OD pair

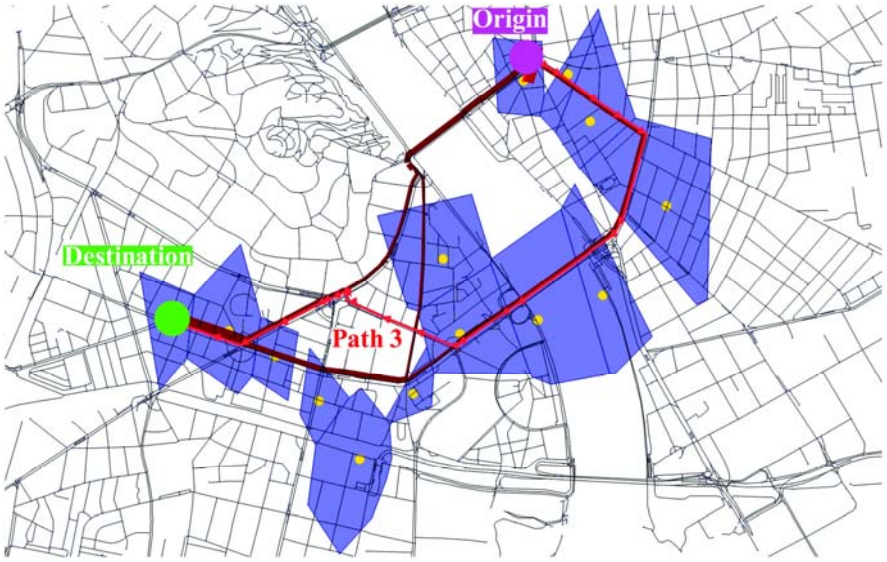


Figure 7
Path 3 of the test OD pair

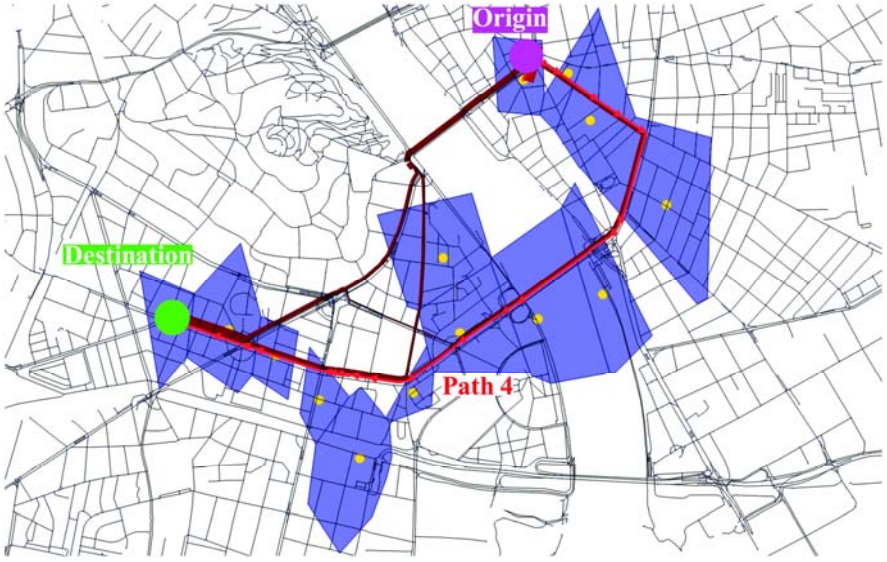


Figure 8
Path 4 of the test OD pair

The assignment result demonstrates the most likely routes of the test OD pair. Table 1 contains the main characteristics of the assigned paths.

Table 1
Distances and travel times of the assigned paths

Path j	Distance (km)	Travel time (min)
1	2.8	7
2	3.5	8
3	4.1	10
4	4.1	8

By assuming that the traveling terminal uses one of these paths, one is able to estimate the route choice of the subscriber. Naturally, the number of potential paths can be increased with the appropriate settings of the traffic assignment. To find the route Eq. 1 must be calculated for all assigned paths ($n = 3$) and cells ($m = 14$). Table 2 contains the calculation results for each assigned route.

Table 2
Sum of deviations and squared deviations for each assigned path

Path j	$\sum_{i=1}^{m} d_{i,j}$	Ratio compared to the lowest value	$D_j = \sum_{i=1}^{m} d_{i,j}^2$	Ratio compared to the lowest value
1	6695	3.4	5130283	8.2
2	4519	2.3	3097741	4.9
3	2907	1.5	1305201	2.1
4	1958	1	627122	1

In our example the final result is obvious (see Fig. 8). Nevertheless, the calculation gives the same result, i.e. Path 4 was chosen by the subscriber as $\min(D_j) = D_4$ was found the lowest among the potential routes.

Additionally, to demonstrate the advantage of the squared evaluation the normal summation ($\sum_{i=1}^{m} d_{i,j}$) is calculated. The ratios of the sums, by taking the lowest value as basis, are also represented for better interpretation. It can be observed that the squared approach reflects the deviations in a more intense way. This property is not apparent in our test case. Nevertheless, it can be advantageous if the applied traffic assignment is configured to produce a more extensive result, i.e. dealing with several potential paths.

Conclusion

The proposed estimation technique was presented through a simple example by measuring a single terminal between one origin and one destination. Nevertheless,

the method can be extended to cover a whole transportation network with multiple OD pairs and traveling terminals. Hence, the technique may constitute a starting step for further potential applications in the traffic engineering field. The future goal of the authors is to investigate the inverse application of the method. If the available mobile signaling events are statistically significant route assignment techniques may be improved. Namely, an advanced dynamic assignment can be provided by using the estimated route choice data of the traveling mobile subscribers.

Moreover, the proposed method can be extended. The route choice estimation may be applied similarly for arbitrary wireless network with regard to the specifics of the given system, e.g. WI-FI, RFID, Bloutooth, etc.

Acknowledgement

The test measurements were carried out by MQA software provided by Nokia Siemens Networks, which is gratefully acknowledged. The work is connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project. This project is supported by the Hungarian Scientific Research Fund (OTKA) through grant No. CNK 78168 and by János Bolyai Research Scholarship of the Hungarian Academy of Sciences which are acknowledged.

References

[1] M. Hata and T. Nagatsu. Mobile Location Using Signal Strength Measurements in a Cellular System. IEEE Transactions on Vehicular Technology, 29(2):245-252, 1980

[2] M. Hellebrandt, R. Mathar, and M. Scheibenbogen. Estimating Position and Velocity of Mobiles in Cellular Radio Networks. IEEE Transaction on Vehicular Technology, 46(1):65-71, 1997

[3] Martin Hellebrandt and Rudolf Mathar. Location Tracking of Mobiles in Cellular Radio Networks. IEEE Transactions on Vehicular Technology, 48:1558-1562, 1999

[4] Y. Zhao. Vehicle Location and Navigation Systems. Artech House Inc, 1997

[5] L. Ygnace, J., C. Drane, Y. B. Yim, and R. de Lacvivier. Travel Time Estimation on the San Francisco Bay Area Network Using Cellular Phones as Probes. Technical report, University of California, Berkeley, 2000

[6] J. White and I. Wells. Extracting Origin Destination Information from Mobile Phone Data. IEE Conference Publications, 2002(CP486):30-34, 2002

[7] V. Astarita, R. L. Bertini, S. d'Elia, and G. Guido. Motorway Traffic Parameter Estimation from Mobile Phone Counts. European Journal of Operational Research, 175(3):1435-1446, 2006

[8] N. Caceres, J. P. Wideberg, and F. G. Benitez. Review of Traffic Data Estimations Extracted from Cellular Networks. IET Intelligent Transport Systems, 2(3):179-192, 2008

[9] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-Time Urban Monitoring Using Cell Phones: A case study in Rome. IEEE Transactions on Intelligent Transportation Systems, 12(1):141-151, 2011

[10] D. Valerio. Road Traffic Information from Cellular Network Signaling. Technical Report FTW-TR-2009-003, Telecommunications Research Center Vienna, 2009

[11] D. Valerio, A. D'Alconzo, F. Ricciato, and W. Wiedermann. Exploiting Cellular Networks for Road Traffic Estimation: A survey and a research roadmap. IEEE 69th Vehicular Technology Conference, pp. 1-5, 2009

[12] A.-E. Baert and D. Semé. Voronoi Mobile Cellular Networks: Topological Properties. In Third International Symposium on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks, pp. 29-35, 2004

[13] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering Individual and Collective Human Dynamics from Mobile Phone Records. Journal of Physics A: Mathematical and Theoretical, 41(22):224015, 2008

[14] J. N. Portela and M. S. Alencar. Cellular Network as a Multiplicatively Weighted Voronoi Diagram. In IEEE Consumer Communications and Networking Conference, pp. 913-917, 2006

[15] J. de Dios Ortúzar and L. G. Willumsen. Modelling Transport. ISBN: 978-0471861102. Wiley, 2001

[16] D. C. Gazis. Traffic Theory. ISBN:1402070950. Springer, 2002

[17] J. J. Brandriss. Estimation of Origin-Destination Flows for Dynamic Traffic Assignment. PhD thesis, MIT, 2001

[18] VISUM 12 User Manual. PTV AG, Karlsruhe, Germany, 2011

Predictive Control Algorithms Verification on the Laboratory Helicopter Model

Anna Jadlovská, Štefan Jajčišin

Department of Cybernetics and Artificial Intelligence, Faculty of Electrotechnics and Informatics, Technical University in Košice
Letná 9, 042 00 Košice, Slovak Republic
anna.jadlovska@tuke.sk, stefan.jajcisin@tuke.sk

Abstract: The main goal of this paper is to present the suitability of predictive control application on a mechatronic system. A theoretical approach to predictive control and verification on a laboratory Helicopter model is considered. Firstly, the optimization of predictive control algorithms based on a state space, linear regression ARX and CARIMA model of dynamic systems are theoretically derived. A basic principle of predictive control, predictor deducing and computing the optimal control action sequence are briefly presented for the particular algorithm. A method with or without complying with required constraints is introduced within the frame of computing the optimal control action sequence. An algorithmic design manner of the chosen control algorithms as well as the particular control structures appertaining to the algorithms, which are based on the state space or the input-output description of dynamic systems, are presented in this paper. Also, the multivariable description of the educational laboratory model of the helicopter and a control scheme, in which it was used as a system to be controlled, is mentioned. In the end of the paper, the results of the real laboratory helicopter model control with the chosen predictive control algorithms are shown in the form of time responses of particular control closed loop's quantities.

Keywords: ARX; CARIMA model; generalized predictive control; state model-based predictive control; optimization; quadratic programming; educational Helicopter model

1 Introduction

The predictive control based on the dynamic systems' models is very popular at present. In the framework of this area, several different approaches or basic principle modifications exist. They can be divided in terms of many different attributes, but mainly in terms of the used model of the dynamic system. Predictive control based on the transfer functions is called generalized predictive control (GPC) [1]. Also predictive control based on the state space dynamic systems models exists [2]. Both approaches use linear models. Until now some

modifications of basic predictive control principle have been created. Some of them, and other important issues like stability are mentioned in [3] or [4].

In this paper, we are engaged in a theoretical derivation of some predictive control methods based on the linear model of controlled system, and in preparing them for subsequent algorithmic design and verification on a real laboratory helicopter model from Humusoft [5], which serves as an educational model for identification and control algorithms verification at the Department of Cybernetics and Artificial Intelligence at the Faculty of Electrotechnics and Informatics at the Technical University in Košice. Particularly, we are concerned with the predictive control algorithm that is based on the state space description of the MIMO (**M**ulti **I**nput **M**ulti **O**utput) system [6], [7] and with generalized predictive algorithm, which is started from linear regression ARX (**A**uto**R**egressive e**X**ogenous) [9] model of SISO (**S**ingle **I**nput **S**ingle **O**utput) systems. Thus it is based on the input-output description of dynamic systems. Moreover, we apply it to the GPC algorithm based on the CARIMA (**C**ontrolled **A**utoRegressive **I**ntegrated **M**oving **A**verage) [8] model of the SISO system. All of mentioned control methods differ, whether in the derivation manner of predictor based on the system's linear model or in computing the optimal control action sequence. This implies that we have to take an individual approach to programming them. We programmed the mentioned predictive control algorithms as Matlab functions, which compute the value of the control action on the basis of particular input parameters. This allowed us to use a modular approach in control. We used these functions in specific control structures, which we programmed as scripts in simulation language Matlab. We carried out communication with a laboratory card connected to helicopter model through Real-Time Toolbox functions [13]. The acquired results will be presented as the time responses of the optimal control action and reference trajectory tracking by output of the Helicopter model.

2 Theoretical Base of Predictive Control

We introduce some typical properties of predictive control in this section. Next we deal with mathematical fundamentals of predictive control algorithm design and their programming as a Matlab functions.

Predictive control algorithms constitute optimization tasks and in general they minimize a criterion

$$J = \sum_{i=N_1}^{N_p} Q_e(i)\left[\hat{y}(k+i) - w(k+i)\right]^2 + \sum_{i=1}^{N_u} R_u(i)\left[u(k+i-1)\right]^2 , \tag{1}$$

where $u(k)$ is a control action, $\hat{y}(k)$ is a predicted value of controlled output and $w(k)$ denotes a reference trajectory. Values N_1 and N_p represent a prediction

horizon. According to [8], the value N_1 should be at least $d+1$, where d is a system transport delay, in our case we suppose $N_1 = 1$. The positive value N_u denotes a control horizon, on which the optimal control action $u(k)$ is computed, whereby $N_u \leq N_p$. If the degrees of freedom of control action reduction is used in the predictive control algorithms, it is valid that $N_u < N_p$ [6]. Values $Q_e(i)$ and $R_u(i)$ constitute weighing coefficients of a deviation between system output and reference trajectory on the prediction horizon and control action on the control horizon. Next we will assume that $Q_e(i)$ and $R_u(i)$ are constant on the entire length of the prediction and control horizon, and thus they do not depend on variable i. In terms of weighing coefficients, their single value is not important, but mainly ratio $\lambda = R_u / Q_e$.

In some cases, the rate of control action $\Delta u(k)$ is used instead its direct value $u(k)$ in the criterion (1), whereby the control obtains an integration character, which results in the elimination of the steady state control deviation in the control process [6].

It is necessary to know the reference trajectory $w(k)$ on the prediction horizon in each sample instant in predictive control algorithms. The simplest reference trajectory example is a constant function with desired value w_0. According to [8], it is the more preferred form of smooth reference trajectory, whose initial value equals to the current system output and comes near to the desired value w_0 through a first order filter. This approach is carried out by equations

$$w(k) = y(k); \quad w(k+i) = \alpha w(k+i-1) + (1-\alpha)w_0, \quad \text{for } i = 1, 2, \ldots, \tag{2}$$

where the parameter $\alpha \in \langle 0;1 \rangle$ expresses the smoothness of reference trajectory. If $\alpha \to 0$ then the reference trajectory has the fastest slope, and, on the contrary, if $\alpha \to 1$ the slowest. In the case when the reference trajectory is unknown, it is customary to use the so-called *random walk* [6], where $w(k+1) = w(k) + \xi(k)$, whereby $\xi(k)$ is a white noise.

Predictive control algorithm design can be divided in two phases:

1) predictor derivation (dynamic system behavior prediction),
2) computing the optimal control by criterion minimization.

The advantage of predictive control consists in the possibility to compose different constraints (of control action, its rate or output) in computing the optimal control action sequence. Most commonly, it is carried out by quadratic programming. In our case we used a *quadprog* function, which is one of functions in the *Optimization Toolbox* in Matlab and computes a vector of optimal values $\boldsymbol{u}$ by formula

$$\min_{\boldsymbol{u}} \frac{1}{2}\boldsymbol{u}^T \boldsymbol{H}\boldsymbol{u} + \boldsymbol{g}^T \boldsymbol{u} \text{ , subject to } \boldsymbol{A}_{con}\boldsymbol{u} \le \boldsymbol{b}_{con} . \tag{3}$$

The basic syntax for using the *quadprog* function to compute the vector $\boldsymbol{u}$ is

$$\boldsymbol{u} = quadprog(\boldsymbol{H}, \boldsymbol{g}, \boldsymbol{A}_{con}, \boldsymbol{b}_{con}) , \tag{4}$$

whereby a form of matrix $\boldsymbol{H}$ (Hessian) and row vector $\boldsymbol{g}^T$ (gradient) depends on the predictive control algorithm used. It is necessary to compose the matrix $\boldsymbol{A}_{con}$ and the vector $\boldsymbol{b}_{con}$ in compliance with required constraints. The detail specification of their structures will be presented next, particularly with each algorithm. If the combination of more constraints is needed, the matrix $\boldsymbol{A}_{con}$ and the vector $\boldsymbol{b}_{con}$ are created by matrices and vectors for concrete constraint, which are organized one after another. The *quadprog* function also permits entering the function's output constraints as the function's input parameters, which abbreviates entering required constraints in matrix $\boldsymbol{A}_{con}$ and vector $\boldsymbol{b}_{con}$.

In the framework of the control process using predictive control algorithms, the so-called *receding horizon computing* is performed [6]. The point is that the sequence of optimal control action $\boldsymbol{u}_{opt} = \left[u_{opt}(k) \quad \cdots \quad u_{opt}(k + N_u - 1) \right]$ is computed on the entire length of control horizon at each sample instant k, but only the first unit $u_{opt}(k)$ is used as the system input $u(k)$.

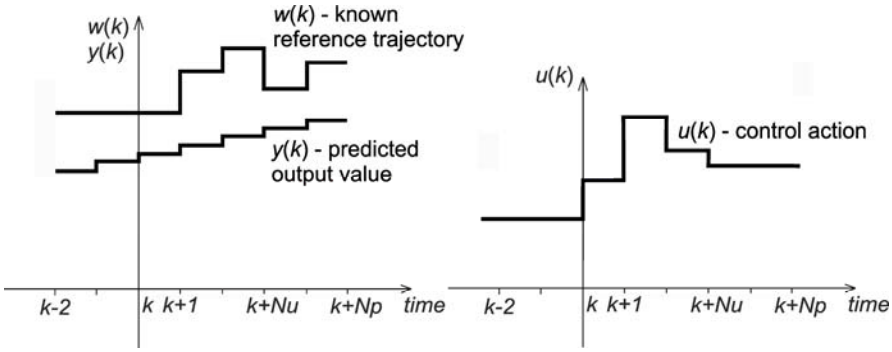


Figure 1
Predictive control principle

As we used the receding horizon principle in control algorithms, computing the optimal control action sequence $\boldsymbol{u}_{opt}$ is evaluated in conformity with Fig. 1. The authors of this paper designed the next procedure, which is carried out within the frame of every control process step k:

step 1: the assigning of the reference trajectory $\boldsymbol{w}$ on the prediction horizon,
step 2: the detection of the actual state $\boldsymbol{x}(k)$ or output $\boldsymbol{y}(k)$ of system in specific sample instant,

step 3: the prediction of system response on the prediction horizon based on the actual values of optimal control action $u_{opt}(k)$ and state $\boldsymbol{x}(k)$ or input $\boldsymbol{u}(k)$ and output $\boldsymbol{y}(k)$ in previous sample instants without influence of next control action, the so-called system free response,

step 4: computing the sequence of optimal control action $\boldsymbol{u}_{opt}$ by the criterion J minimization with known parameters N_1, N_p, N_u, $Q(i)$ and $R(i)$,

step 5: using $u_{opt}(k)$ as a system input.

We implemented the introduced five steps into every type of predictive control algorithm with which we have been concerned, and which are introduced in the next particular parts of this paper.

3 State Space Model-based Predictive Control Algorithm Design

The State-space Model based Predictive Control (SMPC) algorithm predicts a system free response on the basis of its current state. The control structure using the SMPC algorithm is depicted in Fig. 2, where $\boldsymbol{w}$ is a vector of the reference trajectory on the prediction horizon, $\boldsymbol{y}_0$ is a system free response prediction on the prediction horizon, $\boldsymbol{x}(k)$ denotes a vector of current values of state quantities, $\boldsymbol{u}(k)$ represents a vector of control action, $\boldsymbol{d}(k)$ is a disturbance vector and $\boldsymbol{y}(k)$ is a vector of the system outputs, i.e the controlled quantities.

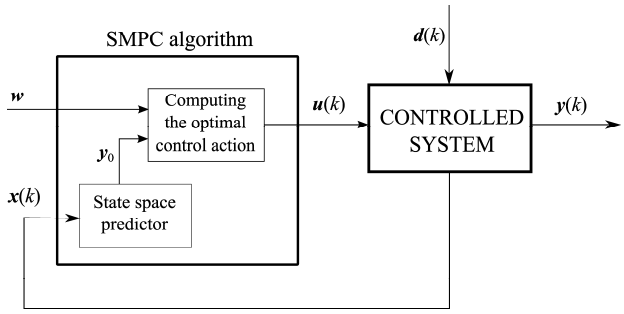


Figure 2
Control structure with SMPC algorithm

The SMPC algorithm belongs to the predictive control algorithms family, which use a state space description of MIMO dynamic systems for system output prediction (provided that there is no direct dependence between the system input and output)

$$\begin{aligned} \boldsymbol{x}(k+1) &= \boldsymbol{A}_d \boldsymbol{x}(k) + \boldsymbol{B}_d \boldsymbol{u}(k) \\ \boldsymbol{y}(k) &= \boldsymbol{C}\boldsymbol{x}(k) \end{aligned}, \tag{5}$$

($\boldsymbol{A_d}$ is a matrix of dynamics with dimension $nx \times nx$, $\boldsymbol{B_d}$ is an input matrix of dimension $nx \times nu$, $\boldsymbol{C}$ is an output matrix of dimension $ny \times nx$, $\boldsymbol{x}(k)$ is a vector of state quantities with length nx, $\boldsymbol{u}(k)$ is a vector of inputs with length nu, $\boldsymbol{y}(k)$ is a vector of outputs with length ny, where variables *nx*, *nu* and *ny* constitute the number of state quantities, inputs and outputs of dynamic system) and compute the sequence of optimal control action $\boldsymbol{u}(k)$ by the minimization of criterion

$$J_{\mathrm{MPC}} = \sum_{i=N_1}^{N_p} Q_e \left[\hat{\boldsymbol{y}}(k+i) - \boldsymbol{w}(k+i)\right]^2 + \sum_{i=1}^{N_u} R_u \left[\boldsymbol{u}(k+i-1)\right]^2 . \tag{6}$$

The coefficients in the criterion (6) have the same meaning as in the criterion (1); however they denote vectors and matrices for multivariable system (5). We also assume equal weighing coefficients for each output/input of the system.

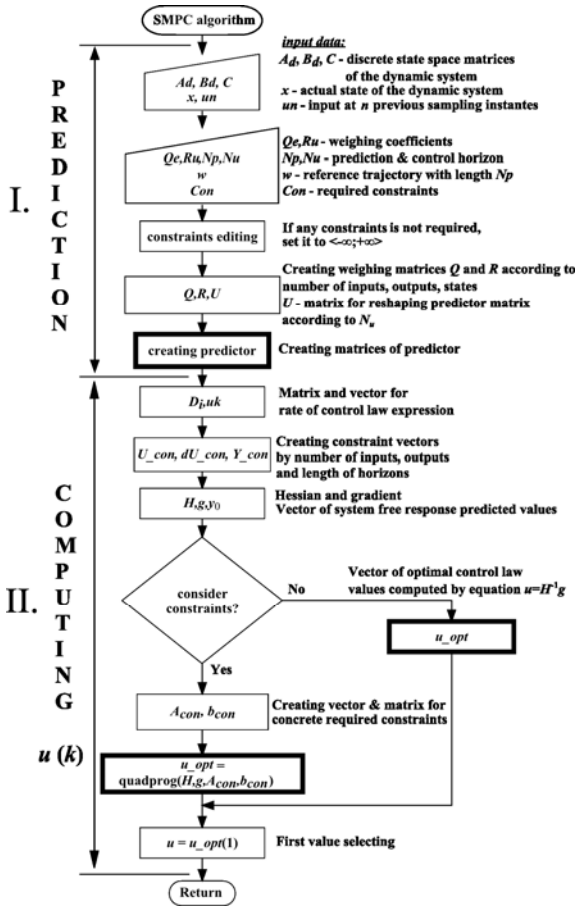


Figure 3
Flow chart of SMPC algorithm function

We programmed the SMPC algorithm as a Matlab function on the basis of the designed flow chart diagram, depicted in Fig. 3. The SMPC algorithm design is divided into two phases in compliance with the procedure mentioned in Section 2.

It is clear from the flow chart that control algorithms offer the optimal control action, computing with and without respect to required constraints of control action value, its rate or output of dynamic system. The mathematical description of two phases of SMPC algorithm design is described in next subsections 3.1 and 3.2.

3.1 Predictor Derivation in SMPC

The state prediction over the horizon N_p can be written according to [6] in form

$$\begin{aligned}
&\hat{\boldsymbol{x}}(k+1) = \boldsymbol{A}_d\hat{\boldsymbol{x}}(k) + \boldsymbol{B}_d\boldsymbol{u}(k)\\
&\hat{\boldsymbol{x}}(k+2) = \boldsymbol{A}_d\hat{\boldsymbol{x}}(k+1) + \boldsymbol{B}_d\boldsymbol{u}(k+1) = \boldsymbol{A}_d^2\,\boldsymbol{x}(k) + \boldsymbol{A}_d\boldsymbol{B}_d\boldsymbol{u}(k) + \boldsymbol{B}_d\boldsymbol{u}(k+1)\\
&\hat{\boldsymbol{x}}(k+3) = \boldsymbol{A}_d\hat{\boldsymbol{x}}(k+2) + \boldsymbol{B}_d\boldsymbol{u}(k+2) = \boldsymbol{A}_d^3\,\boldsymbol{x}(k) + \boldsymbol{A}_d^2\boldsymbol{B}_d\boldsymbol{u}(k) + \boldsymbol{A}_d\boldsymbol{B}_d\boldsymbol{u}(k+1) + \boldsymbol{B}_d\boldsymbol{u}(k+2)\\
&\quad\vdots \qquad\qquad \vdots \qquad\qquad \ddots\\
&\hat{\boldsymbol{x}}(k+N_p) = \boldsymbol{A}_d^{Np}\,\boldsymbol{x}(k) + \boldsymbol{A}_d^{Np-1}\boldsymbol{B}_d\boldsymbol{u}(k) + \;\cdots\; + \;\boldsymbol{B}_d\boldsymbol{u}(k+N_p-1)
\end{aligned} \quad . \tag{7}$$

Then the system output prediction is

$$\begin{aligned}
&\hat{\boldsymbol{y}}(k) \;\;\;\;\; = \boldsymbol{C}\boldsymbol{x}(k)\\
&\hat{\boldsymbol{y}}(k+1) \;= \boldsymbol{C}\boldsymbol{x}(k+1) = \boldsymbol{C}\boldsymbol{A}_d\boldsymbol{x}(k) + \boldsymbol{C}\boldsymbol{B}_d\boldsymbol{u}(k)\\
&\hat{\boldsymbol{y}}(k+2) \;= \boldsymbol{C}\boldsymbol{x}(k+2) = \boldsymbol{C}\boldsymbol{A}_d^2\boldsymbol{x}(k) + \boldsymbol{C}\boldsymbol{A}_d\boldsymbol{B}_d\boldsymbol{u}(k) + \boldsymbol{C}\boldsymbol{B}_d\boldsymbol{u}(k+1)\\
&\quad\vdots \qquad\qquad\;\; \vdots \qquad\qquad\;\; \ddots\\
&\hat{\boldsymbol{y}}(k+N_p) \;= \boldsymbol{C}\boldsymbol{A}_d^{Np}\,\boldsymbol{x}(k) \;+\; \boldsymbol{C}\boldsymbol{A}_d^{Np-1}\boldsymbol{B}_d\boldsymbol{u}(k) + \;\cdots\; + \boldsymbol{C}\boldsymbol{B}_d\boldsymbol{u}(k+N_p-1)
\end{aligned} \quad , \tag{8}$$

that can be written in a matrix form

$$\hat{\boldsymbol{y}} = \boldsymbol{V}\boldsymbol{x}(k) + \boldsymbol{G}\boldsymbol{u} \;, \tag{9}$$

where

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{\boldsymbol{y}}(k+1) & \hat{\boldsymbol{y}}(k+2) & \cdots & \hat{\boldsymbol{y}}(k+N_p) \end{bmatrix}^T,\; \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}(k) & \boldsymbol{u}(k+1) & \cdots & \boldsymbol{u}(k+N_p-1) \end{bmatrix}^T,$$

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{C}\boldsymbol{A}_d \\ \vdots \\ \boldsymbol{C}\boldsymbol{A}_d^{Np} \end{pmatrix},\; \boldsymbol{G} = \begin{pmatrix} \boldsymbol{C}\boldsymbol{B}_d & \boldsymbol{0} & \\ \vdots & \ddots & \ddots \\ \boldsymbol{C}\boldsymbol{A}_d^{Np-1}\boldsymbol{B}_d & \cdots & \boldsymbol{C}\boldsymbol{B}_d \end{pmatrix}.$$

In equation (9) the term $\boldsymbol{V}\boldsymbol{x}(k)$ represents the *free response* $\boldsymbol{y}_0$ and the term $\boldsymbol{G}\boldsymbol{u}$ the *forced response* of system. In the case that the control horizon N_u is considered during computing the optimal control action sequence, it is necessary to multiply the matrix $\boldsymbol{G}$ by matrix $\boldsymbol{U}$ from right: $\boldsymbol{G} \leftarrow \boldsymbol{G}\boldsymbol{U}$, where matrix $\boldsymbol{U}$ has form

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{I} & & \\ & \ddots & \\ & & \boldsymbol{I} \\ & & \vdots \\ & & \boldsymbol{I} \end{pmatrix} \text{ with dimension } \left[nu \cdot N_p \right] \times \left[nu \cdot N_u \right]. \tag{10}$$

The basic mathematical fundamental for the first part of the flow chart diagram depicted in Fig. 3 has been shown in this section.

3.2 Computation of the Optimal Control Action in SMPC

In this section, the mathematical fundamental for the second part of the flow chart diagram depicted in Fig. 3 is derived.

The matrix form of criterion J_{MPC} (6) is

$$J_{\mathrm{MPC}} = \left(\hat{\boldsymbol{y}} - \boldsymbol{w} \right)^T \boldsymbol{Q} \left(\hat{\boldsymbol{y}} - \boldsymbol{w} \right) + \boldsymbol{u}^T \boldsymbol{R} \boldsymbol{u}, \tag{11}$$

where matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ are diagonal with particular dimension and created from weighing coefficients Q_e and R_u ($\boldsymbol{Q} = Q_e \boldsymbol{I}$, $\boldsymbol{R} = R_u \boldsymbol{I}$).

After the predictor (9) substitution into the criterion (11) and multiplication we can obtain the equation

$$J_{\mathrm{MPC}} = \boldsymbol{u}^T \left(\boldsymbol{G}^T \boldsymbol{Q} \boldsymbol{G} + \boldsymbol{R} \right) \boldsymbol{u} + \left[\left(\boldsymbol{V} \boldsymbol{x}(k) - \boldsymbol{w} \right)^T \boldsymbol{Q} \boldsymbol{G} \right] \boldsymbol{u} + \boldsymbol{u}^T \left[\boldsymbol{G}^T \boldsymbol{Q} \left(\boldsymbol{V} \boldsymbol{x}(k) - \boldsymbol{w} \right) \right] + c, \tag{12}$$

from which on the basis of condition of minimum $\frac{\partial J_{\mathrm{MPC}}}{\partial \boldsymbol{u}} \overset{!}{=} \boldsymbol{0}$ and with using equations for vector derivation [6]

$$\frac{\partial}{\partial \boldsymbol{u}} \left(\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{y} \right) = \boldsymbol{H} \boldsymbol{y}, \quad \frac{\partial}{\partial \boldsymbol{u}} \left(\boldsymbol{y}^T \boldsymbol{H} \boldsymbol{u} \right) = \boldsymbol{H}^T \boldsymbol{y}, \quad \frac{\partial}{\partial \boldsymbol{u}} \left(\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u} \right) = \boldsymbol{H} \boldsymbol{u} + \boldsymbol{H}^T \boldsymbol{u}, \tag{13}$$

it is possible to derive an equation for the sequence of optimal control action

$$\boldsymbol{u} = -\boldsymbol{H}^{-1} \boldsymbol{g}, \tag{14}$$

where $\boldsymbol{H} = \boldsymbol{G}^T \boldsymbol{Q} \boldsymbol{G} + \boldsymbol{R}$ and $\boldsymbol{g}^T = \left(\boldsymbol{V} \boldsymbol{x}(k) - \boldsymbol{w} \right)^T \boldsymbol{Q} \boldsymbol{G}$.

It is also possible to ensure the rate of control action $\Delta \boldsymbol{u}$ weighting in the criterion (11) by $\Delta \boldsymbol{u}$ expression with formula $\Delta \boldsymbol{u} = \boldsymbol{D}_i \boldsymbol{u} - \boldsymbol{u}_k$, where

$$\boldsymbol{D}_i = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \\ -\boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{I} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & -\boldsymbol{I} & \boldsymbol{I} \end{pmatrix} \text{ and } \boldsymbol{u}_k = \begin{pmatrix} \boldsymbol{u}(k-1) \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{pmatrix}. \tag{15}$$

Subsequently $\boldsymbol{H} = \boldsymbol{G}^T \boldsymbol{Q}\boldsymbol{G} + \boldsymbol{D}_i^T \boldsymbol{R}\boldsymbol{D}_i$ and $\boldsymbol{g}^T = \left(\boldsymbol{V}\boldsymbol{x}(k) - \boldsymbol{w}\right)^T \boldsymbol{Q}\boldsymbol{G} - \boldsymbol{u}_k^T \boldsymbol{R}\boldsymbol{D}_i$.

It is necessary to use the *quadprog* function (4) for computing the optimal control action sequence, which should be limited by the given constraints, whereby the particular values of matrix $\boldsymbol{H}$ and vector $\boldsymbol{g}$ depend on the control action weighting manner in the criterion (11). It is necessary to compose matrix $\boldsymbol{A}_{con}$ and vector $\boldsymbol{b}_{con}$ in compliance with required constraints:

- for the rate of control action constraints $\Delta u_{\min} \le \Delta u \le \Delta u_{\max}$

$$\boldsymbol{A}_{con} = \begin{pmatrix} \boldsymbol{D}_i \\ -\boldsymbol{D}_i \end{pmatrix}, \ \boldsymbol{b}_{con} = \begin{pmatrix} \boldsymbol{1}\Delta u_{\max} + \boldsymbol{u}_k \\ -\boldsymbol{1}\Delta u_{\min} - \boldsymbol{u}_k \end{pmatrix}, \tag{16}$$

- for the value of control action constraints $u_{\min} \le u \le u_{\max}$

$$\boldsymbol{A}_{con} = \begin{pmatrix} \boldsymbol{I} \\ -\boldsymbol{I} \end{pmatrix}, \ \boldsymbol{b}_{con} = \begin{pmatrix} \boldsymbol{1}u_{\max} \\ -\boldsymbol{1}u_{\min} \end{pmatrix}, \tag{17}$$

- for system output constraints $y_{\min} \le y \le y_{\max}$

$$\boldsymbol{A}_{con} = \begin{pmatrix} \boldsymbol{G} \\ -\boldsymbol{G} \end{pmatrix}, \ \boldsymbol{b}_{con} = \begin{pmatrix} \boldsymbol{1}y_{\max} - \boldsymbol{V}\boldsymbol{x}(k) \\ -\boldsymbol{1}y_{\min} + \boldsymbol{V}\boldsymbol{x}(k) \end{pmatrix}, \tag{18}$$

where $\boldsymbol{I}$ is an unit matrix, $\boldsymbol{1}$ denotes an unit vector, $\boldsymbol{D}_i$ and $\boldsymbol{u}_k$ are the matrix and vector from equation (15), $\boldsymbol{G}$ and $\boldsymbol{V}\boldsymbol{x}(k)$ are from predictor equation (9).

4 Predictive Control Algorithm Based on the ARX Model Design

This algorithm belongs to set of generalized predictive control (GPC) algorithms, i.e. it is based on the input-output description of dynamic systems.

Particularly, this algorithm is based on the regression ARX model

$$A_z(z^{-1})y(k) = B_z(z^{-1})u(k) + \xi(k), \tag{19}$$

where $B_z(z^{-1})$ is m ordered polynomial numerator with coefficients b_i, $A_z(z^{-1})$ is n ordered polynomial denominator with coefficients a_i, $u(k)$ is an input, $y(k)$ is an output of dynamic system and $\xi(k)$ is a system output error or a noise of output measurement [9].

The control structure with GPC algorithm is depicted in Fig. 4, whereby the meaning of particular parameters is the same as in Fig. 2. Additionally, $\boldsymbol{u}_n$ and $\boldsymbol{y}_n$ are vectors of control action values and system output values in n previous samples; n is system's order.

The flow chart, which served as the basis for programming the function of SMPC algorithm (depicted in Fig. 3) is very similar to the flow chart for algorithmic design of GPC algorithm considered in this paper. However, they differ each other in some blocks (steps), which treat the specific data of GPC algorithm.

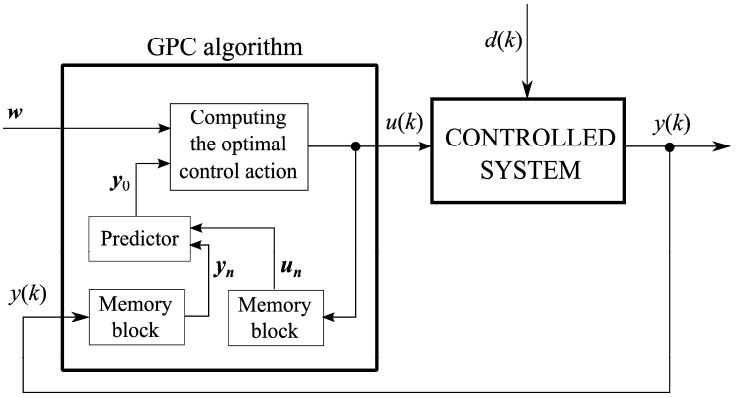


Figure 4
Control structure with GPC algorithm

The next subsections contain the mathematical description of two phases of the GPC algorithm based on the ARX model design.

4.1 Predictor Derivation in GPC Based on the ARX Model

According to [9], provided that $b_0 = 0$ along with $\xi(k) = 0$, we can express the output of dynamic system in sample $k + 1$ from the ARX model (19) by equation

$$y(k+1) = \sum_{i=1}^{n} b_i u(k-i+1) - \sum_{i=1}^{n} a_i y(k-i+1). \tag{20}$$

We are able to arrange (20) into a matrix form:

$$\begin{pmatrix} y(k-n+2) \\ \vdots \\ y(k) \\ y(k+1) \end{pmatrix} = \begin{pmatrix} 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \\ -a_n & \cdots & & -a_1 \end{pmatrix} \begin{pmatrix} y(k-n+1) \\ \vdots \\ y(k-1) \\ y(k) \end{pmatrix} + \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ b_n & \cdots & b_1 \end{pmatrix} \begin{pmatrix} u(k-n+1) \\ \vdots \\ u(k-1) \\ u(k) \end{pmatrix},$$
$$\boldsymbol{X}(k+1) = \boldsymbol{A}\,\boldsymbol{X}(k) + \boldsymbol{B}_0\,\boldsymbol{u}(k) \qquad (21)$$
$$y(k) = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix} \boldsymbol{X}(k)$$
$$y(k) = \boldsymbol{C}\,\boldsymbol{X}(k)$$

which represents a "pseudostate" space model of a dynamic system [10].

By the derivation mentioned in [10] or [12], it is possible to express the predictor from the pseudostate space model as

$$\begin{pmatrix} \hat{y}(k+1) \\ \vdots \\ \hat{y}(k+N_p) \end{pmatrix} = \begin{pmatrix} \boldsymbol{CA} \\ \vdots \\ \boldsymbol{CA}^{Np} \end{pmatrix} \begin{pmatrix} y(k-n+1) \\ \vdots \\ y(k) \end{pmatrix} + \begin{pmatrix} \boldsymbol{CB}_0 & \cdots & 0 \\ \ddots & & \vdots \\ & \boldsymbol{CB}_{Np-1} & \end{pmatrix} \begin{pmatrix} u(k-n+1) \\ \vdots \\ u(k+N_p-1) \end{pmatrix}$$
$$\hat{\boldsymbol{y}} = \bar{\boldsymbol{y}}_0 + \bar{\boldsymbol{G}}\,\bar{\boldsymbol{u}}$$
$$\hat{\boldsymbol{y}} = \bar{\boldsymbol{y}}_0 + \bar{\boldsymbol{G}}_{(:,1:n-1)} \begin{pmatrix} u(k-n+1) \\ \vdots \\ u(k-1) \end{pmatrix} + \bar{\boldsymbol{G}}_{(:,n:n+Np-1)} \begin{pmatrix} u(k) \\ \vdots \\ u(k+N_p-1) \end{pmatrix}, \qquad (22)$$
$$\hat{\boldsymbol{y}} = \boldsymbol{y}_0 + \boldsymbol{G}_{Np}\,\boldsymbol{u}$$

where $\boldsymbol{y}_0$ introduces the *free response* and $\boldsymbol{G}_{Np}\boldsymbol{u}$ represents the *forced response* of the system. Following the control horizon length, it is necessary to create a $N_p \times N_u$ matrix $\boldsymbol{U}$. It is recommended to right multiply $\boldsymbol{U}$ with $\boldsymbol{G}_{Np}$. Thus, we obtain a matrix $\boldsymbol{G} = \boldsymbol{G}_{Np}\boldsymbol{U}$, which ensures that the optimal control action will be considered over the control horizon in computing the optimal control action sequence. The final matrix form of predictor thus will be

$$\hat{\boldsymbol{y}} = \boldsymbol{y}_0 + \boldsymbol{G}\boldsymbol{u}\,. \qquad (23)$$

4.2 Computation of the Optimal Control Action in GPC Based on the ARX Model

The algorithm for SISO system minimizes the criterion

$$J_{\mathrm{ARX}} = \sum_{i=N_1}^{N_p} \left\{ Q_e \left[\hat{y}(k+i) - w(k+i) \right] \right\}^2 + \sum_{i=1}^{N_u} \left\{ R_u \left[u(k+i-1) \right] \right\}^2, \qquad (24)$$

which in contrast to the SMPC algorithm powers also weighing coefficients.

The criterion J_{ARX} (24) has the matrix form

$$J_{\mathrm{ARX}} = \left((\hat{\boldsymbol{y}} - \boldsymbol{w})^T \quad \boldsymbol{u}^T\right) \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{y}} - \boldsymbol{w} \\ \boldsymbol{u} \end{pmatrix}, \tag{25}$$

where the matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ are created from weighing coefficients Q_e and R_u with dimensions $N_p \times N_p$ and $N_u \times N_u$.

According to [10], it is sufficient to minimize only one part:

$$J_{\mathrm{k}} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{y}} - \boldsymbol{w} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{QG} \\ \boldsymbol{R} \end{pmatrix} \boldsymbol{u} - \begin{pmatrix} \boldsymbol{Q}(\boldsymbol{w} - \boldsymbol{y}_0) \\ \boldsymbol{0} \end{pmatrix}. \tag{26}$$

According to [10], the minimization of J_{k} (26) is based on solving the algebraic equations, which are written in a matrix form, when the value of control action $\boldsymbol{u}$ or its rate $\Delta\boldsymbol{u}$ is weighted in the criterion J_{ARX} (25):

$$\begin{pmatrix} \boldsymbol{QG} \\ \boldsymbol{R} \end{pmatrix} \boldsymbol{u} - \begin{pmatrix} \boldsymbol{Q}(\boldsymbol{w} - \boldsymbol{y}_0) \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{0} \quad \text{or} \quad \begin{pmatrix} \boldsymbol{QG} \\ \boldsymbol{RD}_i \end{pmatrix} \boldsymbol{u} - \begin{pmatrix} \boldsymbol{Q}(\boldsymbol{w} - \boldsymbol{y}_0) \\ \boldsymbol{Ru}_k \end{pmatrix} = \boldsymbol{0}, \tag{27}$$

$$\boldsymbol{S} \quad \boldsymbol{u} - \quad \boldsymbol{T} \quad = \boldsymbol{0} \qquad \boldsymbol{S} \quad \boldsymbol{u} - \quad \boldsymbol{T} \quad = \boldsymbol{0}.$$

Since the matrix $\boldsymbol{S}$ is not squared, it is possible to use pseudo-inversion for the optimal control $\boldsymbol{u}$ computing, which is the solution of the system of equations (27)

$$\boldsymbol{u} = (\boldsymbol{S}^T \boldsymbol{S})^{-1} \boldsymbol{S}^T \boldsymbol{T} \tag{28}$$

or according to [11], by the QR-decomposition of matrix $\boldsymbol{S}$, where a transformational matrix $\boldsymbol{Q}_t$ transforms the matrix $\boldsymbol{S}$ to upper triangular matrix $\boldsymbol{S}_t$ as shown in Fig. 4:

$$\begin{aligned} \boldsymbol{Su} &= \boldsymbol{T} \qquad / \times \boldsymbol{Q}_t^T \\ \boldsymbol{Q}_t^T \boldsymbol{Su} &= \boldsymbol{Q}_t^T \boldsymbol{T} \\ \begin{pmatrix} \boldsymbol{S}_t \\ \boldsymbol{0} \end{pmatrix} \boldsymbol{u} &= \begin{pmatrix} \boldsymbol{T}_t \\ \boldsymbol{T}_z \end{pmatrix} \end{aligned} \tag{29}$$

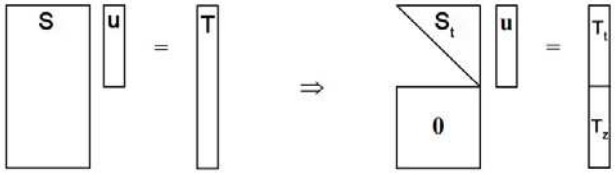


Figure 5
Matrix transformation with QR decomposition

It results from above-mentioned that the optimal control $\boldsymbol{u}$ can also be computed by formula

$$\boldsymbol{u} = \boldsymbol{S}_t^{-1} \boldsymbol{T}_t. \tag{30}$$

The matrix $\boldsymbol{H}$ and vector $\boldsymbol{g}$, which are the input parameters of *quadprog* function (if the optimal control computing with constraints is carried out), have the following form with $\boldsymbol{u}$ or $\Delta\boldsymbol{u}$ weighted in the criterion J_{ARX} (25)

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{G}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{G} + \boldsymbol{R}^T\boldsymbol{R} \\ \boldsymbol{g}^T &= (\boldsymbol{y}_0 - \boldsymbol{w})^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{G} \end{aligned} \quad \text{or} \quad \begin{aligned} \boldsymbol{H} &= \boldsymbol{G}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{G} + \boldsymbol{D}_i^T\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{D}_i \\ \boldsymbol{g}^T &= (\boldsymbol{y}_0 - \boldsymbol{w})^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{G} - \boldsymbol{u}_k^T\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{D}_i \end{aligned} \tag{31}$$

and the matrix $\boldsymbol{A}_{con}$ and vector $\boldsymbol{b}_{con}$ are as well as in previous algorithm given by equations (16), (17), (18), but the system free response is constituted by vector $\boldsymbol{y}_0$ instead of the term $\boldsymbol{V}\boldsymbol{x}(k)$.

5 Predictive Control Algorithm Based on the CARIMA Model Design

Similarly to generalized predictive control (GPC) algorithm based on the ARX model, this algorithm also belongs to the GPC algorithms family, but it is based on the CARIMA model of dynamic systems

$$A_z(z^{-1})y(k) = B_z(z^{-1})u(k-1) + \frac{C_z(z^{-1})}{\Delta}\xi(k)\,, \tag{32}$$

Where, in contrast to the ARX model, $C_z(z^{-1})$ is multi-nominal and $\Delta = 1 - z^{-1}$ introduces an integrator [8].

According to [8], the criterion that is minimized in this GPC algorithm has the form

$$J_{\mathrm{CARIMA}} = \sum_{i=N_1}^{N_p}\left[P(z^{-1})\hat{y}(k+i) - w(k+i)\right]^2 + \lambda\sum_{i=1}^{N_u}\left[\Delta u(k+i-1)\right]^2\,, \tag{33}$$

where in contrast to the previous algorithm, λ is a relative weighing coefficients that expresses a weight ratio between the deviation $y(k) - w(k)$ and the control action $u(k)$. $P(z^{-1})$ provides the same effect as equation (2) in the first part of paper. According to [8], the corresponding first order filter for constant reference trajectory is $P(z^{-1}) = \dfrac{1-\alpha z^{-1}}{1-\alpha}$, where $\alpha \in \langle 0;1\rangle$.

Next, we will restrict ourselves to $\alpha = 0$, i.e. $P(z^{-1}) = 1$.

The next subsections contain the mathematical description of two phases of the GPC algorithm based on the CARIMA model design.

5.1 The Predictor Derivation in GPC Algorithm Based on the the CARIMA Model

According to [8], the output of the dynamic system that is defined by equation (32) in sample instant $k + 1$ is given by the following equation (for notation convenience without z^{-1}):

$$y(k+j) = \frac{B_z}{A_z} u(k+j-1) + \frac{C_z}{\Delta A_z} \xi(k+j) \ . \tag{34}$$

On the basis of the derivation mentioned in [8] and with polynomial dividing

$$\frac{C_z(z^{-1})}{\Delta A_z(z^{-1})} = E_j(z^{-1}) + z^{-1} \frac{F_j(z^{-1})}{\Delta A_z(z^{-1})} \quad \text{and} \quad \frac{B_z(z^{-1})E_j(z^{-1})}{C_z(z^{-1})} = G_j(z^{-1}) + z^{-j} \frac{\Gamma(z^{-1})}{C_z(z^{-1})}$$

or alternatively by solving diophantine equations

$$C_z = E_j \Delta A_z + z^{-j} F_j \quad \text{a} \quad B_z E_j = G_j C_z + z^{-j} \Gamma_j$$

we are able to express the j steps predictor in the matrix form

$$\hat{\boldsymbol{y}} = \boldsymbol{G}\Delta\boldsymbol{u} + \boldsymbol{y}_0 \,, \tag{35}$$

in which the rate of control action $\Delta\boldsymbol{u}$ is present directly.

The form of the matrix $\boldsymbol{G}$ is $\boldsymbol{G} = \begin{pmatrix} g_0 & 0 & \cdots & \cdots & 0 \\ g_1 & g_0 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \\ \vdots & & & g_0 & 0 \\ g_{N_p-1} & & \cdots & & g_0 \end{pmatrix}$,

where coefficients g_i can be obtained by division $B / \Delta A$ and $\boldsymbol{y}_0$ is the *free response* of system. If we take value N_1 into consideration, we will able to remove first $N_1 - 1$ rows of matrix $\boldsymbol{G}$. Moreover, regarding the control horizon, only the first N_u columns of matrix G are necessary for the next calculations. Thus, the reduced matrix $\boldsymbol{G}$ will have dimension $(N_p - N_1 + 1) \times N_u$ [8].

5.2 Computation of the Optimal Control Action in the GPC Algorithm Based on the CARIMA Model

It is possible to express the criterion J_{CARIMA} (33) in the matrix form

$$\begin{aligned} J_{\text{CARIMA}} &= (\hat{\boldsymbol{y}} - \boldsymbol{w})^T (\hat{\boldsymbol{y}} - \boldsymbol{w}) + \lambda \Delta\boldsymbol{u}^T \Delta\boldsymbol{u} = \\ &= (\boldsymbol{G}\Delta\boldsymbol{u} + \boldsymbol{y_0} - \boldsymbol{w})^T \ (\boldsymbol{G}\Delta\boldsymbol{u} + \boldsymbol{y_0} - \boldsymbol{w}) + \lambda \Delta\boldsymbol{u}^T \Delta\boldsymbol{u} = \\ &= c + 2\boldsymbol{g}^T \Delta\boldsymbol{u} + \Delta\boldsymbol{u}^T \boldsymbol{H} \Delta\boldsymbol{u}, \end{aligned} \tag{36}$$

where $\boldsymbol{H} = \boldsymbol{G}^T\boldsymbol{G} + \lambda\boldsymbol{I}$, $\boldsymbol{g}^T = (\boldsymbol{y_0} - \boldsymbol{w})^T \boldsymbol{G}$, c is a constant and $\boldsymbol{I}$ is a unit matrix.

If it is necessary to weigh the value of control action $\boldsymbol{u}$ in the criterion (36); it is possible to obtain $\boldsymbol{H} = \boldsymbol{G}^T\boldsymbol{G} + \lambda \boldsymbol{D}_i^{-T}\boldsymbol{D}_i^{-1}$, $\boldsymbol{g}^T = (\boldsymbol{y_0} - \boldsymbol{w})^T\boldsymbol{G} + \lambda \boldsymbol{u}_k^T\boldsymbol{D}_i^{-T}\boldsymbol{D}_i^{-1}$ on the basis of formula (15) in the first paper.

Following the condition of minimum $\frac{\partial J_{\text{CARIMA}}}{\partial \Delta \boldsymbol{u}} \overset{!}{=} \boldsymbol{0}$, it is easy to express the equation for the optimal control action in disregard of required constraints

$$\Delta \boldsymbol{u} = -\boldsymbol{H}^{-1}\boldsymbol{g} \,. \tag{37}$$

The optimal control computing with regard to the required constraints can be carried out by the *quadprog* function, whereby the matrix $\boldsymbol{A}_{con}$ and vector $\boldsymbol{b}_{con}$ are:

- for the rate of control action constraints $\Delta u_{\min} \le \Delta u \le \Delta u_{\max}$

$$\boldsymbol{A}_{obm} = \begin{pmatrix} \boldsymbol{I} \\ -\boldsymbol{I} \end{pmatrix}, \; \boldsymbol{b}_{obm} = \begin{pmatrix} \mathbf{1}\Delta u_{\max} \\ -\mathbf{1}\Delta u_{\min} \end{pmatrix}, \tag{38}$$

- for the value of control action constraints $u_{\min} \le u \le u_{\max}$

$$\boldsymbol{A}_{obm} = \begin{pmatrix} \boldsymbol{L} \\ -\boldsymbol{L} \end{pmatrix}, \; \boldsymbol{b}_{obm} = \begin{pmatrix} \mathbf{1}u_{\max} - \mathbf{1}u(k-1) \\ -\mathbf{1}u_{\min} + \mathbf{1}u(k-1) \end{pmatrix}, \tag{39}$$

- for system output constraints $y_{\min} \le y \le y_{\max}$

$$\boldsymbol{A}_{obm} = \begin{pmatrix} \boldsymbol{G} \\ -\boldsymbol{G} \end{pmatrix}, \; \boldsymbol{b}_{obm} = \begin{pmatrix} \mathbf{1}y_{\max} - \boldsymbol{y}_0 \\ -\mathbf{1}y_{\min} + \boldsymbol{y}_0 \end{pmatrix}, \tag{40}$$

where $\boldsymbol{I}$ is an unit matrix, $\mathbf{1}$ is an unit vector and $\boldsymbol{L}$ is a lower triangular matrix inclusive of ones. It is necessary to realize that the vector of control action rate $\Delta\boldsymbol{u}$ is the result of optimal control computing in this case.

Equally, as in the previous GPC algorithm, we programmed the GPC algorithm based on the CARIMA model as a Matlab function on the basis of the designed flow chart diagram in Fig. 3, with particular modifications, which are clear from theoretical background of the algorithm.

6 Predictive Control Algorithms Verification on the Helicopter Model

We introduced the basic principle, theoretical background and the manner of forming the algorithms of three different predictive control algorithms in the previous sections. Next we are concerned with using them in a laboratory helicopter model control by Matlab with implemented Real-Time Toolbox functions.

6.1 The Real Laboratory Helicopter Model

The educational helicopter model constitutes a multivariable nonlinear dynamic system with three inputs and two measured outputs, as depicted in Fig. 6.

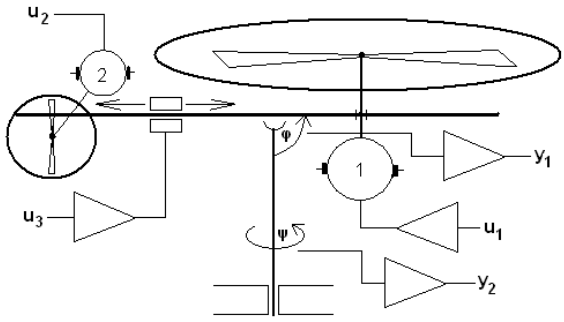


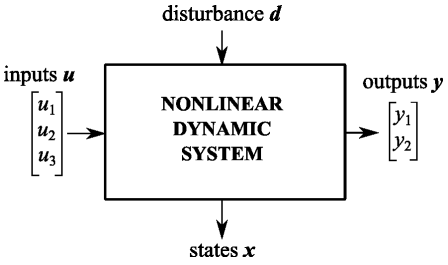
Figure 6
Mechanical system of real laboratory Helicopter model

The model is composed of a body with two propellers, which have their axes perpendicular and are driven by small DC motors; i.e. the helicopter model constitutes a system with two degrees of freedom [5]. The movement in the direction of axis y (elevation = output y_1) presents the first degree of freedom, and the second degree of freedom is presented by the movement in the direction of axis x (azimuth = output y_2). The values of both the helicopter's angular displacements are influenced by the propellers' rotation. The angular displacements (φ – angle for elevation, ψ – angle for azimuth) are measured by incremental encoders.

The DC motors are driven by power amplifiers using pulse width modulation, whereby a voltage introduced to motors (u_1 and u_2) is directly proportional to the computer output. The volatge u_3 serves for controlling the center of gravity, which constitutes a system's disturbance. It is necessary to note that we did not consider this during the design of the control algorithms. The model is connected to the computer by a multifunction card MF614, which communicates with the computer by functions of Real Time Toolbox [13].

The system approach of the real laboratory helicopter model and constraints of the inputs and outputs are shown in Fig. 7.

On the basis of helicopter's mathematic-physical description, mentioned in the manual [5], we can redraw Fig. 7 to Fig. 8, where M_{ep} is the main propeller torque performing in the propeller direction, M_{etp} is torque performing in the turnplate direction, M_{ap} is the auxilliary propeller torque performing in the propeller direction and M_{etp} is torque performing in the turnplate direction.



$u_1 \in \langle -1,1 \rangle$ V – main motor's volatge,

$u_2 \in \langle -1,1 \rangle$ V – side motor's volatge,

$u_3 \in \langle -1,1 \rangle$ V – center of gravity input,

$y_1 \in \langle -45^\circ, 45^\circ \rangle$ – position in elevation,

Figure 7

System approach with technical parameters (constraints) of the real laboratory Helicopter model

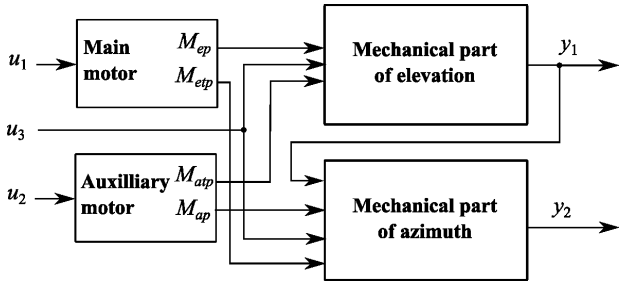


Figure 8

Subsystems of the real laboratory Helicopter model

According to mathematical description and block structure in [5], [7] or [16], it is possible to express the considered system's dynamics by nonlinear differential equations (for simplicity we omitted *(t)* in inputs, states and outputs expression):

$$\dot{x}_1 = -\frac{1}{T_m}\cdot x_1 + \frac{1}{T_m}\cdot u_1 \qquad \dot{x}_2 = -\frac{1}{T_s}\cdot x_2 + \frac{1}{T_s}\cdot u_2$$

$$\dot{x}_3 = \alpha_1 \cdot x_1 \cdot |x_1| + \beta_1 \cdot x_1 + \left(\gamma_2 \cdot x_2 \cdot |x_2| + \delta_2 \cdot x_2\right)\cdot \cos\eta - J_{el}\left(u_3\right)\cdot \delta_{el}\cdot x_3 - M_g\left(u_3\right)\cdot \cos x_4$$

$$\dot{x}_4 = \frac{1}{J_{el}\left(u_3\right)}\cdot x_3 \tag{41}$$

$$\dot{x}_5 = \left(\alpha_2 \cdot x_2 \cdot |x_2| + \beta_2 \cdot x_2\right)\cdot \cos\left(x_4 + \eta\right) + \left(\gamma_1 \cdot x_1 \cdot |x_1| + \delta_1 \cdot x_1\right)\cdot \cos x_4 - J_{az}(u_3).\delta_{az}\cdot x_5$$

$$\dot{x}_6 = \frac{1}{J_{az}\left(u_3\right)\cdot \cos x_4}\cdot x_5$$

$$y_1 = \frac{1}{\pi}\cdot x_4 \qquad y_2 = \frac{1}{\pi}\cdot x_6$$

where x_1 is the rotation speed of main motor, x_2 is the rotation speed of the auxiliary motor, x_3 is the rotation speed of the model in elevation, x_4 is the position of the model in elevation, x_5 is the rotation speed of the model in azimuth, x_6 is the position of the model in azimuth, T_m and T_s are the time constants of the main and auxilliary motors, δ_{el} and δ_{az} is the friction constant in elevation and azimuth. The

moment of gravity M_g, the moment of inertia in elevetaion J_{el}, and the azimuth J_{az} depend on u_3. These dependencies M_g, J_{el}, J_{az} and parameters α_1, α_2, β_1, β_2, γ_1, γ_2, δ_1, δ_2, η are introduced in detail in [5] or [7].

Note that it is also possible to express the model dynamics of the helicopter by nonlinear mathematical description introduced in [7] or [16].

As this paper has considered predictive control algorithms based on the flow chart in Fig. 3 and the utilization of the linear model of dynamic system, it is necessary to carry out the Taylor linearization of equations (41) in an operating point $P \equiv [\boldsymbol{x}_E, \boldsymbol{u}_E]$:

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}_C \boldsymbol{x}(t) + \boldsymbol{B}_C \boldsymbol{u}(t) \\ y(t) &= \boldsymbol{C}_C \boldsymbol{x}(t) \end{aligned}, \quad \text{where} \quad \boldsymbol{A}_C = \left[\frac{\partial \boldsymbol{f}_i}{\partial x_j} \right]_{\substack{\boldsymbol{x}=\boldsymbol{x}_E \\ \boldsymbol{u}=\boldsymbol{u}_E}} \quad \boldsymbol{B}_C = \left[\frac{\partial \boldsymbol{f}_i}{\partial u_j} \right]_{\substack{\boldsymbol{x}=\boldsymbol{x}_E \\ \boldsymbol{u}=\boldsymbol{u}_E}} . \tag{42}$$

For the purpose of linearization, we considered the operating point $P \equiv [\boldsymbol{x}_E, \boldsymbol{u}_E]$, in which angular displacements in elevation and azimuth were zero: $y_1 = 0$, $y_2 = 0$. The forms of matrices $\boldsymbol{A}_C$, $\boldsymbol{B}_C$, $\boldsymbol{C}_C$, describing the helicopter's continuous state space model in operating point P are

$$\boldsymbol{A}_C = \begin{bmatrix} A_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 & 0 & 0 \\ A_{31} & A_{32} & A_{33} & A_{34} & 0 & 0 \\ 0 & 0 & A_{43} & 0 & 0 & 0 \\ A_{51} & A_{52} & 0 & A_{54} & A_{55} & 0 \\ 0 & 0 & 0 & A_{64} & A_{65} & 0 \end{bmatrix} \boldsymbol{B}_C = \begin{bmatrix} B_{11} & 0 \\ 0 & B_{22} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \boldsymbol{C}_C = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ C_{14} & 0 \\ 0 & 0 \\ 0 & C_{26} \end{bmatrix}^T . \tag{43}$$

Note that the experimental identification of the laboratory helicopter model, which resulted in linear models in state-space and input-output description, was solved in [17]. In our case, we used only linear models obtained from the linearization process. The numerical values of particular elements in matrices (43) can be obtained on the basis of numerical values of model parameters in (41), which are supplied with the model from the manufacturer.

Subsequently, it is possible to create a discrete linear model from the continuous with specific sample period T_s. Then we can use the discrete linear model of helicopter dynamic system in the introduced control algorithms.

3.2 Control Structures Programming for Predictive Control Verification on the Helicopter Model

We carried out the control of the real laboratory helicopter model in accordance with the control structure for particular predictive control algorithm, which have been mentioned in this paper.

Unfortunately it is also necessary to note that the steady state deviation between reference trajectory and system output appeared in cases when the SMPC algorithm and the GPC algorithm based on the ARX model were used. Therefore, in order to eliminate this, we inserted a feedforward branch into the control structure, as seen in Fig. 9. A control component in the feedforward branch performed a function which generated steady state values of the main propeller's motor voltage for particular angular displacement. The transient characteristic between the particular angular displacement and the voltage steady state value was obtained experimentally in [15].

Thus, in cases where SMPC and GPC based on the ARX model algorithms were used, it is possible to express the entire control action by equation:

$$\boldsymbol{u}(k) = \boldsymbol{u}_{opt}(k) + \boldsymbol{u}_0(k)\,, \tag{44}$$

where $\boldsymbol{u}_{opt}(k)$ is the optimal control action computed in the function of the predictive control algorithm and $\boldsymbol{u}_0(k)$ is the control action generated by the feedforward controller.

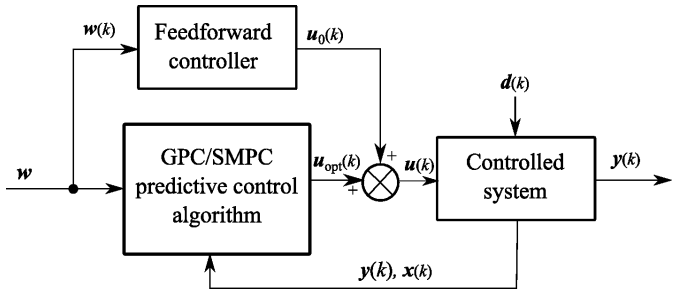


Figure 9
Control structure with feedforward branch

As has already been mentioned, we programmed each used control structures in the form of functions/scripts in Matlab, where the current helicopter's state is obtained by communication with the laboratory card using Real Time Toolbox functions [13] in the control closed loop. The data obtained are utilized by the control algorithm, which results in the particular value of control action. This value is sent back to the laboratory card, thus to the real helicopter model by Real Time Toolbox functions again. Note that we did not use Simulnik functional blocks, but only Real Time Toolbox functions for communication between the laboratory card and the helicopter model (*rtrd* for reading and *rtwr* for writing data to laboratory card).

It is necessary to note that control closed loop, which we considered, is based on the execution of the optimization problem in one sample period. Although it is possible to compute some matrices and vectors in advance, computing the system free response and optimal control action by quadratic programming must be

carried out at each sample instant. In general, optimization tasks are very time-consuming and they require powerful computers. To comply with the defined sample period, we checked the calculation time spent in computing the control action by predictive control algorithm function at every control step. In the case when the calculation time exceeded the sample period, the control algorithms were interrupted. The multifunction card MF614 allowed to use the minimal sample period 1 ms. Unfortunately, in our case, with the given computer we were only able to use 30 ms.

3.3 Results of Algorithm Verification on the Helicopter Model

In this part we present the results of the real laboratory Hhlicopter model control as the time responses of control action and controlled model's outputs. The next table illustrates the settings of variable parameters' values of predictive control algorithms used. If the settings for elevation and azimuth differ from each other, they are written in two rows for particular algorithm in the Tab. 1.

Table 1

Settings of predictive control algorithms' parameters

Algorithms	T_s	N_p	N_u	Q_v	R_v	Constraint	Weighting	Fig.
SMPC	0.03s	40	1	4 400	30 4	$u_e \in \langle 0.4;0.8 \rangle$ $u_a \in \langle -1;1 \rangle$	$\Delta u(k)$	5
2x GPC SISO ARX	0.05s	25 20	1	1 10	2 1	$u_e \in \langle 0.4;0.8 \rangle$ $u_a \in \langle -1;1 \rangle$	$\Delta u(k)$	6
GPC SISO ARX	0.05s	18	1	1	1	$u_e \in \langle 0.4;0.8 \rangle$ $u_a \in \langle -1;1 \rangle$	$\Delta u(k)$	7
GPC SISO CARIMA	0.05s	10	1	3	1	$u_e \in \langle 0.4;0.8 \rangle$ $u_a \in \langle -1;1 \rangle$	$\Delta u(k)$	7

At this point, we wish to note that the real laboratory helicopter model control fulfilled the aim of control with above presented settings of algorithms. However, the results were markedly influenced by small changes in horizons and the weighing coefficients' values. On the other hand, this did not happen in simulation control, which we used as a primary test of the designed algorithms. We carried out the simulation control of the nonlinear model (41) by numerical solving with Runge-Kutta fourth order method in its own Matlab function.

In Fig. 10 are the results of the real laboratory helicopter model with two degrees of freedom control using the SMPC algorithm. As the model's states are not measured, we used the state values estimation by Kalman's predictor, which we designed on the basis of duality principle with LQ control design according to [14]. We used weighing coefficients Q_{est} = 10000 and R_{est} = 0.001 for the

estimator's parameters design. Also, it is necessary to note that the feedforward branch was incorporated in the control structure in compliance with Fig. 9.

The results of MIMO system control are depicted in Fig. 11, too. However, two independent GPC algorithms for SISO systems were used as controllers, instead of one algorithm for the MIMO system. We designed GPC algorithms, which were based on the ARX model especially for elevation and for azimuth control, whereby we neglected mutual interactions and used only relevant states of the system (41). Also, the feedforward branch was incorporated in the control structure.

Fig. 12 illustrates the time responses of the real laboratory helicopter model control only in elevation direction. The model was latched; thus it was impossible to move it in the azimuth direction. The results of the control with the GPC algorithm designed for SISO systems are depicted in the figure, with the GPC algorithm based on the ARX model on left and the GPC algorithm based on the CARIMA model on the ride. It can be seen from figure that control with the GPC algorithm based on the ARX model gets better results than control with the GPC algorithm based on the CARIMA model. However, it must be stated that the feedforward branch was incorporated in the control structure together with GPC algorithm based on the ARX model.

The displayed results were obtained with the rate of control action $\Delta u(k)$ weighted in the criterion. If only the value of control action $u(k)$ was weighted, the time responses were similar to their counterparts, where $\Delta u(k)$ was weighted, but the deviation between system output and reference trajectory appeared.

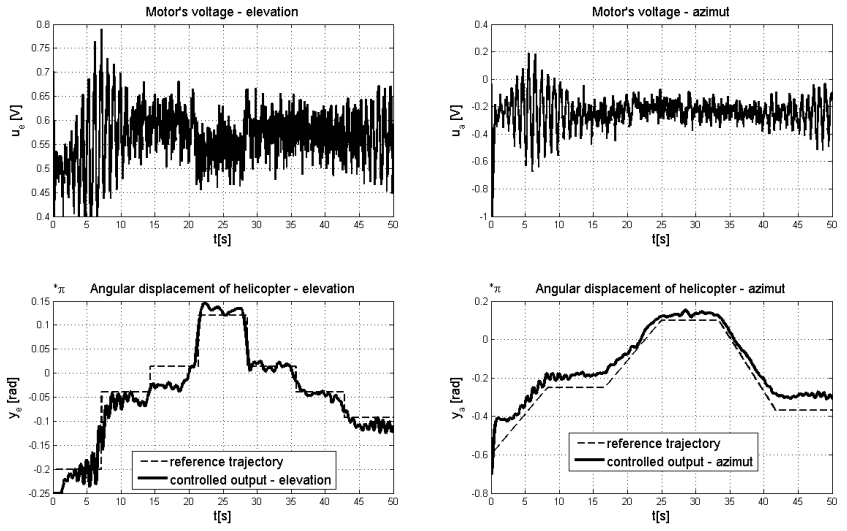


Figure 10
Time responses of Helicopter control with SMPC algorithm

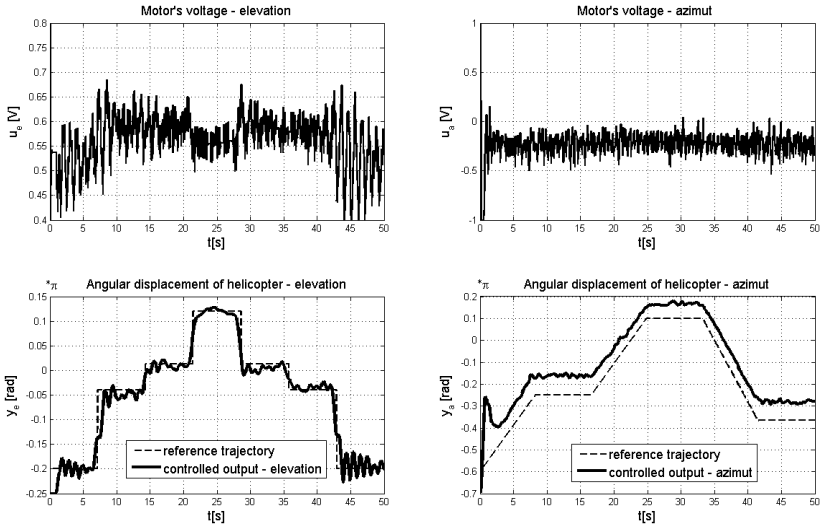


Figure 11
Time responses of Helicopter control with two GPC (ARX) algorithms

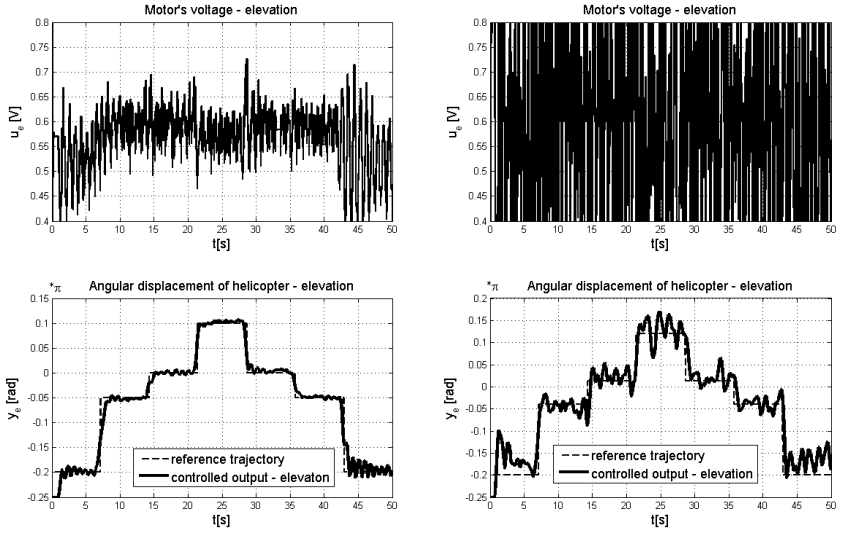


Figure 12
Time responses of Helicopter control in elevation with GPC (ARX) algorithm on left and GPC (CARIMA) algorithm on right

Although constraints of the controlled system input are given by range <-1; +1>, it is necessary to note that we reduced it to <0.4; 0.8> V for the main motor in order to obtain better performance of the control process.

Similar results were obtained by classical PID control in or optimal LQ control of the laboratory helicopter model, which were published in [15].

To accept the applicability of introduced algorithms, the comparison of the obtained results in Fig. 10 – Fig. 12 and the results published in [7] and [16] were useful, too. The time responses of the controlled system output presented in this article are quite similar to the results in [16], where a fuzzy logic controller was used. So it is possible to conclude that predictive control is a particular variant of intelligent control methods.

Conclusions

We have mentioned the theoretical basis of predictive control algorithms and the manner of their implementation to programming language Matlab in this paper. We were engaged in mathematical derivation of state space model based on predictive control algorithms and generalized predictive control algorithms based on the ARX and CARIMA models and their implementation in a control structure.

We have implemented all the mentioned algorithms in Matlab to verify them in a real laboratory helicopter model control in the Laboratory of Cybernetics at the Department of Cybernetics and Artificial Intelligence at Faculty of Electrotechnics and Informatics at the Technical University in Košice.

We have concluded from the obtained results that using GPC algorithms, which are based on the input-output description, seems to be preferable to algorithms based on the state space description of dynamic systems, mainly because it is not possible to measure the states of the helicopter model. Unfortunately, such algorithms comparing in the control of systems, whose states cannot be measured, depend very much on the type and settings of used parameters of the state estimator.

We supported the fact that it is possible to eliminate the deviation between system output and reference trajectory by control with an integration character, in our case by weighting the rate of control action $\Delta \boldsymbol{u}$ in the criterion. Unfortunately, it was valid in the real laboratory model control only when the GPC algorithm based on the CARIMA model was used and the rate of control action $\Delta \boldsymbol{u}$ appeared in the predictor expression. In other cases, when the remaining two mentioned algorithms were used and the rate of control action $\Delta \boldsymbol{u}$ did not appear in the predictor expression, but only its direct value $\boldsymbol{u}$ did, we modified the control structure by including the feedforward branch to control process. In this way it was possible to eliminate the deviation between the system output and reference trajectory.

Also we believe that a reduction in the sample period or an increase in the prediction horizon would improve control results obtained by control with the GPC algorithm based on the CARIMA model. Unfortunately, due to the predictive control algorithms computational demands, especially if the sequence of optimal control action with respect to required constraints was carried out, it was impossible to verify this assumption in our case.

Although we programmed the mentioned GPC algorithms in a broad range for MIMO dynamic systems as well, we must point out that in the laboratory helicopter model control, it was preferable to neglect any mutual interactions between system inputs and outputs and use GPC algorithms for SISO systems extra for each degree of freedom.

For future solutions to the problem of dynamic system control by predictive control algorithms, we suggest verifying the algorithms' extension by a summator of the deviation between the system output and reference trajectory, which will be particularly weighed in the criterion. This solution should include the integration character into the control process.

It is also necessary to note that predictive control insufficiency, which relates to the relatively long calculation time, is a substantial issue in fast mechatronic system control. We suppose it is possible to handle by explicit predictive control, which we also want to verify on the helicopter model. For that purpose we would like to use multi-parametric programming.

On the other hand, we can accept that the mentioned nonlinear mathematical description of the system is not precise enough. So we also plan to use a neural network that would be trained from data measured on a real laboratory model as a nonlinear predictor in nonlinear predictive control. However, it will be hard to handle computational phase, where the iteration optimization task for the minimization of the nonlinear functions must be executed at each sample instant.

We think the best solution can be found in some kind of combination of using a neural network and multi-parametric programming, as this would make it possible to generate nonlinear prediction by neural network and concurrently to compute the corresponding optimal control in advance, thus offline. We assume this way would permit controlling systems with a shorter sample period.

Acknowledgement

This work has been supported by the Scientific Grant Agency of Slovak Republic under project Vega No.1/0286/11 Dynamic Hybrid Architectures of the Multiagent Network Control Systems (50%) and by the project Development of the Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (50%).

References

[1] Clarke, D. W., Mohdani, C., Tuffs, P. S.: Generalized Predictive Control. Part 1 and 2. *Automatica*, Vol. 23, No. 2, pp. 137-160, 1987

[2] Grimble, M. J., Ordys, A. W.: Predictive Control for Industrial Applications. *Annual Reviews in Control*, 25, pp. 13-24, 2001, ISSN 1367-5788

[3] Camacho, E. F., Bordons, C.: Model Predictive Control. *Springer*, 1999

[4] Rossiter, J. A.: Model-based Predictive Control: A Practical Approach. *CRC Press*, 2004, ISBN 0-8493-1291-4

[5] Humusoft: CE150 Helicopter model, Educational Manual, 1996-2004

[6] Roubal, J, Havlena, V.: Range Control MPC Approach For Two-Dimensional System, Proceedings of the 16th IFAC World Congress, Volume 16, Part 1, 2005

[7] Dutka, Arkadiusz S., Ordys, Andrzej W., Grimble, Michael J.: Non-linear Predictive Control of 2 dof helicopter model. Proceeding of the 42nd IEEE Conference on Decision and Control, pp. 3954-3959, Hawai USA, 2003

[8] Fikar, M.: Predictive Control – An Introduction. Slovak Technical University - FCHPT, Bratislava 1999

[9] Belda, K., Böhm, J.: Adaptive Predictive Control for Simple Mechatronic Systems. *Proceedings of the WSEAS CSCC & EE International Conferences*. WSEAS Press, Athens, Greece 2006, pp. 307-312

[10] Belda, K., Böhm, J.: Adaptive Generalized Predictive Control for Mechatronic Systems. *WSEAS Transactions on Systems*. Volume 5, Issue 8, August 2006, Athens, Greece 2006, pp. 1830-1837

[11] Belda, K.: Control of Parallel Robotic Structures Driven by Electromotors. (Research Report). Czech Technical University, FEE, Prague 2004, 32 pp.

[12] Jajčišin, Š.: Application Modern Methods in Control of Non-linear Educational Models (in Slovak). Diploma thesis (Supervisor: doc. Ing. Anna Jadlovská, PhD) TU-FEI, Košice 2010

[13] Humusoft: Real-Time Toolbox, User's manual, 1996-2002

[14] Krokavec, Dušan, Filasová, Anna: Diskrétne systémy. Košice: Technical University-FEI, 2006, ISBN 80-8086-028-9

[15] Jadlovská, A., Lonščák, R.: Design and Experimental Verification of Optimal Control Algorithm for Educational Model of Mechatronic System (in Slovak) In: Electroscope – online journal for Electrotechnics, Vol. 2008, No. I, ISSN 1802-4564

[16] Velagic, Jasmin, Osmic, Nedim: Identification and Control of 2DOF Nonlinear Helicopter Model Using Intelligent Methods. *Systems Man and Cybernetics* (*SMC*), 2010 IEEE International Conference, pp. 2267-2275, 2010

[17] Dolinský, Kamil, Jadlovská, Anna: Application of Results of experimental Identification in Control of Laboratory Helicopter Model. Advances in Electrical and Electronic Engineering, scientific reviewed Journal published in Czech Republic, Vol. 9, Issue 4, 2011, pp. 157-166, ISSN 1804 3119