

# Ensembles of Fuzzy Cognitive Map Classifiers Based on Quantum Computation

Nan Ma<sup>1</sup>, Hamido Fujita<sup>2</sup>, Yun Zhai<sup>3</sup>, Shupeng Wang<sup>4</sup>

<sup>1</sup>College of Information Technology, Beijing Union University, Beisihuan East 97, 100101 Beijing, China, xxtmanan@buu.edu.cn

<sup>2</sup>Iwate Prefectural University, Takizawa Sugo 152, 020-0693 Iwate, Japan, hfujita-799@acm.org

<sup>3</sup>E-Government Research Center, Chinese Academy of Governance, Changchunqiao 6, 100089 Beijing, China, yunfei\_2001\_1@aliyun.com

<sup>4</sup>Institute of Information Engineering, China Academy of Sciences, minzhuang 89, 100093 Beijing, China, wangshupeng@iie.ac.cn

---

*Abstract: Fuzzy cognitive maps (FCMs) have been widely employed in the dynamic simulation and analysis in the complex systems. While a novel classifier model based on FCMs (FCMCM) was proposed in our former work, the obvious bottleneck of the genetic leaning algorithm used in FCMCM is its irksome efficiency, in particular, low speed in cross over and mutation delay in global convergence. Moreover the lack of the necessary robustness of a single FCMCM limits its generalization. To this end, a quantum computation based ensemble method FCMCM\_QC is proposed to address the scalability problem, which employs a novel evolutionary algorithm inspired by quantum computation. The FCMCM\_QC effectively uses the concept and principle of quantum computation to facilitate the computational complexity of genetic optimization for the FCMCM and reasonably selects classifiers with better performance for efficient ensembles. The experimental studies demonstrate the quality of the proposed FCMCM\_QC in generally used UCI datasets, and the simulation results prove that the FCMCM\_QC does enhance the speed of the convergence with high efficiency and good quality.*

*Keywords: fuzzy cognitive maps; classifier model; quantum computation*

---

## 1 Introduction

FCMs were proposed by Kosko to represent the causal relationship between concepts and analyze inference patterns [1]. FCMs represent knowledge in a symbolic manner and relate states, variables, events, outputs and inputs in a cause and effect approach. FCMs are illustrative causative representations of the

description and modeling of complex systems where the soft computation methodology of FCMs has been improved and enhanced using a new construction algorithm, and are implemented for modeling complex systems. FCMs are interactive structures of concepts, each of which interacts with the rest showing the dynamics and different aspects of the behavior of the system [2]. Recently, some scholars had made great efforts to explore the classification issue with FCMs, and had achieved some primitive results. Peng, Yang and Liu constructed a simple FCMs classifier and verified its validity by simulating the classification process with FCMs, i.e., the classification process is regarded as a status transition process of fuzzy cognitive map [3]. Zhu, Mendis and Gedeon put the FCMs to human perception pattern recognition to find out internal relevance between the visual behavior and cognitive process [4]. Abdollah, Mohammad, Mosavand Shahriar proposed a classification method of intraductal breast lesions [5]. Ma, Yang and Zhai constructed a novel classifier model FCMCM based on fuzzy cognitive map, which consists of model structure, activation functions, inference rules and learning algorithms [6].

Although the methods above extend the use of FCMs to the classification process, some problems are still pending: these methods do not put forward a full set of ensemble scheme for multiple classifiers; furthermore, the lack of the necessary robustness of a single FCMs classifier limits its generalization. To this end, a new ensemble model EFCMC<sub>QC</sub> is proposed in which every FCMCM employs a novel evolutionary algorithm inspired by quantum computation to provide an efficient solution to resolve these stated issues.

This paper is organized as follows. Section 2 presents the formalization representation of FCMs and introduces the research directions in using quantum computation as ensemble classifiers. In section 3, the ensemble of fuzzy cognitive map classifiers based on quantum computation, i.e., EFCMC<sub>QC</sub>, is implemented in the classification process for the ensemble classifiers. In Section 4 the performance of FCMCM\_QC model is evaluated with other traditional classifiers using some well selected UCI datasets. The paper ends with a summary and an outline of future research work on the application of FCMCM\_QC in dynamic and real time systems using the proposed framework.

## 2 Theory Basis and Formalization Representation

### 2.1 Basic Concepts of FCM

For two different nodes  $c_i$  and  $c_j$  in the FCM, if the value  $x_i$  of the node  $c_i$  changes, the value  $x_j$  of the node  $c_j$  changes consequently. It can be said that

the nodes  $c_i$  and  $c_j$  have the causality relationship. The arc from the node  $c_i$  to  $c_j$  is called a directed arc. The node  $c_i$  is called the reason node, and the node  $c_j$  is called the result node.

Let  $C = \{c_1, c_2, \dots, c_N\}$  be a finite set of vertices in FCM, where  $N$  is the number of nodes, for two any nodes  $c_i$  and  $c_j$ , and the finite set  $E = \{e_{11}, e_{12}, \dots, e_{1N}, e_{21}, e_{22}, \dots, e_{2N}, \dots, e_{N1}, e_{N2}, \dots, e_{NN}\}$  in the FCM, each arc has a corresponding weight  $w_{ij}$  indicating the influence of  $c_i$  to  $c_j$ .

For any nodes  $c_i$  and  $c_j$  in FCM, if there exists a directed relation, the interval  $[-1, 1]$  can be used to describe the influence degree, i.e.,  $w_{ij} \in E$ , and  $w_{ij}$  is the weight of  $c_i$  to  $c_j$ .

This paper introduces a learning method, which uses a real-coded quantum computation algorithm to develop FCM connection matrix based on historical data consisting of one sequence of state vectors. With this regard, it is advantageous to identify main differences between the approach taken here and those already reported in the literature. A concise comparative summary of the learning scenarios is stated in Table 1. This table includes the methods considering essential design factors such as the algorithm, learning goal, type of data used, type of transformation function, the node of FCMs and types of proposed learning methods.

Table 1  
Comparative summary of the learning scenarios

Algorithm	Reference	Learning Goal	Type of data used	FCMs type		Learning type
				Transform function	# nodes	
DHL	Dickerson, Kosko (1994) [7]	Con.matrix	single	N/A	N/A	Hebbian
BDA	Vazquez (2002) [8]	Con.matrix	single	binary	5, 7, 9	modified Hebbian
NHL	Papageorgiou (2003) [9]	Con.matrix	single	continuous	5	modified Hebbian
GS	Koulouriotis (2001) [10]	Con.matrix	multiple	continuous	7	genetic
GA	Georgopoulos (2009) [11]	Con.matrix	multiple	continuous	N/A	genetic
BB-BC	E. Yesil	Con.matrix	single	continuous	5	BB-BC

	(2010) [12]					
RCGA	Wojciech Stach (2010) [13]	Con.matrix	single	continuous	5	RCGA
PSO	Koulouriotis (2003) [45]	initial vector	multiple	continuous	5	particle swarm optimization
QC	This paper	Con.matrix	single	continuous	matrix	quantum computation

Note: Single – historical data consisting of one sequence of state vectors, Multiple – historical data consisting of several sequences of state vectors, for different initial conditions.

## 2.2 Ensemble Classifiers

### 2.2.1 Main Idea of Ensemble Classifiers

The main idea behind the ensemble classifiers is to weigh several individual classifiers, and combine them to obtain a classifier outperforming every one of them. The resulting classifier (hereafter referred to as an ensemble) is generally more accurate than any of the individual classifiers making up the ensemble.

The typical ensemble classifiers for classification problems include the following building blocks:

**Training set**—A labelled dataset used for ensemble learning. The training set, most frequently, is represented using attribute-value vectors. We use the notation  $A$  to denote the set of input attributes containing attributes:  $A = \{a_1, \dots, a_i, \dots, a_n\}$  and  $y$  to represent the class variable.

**Base Inducer**—The inducer is an induction algorithm that obtains a training set and forms a classifier representing the generalized relationship between the input attributes and the target attribute. Let  $I$  represent an inducer. Then a classifier  $C$  is represented using the notation  $C = I(A)$  induced by  $I$  on a training set  $A$ .

**Diversity Generator**—The diversity generator is responsible for generating classifiers with diverse classification performance.

**Combiner**—The combiner is responsible for combining various classifiers.

The motivation is to devise a cost-effective ensemble method, SD-EnClass, which is not influenced by the base classifiers and shows consistently improved detection rates compared to the base classifiers in the combination (Table 2).

Table 2  
Characteristic of different ensemble classifiers

Algorithm	Reference	Training approach	Classifiers	Decision fusion	Advantage	Weakness
Bagging	Breiman, (1996) [14]	re-sampling	unstable learner trained over re-sampled sets outputs different models	majority voting	simple and easy to understand and implement	accuracy value lower than other ensemble approaches
Roughly Balanced Bagging	Shohei, (2008) [15]					
Boosting	Freund, (1995) [16]	re-sampling	weak learner re-weighted in every iteration	weighted majority voting	performance of the weak learner boosted manifold	degrades with noise
adaboost	Schapiro, (1997) [17]					
AdaBoostM1	Freund, (1996) [18]					
AdaBoostMH	Schapiro, (1999) [19]					
stack generalization	Wolpert, (1992) [20]	re-sampling and k-folding	diverse base classifiers	meta-classifier	good performance	storage and time complexity

### 2.2.2 Combining Methods for Ensemble Classifiers

The way of combining more classifiers may be divided into two main categories, i.e., simple multiple classifiers combinations and meta-combiners. The former are best suited for problems where the individual classifiers perform the same task and have comparable performance. Such combiners, however, are more vulnerable to outliers and to unevenly performing classifiers. On the other hand, the latter are theoretically more powerful but are susceptible to all the problems associated with such added learning as over-fitting, long training time, etc. Here simple multiple classifier combination mechanisms are summarized as follows while more details about the meta-combiners can be found in many references [21-23].

(1) Dempster-Shafer. The idea of using the Dempster-Shafer theory of evidence (Buchanan and Shortliffe, 1984) for combining models is suggested by Shilen[24-25]. This method uses the notion of basic probability assignment defined for a certain class  $ci$  given the instance  $x$ .

(2) Naïve Bayes. Naïve Bayes idea for combining classifiers is extended by using Bayes' rule.

(3) Entropy Weighting. The idea in this combining method is to give each classifier a weight that is inversely proportional to the entropy of its classification vector.

(4) Density-based Weighting. When the various classifiers were trained using datasets obtained from different regions of the instance space, it might be useful to weight the classifiers according to the probability of sampling  $x$  by classifier  $M_k$ .

(5) Distribution Summation. This combining method was presented by Clark and Boswell [26]. The idea is to sum up the conditional probability vector obtained from each classifier. The selected class is chosen according to the highest value in the total vector.

### 2.2.3 Current use of Ensemble Classifiers

The ensemble classifiers are suitable in such fields as: finance [27], bioinformatics [28], healthcare [29], manufacturing [30], geography [31], predicting protein fold patterns [32], early diagnosis of alzheimer's disease [33], microarray cancer [34], etc.

## 2.3 Concepts for Quantum Computation in Ensemble Classifiers

### 2.3.1 Quantum Bit in Ensemble Classifiers (ECQ-bit)

In 1980 Richard Feynman showed the possibility of the use of quantum effects in data processing. Later in 1994, Shor demonstrated that quantum computation (QC) can solve efficiently NP-hard problem [35]. Shor described a polynomial time quantum algorithm for factoring numbers. In 1996 Grover [36] presented a quadratic algorithm for database search.

In today's digital computers, the smallest information unit is one bit being either in the state "1" or "0" at any given time. In contrast, the basic concepts of quantum computation algorithm (QC) are well addressed by Han and Kim [37] and the main idea of QC is based on the concepts of quantum bits and superposition of states. The smallest unit of information stored in a quantum computer is called a quantum bit or Q-bit, which may be in the "0" state, in the "1" state, or in any superposition of the two. Obviously, QC uses a new representation to store more information in the concept of Q-bit.

In a binary classification problem with the Ensemble classifiers, just as the Q-bit, the ECQ-bit can also be described by the "0" state and the "1" state, where the former represents that one classifier is not selected, while the latter represents that one classifier is just selected.

As QC develops, the ECQ-bit chromosome converges to a single state, i.e., 0 or 1, and the property of diversity disappears gradually. The state of ECQ-bit can be represented as QC-bit [37]:

$$|\varphi\rangle = \alpha|0\rangle + \beta|1\rangle = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad (1)$$

Where  $\alpha$  and  $\beta$  are complex numbers that specify the probability amplitudes of the corresponding states,  $|\alpha|^2$  and  $|\beta|^2$  denote the probability that the ECQ-bit is found in the “1” state and “0” state respectively, i.e., one classifier is selected or not in the ensemble classifiers. Normalization of the state to the unity satisfies

$$|\alpha|^2 + |\beta|^2 = 1. \quad (2)$$

### 2.3.2 ECQ-Bit Individual and Quantum Gates

As mentioned above, in an ECQ-bit individual of ECQ-bits the resulting state space has  $2^m$  dimensions. Consequently, the exponential growth of the state space with the number of particles suggests a possible exponential speed-up of computation on quantum computers over the classical computers. The measurement of an ECQ-bit individual projects the quantum state of ECQ-bit individual onto one of the single states. The result of a measurement is probabilistic and the process of measurement changes the state to what is known as quantum collapse, i.e., “0” or “1”.

Just as the process of Quantum computation, the state of an ECQ-bit individual can be changed by a quantum gate. A quantum gate is a reversible gate and can be represented as a unitary operator  $U$  acting on the quantum states satisfying  $U^+ U = U U^+$ , which is the Hermitian adjoint of  $U$ . There are several commonly used quantum gates, such as the NOT gate, the rotation gate, the controlled NOT gate and the Hadamard gate [38].

Here the rotation gate and the NOT gate are employed as the quantum gates respectively. The operation of the rotation gate  $R_i$  on each  $q_j^i$  in a population consisting of  $m$  members is defined as follows:

$$\begin{pmatrix} \alpha_j^{t+1} \\ \beta_j^{t+1} \end{pmatrix} = R_i^t \begin{pmatrix} \alpha_j^t \\ \beta_j^t \end{pmatrix}, \quad j=1,2,\dots,m. \quad (3)$$

where  $m$  is the number of ECQ-bits in the  $i$ th ECQ-bit individual and the rotation gate is defined as:

$$R_i^t = \begin{pmatrix} \cos(\Delta\theta_i) & -\sin(\Delta\theta_i) \\ \sin(\Delta\theta_i) & \cos(\Delta\theta_i) \end{pmatrix}, \quad i=1,2,\dots,n. \quad (4)$$

### 3 Ensemble of Fuzzy Cognitive Map Classifiers Based on Quantum Computation (EFCMCQC)

#### 3.1 Generating Base FCM Classifiers

Just as the bagging algorithm [14], for the given training set  $D$  of size  $n$ ,  $m$  new training sets are generated, each of size  $n'$ , by sampling from  $D$  uniformly and with replacement. Then  $m$  FCM classifiers are generated respectively with these  $m$  training sets as Ma [6].

#### 3.2 Ensemble Algorithm

The pseudo code algorithm for ensemble of FCM classifiers based on quantum computation (EFQC) is described as follows:

```

1 Begin
2  $t=0$ 
3   Initialize  $Q(t)$  of FCM classifiers
4   Observe  $P(t)$  by observing the states of  $Q(t)$ 
5   Repair  $P(t)$ 
6   Evaluate  $P(t)$ 
7   Store the best solutions among  $P(t)$  into  $B(t)$  and best solution  $b$  among  $B(t)$ 
8   While ( $t < \text{MAXGEN}$ )
     Begin
        $t = t + 1$ 
       Observe  $P(t)$  by observing the states of  $Q(t-1)$ 
       Evaluate  $P(t)$ 
       Update  $Q(t)$  using quantum rotation gate or NOT gate
       Store the best solutions among  $P(t)$  and  $B(t-1)$  into  $B(t)$  and the best solution  $b$  among  $B(t)$ 
     end
9   Take all classifiers  $h_i$  with value 1 of the ECQ-bit individual and then classify the training set to Generate

$$h_i : X \rightarrow \{-1, 1\}$$


```



10 Compute the classification error  $\varepsilon$  with the distribution  $D$  and the weight  $w_i$  for the  $i$ th classifier

$$\varepsilon = P_D(h_i(x_i) \neq y_i) = \frac{\sum_i w_i^t \cdot I(f_i(x_i) \neq y_i)}{\sum_i w_i^t}$$

$$\text{where } w_i = \frac{1}{2} \ln \left( \frac{1-\varepsilon}{\varepsilon} \right)$$

11 Generate the ensemble classifiers  $H_{final}(x)$  and compute the classification result for the instance  $x$

$$H_{final}(x) = \text{sign} \left( \sum_{i=1}^m w_i \cdot h_i(x) \right)$$

12 End

In step 3 and step 4, the  $Q(t)$  and  $P(t)$  are defined as reference[46]. When running the step 3, in formula 2 and the function initialize  $Q(t)$ , the value  $\alpha$  and  $\beta$  are initialized to the same probability amplitude  $1/\sqrt{2}$ , so that all classifiers are selected with the same probability at the beginning. As for the method to measure the population, a random  $\eta$  is produced  $\eta \in [0,1]$ , the measurement result can be calculated as follows.

$$Mr = \begin{cases} 1 & \eta \geq |\alpha_i|^2 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

In the step 6, the  $P(t)$  is evaluated, and such fitness function is applied as following.

$$\text{Fitness}(Q(t)) = \frac{1}{n} \sum_{i=1}^n \text{precision}(H_{final}(x_i)) \quad (6)$$

Where  $\text{Fitness}(Q(t))$  represents the average precision of the instances with the ensemble classifier.

In the step 8, when the  $Q(t)$  is updated with the quantum rotation gate, the look-up table as shown in Table 3 is employed.

In Table 3,  $x_i$  and  $b_i$  are the current and the best solution,  $\alpha_i$  and  $\beta_i$  are the probability amplitude of  $x_i$  and  $b_i$  respectively.

Table 3  
Lookup table of the rotation angle

$x_i$ $\Delta\theta_i$	$b_i$	$f(x) < f(b)$	$\alpha_i$				$\beta_i$	
			$\alpha_i\beta_i > 0$	$\alpha_i\beta_i < 0$	$\alpha_i = 0$	$\beta_i = 0$	$\alpha_i = 0$	$\beta_i = 0$
0	0	False	0	0	0	0	0	0
0	0	True	0	0	0	0	0	0
0	1	False	0	0	0	0	0	0
0	1	True	$0.05\pi$	-1	+1	$\pm 1$	0	0
1	0	False	$0.01\pi$	-1	+1	$\pm 1$	0	0
1	0	True	$0.025\pi$	+1	-1	0	$\pm 1$	$\pm 1$
1	1	False	$0.005\pi$	+1	-1	0	$\pm 1$	$\pm 1$
1	1	True	$0.025\pi$	+1	-1	0	$\pm 1$	$\pm 1$

In the step 10, a learner  $L$  finds a weak hypothesis  $h_t$ : given the training set and  $D_t$ . Each round,  $t=1, \dots, T$ , base learning algorithm accepts a sequence of training examples ( $S$ ) and a set of weights over the training example  $D_t(i)$ . Initially, all weights are set equally, but each round the weights of incorrectly classified examples are increased so that those observations that the previous classifier poorly predicts receive greater weight on the next iteration.

In the step 11, the combined hypothesis  $H$  is a weighted majority vote of the  $m$  weak hypotheses. Each hypothesis  $h_t$  has a weight  $w_t$ .

## 4 Experiments and Results

### 4.1 UCI Datasets

The characteristics of the data sets used to evaluate the performance of the proposed ensemble techniques are given below in Table 4. The data sets are obtained from the University of Wisconsin Machine Learning repository as well as the UCI data set repository [39]. These data sets were specially selected such that (a) all come from real-world problems, (b) have varied characteristics, and (c) have shown valuable application for many data mining algorithms.

Table 4  
Specification of UCI data sets

Data set	Number of attributes	Number of the classes	Total instances
breast-cancer-w	9	2	699
glass	9	6	214
hypo	7	5	3772
iris	4	3	159
labor	8	2	57
letter	16	26	20000
satellite	36	6	6435
sick	7	2	3772
splice	60	3	3190
vehicle	18	4	846

## 4.2 Methodology

To conduct the evaluation, all the previously mentioned data sets have a reasonable number of observations and they were all partitioned into ten-fold cross validation sets randomly. Each fold was used as an independent test set in turn, while the remaining nine folds were used as the training set.

Boosting [16] and Bagging [14] are two relatively new but widely used methods for generating ensembles. To better evaluate the performance of the proposed ensemble classifiers, it is compared with the Bagging and Boosting ensembles with the neural networks [40] and the C4.5 [41] as the base classifiers. Cross validation folds are performed independently for each algorithm. Parameter details for the neural networks are as follows: a learning rate of 0.15, a momentum term of 0.9, and weights are initialized randomly to be between -0.5 and 0.5. The number of hidden units based on the number of input and output units are selected. For the decision trees, the C4.5 and pruned trees are used which empirically produce better performance. We employed the Rotation gate and the NOT gate with the EFCMC<sub>QC</sub> respectively. Results in Table 5 are the average error rate performed with ten cross validation folds by the Neural Network, C4.5 and EFCMC<sub>QC</sub> respectively.

### Experiment 1: Error Rates

Table 5 summarizes the average test error rates obtained when applying proposed EFCMC<sub>QC</sub> in the the UCI data sets. It contains also the average test error rates of Bagging and Boosting ensembles with the neural networks and the C4.5 as the base classifiers for comparison. It can be noticed that our ensembles (EFCMC<sub>QC</sub>), not only in the smallest data set labor (57 samples only) but the biggest data set

letter(as seen in Table 4, the number of samples is up to 20000), gives the most dominated superiority.

Table 5  
Test set error rates

Data Set	Neural Network			C4.5			EFCMC <sub>QC</sub>			
	Standar rd	Bagg g	Ada_Bo ost	Standar d	Bagg ing	Ada_Bo ost	Rotation gate	NOT gate	Rank	
									Rotation gate	NOT gate
breast-cancer-w	3.4	3.4	4.0	5.0	3.7	3.5	3.1	3.4	1	2
glass	38.6	33.1	31.1	27.8	24.4	25.7	15.9	15.7	2	1
hypo	6.4	6.2	6.2	2.5	2.4	2.4	2.8	2.9	4	5
iris	4.3	4.0	3.9	5.2	4.9	5.6	3.4	3.6	1	2
labor	6.1	4.2	3.2	16.5	13.7	11.6	5.7	6.6	3	5
letter	18.0	10.5	4.6	14.0	7.0	3.9	3.0	3.3	1	2
satellite	13.0	10.6	10.0	13.8	9.9	8.4	6.7	7.8	1	2
sick	5.9	5.7	4.5	1.3	1.2	1.0	0.9	1.2	1	3
splice	4.7	3.9	4.2	5.9	5.4	5.3	3.9	4.8	1	5
vehicle	24.9	20.7	19.7	29.4	27.1	22.9	18.8	19.0	1	2
Average	12.53	10.23	9.14	12.14	9.97	9.03	6.42	6.83	1.5	2.8
Rank	8	6	4	7	5	3	1	2	--	--

## Experiment 2: Average test-set error over with ensemble size

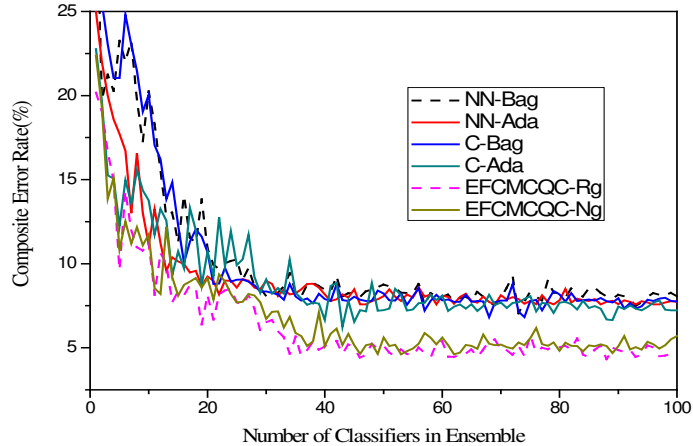


Figure 1

Average test-set error over all 10 UCI data sets used in our studies for ensembles incorporating from one to 100 decision trees or neural networks.

Hansen and Salamon [42] had revealed that ensembles with as few as ten members are adequate to sufficiently reduce test-set error. Although it would be fitful for the earlier proposed ensembles, based on a few data sets with decision trees, Schapire [17] recently suggested that it is possible to further reduce test-set error even after ten members have been added to an ensemble. To this end, the proposed EFCMC<sub>QC</sub> are applied with Bagging and Boosting ensembles to further investigate the performance variety relative to the size of an ensemble.

Figure 1 illustrates the composite error rate over all of ten UCI data sets for EFCMC<sub>QC</sub> with Bagging and Boosting ensembles using up to 100 classifiers. The comparison results show that most of the methods produce similarly shaped curves. Figure 3 also provides numerous interesting insights. The first is that much of the reduction in error due to adding classifiers to an ensemble comes with the first few classifiers as expected. The second is that some variations exist in reality with respect to where the error reduction finally flattens. The last but the most important is that for both Bagging and Ada\_Boosting applied to neural networks and the C4.5, much of the reduction in error appears to have occurred after approximately 40 classifiers, which is partly consistent with the conclusion of Breiman [14]. To our surprise, at 30 classifiers the error reduction for the proposed EFCMC<sub>QC</sub> appears to have nearly asymptote to a plateau. Therefore, the results reported in this paper suggest that our method reaches convergence with the smallest ensemble size.

### Experiment 3: Robustness of EFCMC<sub>QC</sub>

To further test the EFCMC<sub>QC</sub> performance, the different noise ratio is employed as Dietterich [43] in the four datasets satellite, sick, splice and vehicle. The four different noise ratios  $\gamma=0\%$ ,  $\gamma=5\%$ ,  $\gamma=10\%$  and  $\gamma=15\%$  are implied in the ten standard 10-fold cross validation, and results are described as follows:

Table 6  
CV result on the satellite dataset

Noise rate	Algorithm/Classifier	EFCMC <sub>QC</sub>	NN-Bag	NN-Ada	C4.5-Ada
$\gamma=0\%$	C4.5- Bag	2- 4- 4	3- 3- 4	3- 2- 5	2- 4- 4
	C4.5-Ada	3- 3- 4	5- 3- 2	4- 4- 2	
	NN-Ada	2- 4- 4	4- 3- 3		
	NN-Bag	3- 3- 4			
$\gamma=5\%$	C4.5- Bag	3- 2- 5	4- 4- 2	4- 5- 1	3- 2-5
	C4.5-Ada	3- 2- 5	5- 2- 3	5- 2- 3	
	NN-Ada	2- 3- 5	3- 4- 3		
	NN-Bag	2- 2- 6			
$\gamma=10\%$	C4.5- Bag	3- 2- 5	4- 3- 3	7- 1- 2	5- 2-3
	C4.5-Ada4	2- 2- 6	3- 4- 3	5- 2- 3	

$\gamma=15\%$	NN-Ada	2- 1- 7	2- 5- 3		
	NN-Bag	3- 2- 5			
	C4.5- Bag	3- 3- 4	4- 2- 4	7- 0-3	6- 2-3
	C4.5-Ada	0- 3- 7	2- 4- 4	5- 3- 2	
	NN-Ada	0- 2- 8	2- 1- 7		
	NN-Bag	3- 3- 4			

Table 7  
CV result on the sick dataset

Noise rate	Algorithm/ Classifier	EFCMC QC	NN- Bag	NN- Ada	C4.5- Ada
$\gamma=0\%$	C4.5- Bag	3- 1- 6	5- 3- 2	5- 2- 3	3- 2-5
	C4.5-Ada	4- 1- 5	6- 2- 2	6- 1- 3	
	NN-Ada	2- 1- 7	5- 1- 4		
	NN-Bag	2- 0- 8			
$\gamma=5\%$	C4.5- Bag	4- 1- 5	5- 3- 2	6- 2- 2	3- 4-3
	C4.5-Ada	3- 1- 6	4- 4- 2	5- 2- 3	
	NN-Ada	2- 0- 8	4- 3- 4		
	NN-Bag	3- 1- 6			
$\gamma=10\%$	C4.5- Bag	4- 3- 3	4- 3- 3	7- 2- 1	5- 3-2
	C4.5-Ada	2- 2- 6	3- 4- 3	5- 3- 2	
	NN-Ada	1- 1- 8	3- 3- 4		
	NN-Bag	3- 3- 4			
$\gamma=15\%$	C4.5- Bag	4- 3- 3	4- 2- 4	8- 1- 1	7- 1-2
	C4.5-Ada	2- 1- 7	3- 2- 5	5- 2- 3	
	NN-Ada	0- 1- 9	6- 3- 1		
	NN-Bag	3- 6- 1			

Table 8  
CV result on the splice dataset

Noise rate	Algorithm/ Classifier	EFCMC QC	NN- Bag	NN- Ada	C4.5- Ada
$\gamma=0\%$	C4.5- Bag	2- 0- 8	3- 1- 6	3- 2- 5	4- 2- 4
	C4.5-Ada	3- 1- 6	3- 1- 6	3- 4- 3	
	NN-Ada	3- 2- 5	4- 1- 5		
	NN-Bag	3- 3- 4			
$\gamma=5\%$	C4.5- Bag	3- 0- 7	3- 1- 6	4- 4- 2	5- 3-2
	C4.5-Ada	3- 0- 7	3- 0- 7	2- 4- 4	
	NN-Ada	3- 3- 4	3- 3- 4		
	NN-Bag	3- 3- 4			
$\gamma=10\%$	C4.5- Bag	4- 1- 5	3- 3- 4	5- 3- 2	5- 4-1
	C4.5-Ada	2- 0- 8	2- 1- 7	3- 3- 4	
	NN-Ada	3- 2- 5	3- 3- 4		

	NN-Bag	3- 3- 4	1	3	4
$\gamma=15\%$	C4.5- Bag	4- 2- 4	3- 2- 5	3- 5- 2	5- 4- 1
	C4.5-Ada	1- 0- 9	0- 1- 9	2- 2- 6	
	NN-Ada	2- 2- 6	2- 3- 5		
	NN-Bag	4- 3- 3			

Table 9  
CV result on the vehicle dataset

Noise rate	Algorithm/Classifier	EFCMC <sub>QC</sub>	NN-Bag	NN-Ada	C4.5-Ada
$\gamma=0\%$	C4.5- Bag	0- 0- 10	2- 1- 7	0- 2- 8	3- 1- 6
	C4.5-Ada	2- 0- 8	3- 1- 7	2- 2- 6	
	NN-Ada	3- 1- 6	4- 3- 3		
	NN-Bag	2- 1- 7			
$\gamma=5\%$	C4.5- Bag	2- 0- 8	2- 0- 8	3- 2- 5	4- 4- 2
	C4.5-Ada	2- 0- 8	1- 1- 8	2- 2- 6	
	NN-Ada	3- 0- 7	3- 4- 3		
	NN-Bag	4- 3- 3			
$\gamma=10\%$	C4.5- Bag	4- 0- 6	2- 3- 5	4- 5- 1	5- 4- 1
	C4.5-Ada	3- 0- 7	1- 0- 9	2- 2- 6	
	NN-Ada	3- 1- 6	3- 1- 6		
	NN-Bag	4- 3- 3			
$\gamma=15\%$	C4.5- Bag	4- 0- 6	3- 3- 4	4- 5- 1	6- 3- 1
	C4.5-Ada	2- 0- 8	1- 0- 9	3- 2- 5	
	NN-Ada	3- 0- 7	3- 1- 6		
	NN-Bag	4- 3- 3			

From Table 6 to Table 9 it shows the average results of the ten-fold cross validation for the EFCMC<sub>QC</sub> and other ensembles. Every number, e.g., 2-4-4 in the first place of Table 6, represents the times of performance results of the C4.5-Bag better than that of the EFCMC<sub>QC</sub> is 2, equal to the EFCMC<sub>QC</sub> is 4, and worse than the EFCMC<sub>QC</sub> is 4.

These results also suggest that (a) all the ensembles can, at least to a certain extent, be inevitably affected by the noise. (b) the EFCMC<sub>QC</sub> has more resistant not only to the ensembles with C4.5 but to the neural network, especially with greater noise ratio. The conclusion generally happened on the same dataset (e.g., satellite, sick, sick and vehicle) for all three ensemble methods. (3) for a problem with noise, focusing on misclassified examples would, to a great degree, cause a classifier tend to focus on boosting the margins of (noisy) examples that would in fact be misleading in overall classification which is obvious especially when the noise ratio becomes bigger. This phenomenon is consistent with the conclusion of David and Richard [44]. And (4) with the noise ratio becoming bigger, the

superiority of the EFCMC<sub>QC</sub> in classification gradually becomes apparent. The bigger the noise ratio is, the more obvious its advantage is.

### **Conclusions and Future Directions**

The paper has discussed relevant work, and proposed a novel ensemble strategy for the FCM classifiers, based on a novel evolutionary algorithm inspired by quantum computation for the FCMs. In this study, a comprehensive ensemble EFCMC<sub>QC</sub> was then developed for the design of fuzzy cognitive maps classifier. It has been demonstrated how the EFCMC<sub>QC</sub> helps construct classifier addressing the classification issue on a basis of numeric data. The feasibility and effectiveness of the EFCMC<sub>QC</sub> were showed and quantified via series of numeric experiments with the UCI datasets. The follow-up results show that the proposed ensemble method is very effective and precede such traditional classifiers ensemble model such as the bagging and boosting. The future work will concern the use of the learning method in a context of more practical applications such as classification in dynamic systems and complex circumstances. Consummating the proposed EFCMC<sub>QC</sub> and exploring applying area will be investigated.

### **Acknowledgment**

This research is partially supported by the National Natural Science Foundation of China under grants 61300078 and 61175048.

### **References**

- [1] B. Kosko, "Fuzzy Cognitive Maps", *Journal of Man-Machine Studies*, Vol. 24, No. 1, pp. 65-75, 1986
- [2] D.-S. Chrysostomos and P.-G. Peter, "Modeling Complex Systems Using Fuzzy Cognitive Maps", *IEEE Transactions on Systems Man and Cybernetics- Part A Systems and Humans*, Vol. 34, No. 1, pp. 155-166, 2004
- [3] Z. Peng, B.-R. Yang, C.-M. Liu, "Research on One Fuzzy Cognitive Map Classifier", *Application Research of Computers*, Vol. 26, No. 5, pp. 1757-1759, 2009
- [4] D.-Y. Zhu, B. S-. Mendis and T. Gedeon, "A Hybrid Fuzzy Approach for Human Eye Gaze Pattern Recognition", In *Proceedings of International Conference on Neural Information Processing of the Asia-Pacific Neural Network Assembly*, Berlin, Germany, pp. 655-662, 2008
- [5] A. Abdollah, R. Mohammad, B. Mosavi and F. Shahriar, "Classification of Intraductal Breast Lesions Based on the Fuzzy Cognitive Map", *Arabian Journal for Science and Engineering*, Vol. 39, No. 5, pp. 3723-3732, 2014
- [6] N. Ma, B.-R. Yang, Y. Zhai, G.-Y. Li and D.-Z. Zhang, "Classifier Construction Based on the Fuzzy Cognitive Map", *Journal of University of Science and Technology Beijing*, Vol. 34, No. 5, pp. 590-595, 2012



- [7] J. Dickerson and B. Kosko, "Virtual Worlds as Fuzzy Cognitive Maps", Presence, Vol. 3, No. 2, pp. 173-189, 1994
- [8] A. Vazquez, "A Balanced Differential Learning Algorithm in Fuzzy Cognitive Maps", In: Proceedings of the 16<sup>th</sup> International Workshop on Qualitative Reasoning, Barcelona, Spain, pp. 10-12, 2002
- [9] E. Papageorgiou, C.-D. Stylios and P.-P. Groumpos. "Fuzzy Cognitive Map Learning Based on Nonlinear Hebbian Rule", In Australian Conference on Artificial Intelligence, Perth, Australia, pp. 256-268, 2003
- [10] D.-E. Koulouriotis, I.-E. Diakoulakis and D.-M. Emiris, "Learning Fuzzy Cognitive Maps Using Evolution Strategies: a Novel Schema for Modeling and Simulating High-Level Behaviour", In IEEE Congress on Evolutionary Computation (CEC2001), Seoul, Korea, pp. 364-371, 2001
- [11] V.-C. Georgopoulos and C.-D. Stylios, "Diagnosis Support using Fuzzy Cognitive Maps combined with Genetic Algorithms", 31<sup>st</sup> Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, pp. 6226-2669, 2009
- [12] Y. Engin and U. Leon, "Big Bang-Big Crunch Learning Method for Fuzzy Cognitive Maps", International Scholarly and Scientific Research & Innovation, Vol. 4, No. 11, pp. 693-702, 2010
- [13] W. Stach, L. Kurgan, and W. Pedrycz, "Expert-based and Computational Methods for Developing Fuzzy Cognitive Maps", Fuzzy Cognitive Maps, STUDEFUZZ 247, pp. 23-41, 2010
- [14] L. Breiman, "Bagging Predictors", Machine Learning, Vol. 24, No. 2, pp. 123-140, 1996
- [15] S. Hido and H. Kashima, "Roughly Balanced Bagging for Imbalanced Data", In: 2008 SIAM International Conference on Data Mining, pp. 143-152, 2008
- [16] Y. Freund, "Boosting a Weak Learning Algorithm by Majority", Information and Computation, Vol. 121, No. 2, pp. 256-285, 1995
- [17] R. E. Schapire, Y. Freund, P. Bartlett and W. Lee, "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods", In Proceedings of the Fourteenth International Conference on Machine Learning, Nashville, USA, 1997
- [18] Y. Freund and R.-E. Schapire, "Experiments with a New Boosting Algorithm", International Conference on Machine Learning, pp. 148-156, 1996
- [19] E.-R. Schapire and Y. Singer, "Improved Boosting Algorithms Using Confidence-rated Predictions", Machine Learning, Vol. 37, No. 3, pp. 297-336, 1999

- [20] D. Wolpert, "Staked Generalization", *Neural Networks*, Vol. 5, No. 2, pp. 241-259, 1992
- [21] S. Džeroski and B. Ženko, "Is Combining Classifiers with Stacking Better than Selecting the Best One?", *Machine Learning*, Vol. 54, No. 3, pp. 255-273, 2004
- [22] P.-K. Chan and S.-J. Stolfo, "Toward Parallel and Distributed Learning by Metalearning", In *AAAI Workshop in Knowledge Discovery in Databases*, Washinton D.C., USA, 1993
- [23] P.-K. Chan and S.-J. Stolfo, "On the Accuracy of Meta-learning for Scalable Data Mining", *Intelligent Information Systems*, Vol. 8, pp. 5-28, 1997
- [24] S. Shilen, "Multiple Binary Tree Classifiers", *Pattern Recognition*, Vol. 23, No. 7, pp. 757-763, 1990
- [25] S. Shilen, "Nonparametric Classification Using Matched Binary Decision Trees", *Pattern Recognition Letters*, Vol. 13, pp. 83-87, 1992
- [26] P. Clark and R. Boswell, "Rule Induction with CN2: Some Recent Improvements", *Proceedings of the 5<sup>th</sup> European Working Session on Learning*, Springer-Verlag, pp. 151-163, 1989
- [27] W. Leigh, R. Purvis and J.-M. Ragusa, "Forecasting the NYSE Composite Index with Technical Analysis, Pattern Recognizer, Neural Networks, and Genetic Algorithm: a Case Study in Romantic Decision Support", *Decision Support Systems*, Vol. 32, No. 4, pp. 361-377, 2002
- [28] A. -C. Tan, D. Gilbert and Y. Deville, "Multi-class Protein Fold Classification using a New Ensemble Machine Learning Approach", *Genome Informatics*, Vol. 14, pp. 206-217, 2003
- [29] P. Mangiameli, D. West and R. Rampal, "Model Selection for Medical Diagnosis Decision Support Systems", *Decision Support Systems*, Vol. 36, No. 3, pp. 247-259, 2004
- [30] O. Maimon, and L. Rokach, "Ensemble of Decision Trees for Mining Manufacturing Data Sets", *Machine Engineering*, Vol. 4, No. 2, pp. 32-57, 2001
- [31] L. Bruzzone, R. Cossu and G. Vernazza, "Detection of Land-Cover Transitions by Combining Multidate Classifiers", *Pattern Recognition Letters*, Vol. 25, No. 13, pp. 1491-1500, 2004
- [32] W. Chen, X. Liu, Y. Huang, Y. Jiang, Q. Zou and C. Lin, "Improved Method for Predicting Protein Fold Patterns with Ensemble Classifiers", *Genetics and Molecular Research*, Vol. 11, No. 1, pp. 174-181, 2012
- [33] F. Saima, A.-F. Muhammad and T. Huma, "An Ensemble-of-Classifiers Based Approach for Early Diagnosis of Alzheimer's Disease: Classification

- Using Structural Features of Brain Images”, Computational and Mathematical Methods in Medicine Volume, Vol. 2014, 2014
- [34] N. Sajid and K.-B. Dhruva, “Classification of Microarray Cancer Data Using Ensemble Approach”, Network Modeling Analysis in Health Informatics and Bioinformatics, Vol. 2, pp. 159-173, 2013
- [35] P.-W. Shor, “Algorithms for Quantum Computation: Discrete Logarithms and Factoring”, In Proceedings 35<sup>th</sup> Annu. Symp. Foundations Computer Science, Sante Fe, USA, pp. 124-134, 1994
- [36] L.-K. Grover, “A Fast Quantum Mechanical Algorithm for Database Search”, In Proceedings 28<sup>th</sup> ACM Symp, Theory Computing, Philadelphia PA, USA, 1996
- [37] K. Han and J. Kim, “Quantum-Inspired Evolutionary Algorithm for a Class of Combinatorial Optimization”, IEEE Transactions on Evolutionary Computation, Vol. 6, No. 6, pp. 580-593, 2002
- [38] Y. Kim, J.-H. Kim and K.-H. Han, “Quantum-inspired Multi Objective Evolutionary Algorithm for Multiobjective 0/1 Knapsack Problems”, In 2006 IEEE Congress on Evolutionary Computation, Vancouver, Canada, 9151-9156, 2006
- [39] C. Blake, E. Keogh and C.-J. Merz, UCI Repository of Machine Learning Databases. Irvine: Department of Information and Computer Science, University of California.  
<http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1998
- [40] D. Rumelhart, G. Hinton and R. Williams, “Learning Internal Representations by Error Propagation”, In Rumelhart, D., & McClelland, J. (Eds.), Parallel Distributed Processing: Explorations in the microstructure of cognition, Volume 1: Foundations Cambridge MA:MIT Press, pp. 318-363, 1986
- [41] J. Quinlan, “Introduction of Decision Trees”, Machine Learning, Vol. 1, No. 1, pp. 84-100, 1986
- [42] L. Hansen and P. Salamon, “Neural Network Ensembles”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, pp. 993-1001, 1990
- [43] T.-G. Dietterich, “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization”, Machine Learning, Vol. 40, No. 2, pp. 139-157, 2000
- [44] R.-L. Richard and P. -R. David, “Voting Correctly”, The American Political Science Review, Vol. 91, No. 3, pp. 585-598, 1997
- [45] K.-E. Parsopoulos, P. Groumpos and M.-N. Vrahatis, “A First Study of Fuzzy Cognitive Maps Learning Using Particle Swarm Optimization”, In

Proceedings of the IEEE 2003 Congress on Evolutionary Computation,  
Canberra, Australia, pp. 1440-1447, 2003

# Using Train Interconnection for Intra-train Communication via CAN

**Tamás Bécsi, Szilárd Aradi**

Budapest University of Technology and Economics, Department of Control for Transportation and Vehicle Systems, Stoczek u. 2, H-1111 Budapest, Hungary, email: becsi.tamas@mail.bme.hu, aradi.szilard@mail.bme.hu

**Péter Gáspár**

Computer and Automation Research Institute, Hungarian Academy of Sciences, Kende u. 13-17, H-1111 Budapest, Hungary, e-mail: gaspar.peter@sztaki.mta.hu

---

*Abstract: This paper presents the possibilities for using currently installed interconnection cables for extended intra-train communication purposes and proposes a CAN-based communication solution for the problem. The implementation of such a communication technique raises feasibility issues due to the non-standard physical media and the non-fixed network topologies, such as, the determination of an achievable bandwidth, the dynamic termination of the network and the enumeration of units. As a basis of on-board passenger information and telemetry solutions for the operator, this solution could be a cost effective way to improve the quality of service of rail transportation. This paper describes the specialties of the vehicular environment and the proposed network model; it then presents a network lookup algorithm for automated enrolling and ordering the needed train units. Laboratory and field measurements are presented to validate the feasibility of this solution.*

---

*Keywords: rail transport; vehicle; on-board communication; CAN Network; UIC558*

---

## 1 Introduction

Improving quality of service and reducing maintenance costs, in the area of rail services, are essential to preserve a competitive environment for this transportation mode over the various alternatives. As a consequence of these needs, on-line communication, telemetry and fleet management have gained a significant role in today's modern rail transport. Such systems utilize high-level integration of communication, data management and control systems. Another possibility could include the involvement of intra-train communications within the

fleet management systems. However, installing new physical communication media in all of the existing units can be very expensive. These facts lead to the idea of using the existing network for the extended communication needs. Multiple train communication networks exist for general train control purposes, such as remote traction control to handle push-pull train operations, door and light control or the audio channels. However, a demand for extended services of intra-train communications has risen. These extended systems can serve several purposes:

- Such systems can provide train specific data to the passenger information system improving passenger satisfaction by providing real-time information about the state of the journey, i.e. current delay, estimated arrival, connections etc. This information could be displayed in passenger cars via various kinds of displays, which obtain data from the on-board unit of the driver's stand.
- Another benefit is the collection of train information for the operator. The telemetry or the automatic enrolment of train units can be centralized at the train level with a closed communication solution, where the only connection to a central data center of the operator is managed by the on-board unit (OBU) of the driver's station.
- The system can improve the safety of railroad operations by the remote monitoring of various mechanical elements of a railcar. [3]

Although a standardized extension of the Train Communication Network (TCN) [7] [8] [9] exists in newer passenger and traction units, which provides additional channels and protocols for intra-train communication, the penetration is low and the need of addressing this problem in current units that are intended to be in service for a long term, is growing at the operator level. There have been various attempts to implement such a communication system on existing interconnections [5] [6].

Naturally, one must examine the pin allocation of the connectors for the current system and recognize wire pairs that do not carry safety-critical information, in order to maintain the safety level of train operation.

Our previous work deals with the theoretical aspects of this communication, via a non-standard physical layer, the non-fixed topology of the network, network length and speed [2].

This paper proposes a solution that uses CAN communications with the interconnection media. Section 2 briefly describes the specialties of the vehicular environment and the proposed network model. Section 3 presents a network lookup algorithm for automated enrolment and ordering of the train units. Section 4, laboratory and field measurements are presented that validate the feasibility of our solution.

## 2 Digital Data Transfer via Existing Interconnection

As a consequence, of the existing conditions of the current installation interconnections, the network topology of the train can only form two kinds of basic topologies: the chain and the bus topology as shown in Figure 2. Of these two, bus topology has been chosen in our research, since the implementation of chain topology frequently breaks the continuity of existing interconnections, which is undesirable. In our scope, the application of UIC 558 type 18 pin interconnection [1] is examined.

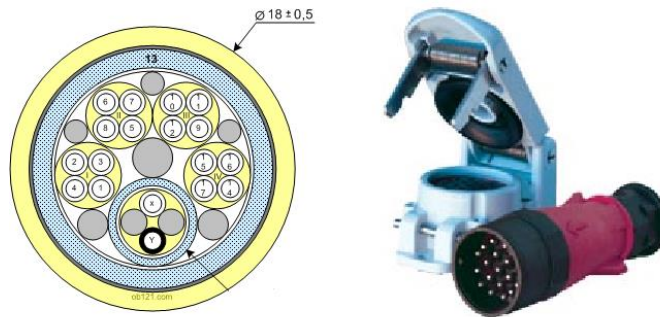
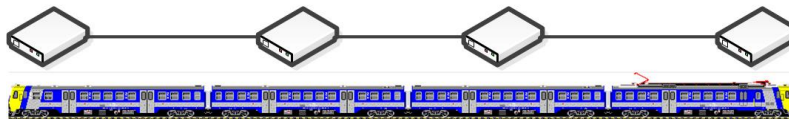


Figure 1

UIC 558 Series Cable and Connector (source: [www.hertfordcontrols.co.uk](http://www.hertfordcontrols.co.uk), [www.ob121.com](http://www.ob121.com))

The function allocation of the UIC 558 wire is shown in Table 1. Theoretically, one could implement digital signal transfer on any wire within interconnection, wires 9-18 transfer safety critical functions and therefore, the utilization of only the audio-related wires is desired.

Chain Topology



Bus topology

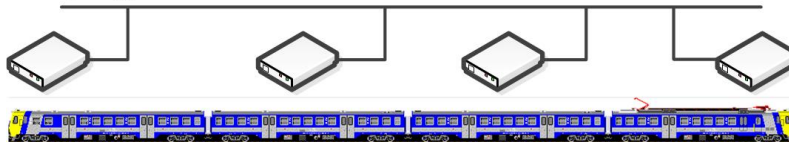


Figure 2

The chain and bus topology

The network topology of the train is not fixed, since the individual units can form different configurations of, length and in addition, the traction or control units can take any position within the train. These features of a modular network pose several problems:

- The switchable galvanic separation of communication units from the media
- The priority of the original function
- The adequate or even adaptive termination of wire pairs
- The determination of an achievable network speed
- The enumeration enrolment and ordering of units currently on the network

Since our previous work deals with theoretical electronics and data transfer-related problems, the network build, enrolment issues and validation measurements are presented in the following sections.

Table 1  
Function allocation of the UIC 558 cables

Wire Pair		Function
1	2	Audio towards train loudspeaker
3	4	Voice communication towards:
3 (-)	4 (+)	- the driver
3 (+)	4 (-)	- the switchboard
5 (+)	6 (-)	Switch on loudspeaker amplifiers
7 (+)	8 (-)	Priority of announcement command
7 (X*)	7 (Y*)	TCN* Communication
9 (+)	12 (-)	Remote command of door closing
10 (+)	12 (-)	Remote command: light on
11 (+)	12 (-)	Remote command: light off
14 (+)	12 (-)	Unlock right doors command
15 (+)	12 (-)	Unlock left doors command
16 (+)	12 (-)	Remote control of door locked condition
17 (X)	18 (Y)	TCN Communication
S		Shield of wires 17-18
13		Common shield for all wires

### 3 Network Look-Up Algorithm

As mentioned before, knowing the order of the units in the train, is essential since the correct positioning of the termination, in the CAN network, highly increases the optimum bandwidth. On the other hand, the determination of the order of the



devices in the network also serves an operational purpose, by enabling an adequate automatic enumeration, of the train's units. For this task, a suitable network lookup algorithm is needed.

As a consequence of the given conditions, the train has a bus network topology for the CAN communications, as shown in Figure 3. It can be seen that all on-board units, of the train's unit, connect to a common two-wire bus provided by the interconnection cables. The problem is that the topology is not constant, since the composition of the train can be freely manipulated by the operator.

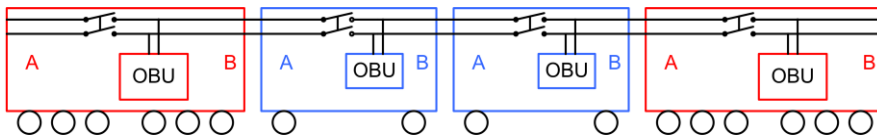


Figure 3

Network topology scheme

To enable the determination of the order of the units, the interconnection cable has a circuit breaker for all OBUs. Since the circuit breaker also prevents the original function of the wire pair from operating, it must be designed to a high safety standard and have the ability to close, whenever the original functionality is needed. These switches are normally in closed state and are only open during the execution of the network lookup algorithm.

The basic concept behind the algorithm is the following:

At the start, all CAN devices utilize a weak termination (example,  $1500\Omega$ ) enabling approximately 1 kbit/s bandwidth on the network, even with a relatively long bus. With this bus speed, the number, position and orientation of each unit can be determined. With this knowledge the network can be adequately terminated at the ends of the train.

Since the media access of the CAN network is based on arbitration, there is no master role on the network. Thus to control the network lookup algorithm a dedicated OBU must be chosen for the task. This way, communications on the network, has the following stages (see Figure 4):

- Normal communication
- Network Lookup
- Disconnected (Media used by original function)

Since the wire pair chosen for communication is rarely used, it is considered to be normal state, when all switches are closed and all OBUs are communicating on the network. This state can be interrupted in two ways: one is when the original function needs the media line and the other is when a new OBU appears on the network.

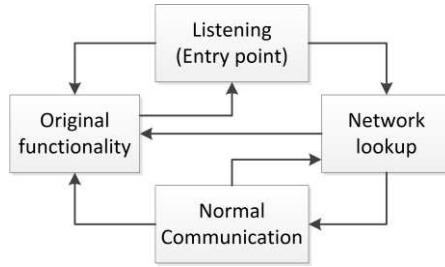


Figure 4

Network state diagram

When the original function needs the media, the OBUs leave their switches closed, but disconnect their transceivers from the network.

When a new unit ID appears on the network the network lookup algorithm starts.

First, all OBUs send their unique identifiers on the network. The one with the lowest value is chosen as the master for communication.

During the network lookup phase, all units, including the master, continuously send their IDs on the network, and their order is determined by their visibility, depending on the state of the circuit breakers.

The master systematically sends a message to each OBU (including to itself) to temporarily open its circuit breaker. This message is available for any unit on the network, so all OBUs know which unit opens the media separating the network into two halves. For this short time, all OBUs receive only the IDs of those units which are on the same side of the network. This way, all OBUs will have a set of visible IDs for each opened breaker.

Figure 5 shows a simple example for this systematic enumeration. It shows the algorithm, from the point of view, of the left-side unit with two additional units having different orientations. It can be seen that with any combination of the orientations, of the two units, their visibility is different according to the state of the switches and so both their order and orientation can be determined.

After acquiring a complete set of visible OBUs, for all opened switches, every OBU can determine the order and orientation with two easy steps:

- First it determines the orientation of each OBU, by examining the set generated, when the given unit's circuit breaker was opened. If the unit's ID is contained in this set, it means that its transceiver is closer than its circuit breaker.
- Second the unit excludes all visible unit IDs from their sets, and orders the set with a sorting algorithm, that gives the order of the units. Naturally, with OBUs in the middle of the train, the units must be handled differently on their two respective sides.

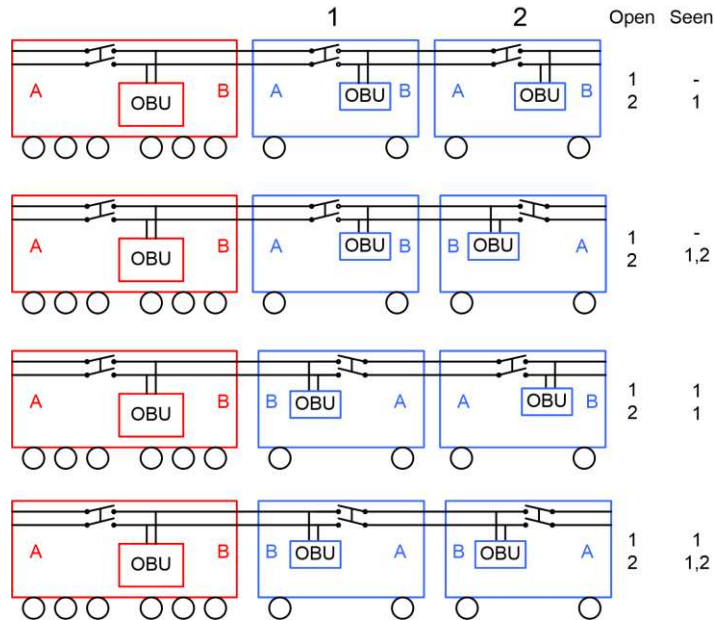


Figure 5  
Network lookup example

Table 2 shows an example with a unit train, from the view of the left side end of the network. For the simplicity of the example the OBUs' order and their IDs' are identical and the "Spin" column shows the orientation of the unit. Spin value 0 means that the circuit breaker is on the left of the CAN transceiver and a value of 1 means it is on the right. It is shown that only the IDs of the units with the spin of 1 (3, 4, 8 and 9) appear in their corresponding sets. By excluding these identifiers the cardinality of set, determines the distance from the left end.

Table 2  
Example network lookup results of a nine unit train

Spin	ID	Visible
0	1	
0	2	1
1	3	1,2,3*
1	4	1,2,3,4*
0	5	1,2,3,4
0	6	1,2,3,4,5
0	7	1,2,3,4,5,6
1	8	1,2,3,4,5,6,7,8*
1	9	1,2,3,4,5,6,7,8,9*
	None	1,2,3,4,5,6,7,8,9

Let us define the set of all visible identifiers as  $A$  and the set of the IDs from the point of view of the unit with the ID  $i$  when the circuit breaker of the  $j^{\text{th}}$  unit is opened as  $A_i^j$ . If we state that the circuit breaker of the  $i^{\text{th}}$  unit is on the left of its transceiver then the units on its right side are:

$$U_i^{\text{right}} = A_i^i \quad (1)$$

and the units on its left:

$$U_i^{\text{left}} = A, \quad A_i^i \quad (2)$$

The determination of the orientation of any unit depends on its position according to unit  $i$ , see Table 3:

Table 3  
Determination of the orientation of each unit

	$j \in U_i^{\text{left}}$	$j \in U_i^{\text{right}}$
$j \in A_i^j$	Right (1)	Left (0)
$j \notin A_i^j$	Left (0)	Right (1)

The order can be determined from the distance of each unit, which can be formally written as:

$$D_i^j = \begin{cases} |A_i^j, U_i^{\text{left}}, \{j\}| & \text{if } j \in U_i^{\text{right}} \\ |A_i^j, U_i^{\text{right}}, \{j\}| & \text{if } j \in U_i^{\text{left}} \end{cases} \quad (3)$$

Another issue is the switching between different data transfer rates, on the CAN network. Generally, when one transceiver with a different bandwidth setting connects to a CAN network it automatically detects the messages as erroneous, disrupts the data transfer and begins to send error frames to the network. This undesired phenomenon can be avoided by using the "listening" or the "listen-only" modes of the CAN devices [4]. After start-up, the OBUs can initiate their CAN controller to the listen-only mode and listen alternately in the two possible frequencies and detect the speed of the network. If no communication is taking place, they start with low bandwidth.

Special cases exist when network connection and disconnection events occur:

- 1 If any disturbance occurs on the network resulting in erroneous network operation, the devices can detect it and fall back to the listening/low bandwidth mode.

- 2 When one or more units disconnect, their periodic identifiers stop appearing on the network. In this case, the device with the currently highest ID priority sends a signal to perform the network lookup.
- 3 The same algorithm occurs when new identifiers appear on the network.
- 4 When two working networks are connected the termination resistance halves because of their parallel configuration. This is the operational boundary of the CAN network. Depending on the current conditions this can result in either an erroneous or a correctly working network. If the communication on the network remains, this case leads to case 2. When this connection disrupts the communication, it leads to case 1.
- 5 When the train units are disconnected the termination could be ineligible leading to case 1.
- 6 When a device restarts, it stops sending its identifier for a short period of time. If the identifier reappears in a previously defined time there is no need for network lookup.

## 4 Experimental Results

Two tests have been carried out to validate the feasibility of the theory. Since the proposed solution utilizes a non-standard termination of the CAN network, the achievable data transfer rates were tested under laboratory conditions. The physical media of the network is an unshielded twisted pair (UTP Cat5) cable. On one hand, the wire pairs of the interconnection cables are also twisted, although their electrical parameters are different. So, the purpose of the laboratory tests, were the determination of the effect of the non-standard termination.

Along with the standard termination, where  $120\Omega$  terminator resistors were applied at either end of the network, three other terminator resistor configurations were tested:

- $120\Omega$  resistor at one end (120/-)
- No termination (-/-)
- $1500\Omega$  resistors at both ends (1500/1500)

Figure 6 shows the results of the test. This measurement has clearly shown that the CAN network operates with a non-standard termination, although on a lower bandwidth.

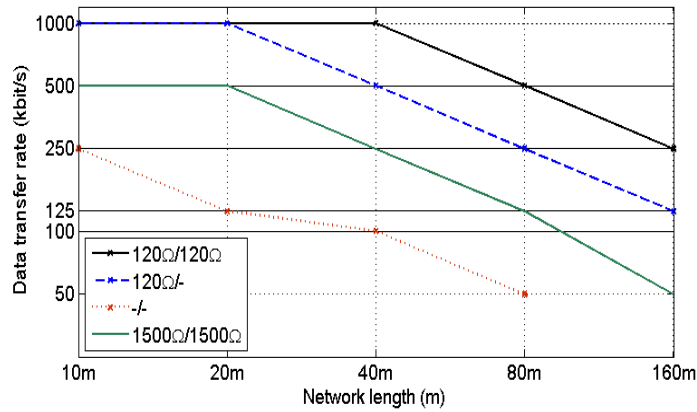


Figure 6

Achievable CAN bandwidth with different termination values

To examine the behavior of the concept under real conditions, a field test was organized with the support of the Hungarian State Railways. The train consisted of 10 - Bhv type passenger cabs and 1 - Bdt type control car, with all units having the same length of 23.740m. The control car was placed in the middle of the train and the units were connected by an 18-pole UIC558 type interconnection cables. First, as a proof of concept, basic measurements were conducted on a stopped train. After that, the validation of operability, the network reliability tests were performed under a real-world environment test, while the train was moving.

Ten units were involved in the test, all having switchable 1500Ω resistors and two 120Ω switchable resistors installed in the 1st and 10th unit. The wire pair used for the communication was:

- 5. Switch on loudspeaker amplifiers (+) for CANH
- 7. Priority of announcement command (+) for CANL

Two CAN analyzers were installed in the 1<sup>st</sup> and 10<sup>th</sup> car to generate and measure network load. The distance between the two can devices was approximately 230 meters. CAN frames with extended IDs and 8 byte length data field were sent on the network, with different data values, for example (0x55) for alternating bits and (0x00) for long unchanged state, where bit-stuffing occurs. The average network load was at least 60% for each measurement, with the maximum value of 80%.

Table 4 summarizes the measurement results. The same measurements were made between the 1<sup>st</sup> and 2<sup>nd</sup> and the 1<sup>st</sup> and 5<sup>th</sup> units. It can be stated, as a general conclusion, that the CAN technology is sufficiently robust for communication purposes, via train interconnection, even with high length and non CAN-standard cabling and termination.

Table 4  
Actual achievable CAN data transfer rates using UIC 558 interconnection

CAN Termination	Bus speed [kbit/s]	Average Bus Load	Ratio of erroneous frames	Network Operable
1500Ω in every unit	100	69	0	Yes
1500Ω in every unit	125	-	-	No
120Ω at two ends	125	68	0	Yes
120Ω at two ends	250	-	-	No
120Ω/1500Ω at two ends	125	68	0.019	Yes
120Ω/1500Ω at two ends	250	-	-	No
1500Ω at two ends	20	60	0	Yes
1500Ω at two ends	50	-	-	No

Based on the measurements it can be assumed that CAN communications with a 20-50kbit/s bandwidth, can be safely applied on the existing train interconnection solutions of railways.

### Conclusions

This paper has proposed the utilization of a UIC558 type interconnection for extended communications purposes. The problems arising from this method are diverse, due to the safety constraints of the existing communication media. The adequate design of the non-standard physical layer, for the CAN communications, the switching between network states, the handling of the network control and the enrolment algorithm all need an optimal approach. The benefits that can be obtained from these solutions are significant. In addition to the current railway-specific applications, the measurements of the non-standard CAN network can be applied in any other field, with ordered electrical constraints, and the results can be utilized in other systems.

The laboratory and field measurements have proven the feasibility of our solution, from the view, of an achievable stable data transfer. It can be stated, that this solution is feasible with long train units, without using any repeaters or signal amplifiers. The network lookup algorithm presented in this paper is sufficiently fast, for the needs of the application, since the slower component, when the circuit breakers are operated, has a linear  $O(n)$ , while the sorting algorithm proceeding, has a quadratic  $O(n^2)$  (worst-case) complexity.

### Acknowledgement

The research has been conducted as part of the projects TÁMOP-4.2.2.A-11/1/KONV-2012-0012: 'Basic research for the development of hybrid and electric vehicles' and TÁMOP-4.2.2.C-11/1/KONV-2012-0012: 'Smarter Transport - IT for co-operative transport system'. The projects are supported by the Hungarian Government and co-financed by the European Social Fund.

## References

- [1] MAV-MI UIC 558:1999, Hungarian State Railways Technical Guideline, 1999
- [2] Aradi, Sz., Bécsi, T., and Gáspár, P. Development of Vehicle On-Board Communication System for Harsh Environment. *Acta Technica Jaurinensis*, 6(3):53-63
- [3] Edwards, M. C., Donelson, J., Zavis, W. M., Prabhakaran, A., Brabb, D. C., Jackson, A. S., Improving Freight Rail Safety with On-Board Monitoring and Control Systems, Rail Conference, 2005. Proceedings of the 2005 ASME/IEEE Joint , Vol., No., pp. 117-122, 16-18 March 2005
- [4] Koppe, U. Automatic Baudrate Detection in Canopen Networks. Proceedings of the 9<sup>th</sup> International CAN Conference, 2003
- [5] Rodriguez-Morcillo, C., Alexandres, S. and Munoz, J. D. Broadband System to Increase Bitrate in Train Communication Networks. *Computer Standards & Interfaces*, 31(2):261-271, 2009
- [6] Russo, D., Gatti, A., Ghelardini, A., Mancini, G., Verduci, A., Amato, D., Battani, R.: Power Line Communication: a New Approach for Train Passenger Information Systems. 8<sup>th</sup> World Congress on Railway Research Proceedings, 1(1), 1993
- [7] Schäfer, C., Hans, G. IEC 61375-1 and UIC 556-International Standards for Train Communication. In *Vehicular Technology Conference Proceedings, VTC 2000-Spring Tokyo*, 2000 IEEE 51<sup>st</sup>, Volume 2, pp. 1581-1585, 2000
- [8] Zeltwanger H. Canopen in the Application Field of Rail Vehicles. *Railway Pro*, 2010
- [9] Zeltwanger, H. Canopen on Track. *CAN in Automation e. V.*, 2012



# Fuzzy Brain Emotional Cerebellar Model Articulation Control System Design for Multi-Input Multi-Output Nonlinear

Chang-Chih Chung<sup>1</sup>, Chih-Min Lin<sup>2</sup>

<sup>1</sup>Department of Electrical Engineering, Yuan Ze University, Chung-Li, Tao-Yuan 320, Taiwan, Republic of China

<sup>2</sup>School of Information Science and Engineering, Xiamen University, China, and Department of Electrical Engineering and Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Chung-Li, Tao-Yuan 320, Taiwan, Republic of China, E-mail: cml@saturn.yzu.edu.tw

---

*Abstract: A brain emotional cerebellar model articulation controller (BECMAC) is developed, which is a mathematical model that approximates the judgmental and emotional activity of a brain. A fuzzy inference system is incorporated into the BECMAC, to give the novel fuzzy brain emotional cerebellar model articulation controller (FBECMAC) that is also proposed in this paper. The developed FBECMAC has the benefit of fuzzy inference and judgment and emotional activity, and it is used to control multi-input multi-output nonlinear systems. A 3-dimensional (3D) chaotic system and a mass spring damper mechanical system are simulated, to illustrate the effectiveness of the proposed control method. A comparison between the proposed FBECMAC and other controller shows that the proposed controller exercises better control than the other controllers.*

*Keywords: brain emotional cerebellar model articulation controller; fuzzy system; chaotic system; mass spring damper mechanical system*

---

## 1 Introduction

In 1992, LeDoux found that, in the human brain, the association between a stimulus and its emotional consequence occurs in the amygdala [1]. The brain has an orbitofrontal cortex and an amygdala; the former is a sensory neural network and the latter is an emotional neural network [2]. Many works studied the form of the amygdala to determine the usefulness for a neural network control system. In recent years, brain emotional learning controllers (BELC) have been used for control systems by several studies [3-7]. A brain emotional learning controller has two systems: an emotional system and a neural network judgment system.

A cerebellar model articulation controller (CMAC) is a network model that uses an associated memory network [8]. It has better computation and adaptation ability than a neural network. CMAC approximation can be tuned for greater accuracy, even for complex nonlinear system. Therefore, CMAC has been the subject of more studies, because it is more generally applicable to various nonlinear systems and can learn rapidly. Chiang and Lin introduced a Gaussian-based CMAC with faster convergence [9, 10]. This gives more option for the researcher to experiment with a Gaussian-based CMAC to control various nonlinear systems. The enhanced performance of a CMAC over a neural network has been demonstrated in some studies [11, 12]. By constructing a BELC using a CMAC, a new brain emotion network, called a brain emotion CMAC (BECMAC), is proposed. This improves the learning ability of a conventional BELC.

A fuzzy inference system mimics the human reasoning process and is widely used successfully in various fields. Initially, the control system algorithms required a detailed system model. However, in recent years, fuzzy control systems have used fuzzy inference rules to control systems, without the need for detailed mathematical models [13, 14].

This paper incorporates a fuzzy inference system with a BECMAC, to produce a novel fuzzy brain emotional cerebellar model articulation controller (FBECMAC). This controller has the benefits of a fuzzy system, because there are fuzzy inference rules, and of BECMAC, because the controller learns more completely. This FBECMAC is then used to control multi-input multi-output nonlinear systems, to illustrate its effectiveness.

## 2 Problem Formulation

A class of  $n$ -th order multi-input multi-output nonlinear systems is described by the following equation:

$$\mathbf{x}^{(n)}(t) = \mathbf{f}(\mathbf{x}(t)) + \mathbf{G}(\mathbf{x}(t))\mathbf{u}(t) + \mathbf{d}(t) \quad (1)$$

where  $\mathbf{u}(t) = [u_1(t), u_2(t), \dots, u_m(t)]^T \in \mathfrak{R}^m$  and  $\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_m(t)]^T \in \mathfrak{R}^m$ . The former is a control input and the latter represents the state vectors of the system.  $\mathbf{d}(t) = [d_1(t), d_2(t), \dots, d_m(t)]^T \in \mathfrak{R}^m$  denotes the unknown bounded external disturbance and  $m$  is the number of system inputs and outputs.  $\mathbf{x}(t) = [\mathbf{x}^T(t), \dot{\mathbf{x}}^T(t), \dots, \mathbf{x}^{(n-1)T}(t)]^T \in \mathfrak{R}^{mn}$  and it is assumed that it is measureable. It is also true that  $\mathbf{f}(\mathbf{x}(t)) \in \mathfrak{R}^m$  and  $\mathbf{G}(\mathbf{x}(t)) \in \mathfrak{R}^{m \times m}$  and that these are smooth nonlinear uncertain functions, which are assumed to be bounded, but not exactly known.

If modeling uncertainties and external disturbance are neglected, the nominal system for (1) is:

$$\mathbf{x}^{(n)}(t) = \mathbf{f}_0(\mathbf{x}(t)) + \mathbf{G}_0 \mathbf{u}(t) \quad (2)$$

where  $\mathbf{f}_0(\mathbf{x}(t)) \in \mathfrak{R}^m$  and  $\mathbf{G}_0 = \text{diag}(g_{01}, g_{02}, \dots, g_{0m}) \in \mathfrak{R}^{m \times m}$  are the nominal parts of  $\mathbf{f}(\mathbf{x}(t))$  and  $\mathbf{G}(\mathbf{x}(t))$ , respectively. Without loss of generality, it is assumed that the constants  $g_{0i} \geq 0$  for  $i = 1, \dots, m$ . It is also assumed that the nonlinear system (2) is controllable and that  $\mathbf{G}_0^{-1}$  exists. If there are modeling uncertainties and external disturbances, the nonlinear system (1) can be reformulated as

$$\mathbf{x}^{(n)}(t) = \mathbf{f}_0(\mathbf{x}(t)) + \mathbf{G}_0 \mathbf{u}(t) + \mathbf{I}(\mathbf{x}(t), t) \quad (3)$$

where  $\mathbf{I}(\mathbf{x}(t), t)$  is referred to as the lumped uncertainty, which includes the system uncertainties and the external disturbances.

The control problem is the design of a proper control system, wherein the system output,  $\mathbf{x}(t)$ , can track a desired trajectory vector,  $\mathbf{x}_r(t) = [x_{r1}(t), x_{r2}(t), \dots, x_{rm}(t)]^T \in \mathfrak{R}^m$ .

The tracking error is defined as

$$\mathbf{e}(t) \triangleq \mathbf{x}_d(t) - \mathbf{x}(t) \in \mathfrak{R}^m \quad (4)$$

and the system tracking error vector is defined as

$$\underline{\mathbf{e}}(t) \triangleq [\mathbf{e}^T(t), \dot{\mathbf{e}}^T(t), \dots, \mathbf{e}^{(n-1)T}(t)]^T \in \mathfrak{R}^{mn} \quad (5)$$

If the nominal functions,  $\mathbf{f}_0(\mathbf{x}(t))$ ,  $\mathbf{G}_0$  and the lumped uncertainty,  $\mathbf{I}(\mathbf{x}(t), t)$ , are exactly known, an ideal controller can be designed as:

$$\mathbf{u}^* = \mathbf{G}_0^{-1} [\mathbf{x}_d^{(n)} - \mathbf{f}_0(\mathbf{x}) - \mathbf{I}(\mathbf{x}, t) + \underline{\mathbf{K}}^T \underline{\mathbf{e}}] \quad (6)$$

where  $\underline{\mathbf{K}} = [\mathbf{K}_n, \dots, \mathbf{K}_2, \mathbf{K}_1]^T \in \mathfrak{R}^{mn \times m}$  is the feedback gain matrix, which contains real numbers, and  $\mathbf{K}_i = \text{diag}(k_{i1}, k_{i2}, \dots, k_{im}) \in \mathfrak{R}^{m \times m}$  is a nonzero positive constant diagonal matrix.

Substituting the ideal controller (6) into (3) gives the error dynamic equation:

$$\mathbf{e}^{(n)} + \underline{\mathbf{K}}^T \underline{\mathbf{e}} = \mathbf{0} \quad (7)$$

In (7), if  $\underline{\mathbf{K}}$  is chosen to correspond to the coefficients of a Hurwitz polynomial, then  $\lim_{t \rightarrow \infty} \|\underline{\mathbf{e}}\| = 0$ . However, the lumped uncertainty,  $\mathbf{I}(\mathbf{x}(t), t)$ , is generally

unknown for practical applications, so the ideal controller,  $u^*$ , in (6) is not possible. Therefore, a fuzzy BECMAC that mimics this ideal controller is proposed in the next section.

### 3 A Fuzzy Brain Emotion Cerebellar Model Articulation Control System

#### 3.1 The Fuzzy Brain Emotional Cerebellar Model Articulation Controller

A fuzzy BECMAC (FBECMAC) control system can be classified as a supervised network. A FBECMAC is not very complex in operation and has fast convergence, so it is applicable to many nonlinear control systems. The proposed FBECMAC is shown in Fig. 1, and it has two systems; the first system is a propinquity amygdala system, similar to that in a mammalian brain; and the second system is a propinquity cerebellar system, which is also similar to that in a mammalian brain. In this novel inference system, two fuzzy rule bases are proposed for the BECMAC.

The fuzzy amygdala system is designed as:

If  $I_1$  is  $\lambda_1$  and ...  $I_i$  is  $\lambda_i$ , ..., and  $I_m$  is  $\lambda_m$  then

$$a_q = z_{iq} \text{ for } i = 1, 2, \dots, m, q = 1, 2, \dots, p \quad (8)$$

where  $z_{iq}$  is the amygdala's weight and  $a_q$  is the amygdala's output.

The fuzzy cerebellar model articulation system is designed as:

If  $I_1$  is  $\phi_j$  and ...  $I_i$  is  $\phi_j$ , ..., and  $I_m$  is  $\phi_{mj}$  then

$$o_q = w_{jq} \text{ for } i = 1, 2, \dots, m, j = 1, 2, \dots, n, q = 1, 2, \dots, p \quad (9)$$

where  $w_{jq}$  is the prefrontal weight and  $o_q$  is the prefrontal output.

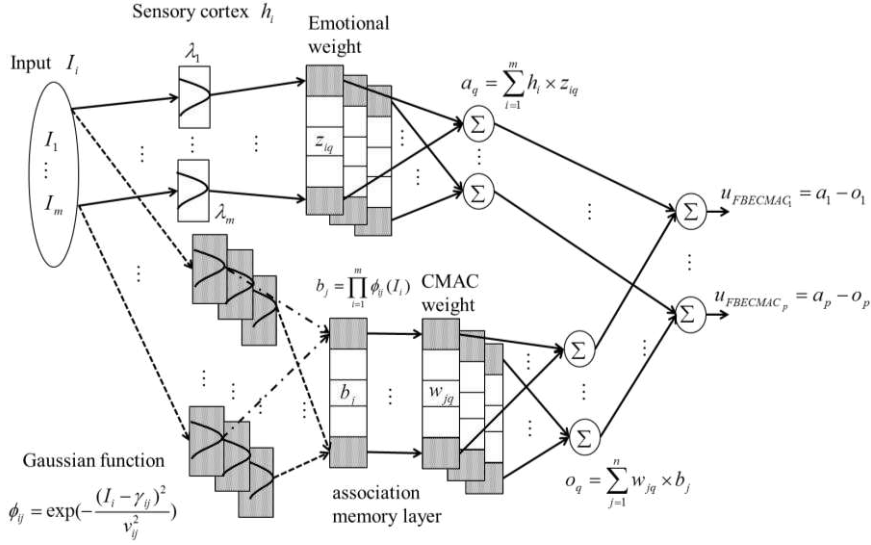


Figure 1

The emotional cerebellar model articulation controller

Fuzzy amygdala systems have two layers. The first layer is a Gaussian function and the second layer is a weight layer:

$$h_i = I_i \times \lambda_i, \quad i = 1, 2, \dots, m \quad (10)$$

where  $h_i$  is the amygdala system's input to the sensory cortex output,  $I_i$  is the controller's input and  $\lambda_i$  is the Gaussian function, which is denoted as:

$$\lambda_i = \exp\left(-\frac{(I_i - \delta_i)^2}{\sigma_i^2}\right) \quad (11)$$

where  $\delta_i$  is a mean and  $\sigma_i$  is a variance.

$$a_q = \sum_{i=1}^m h_i \times z_{iq} \quad (12)$$

where  $z_{iq}$  is amygdala system weight.

The fuzzy cerebellar model articulation system has three layers. The first layer is a Gaussian function:

$$\phi_{ij} = \exp\left[-\frac{(I_i - \gamma_{ij})^2}{v_{ij}^2}\right], \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, n \quad (13)$$

where  $\gamma_{ij}$  is the mean and  $v_{ij}$  is the variance. The second layer is an association memory layer:

$$b_j = \prod_{i=1}^m \phi_{ij}(I_i) \quad (14)$$

The third layer is a weight layer:

$$o_q = \sum_{j=1}^n w_{jq} \times b_j \quad (15)$$

where  $w_{jq}$  is a weight.

$$u_{FBECMAC_q} = a_q - o_q, \quad q = 1, 2, \dots, p \quad (16)$$

The amygdala's system's updated weights,  $\Delta z_{iq}$ , are given by

$$\Delta z_{iq} = \eta_z [h_i \times (\max[0, d_q - a_q])] \quad (17)$$

where  $\eta_z$  is the learning rate. In (17),  $d_q$  is a parameter adjustment, given by:

$$d_q = (\sum_{i=1}^m \beta_{iq} \times I_i) + (c_q \times u_{FBECMAC_q}) \quad (18)$$

where  $\beta_{iq}$  and  $c_q$  are the gains.

The updating law for the amygdala's system's weight is given by

$$z_{iq}(t+1) = z_{iq}(t) + \Delta z_{iq} \quad (19)$$

The fuzzy CMAC hypercube weight,  $w_{jq}$ , and the mean,  $m_{ij}$ , and variance,  $v_{ij}$ , of the Gaussian function are updated using the following equation:

$$w_{jq}(t+1) = w_{jq}(t) + \Delta w_{jq} \quad (20)$$

$$\gamma_{ij}(t+1) = \gamma_{ij}(t) + \Delta \gamma_{ij} \quad (21)$$

$$v_{ij}(t+1) = v_{ij}(t) + \Delta v_{ij} \quad (22)$$

An integrated error function is defined as

$$s(\underline{e}, t) \equiv \underline{e}^{(n-1)} + \underline{K}_1 \underline{e}^{(n-2)} + \dots + \underline{K}_n \int_0^t \underline{e}(\tau) d\tau \quad (23)$$

where  $s(\underline{e}, t) = [s_1(t), s_2(t), \dots, s_m(t)]^T \in \Re^m$ .

Substituting (3) into (23) yields

$$\dot{s}(\underline{e}, t) = \underline{x}_d^{(n)} - \underline{f}_0(\underline{x}) - \underline{G}_0 \underline{u}(t) - \underline{I}(\underline{x}(t), t) + \underline{K}^T \underline{e} = \underline{e}^{(n)} + \underline{K}^T \underline{e} \quad (24)$$

If  $(1/2)\underline{s}^T(\underline{e}, t)\underline{s}(\underline{e}, t)$  is chosen as a cost function, then its derivative is  $\underline{s}^T(\underline{e}, t)\dot{s}(\underline{e}, t)$ .

### 3.2 The Robust Feedback Control System

Since the FBECMAC cannot completely mimic an ideal controller, the approximation error induces a tracking error in the control system, so a robust controller is required, in order to make the control system stable. Thus, the control system is composed of a FBECMAC controller and a robust controller.

The proposed a FBECAMC control system for a nonlinear system is shown in Fig. 2.

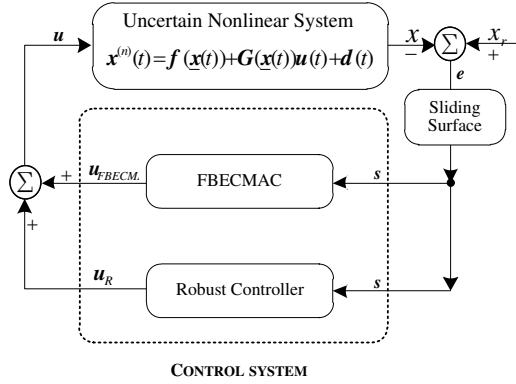


Figure 2

An intelligent control system for nonlinear systems

The control system is defined as:

$$\mathbf{u} = \mathbf{u}_{FBECMAC} + \mathbf{u}_R \quad (25)$$

Substituting (25) into (24) and multiplying both sides by  $\mathbf{s}^T(\underline{\mathbf{e}}, t)$  yields

$$\begin{aligned} \mathbf{s}^T(\underline{\mathbf{e}}, t) \dot{\mathbf{s}}(\underline{\mathbf{e}}, t) = & -\mathbf{s}^T(\underline{\mathbf{e}}, t) \mathbf{f}_0(\underline{\mathbf{x}}(t)) - \mathbf{s}^T(\underline{\mathbf{e}}, t) \mathbf{G}_0 [\mathbf{u}_{FBECMAC} + \mathbf{u}_R(t)] \\ & + \mathbf{s}^T(\underline{\mathbf{e}}, t) (\mathbf{x}_d^{(n)} - \mathbf{I}(\underline{\mathbf{x}}(t), t) + \mathbf{K}^T \underline{\mathbf{e}}) \end{aligned} \quad (26)$$

The fuzzy CMAC training algorithms in (20), (21) and (22) perform error back propagation, using the following chain-rule algorithm:

$$\Delta w_{jq} = -\eta_w \frac{\partial \mathbf{s}^T \dot{\mathbf{s}}}{\partial w_{jq}} = -\eta_w \frac{\partial \mathbf{s}^T \dot{\mathbf{s}}}{\partial \mathbf{u}_{FBECMAC}} \frac{\partial \mathbf{u}_{FBECMAC}}{\partial w_{jq}} = \eta_w \cdot s_i g_{0i}(x) \cdot \left[ \prod_{i=1}^n (\phi_{ij}(I_i)) \right] \quad (27)$$

$$\begin{aligned} \Delta \gamma_{ij} &= -\eta_m \frac{\partial \mathbf{s}^T \dot{\mathbf{s}}}{\partial m_{ij}} = -\eta_m \frac{\partial \mathbf{s}^T \dot{\mathbf{s}}}{\partial \mathbf{u}_{FBECMAC}} \frac{\partial \mathbf{u}_{FBECMAC}}{\partial \phi_{ij}} \frac{\partial \phi_{ij}}{\partial \gamma_{ij}} \\ &= \eta_m \cdot s_i g_{0i}(x) \cdot \sum_{q=1}^p w_{jq} \cdot \left[ \prod_{i=1}^n (\phi_{ij}(I_i)) \right] \left[ \frac{2(I_i - \gamma_{ij})}{v_{ij}^2} \right] \end{aligned} \quad (28)$$

$$\begin{aligned}
\Delta v_{ij} &= -\eta_v \frac{\partial s^T \dot{s}}{\partial v_{ij}} = -\eta_v \frac{\partial s^T \dot{s}}{\partial u_{FECMAC}} \frac{\partial u_{FECMAC}}{\partial \phi_{ij}} \frac{\partial \phi_{ij}}{\partial v_{ij}} \\
&= \eta_v \cdot s_i g_{0i}(x) \cdot \sum_{q=1}^p w_{jq} \cdot \left[ \prod_{i=1}^n (\phi_{ij}(I_i)) \right] \left( \frac{2(I_i - \gamma_{ij})^2}{v_{ij}^2} \right)
\end{aligned} \tag{29}$$

An approximation error between the FBECMAC and the ideal controller is navoidable, so an ideal controller is formulated as the summation of the FBECMAC and the approximation error:

$$\mathbf{u}^*(t) = \mathbf{u}_{FBECMAC} + \boldsymbol{\varepsilon}(t) \tag{30}$$

where  $\boldsymbol{\varepsilon}(t) = [\varepsilon_1(t), \varepsilon_2(t), \dots, \varepsilon_m(t)]^T \in \Re^m$  denotes the approximation error. It is assumed that  $\|\boldsymbol{\varepsilon}\| \leq E$ , where  $E$  is an unknown bound and  $\|\cdot\|$  is an induced norm.  $\hat{E}$  is defined as an estimate of  $E$ , and  $\tilde{E} = E - \hat{E}$ .

From (6) and (24) and after some straightforward manipulations, it is seen that

$$\mathbf{e}^{(n)} + \underline{\mathbf{K}}^T \underline{\mathbf{e}} = \mathbf{G}_0 [\mathbf{u}^* - \mathbf{u}_{FBECMAC} - \mathbf{u}_R(t)] = \mathbf{G}_0 [\boldsymbol{\varepsilon} - \mathbf{u}_R] = \dot{\mathbf{s}}(\underline{\mathbf{e}}, t) \tag{31}$$

Then the following theorem guarantees the stability of the feedback control system.

**Theorem 1:** For the  $n$ th-order nonlinear systems represented by (3), the FBECMAC control system is designed as in (25), where  $\mathbf{u}_{FBECMAC}$  is given in (16). The on-line parameter adaptation algorithms are given as (19)-(22) and the updating laws are given as (17) and (27)-(29), and the robust controller is designed as follows:

$$\mathbf{u}_R = \hat{E} \tanh\left(\frac{\mathbf{s}(t)}{\varsigma}\right) \tag{32}$$

where  $\tanh(\cdot)$  is a hyperbolic tangent function,  $\hat{E}$  is the estimated value of the approximation error bound and  $\varsigma$  is a positive parameter, such that:

$$\dot{\hat{E}} = \eta_\alpha [s^T(\underline{\mathbf{e}}, t) \mathbf{G}_0 \tanh\left(\frac{\mathbf{s}(t)}{\varsigma}\right) - \xi(\hat{E} - E_0)] \tag{33}$$

The feedback control system is then robustly stable.

**Proof:** The Lyapunov function is defined as:

$$V(\mathbf{s}(\underline{\mathbf{e}}, t)) = \frac{1}{2} \mathbf{s}^T(\underline{\mathbf{e}}, t) \mathbf{s}(\underline{\mathbf{e}}, t) + \frac{1}{2} \frac{\tilde{E}^2}{\eta_\alpha} \tag{34}$$

The derivative of the Lyapunov function and (30) and (31) yield:



$$\begin{aligned}
\dot{V}(s(\underline{e}, t)) &= \mathbf{s}^T(\underline{e}, t) \dot{\mathbf{s}}(\underline{e}, t) - \frac{\tilde{E} \dot{\hat{E}}}{\eta_\alpha} \\
&= \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 [\boldsymbol{\varepsilon}(t) - \mathbf{u}_R] - \frac{\tilde{E} \dot{\hat{E}}}{\eta_\alpha}
\end{aligned} \tag{35}$$

The robust controller is designed as (32) and (33), so (35) can be rewritten as:

$$\begin{aligned}
\dot{V}(s(\underline{e}, t)) &= \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 [\boldsymbol{\varepsilon}(t) - (\hat{E} \tanh(\frac{\mathbf{s}(\underline{e}, t)}{\varsigma}))] - \frac{\tilde{E}}{\eta_\alpha} \eta_\alpha [\mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 \tanh(\frac{\mathbf{s}(\underline{e}, t)}{\varsigma}) - \xi(\hat{E} - E_0)] \\
&= \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 \boldsymbol{\varepsilon}(t) - \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 \hat{E} \tanh(\frac{\mathbf{s}(\underline{e}, t)}{\varsigma}) - \tilde{E} \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 \tanh(\frac{\mathbf{s}(\underline{e}, t)}{\varsigma}) + \tilde{E} \xi(\hat{E} - E_0) \\
&= \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 \boldsymbol{\varepsilon}(t) - \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 \tanh(\frac{\mathbf{s}(\underline{e}, t)}{\varsigma}) [\tilde{E} + \hat{E}] + \tilde{E} \xi(\hat{E} - E_0) \\
&\leq \|\mathbf{s}^T(\underline{e}, t)\| \|\mathbf{G}_0\| \|\boldsymbol{\varepsilon}(t)\| - E \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 \tanh(\frac{\mathbf{s}(\underline{e}, t)}{\varsigma}) + \tilde{E} \xi(\hat{E} - E_0) \\
&\leq E \|\mathbf{s}^T(\underline{e}, t)\| \|\mathbf{G}_0\| \|\boldsymbol{\varepsilon}(t)\| - \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 \tanh(\frac{\mathbf{s}(\underline{e}, t)}{\varsigma}) + \tilde{E} \xi(\hat{E} - E_0)
\end{aligned} \tag{36}$$

It is seen that the following inequality exists for any  $\varsigma > 0$ :

$$0 \leq \|\mathbf{s}^T(\underline{e}, t)\| \|\mathbf{G}_0\| \|\boldsymbol{\varepsilon}(t)\| - \mathbf{s}^T(\underline{e}, t) \mathbf{G}_0 \tanh(\frac{\mathbf{s}(\underline{e}, t)}{\varsigma}) \leq \mathcal{G}_\varsigma \tag{37}$$

where  $\mathcal{G}$  is a constant that satisfies  $\mathcal{G} = \exp(-(\mathcal{G} + 1))$ . Using inequality (37), (36) can be rewritten as:

$$\begin{aligned}
\dot{V}(s(\underline{e}, t)) &\leq \xi \tilde{E} (\hat{E} - E_0) + E \mathcal{G}_\varsigma \\
&\leq -\frac{1}{2} \xi [(E - \hat{E})^2 - (E - E_0)^2 + (\hat{E} - E_0)^2] + E \mathcal{G}_\varsigma \\
&\leq -\frac{1}{2} \xi \tilde{E}^2 + \frac{1}{2} \xi (E - E_0)^2 + E \mathcal{G}_\varsigma
\end{aligned} \tag{38}$$

Using the Lyapunov function (34), (38) can be rewritten as:

$$\dot{V} \leq -\varpi V + \psi \tag{39}$$

where  $\varpi$  and  $\psi$  are positive constants given by

$$\varpi = \xi \eta_\alpha \tag{40}$$

$$\psi = E\mathcal{G}\zeta + \frac{1}{2}\xi(E - E_0)^2 \quad (41)$$

Since  $\frac{\psi}{\varpi} > 0$  and the solution of the differential inequality satisfies

$$0 \leq V(t) \leq \frac{\psi}{\varpi} + [V(0) - \frac{\psi}{\varpi}]e^{-\varpi t} \quad (42)$$

where  $V(0)$  is the initial value of  $V$ , then  $s$  and  $E$  are uniformly ultimately bounded, according to the extensions of the Lyapunov theory [15]. From (42), it is true that:

$$\frac{1}{2}s^2 \leq \frac{\psi}{\varpi} + [V(0) - \frac{\psi}{\varpi}]e^{-\varpi t} \leq \frac{\psi}{\varpi} + V(0)e^{-\varpi t} \quad (43)$$

so

$$s^2 \leq 2[\frac{\psi}{\varpi} + V(0)e^{-\varpi t}] \quad (44)$$

which implies that, given  $\rho > \sqrt{2\psi/\varpi}$ , there exists a finite time,  $T$ , such that for all  $t \geq T$ , the tracking index satisfies:

$$|s(e, t)| < \rho \quad (45)$$

where  $\rho$  is the size of a small residual set that depends on the control system approximation error and the controller parameters and  $\rho$  is a positive constant.  $\rho$  is chosen to be small and the finite time is long, so that there is precise tracking of the error.

## 4 Simulation Results

Two uncertain nonlinear systems, a chaotic system and Mass-spring-damper mechanical system, are studied, in order to illustrate the effectiveness of proposed design.

### 4.1 A Chaotic System

For a general master-slave unified chaotic system, the master system is given as [16]:

$$\dot{x}_1(t) = (25\theta + 10)(x_2(t) - x_1(t))$$

$$\dot{x}_2(t) = (28 - 35\theta)x_1(t) - x_1(t)x_3(t) + (29\theta - 1)x_2(t) \quad (46)$$

$$\dot{x}_3(t) = x_1(t)x_2(t) - \left(\frac{8+\theta}{3}\right)x_3(t)$$

where  $x_i$   $i=1, 2, 3$  are the system state variables of the master system and  $\theta=[0 \sim 1]$ , where  $\theta=[0 \sim 0.8]$ , the system is known to be a generalized Lorenz system. When  $\theta=0.8$ , the system is called a Lu system, and when  $\theta=(0 \sim 0.8]$ , the system is called a Chen system. Figure 3 shows the state trajectories for  $\theta=0$ , with the initial condition:  $x_1(0)=3$ ,  $x_2(0)=5$ ,  $x_3(0)=7$

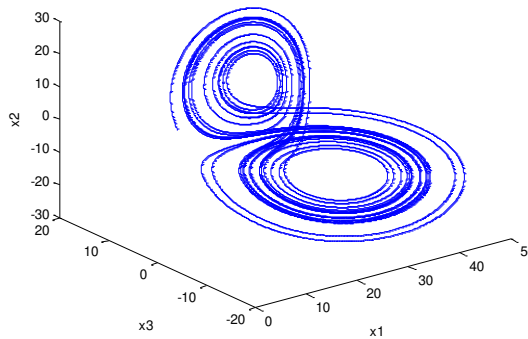


Figure 3  
The unified chaotic system

It is assumed that  $\pi_1 = (25\theta + 10)$ ,  $\pi_2 = (28 - 35\theta)$ ,  $\pi_3 = (29\theta - 1)$  and  $\pi_4 = \left(\frac{8+\theta}{3}\right)$ ,

so (46) can be rewritten as:

$$\begin{aligned} \dot{x}_1(t) &= \pi_1(x_2(t) - x_1(t)) \\ \dot{x}_2(t) &= \pi_2 x_1(t) - x_1(t)x_3(t) + \pi_3 x_2(t) \\ \dot{x}_3(t) &= x_1(t)x_2(t) - \pi_4 x_3(t) \end{aligned} \quad (47)$$

That is, the master system can be expressed as:

$$\dot{\mathbf{x}}(t) = \boldsymbol{\pi}(\mathbf{x}(t)) \quad (48)$$

where  $\boldsymbol{\pi} = [\pi_1, \pi_2, \pi_3]^T$   $\mathbf{x}(t) = [x_1(t), x_2(t), x_3(t)]^T$ .

The slave system is given as:

$$\begin{aligned} \dot{y}_1(t) &= \pi_1(y_2(t) - y_1(t)) + \chi_1(t) + u_1(t) \\ \dot{y}_2(t) &= \pi_2 y_1(t) - y_1(t)y_3(t) + \pi_3 y_2(t) + \chi_2(t) + u_2(t) \end{aligned} \quad (49)$$

$$\dot{y}_3(t) = y_1(t)y_2(t) - \varphi_4 y_3(t) + \chi_3(t) + u_3(t)$$

where  $y_i, i = 1, 2, 3$  are the system state variables of the slave system,  $x_i, i = 1, 2, 3$  are the external disturbances and  $u_i, i = 1, 2, 3$  are the control inputs.

This slave system can be also expressed as:

$$\dot{\mathbf{y}}(t) = \boldsymbol{\pi}(\mathbf{y}(t)) + \boldsymbol{\zeta}(t) + \mathbf{u}(t) \quad (50)$$

where  $\mathbf{y}(t) = [y_1(t), y_2(t), y_3(t)]^T$ ,  $\mathbf{x}(t) = [x_1(t), x_2(t), x_3(t)]^T$  and  $\mathbf{u}(t) = [u_1(t), u_2(t), u_3(t)]^T$ .

If the error states between the master system and the slave system are defined as:

$$\begin{aligned} e_1(t) &= y_1(t) - x_1(t) \\ e_2(t) &= y_2(t) - x_2(t) \\ e_3(t) &= y_3(t) - x_3(t) \end{aligned} \quad (51)$$

From (47) and (49) gives the error dynamics as:

$$\begin{aligned} \dot{e}_1(t) &= \pi_1(e_2(t) - e_1(t)) + x_1(t) + u_1(t) \\ \dot{e}_2(t) &= \pi_2 e_1(t) + \pi_3 e_2(t) + x_1(t)x_3(t) - y_1(t)y_3(t) + x_2(t) + u_2(t) \\ \dot{e}_3(t) &= y_1(t)y_2(t) - x_1(t)x_2(t) - \pi_4 e_3(t) + x_3(t) + u_3(t) \end{aligned} \quad (52)$$

This can be also rewritten as:

$$\dot{\mathbf{e}}(t) = \mathbf{A}\mathbf{e}(t) + \mathbf{f}(t) + \mathbf{x}(t) + \mathbf{u}(t) \quad (53)$$

where  $\mathbf{e}(t) = [e_1(t), e_2(t), e_3(t)]^T$  is the state error vector,  $\mathbf{A} = \begin{bmatrix} -\pi_1 & \pi_1 & 0 \\ \pi_2 & \pi_3 & 0 \\ 0 & 0 & -\pi_4 \end{bmatrix}$ ,

$$\text{and } \mathbf{f}(t) = \begin{bmatrix} 0 \\ x_1(t)x_3(t) - y_1(t)y_3(t) \\ -x_1(t)x_2(t) + y_1(t)y_2(t) \end{bmatrix}.$$

In order to illustrate the effectiveness of the proposed FBECMAC control system, it is compared with the fuzzy neural network based controller in [16], The control

parameters are selected as  $k_1 = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.2 \end{bmatrix}$ ,  $c_q = 0.8, q = 1, 2, 3$ ,  $\eta_z = 5 \times 10^{-4}$ ,

$\eta_w = 4$ ,  $\eta_m = 0.7$ ,  $\eta_v = 5 \times 10^{-3}$ ,  $\varsigma = 1$  and  $\xi = 2$  and the other parameters are

random values. The fuzzy neural network (FNN) control chaotic system is shown in Fig. 4. The tracking error for the FNN control system is shown in Fig. 5. The FBECMAC control chaotic system is shown in Fig. 6 and the tracking error for the FBECMAC control system is shown in Fig. 7. A comparison of the simulation results shows that the proposed FBECMAC control system achieves better control than a FNN control system.

## 4.2 A Mass-Spring-Damper Mechanical System

A mass-spring-damper mechanical system is shown in Fig. 8. The dynamic equations for this mechanical system are expressed as [16, 17]:

$$\begin{aligned}\tau_1 \ddot{x}_1(t) &= -f_{K1}(x) - f_{B1}(x) + f_{K2}(x) + f_{B2}(x) + u_1(t) + \Delta u_{12}(t) + \delta_1(t) \\ \tau_2 \ddot{x}_2(t) &= -f_{K2}(x) - f_{B2}(x) + u_2 + \Delta u_{21} + \delta_2(t)\end{aligned}\quad (54)$$

where  $\tau_1$  and  $\tau_2$  are the masses in the system and  $\mathbf{x}(t) = [x_1(t), x_2(t), \dot{x}_1(t), \dot{x}_2(t)]^T$  are the positions and the velocities of the mechanical system. The spring forces are  $f_{K2}(\underline{x}) = k_{20}(x_2 - x_1) + \Delta k_2(x_2 - x_1)^3$  and  $f_{K1}(\underline{x}) = k_{10}(x_1 - x_2) + \Delta k_1(x_1 - x_2)^3$  and the frictional forces are  $f_{B1}(x) = b_{10}\dot{x}_1 + \Delta b_1\dot{x}_1^2$  and  $f_{B2}(x) = b_{20}(\dot{x}_2 - \dot{x}_1) + \Delta b_2(\dot{x}_2 - \dot{x}_1)^2$ . The parameters for the system are given as  $\tau_1 = 1$ ,  $\tau_2 = 0.8$ ,  $k_{10} = 3$ ,  $k_{20} = 4$ ,  $b_{10} = 2$ ,  $b_{20} = 2.2$ ,  $\Delta k_1 = 0.5$ ,  $\Delta k_2 = 0.5$ ,  $\Delta b_1 = 0.5$ ,  $\Delta b_2 = 0.5$ ,  $\Delta u_{12} = 0.2u_2$ ,  $\Delta u_{21} = 0.25u_1$ ,

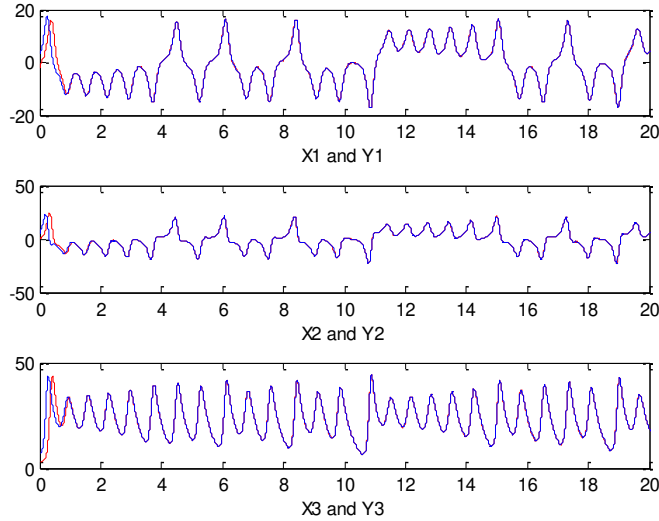


Figure 4

The FNN control for the chaotic system: (A)  $x_1$  and  $y_1$  (B)  $x_2$  and  $y_2$  (C)  $x_3$  and  $y_3$

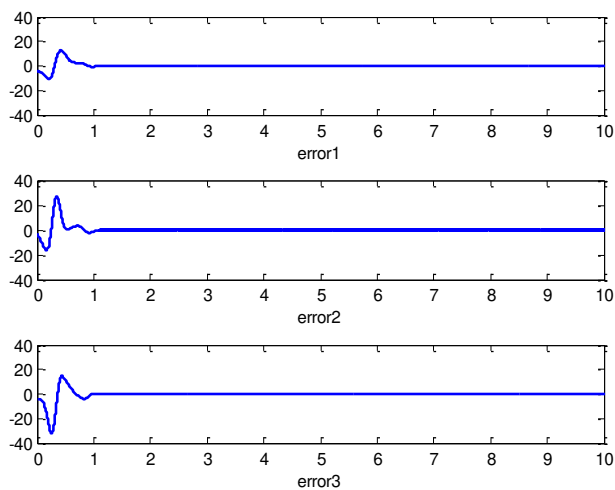


Figure 5  
The tracking error for the FNN control system

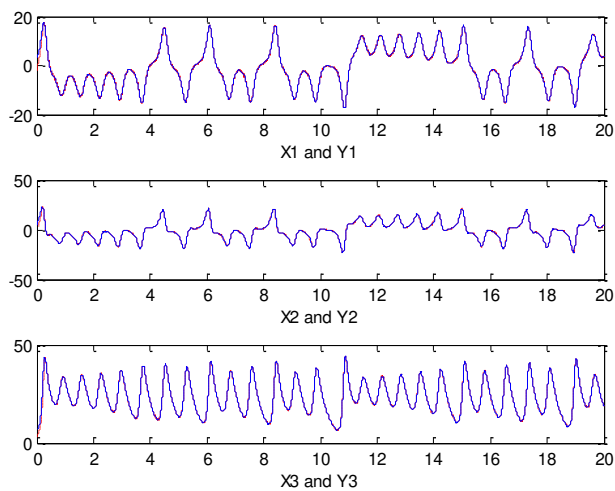


Figure 6  
The FBECMAC control for the chaotic system: (A)  $x_1$  and  $y_1$  (B)  $x_2$  and  $y_2$  (C)  $x_3$  and  $y_3$

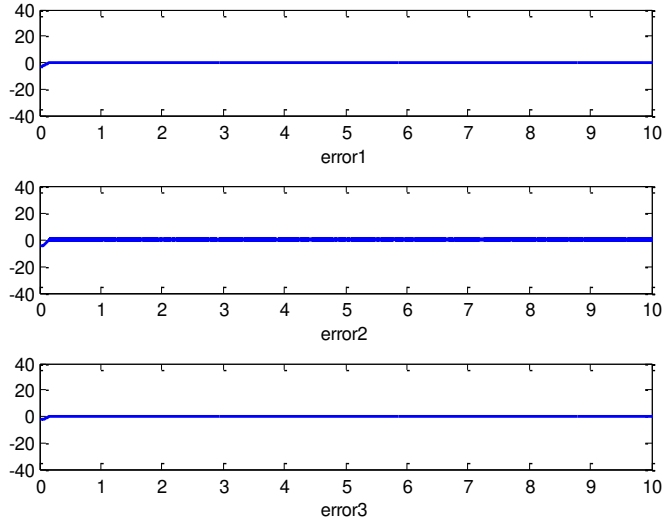


Figure 7

The tracking error for the FBECMAC control system

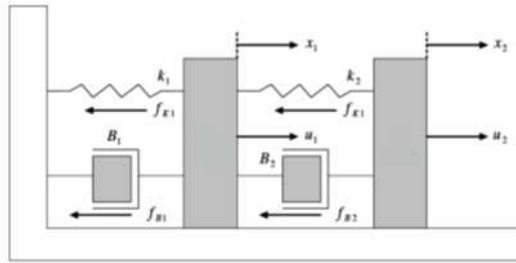


Figure 8

A mass-spring-damper mechanical system

$$d_1(t) = 2\exp(-0.2t) \text{ and } d_2(t) = -2\exp(-0.1t).$$

Consequently, the dynamic equation for the mass-spring-damper mechanical system can be rewritten as:

$$\dot{\underline{x}}(t) = \underline{f}(\underline{x}) + \underline{G}(\underline{x})\underline{u}(t) + \underline{d}(t) \quad (55)$$

where

$$\underline{f}(\underline{x}) = \left[ \frac{-f_{K1}(\underline{x}) - f_{B1}(\underline{x}) + f_{K2}(\underline{x}) + f_{B2}(\underline{x})}{\tau_1}, \frac{-f_{K2}(\underline{x}) - f_{B2}(\underline{x})}{\tau_2} \right]$$

$$\mathbf{G}(\underline{x}) = \begin{bmatrix} \frac{1}{\tau_1} & \frac{0.2}{\tau_1} \\ \frac{0.25}{\tau_2} & \frac{1}{\tau_2} \end{bmatrix} \quad \text{and} \quad \mathbf{u}(t) \triangleq [u_1(t), u_2(t)]^T \quad \text{denotes the control input and}$$

$$\mathbf{d}(t) \triangleq \left[ \frac{d_1(t)}{\tau_1}, \frac{d_2(t)}{\tau_2} \right]^T \quad \text{denotes the external disturbance. The desired trajectories}$$

come from the reference model outputs. The reference model is chosen as  $\ddot{x}_{di}(t) = -16x_{di}(t) - 4\dot{x}_{di}(t) + 12r_i$ ,  $i = 1, 2$ . The initial conditions for the mechanical system and the reference model are given as  $x_1(0) = 1$ ,  $\dot{x}_1(0) = 0$ ,  $x_2(0) = 1$ ,  $\dot{x}_2(0) = 0$ ,  $x_{d1}(0) = 0$ ,  $\dot{x}_{d1}(0) = 0$  and  $\dot{x}_{d2}(0) = 0$ . The control parameters are selected as  $k_1 = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}$ ,  $c_q = 0.8$ ,  $q = 1, 2, 3$ ,  $\eta_z = 0.3$ ,  $\eta_w = 0.01$ ,  $\eta_m = 0.2$ ,  $\eta_v = 0.2$ ,  $\varsigma = 1.1$  and  $\xi = 1.5$  and the other parameters are random values. The reference inputs are  $r_1 = \frac{\pi}{3}(0.9\sin(\frac{t}{2}) + 0.1\sin(2t))$  and  $r_2 = \pi(0.4\sin(t) + 0.1\sin(3t))$ .

The FNN control for the mass-spring-damper mechanical system is shown in Fig. 9 and the tracking error is shown in Fig. 10. The FBECMAC control for the mass-spring-damper mechanical system is shown in Fig. 11 and the tracking error is shown in Fig. 12. These simulations also demonstrate the better control of the FBECMAC control system.

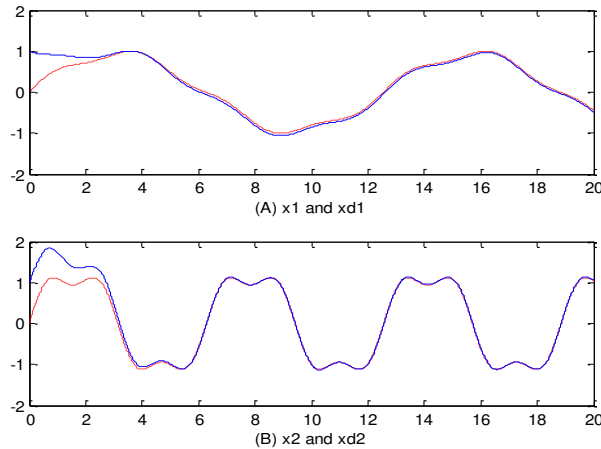


Figure 9

The FNN control for the mass-spring-damper mechanical system



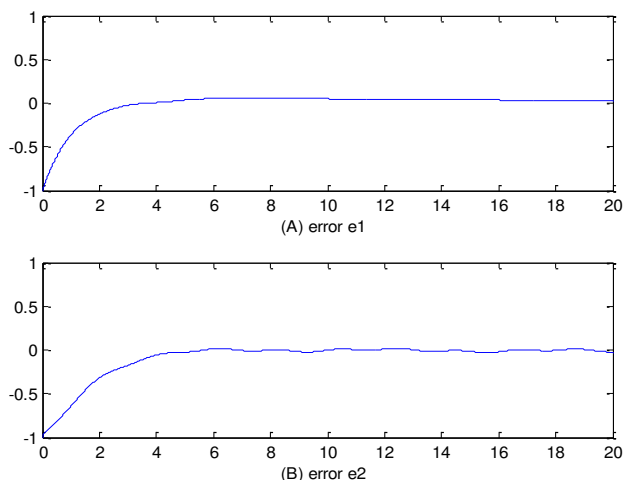


Figure 10

The tracking error for the FNN control system

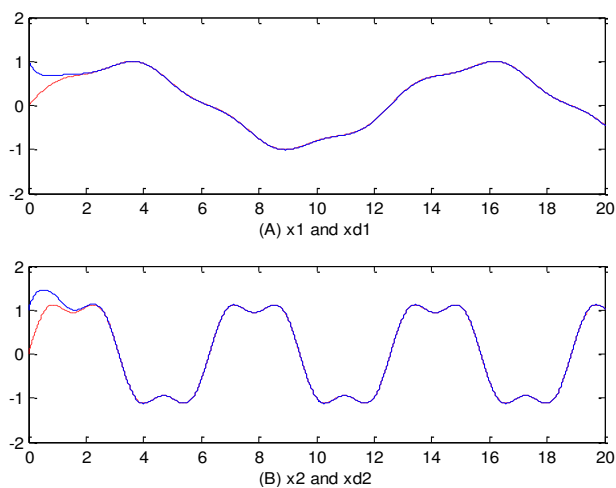


Figure 11

The FBECMAC control for the mass-spring-damper mechanical system

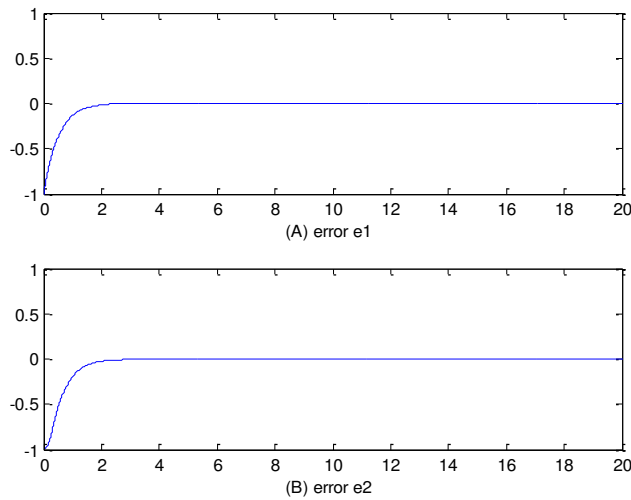


Figure 12

The tracking error for the FBECMAC control system

## Conclusion

This study successfully proposes an efficient FBECMAC control system, which has the benefits of a fuzzy inference system and a brain emotional CMAC. The controller is then used to control nonlinear systems. The stability analysis is also presented in the feedback control system. The proposed FBECMAC reduces the tracking error, even if the systems are subjected to external disturbances. The results of the comparison also show that the tracking error converges faster in the FBECMAC than that in a fuzzy neural network control system.

## Acknowledgment

This work was supported by the National Science Council of the Republic of China under Grant NSC-95-2221-E-155-014-MY3.

## Reference

- [1] J. E. LeDoux, "The Amygdala: Neurobiological Aspects of Emotion," Wiley-Liss, New York, pp. 339-351, 1992
- [2] C. Balkenius and J. Moren, "Emotional Learning: A Computational Model of The Amygdala," *Cybernetics and Systems*, Vol. 32, No. 6, pp. 611-636, 2001
- [3] J. Moren. "Emotion and Learning-A Computational Model of the Amygdala," PhD dissertation, Lund University, 2002
- [4] M. Valikhani and C. Sourkounis, "A Brain Learning-based Intelligent Controller (BELBIC) for DFIG System," *2014 International Symposium on Power Electronics, Electrical Drives, Automation and Motion*, pp. 713-718,

2014

- [5] M. A. Sharbafi, C. Lucas and R. Daneshavar, "Motion Control of Omni-Directional Three-Wheel Robots by Brain-Emotional-Learning-based Intelligent Controller," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 40, No. 6, pp. 630-638, 2010
- [6] M. A. Rahman, R. M. Milasi, C. Lucas, B. N. Araabi and T. S. Radwan, "Implementation of Emotional Controller for Interior Permanent-Magnet Synchronous Motor Drive," *IEEE Transactions on Industry Applications*, Vol. 44, No. 5, pp. 1466-1476, 2008
- [7] H. A. Zarchi, E. Daryabeigi, G. R. A. Markadeh and J. Soltani, "Emotional Controller (BELBIC) Based DTC for Encoderless Synchronous Reluctance Motor Drives," *2011 2<sup>nd</sup> Power Electronics, Drive Systems and Technologies Conference*, pp. 478-483, 2011
- [8] J. S. Albus, "A New Approach to Manipulator Control: The Cerebellar Model Articulation Controller (CMAC)," *Journal of Dynamic System, Measurement and Control*, Vol. 97, No. 3, pp. 220-227, 1975
- [9] C. M. Lin, C. F. Hsu and C. M. Chung, "RCMAC-based Adaptive Control Design for Brushless DC Motors," *Neural Computing and Applications*, Vol. 18, No. 7, pp. 781-790, 2009
- [10] P. E. M. Almedia and M. G. Simoes, "Parametric CMAC Networks Fundamentals and Applications of a Fast Convergence Neural Structure," *IEEE Transactions Industrial Applications*, Vol. 39, No. 5, pp. 1551-1557, 2003
- [11] C. M. Lin and T. Y. Chen, "Self-Organizing CMAC Control for a Class of MIMO Uncertain Nonlinear System," *IEEE Transactions on Neural Network*, Vol. 20, No. 9, pp. 1377-1384, 2009
- [12] C. M. Lin, Y. F. Peng and C. F. Hsu, "Robust Cerebellar Model Articulation Controller Design for Unknown Nonlinear Systems," *IEEE Transactions on Circuits System*, Vol. 51, No. 7, pp. 354-358, 2004
- [13] L. Bessissa, L. Boukezzi and D. Mahi "A Fuzzy Logic Approach to Model and Predict HV Cable Insulation Behaviour under Thermal Aging," *Journal of Applied Science, Acta Polytechnica Hungarica*, Vol. 11, No. 3, pp. 107-123, 2014
- [14] C. B. Regaya, A. Zaafouri and A. Chaari, "A New Sliding Mode Speed Observer of Electric Motor Drive Based on Fuzzy-Logic," *Journal of Applied Science, Acta Polytechnica Hungarica*, Vol. 11, No. 3, pp. 220-232, 2014
- [15] J. J. E. Slotine and W. P. Li, "Applied Nonlinear Control," *Englewood Cliffs, NJ, USA:Prentice-Hall*, 1991

- [16] C. M. Lin and C. F. Hsu, "Supervisory Recurrent Fuzzy Neural Network Control of Wing Rock for Slender Delta Wings," *IEEE Trans. Fuzzy Systems*, Vol. 12, No. 5, pp. 733-742, 2004
- [17] Y. C. Chang, "Robust  $H_{\infty}$  Control for a Class of Uncertain Nonlinear Time-Varying System and Its Application," *IEE Proceedings, Control Theory and Applications*, Vol. 151, No. 5, pp. 601-609, 2004

# Validating Rule-based Algorithms

**László Lengyel**

Department of Automation and Applied Informatics  
Budapest University of Technology and Economics  
Magyar tudósok körútja 2, 1117 Budapest, Hungary  
lengyel@aut.bme.com

---

*Abstract: A rule-based system is a series of if-then statements that utilizes a set of assertions, to which rules are created on how to act upon those assertions. Rule-based systems often construct the basis of software artifacts which can provide answers to problems in place of human experts. Such systems are also referred as expert systems. Rule-based solutions are also widely applied in artificial intelligence-based systems, and graph rewriting is one of the most frequently applied implementation techniques for their realization. As the necessity for reliable rule-based systems increases, so emerges the field of research regarding verification and validation of graph rewriting-based approaches. Verification and validation indicate determining the accuracy of a model transformation / rule-based system, and ensure that the processing output satisfies specific conditions. This paper introduces the concept of taming the complexity of these verification/validation solutions by starting with the most general case and moving towards more specific solutions. Furthermore, we provide a dynamic (online) method to support the validation of algorithms designed and executed in rule-based systems. The proposed approach is based on a graph rewriting-based solution.*

*Keywords: verification/validation of rule-based systems; graph rewriting-based model transformations; dynamic verification of model transformations*

---

## 1 Introduction

Rule-based systems [1] [2] provide an adaptable method, suitable for a number of different problems. Rule-based systems are appropriate for fields, where the problem area can be written in the form of *if-then* rule statements and for which the problem area is not extremely too great. In case of too many rules, the system may become difficult to maintain and can result in decreased performance speeds.

A classic example of a rule-based system is a domain-specific expert system that uses rules to make deductions or narrow down choices. For example, an expert system might help a doctor choose the correct diagnosis based on a dozen symptoms, or select tactical moves when playing a game. Rule-based systems can

be used in natural language processing or to perform lexical analysis to compile or interpret computer programs. Rule-based programming attempts to derive execution instructions from a starting set of data and rules. This is a more indirect method than that employed by an imperative programming language, which lists execution steps sequentially.

As rule-based systems are being applied to many diverse scenarios, there is a need for methods that support the verification and validation (V&V) of the algorithms performed by these systems. In this paper, we discuss the concept of taming the complexity of the verification/validation solutions. Furthermore, we introduce a dynamic (online) approach to address the V&V of rule-based systems.

V&V of a rule-based system is the process of ensuring that the rules meet specifications and fulfill their intended purpose. Based on [3], we use the following definitions: *Verification* is the process of evaluating the rule definitions to determine whether the imposed specification is fulfilled. *Validation* is the process of evaluating the rules, either during or following the rule execution, to determine whether it satisfies the end-user requirements. In other words, validation is intended to answer the question: “Is this the system we intended to create from the users perspective?” (Is this product specified according to the user's actual needs?) Verification provides answers to the question: “Is the system built in accordance with the design?” (Does the product conform to the specifications?)

During the analysis of a rule-based system, our goal is to prove that (i) certain properties hold for the output, if the input is valid, or (ii) to provide the criteria that must be satisfied by the input in order to guarantee the desired properties for the output. The analysis of a rule-based system is said to be *static* when the implementation of the rules and the language definition of the input and output models are used during the analysis process without considering the specific input. In the case of the *dynamic* approach, we analyze the rule-based system for a specific input, and then check whether certain properties hold for the output during or after the successful application of the rules. The static technique is more general and poses more complex challenges. The goal of static analysis is to determine whether the rule-based system itself meets various, specific requirements.

The rest of this paper is organized as follows. Section 2 discusses the concept of taming the complexity of verification/validation solutions. We start with the most general case, static methods, and work toward the most specific, the dynamic solution. Section 3 introduces a method, which makes possible to dynamically validate rule-based systems. Section 4 compares our solution with the related V&V approaches and further highlights the relevance of the suggested approach. Finally, concluding remarks are elaborated.

## 2 Taming Verification/Validation Complexity

Several static approaches provide formalism and verify that the semantics are preserved or guaranteed during the transformation of a model, e.g. approaches provided by Asztalos et al. [4], Biermann et al. [5], Bisztray and Heckel [6], Cabot et al. [7], or Schatz [8].

The approach of Asztalos et al. focuses on the static analysis of special model processing programs. This approach provides the theoretical basis for a possible verification framework. It applies a final formula that describes the properties that remain true at the end of the transformation. It is possible to derive either proof or refutation of a verifiable property from this final formula. The approach provides predefined components to deduct the desired properties.

In the approach, presented by Bisztray and Heckel, to understand and control the semantic consequences, Communicating Sequential Processes (CSP) are applied to capture the behavior of processes both before and after the transformation. The approach verifies semantic properties of the transformations at the level of rules, such that every application of a rule has a known semantic effect.

In the approach of Bierman et al., model transformations are defined as a special kind of typed graph transformations. The solution implements a formal approach to validate various functional behaviors and consistencies of model transformations.

There exist noticeable differences between the complexity of static and dynamic V&V approaches. The static technique is more general, because its responsibility is to determine if the rule-based system itself meets certain requirements. Contrarily, in the case of the dynamic approach, the transformation is analyzed based on a single specific input.

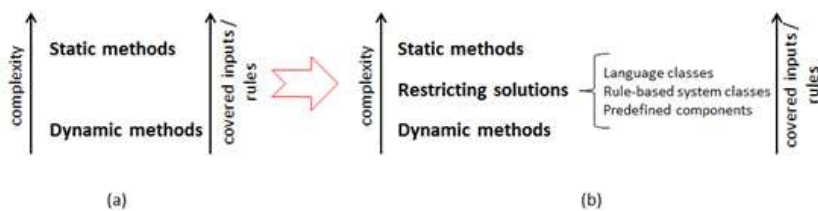


Figure 1

Taming the complexity of verification/validation

It is common knowledge that the algorithmic complexity of V&V can be very challenging, if not altogether hopeless in a general sense. However, practical cases do not require generality. The complexity-related questions are as follows: Does the problem contain specific subclasses that are solvable, yet practically relevant? Is it necessary to analyze the algorithm of the rule-based system? Or will it suffice to verify the system for a certain class of possible input models?

Our classification defines that the static approach is the most general and the dynamic is the most specific of V&V methods (Figure 1a). In order to reduce V&V complexity, we classify the V&V approaches to address complexity. In order to accomplish this, we begin with the general case (static verification) and create more specific cases (dynamic verification). Between these two extreme approaches, we identify several complexity-related restricting solutions (Fig. 1b). These methods do not attempt to prove the semantic correctness for one or all possible inputs (prove the properties of the rule-based system), but instead take a class of input types into consideration. We have identified the following complexity categories:

- A. Static methods
- B. Restricting solutions:
  - 1. Language classes
  - 2. Rule-based system classes
  - 3. Predefined components
- C. Dynamic methods

**A. *Static methods.*** Model checker tools (e.g. Augur [9], CheckVML [10], or GROOVE [11]) apply static methods during the verification.

**B1. *Restricting solutions / Language classes.*** This approach defines a class of input models. Based on a metamodel, a language class is defined by additional metamodel constraints or the simplification of the metamodel, i.e. through the elimination of some domain concepts. As a result, the modified language contains only a restricted class of original models, therefore, the complexity of the processing transformation decreases along with the complexity of the transformation verification. Examples of language classes are provided in OMG Query/View/Transformation Specification [12]. The Annex A of the QVT specification introduces two language classes: *Simple UML Metamodel* and *Simple RDBMS Metamodel*. These domains provide limited language elements and attributes that are suitable to define the required models, but do not provide additional, unnecessary language constructions. For instance, a *Table* containing *Keys* can be modeled, but the *Key* type does not provide attributes to specify further details. Another example regarding language classes is the limitation of the multiplicity to *1* or *0..1*. A third example presents itself when only finite input models are permitted. A sample language class also introduced in the next section: *DomainServers* (Figure 2).

**B2. *Restricting solutions / Rule-based system classes.*** This approach restricts the rule specification language itself. We modify the metamodel of the rule specification language in order to allow for rule-based system definitions with



specific properties. An example of a rule-based system class is one in which rule chains are allowed (successively applying several rules in a predefined order) but loops are forbidden. Proving the termination of such rule-based system requires reasonably less complexity than in the general case, when loops are permitted. For example in the field of layered grammars [13], Bottoni et al. [14] developed a termination criterion which ensures that the creation of all objects of a certain type should precede the deletion of an object of the same type. Therefore, the layer deleting an object of a given type should not create such an object, nor should the subsequent rules. This means the productions in a deletion layer will terminate. Therefore, the termination analysis of transformations satisfying this criterion requires less complexity than the general case.

**B3. Restricting solutions / Predefined components.** In this case, the verification procedure is constructed from predefined components. We can state facts about the components that treat the verification process as axioms: therefore, the results of other tools or human analysis can be also utilized. Applying these predefined components, we can deduce what output model properties are provided by the given transformation for the provided input domain. For example, the formal language, developed by Asztalos et al. [4], is able to express a set of model transformation properties. The language is appropriate to specify both the properties of the output models and the properties of the relation between the input and output model pairs. In most cases, the proofs within the class of predefined components are conducted by dedicated checker tools (e.g. GROOVE [11] or CheckVML [10]) or through human analysis.

**C. Dynamic methods.** Examples of dynamic methods are provided by Lengyel [15]. In their approach, the validation of the rule-based system is achieved with constraints assigned to the rules as pre- and postconditions. A similar approach was developed by Narayanan and Karsai [16], in which the semantic equivalence between inputs was guaranteed via bi-simulation checks on the execution log of the transformation.

In applying these restricting solutions, i.e. working with language classes, rule-based system classes, or predefined components, we ensure that (i) the verification of the rule-based system requires less complexity than the classical static verification and (ii) the verification results are valid not only for a specific input, but for a class of input models or rule-based systems. The next section discusses a dynamic validation method.

### 3 Dynamically Validated Rule-based Systems

There are several model transformation approaches ranging from relational specifications [17] and graph transformation techniques [18], to algorithmic techniques for the implementation of a model transformation. The following provides our categorization of these approaches: [19] traversal-based and direct manipulation approaches [20], template-based approaches (e.g., OCL [21], XPath, or T4 Text Templates), relational approaches (e.g., Query, Views, Transformations (QVT) [12]), graph rewriting-based approaches (e.g., AGG [22], AToM<sup>3</sup> [23], GReAT [24], TGGs [25], VIATRA2 [26], and VMTS [27]), structure-driven approaches (e.g., OptimalJ and QVT), and hybrid approaches that combine two or more of the previous categories (e.g., ATL [28]).

Rule-based systems are often realized based on the graph rewriting-based approach. Therefore, our focus is on the V&V of the graph transformations.

Graph rewriting-based transformations [29] have their roots in classical approaches to rewriting, such as Chomsky grammars and term rewriting [30]. There are many other representations of this, which will be addressed later. In essence, a rewriting rule is composed of a left-hand side (LHS) pattern and a right-hand side (RHS) pattern. Operationally, a graph transformation from a graph  $G$  to a graph  $H$  is mainly conducted based on the following three steps:

- a. Choose a rewriting rule.
- b. Find an occurrence of the LHS in host graph  $G$  satisfying the application conditions of the rule.
- c. Finally, replace the subgraph matched in  $G$  by the RHS.

Graph transformations define the transformation of models. The LHS of a rule defines the pattern to be found in the host model; therefore, the LHS is considered the positive application condition (PAC). However, it is often necessary to specify what pattern should not be present. This is referred to as negative application condition (NAC) [31]. Besides NACs, some approaches [22] [26] use other constraint languages, e.g., OCL, Java, C# or Python to define the execution conditions.

The ordering of rules can be achieved by explicit control structures or can be implicit due to the nature of their rule specifications. Moreover, several rules may be applicable simultaneously. Blostein et al. [32] have classified graph transformation organization into four categories. (i) An unordered graph-rewriting system simply consists of a set of graph-rewriting rules. Applicable rules are selected non-deterministically until none are any longer applicable. (ii) A graph grammar consists of rules, a start graph and terminal states. Graph grammars are used for generating language elements and language recognition. (iii) In ordered graph-rewriting systems, a control mechanism explicitly orders the rule

application of a set of rewriting rules (e.g. priority-based, layered/phased, or those containing an explicit control flow structure). (iv) In event-driven graph-rewriting systems, rule execution is triggered by external events. This approach has recently seen a rise in popularity [33].

Controlled (or programmed) graph transformations impose a control structure over the transformation rules to maintain a more strict order of execution in a sequence of rules. The control structure primitives of graph transformation may provide the following properties: atomicity, sequencing, branching, looping, non-determinism, recursion, parallelism, back-tracking and/or hierarchy [15] [30].

Some examples of control structures are as follows: AGG [22] uses layered graph grammars. The layers fix the order in which rules are applied. The control mechanism of AToM<sup>3</sup> [23] is a priority-based transformation flow. Fujaba [34] uses story diagrams to define model transformations. The control structure language of GReAT [24] uses a data flow diagram notation. GReAT also has a test rule construction; a test rule is a special expression that is used to change the control flow during execution. VIATRA2 [26] applies abstract state machines (ASM). VMTS [27] uses stereotyped UML activity diagrams to further specify control flow structures. In [29], a comparative study is provided that examines the control structure capabilities of the tools AGG, AToM<sup>3</sup>, VIATRA2, and VMTS.

In the case of rule-based systems, the application order of the rules is supported by a conflict resolution strategy. The strategy may be determined by the actual area or may simply be a matter of preference. In any case, it is vital as it controls which of the applicable rules are fired and thus the behavior of the entire system. The most common strategies are as follows:

- a. *First applicable*: If the rules are in a specified order, firing the first applicable rule allows for control over the order in which rules are fired.
- b. *Random*: Though it does not provide the predictability or control of the first-applicable strategy, it does have certain advantages. For one, its unpredictability is an advantage in some circumstances (e.g., in games). A random strategy simply chooses a single random rule to fire from the conflict set. Another possibility for a random strategy is a fuzzy rule-based system in which each rule has a factored probability, i.e., some rules are more likely to fire than others.
- c. *Least recently used*: Each of the rules is accompanied by a time or step stamp, which marks the time of its last usage. This maximizes the number of individual rules that are fired at least once. This strategy is perfect when all rules are needed for the solution of a given problem.
- d. *Best rule*: Each rule is given a weight, which specifies its comparative consideration to the alternatives. The rule with the most preferable outcomes is chosen based on this weight.

### 3.1 An Example

Rules can be made more relevant to software engineering models if the transformation specifications allow the assigning of validation constraints to the transformation rules.

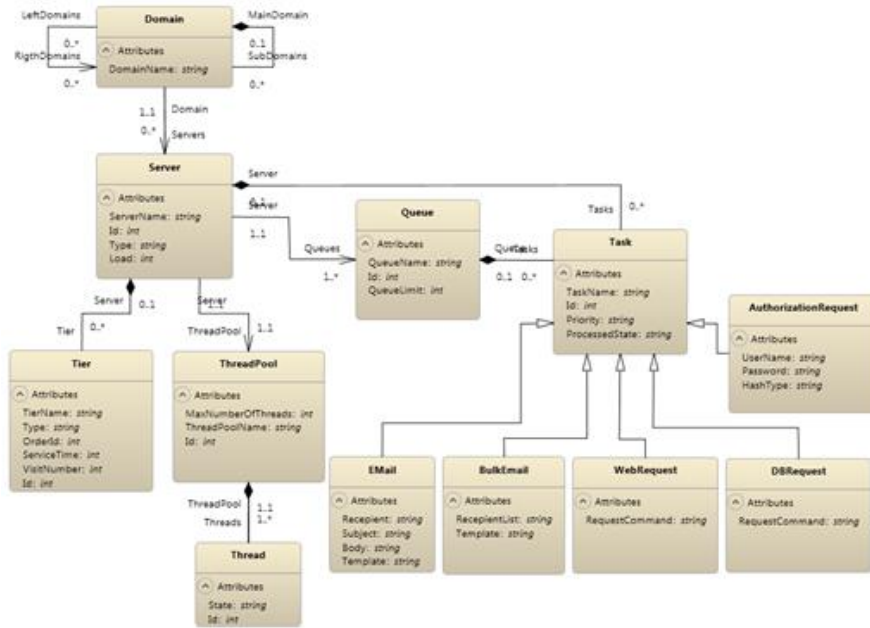


Figure 2

The *DomainServers* metamodel

Figure 2 depicts the metamodel of a domain-specific language. This language defines that an instance model contains *Domain* objects. A domain can contain sub-domains and domains can also be linked to each other. A domain has *Server* objects. A server must belong to a domain. A server contains a *ServerName*, *Id*, *Type* (enum attribute with values *Web*, *Database*, *Mail*, and *Gateway*), and *Load* attributes. *Servers* contain sequentially ordered *Tiers*. A tier has the following attributes: *TierName*, *Id*, *Type* (enum attribute with values *CPU* and *I/O*), *OrderId*, *ServiceTime*, and *VisitNumber*. Each server has exactly one *ThreadPool* element. A *ThreadPool* is comprised of *ThreadPoolName*, *Id*, and *MaxNumberOfThreads* attributes. The *ThreadPool* contains *Threads*. Each thread has *Id* and *State* (enum attribute with values *Ready* and *Occupied*) attributes. *Servers* have one or more *Queues*. A queue has *QueueName*, *Id*, and *QueueLimit* attributes. A queue must belong to a server. Furthermore, servers and queues can contain *Tasks*. *Tasks* assigned to servers are under processing, while tasks in a queue are in waiting state. A task has the following attributes: *TaskName*, *Id*, *Priority* (enum attribute

with values *Normal*, *High*, and *Urgent*), and *ProcessedState* (enum attribute with values *Waiting*, *Processing*, and *Complete*). There are also more specific task types inherited from *Task*: *Email*, *BulkEmail*, *WebRequest*, *DBRequest*, and *AuthorizationRequest*. Each of these metamodel elements also includes further attributes.

Figure 3 introduces a control flow model of a rule-based system. The processing has three transformation rules. The rule *CheckServerLoad* selects a *Server*, which *Load* is over 80%. If there is no such server, then the transformation terminates. Otherwise, a new server node, with a *ThreadPool* and a *Queue* node, is inserted into the domain. Next, the transformation rule, *RearrangeTasks*, rearranges tasks from the queue of the overloaded server to the queue of the new server. The rule *RearrangeTasks* is executed in *Exhaustive* mode: the rule is continuously applied while the *Load* of the overloaded servers is over 70%, and the *Load* of the new server remains under 70%. The transformation is executed in a loop. This means, after easing the load of one server, the process continues and therefore, the transformation can insert additional, new servers.

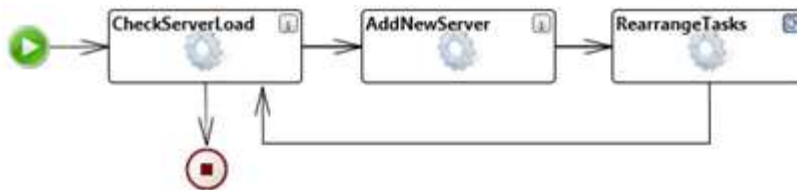


Figure 3

Example model transformation: *LoadBalancing*

Figure 4 depicts two example rules: *AddNewServer* and *RearrangeTasks*. The figure follows a compact notation, containing no separated LHS and RHS pattern. The colors code the following: black nodes and edges denote unmodified elements, blue ones indicate newly created elements, and red ones mark the elements deleted by the rule. The transformation rule *AddNewServer* gets the *Domain* type node as a parameter and creates the new *Server* with a *ThreadPool*, two *Threads*, a *Tier*, and a *Queue*. The transformation rule, *RearrangeTasks*, receives the two servers with their queues as parameters and performs the rearrangement as a single task. The rule is executed in *Exhaustive* mode, which enables several tasks to be moved between the queues.

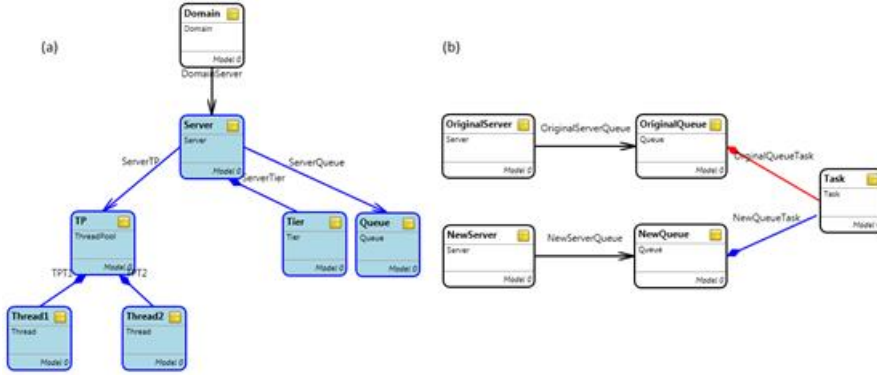


Figure 4

Example model transformation rules: (a) *AddNewServer* and (b) *RearrangeTasks*

Some example constraints assigned to the rules are as follows:

**context** Server **inv** serverCardinality:

```
Server.allInstances()->count() < 40
```

**context** Queue **inv** queueCardinality:

```
Queue.allInstances()->count() >= Server.allInstances()->count()
```

The constraints *serverCardinality* and *queueCardinality* define the number of specific type elements in the model. These are cardinality issues related to the whole model.

**context** Queue **inv** queueLimit:

```
QueueLimit < 1500
```

The constraint *queueLimit* is an attribute value constraint that maximizes *QueueLimit* attribute of *Queue* type nodes.

**context** Server **inv** largeThreadPools:

```
Server.allInstances()->forall(s | s.ThreadPool.Threads->count() <= 50 OR
  (s.Tiers->exists(t | t.Type = Type::CPU) AND
    s.Tiers->exists(t | t.Type = Type::I/O)))
```

The constraint *largeThreadPools* defines that for each server, if the number of threads in the *ThreadPool* exceeds 50, then separated CPU and I/O tiers are employed.

The presented constraints are assigned to the rules and guarantee our requirements. After a successful rule execution, the conditions hold and the output is valid. The fact that the successful execution of the rule guarantees the valid output cannot be achieved without these validation constraints.

### 3.2 Validating Rule-based Systems

The objective of our research activities is to support the V&V of algorithms performed by rule-based systems. The requirements, assigned to the rules are both input and output related requirements, i.e. we define certain pre- and postconditions that should hold before and after the execution of the rule. In several cases rules do not contain certain node or edge types that are about to be included into our V&V requirements. These requirements may relate to a temporary (during the processing) or a final (following the processing) state of the input or generated models. Moreover, several different directions can be followed; e.g. we can assert additional requirements to the input and output models (metamodel constraints), or the rule-based system can be extended with the use of appropriate testing and validating rules.

Dynamic validation covers both the attribute value and the structure validation, which can be expressed in first-order logic extended with traversing capabilities. Example languages that currently applied for defining attribute value and interval conditions are Object Constraint Language (OCL), C, Java, and Python. These conditions and requirements are pre- and postconditions of a transformation rule.

*Definition (Precondition).* A precondition assigned to a rule is a Boolean expression that must be true at the moment of rule firing.

*Definition (Postcondition).* A postcondition assigned to a rule is a Boolean expression that must be true after the completion of a rule.

If a precondition of a rule is not true, then the rule fails without being fired. If a postcondition of a rule is not true after the execution of the rule, the rule fails.

Regarding pre- and postconditions the execution of a rule is as follows (Figure 5):

- a. Finding the match according to the LHS structure.
- b. Validating the constraints defined in LHS on the matched parts of the input model.
- c. If a match satisfies all constraints (preconditions), then executing the rule, otherwise the rule fails.
- d. Validating the constraints defined in RHS on the modified/generated model. If the result of the rule satisfies the postconditions, then the rule was successful, otherwise the rule fails.

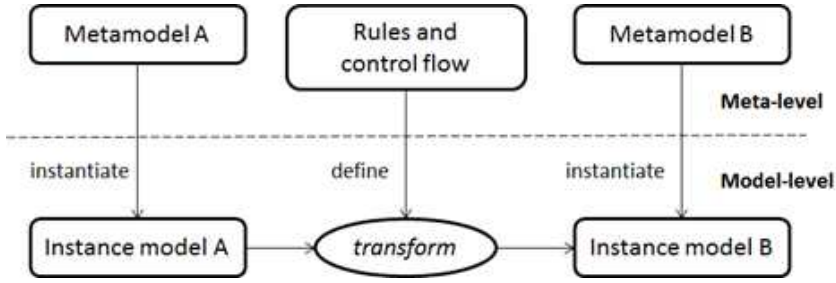


Figure 5

The transformation process

A direct corollary is that an expression in LHS is a precondition to the rule, and an expression in RHS is a postcondition to the rule. A rule can be executed if and only if all conditions enlisted in LHS are true. Also, if a rule finished successfully, then all conditions enlisted in RHS must be true.

*Statement 1.* If a finite sequence of rules is specified properly with the help of validation constraints, and the sequence of rules has been executed successfully for the input model, then the modified/generated output model is in accordance with the expected result that is described by the finite sequence of transformation rules refined with the constraints [7].

*Definition (Low-level construct).* Pre- and postconditions defined as constraints and propagated to the rules are low-level constructs.

*Definition (High-level construct).* Validation, preservation and guarantee properties are high-level constructs.

*Definition (Validated rule execution).* A rule execution is validated if it satisfies a set of high-level constructs.

To summarize, high-level constructs define the requirements on a higher abstraction level, e.g. servers should not be overloaded. Low-level constructs are the appropriate constraints assigned to the appropriate rules. These constraints assist in achieving the required conditions.

This method can be followed in Figure 4. Finding the structural match the preconditions are validated, and after performing the rule execution, postconditions are validated. Both of the validation should be successful in order for the whole rule to be successful.

With this method the required properties can be defined on low-level, i.e. on the level of rules. In summary, we can say that the presented dynamic approach supports that if the execution of a rule finishes successfully, the generated output is valid and fulfills the required conditions. The validation of the rule-based system is achieved with constraints assigned to the rules as pre- and postconditions.



*Statement 2.* Rule-based systems can be validated with the presented dynamic validation method.

*Statement 3.* Taming verification complexity can be applied for rule-based systems.

## 4 Related Work

In order to underpin the relevance of current results we compiled a collection of challenging transformations requiring V&V. The provided methods support different domain-specific languages-based [35] model-driven approaches.

Giese et al. [36] points out the challenge in using model-driven software development (MDD). The problem is the lack of verified transformations, especially in the area of safety-critical systems. The verification of critical safety properties on the model level is useful only if the automatic code generation is guaranteed to be correct, i.e. the verified properties are guaranteed to hold true for the generated code as well. This means it is necessary to pay special attention to checking for semantic equivalence, at least to a moderate level, between the model specification and the generated code.

In the field of developing safety-critical systems, model analysis possesses advantages over pure testing of implemented systems. For example, important required safety properties of a system under development could be verified on the model level rather than trying to systematically test for the absence of failures.

Narayanan and Karsai [16] have summarized that in the development of a model-based software, a complete design and analysis process involves designing the system using the design language, converting it into the analysis language and performing the verification on the analysis model. They established that graph transformations were a powerful and convenient method increasingly being used to automate this conversion. In such a scenario, the transformation must ensure that the analysis model preserves the semantics of the design model. They concluded that methods are required to verify that the semantics used during the analysis are indeed preserved across the transformation.

de Lara and Taentzer [37] discussed the need for verified and validated model processing in the field of Multi-Paradigm Modeling (MPM) [38]. Software systems have components that may require descriptions using different notations, due to different characteristics. For the analysis of certain properties of the system as a whole, or its simulation, we transformed each component into a common single formalism, in which appropriate analysis or simulation techniques are available.

Varró [39] went on to state that due to the increasing complexity of IT systems and modeling languages, conceptual, human design errors will occur in any model on any high level of the formal modeling paradigm. Accordingly, the use of formal specification techniques alone does not guarantee the functional correctness and consistency of the system under design. Therefore, automated formal verification tools are required to verify the requirements fulfilled by the system model. As the input language of model checker tools is too basic for direct use, model transformations are applied to project behavioral models into the input languages of the model-checking tools.

In conclusion, it is important to understand that model transformations and rule-based systems themselves can be erroneous; therefore, uncovering solutions to make model transformations and rule-based systems free of conceptual errors is essential.

## **Conclusions**

Rule-based systems can effectively automate problem-solving standards. Such systems provide a method for capturing and refining human expertise, and affirm their relevance to the industry. Instead of representing knowledge in a relatively declarative way, i.e., numerous things that are known to be true, rule-based systems represent knowledge in terms of a collection of rules that tell what should be done, i.e., what can be concluded from different situations?

The motivation of the current work was to support the verification/validation of rule-based systems. In this paper, we have introduced the concept of taming verification complexity. We have seen that the static validation method is more general and raises challenges that are more complex. We have discussed the possibilities of reducing the complexity of V&V and have introduced different restricting solutions. Finally, we have presented the dynamic approach, in which the rule-based system is validated for a specific input model.

Then, we have introduced a method, which facilitates to apply the graph rewriting-based dynamic (online) validation results in the field of rule-based systems. The solution facilitates to validate single rules, rule chains, and in effect transformations as a whole. The validation is driven by the pre- and postconditions assigned to these rules.

Our current research activities concentrate on trace-based verification/validation approaches. In these cases, constraints are validated based on the trace files, following the execution. This is the difference between the trace-based approach and the currently presented dynamic approach.

## **Acknowledgement**

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013) organized by VIKING Zrt. Balatonfüred.

## References

- [1] Hayes-Roth, F.: Rule-based systems, *Communication of ACM* 28, 9, 921-932 (1985), DOI=10.1145/4284.4286 <http://doi.acm.org/10.1145/4284.4286>
- [2] Williams, T. and Bainbridge, B.: *Rule-based Systems*, In *Approaches to Knowledge Representation: an Introduction*, Research Studies Press Ltd., Taunton, UK, UK 101-115 (1988)
- [3] IEEE Standard Glossary of Software Engineering Terminology, 610.12-1990 (1990)
- [4] Asztalos, M., Lengyel, L., Levendovszky, T.: A Formal Framework for Automated Verification of Model Transformations, *Software Testing, Verification and Reliability*, 23:(5), 405-435 (2013)
- [5] Biermann, E., Ermel, C., Taentzer, G.: Formal Foundation of Consistent EMF Model Transformations by Algebraic Graph Transformation, *Software and Systems Modeling (SoSyM)*, Springer, 1-24 (2011)
- [6] Bisztray, D., Heckel, R., Ehrig, H.: Verification of Architectural Refactorings by Rule Extraction, In *Fundamental Approaches to Software Engineering*, LNCS, Vol. 4961, Springer, 347-361 (2008)
- [7] Cabot, J., Clariso, R., Guerra, E., de Lara, J.: V&V of Declarative Model-to-Model Transformations through Invariants, *J. Syst. Softw.*, Vol. 83(2), 283-302 (2010)
- [8] Schatz B.: Formalization and Rule-based Transformation of EMF Ecore-based Models, *Software Language Engineering: First International Conference, SLE 2008, France*, 227-244 (2008)
- [9] Augur website, <http://www.ti.inf.uni-due.de/research/augur/index.html>
- [10] Rensink, A., Schmidt, A., Varró, D.: Model Checking Graph Transformations: A Comparison of Two Approaches. *Proceedings of the ICGT 2004: Second International Conference on Graph Transformation*, LNCS, Vol. 3256, Springer, Rome, Italy, 226-241 (2004)
- [11] GROOVE: GRaphs for Object-oriented VERification Website, <http://groove.sourceforge.net/groove-index.html>
- [12] OMG Query/View/Transformation (QVT) Specification, Meta Object Facility 2.0 Query/Views/Transformation Specification, OMG doc. ptc/07-07-07, 2007, <http://www.omg.org/>
- [13] Ehrig, H., Ehrig, K., de Lara, J., Taentzer, G., Varró, D., Varró-Gyapay, Sz.: Termination Criteria for Model Transformation, *FASE 2005*, LNCS, 49-63 (2005)
- [14] Bottoni, P., Taentzer, G., Schürr, A.: Efficient Parsing of Visual Languages based on Critical Pair Analysis and Contextual Layered Graph

- Transformation, Proceedings of the Visual Languages 2000 IEEE Computer Society, 59-60 (2000)
- [15] Lengyel, L.: Online Validation of Visual Model Transformations, PhD thesis, Budapest University of Technology and Economics, Department of Automation and Applied Informatics (2006)
  - [16] Narayanan, A., Karsai, G.: Towards Verifying Model Transformations, ENTCS, Vol. 211, 191-200 (2008)
  - [17] Akehurst, D., Kent, S.: A Relational Approach to Defining Transformations in a Metamodel, In UML 2002 - The Unified Modeling Language, 5<sup>th</sup> International Conference, Dresden, Germany, LNCS, Vol. 2460, Springer-Verlag, 243-258 (2002)
  - [18] Ehrig, H., Engels, G., Kreowski, H.-J., Rozenberg, G. editors: Handbook on Graph Grammars and Computing by Graph Transformation: Application, Languages and Tools, Vol. 2, World Scientific, Singapore (1999)
  - [19] Mens, T., v. Gorp, P.: A Taxonomy of Model Transformation, Electronic Notes in Theoretical Computer Science, Vol. 152, Proceedings of the International Workshop on Graph and Model Transformation (GraMoT 2005), 125-142 (2006)
  - [20] Vajk, T., Kereskényi, R., Levendovszky, T., Lédeczi, Á.: Raising the Abstraction of Domain-Specific Model Translator Development, 16<sup>th</sup> Annual IEEE International Conference and Workshop on the Engineering of Computer Based Systems, USA, 31-37 (2009)
  - [21] OMG Object Constraint Language (OCL) Specification, Version 2.2, OMG document formal/2010-02-01, 2010, <http://www.omg.org/>
  - [22] AGG: The Attributed Graph Grammar System website, <http://tfs.cs.tu-berlin.de/agg>
  - [23] AToM<sup>3</sup>: A Tool for Multi-paradigm, Multi-formalism and Meta-modeling website, <http://atom3.cs.mcgill.ca>
  - [24] GReAT: Graph Rewriting and Transformation website, <http://www.isis.vanderbilt.edu/tools/GReAT>
  - [25] Schürr, A.: Specification of Graph Translators with Triple Graph Grammars, Proceedings of the WG94 international workshop on graph-theoretic concepts in computer science, LNCS, Vol. 903, Springer, Berlin Heidelberg New York, 151-163 (1994)
  - [26] VIATRA2 (VIsual Automated model TRAnsformations) framework website, <http://eclipse.org/gmt/VIATRA2>
  - [27] VMTS: Visual Modeling and Transformation System website, <http://www.aut.bme.hu/vmts>
  - [28] ATL: ATLAS Transformation Language website, <http://eclipse.org/atl/>

- [29] Taentzer, G., Ehrig, K., Guerra, E., de Lara, J., Lengyel, L., Levendovszky, T., Prange, U., Varró D., Varró-Gyapay, Sz.: Model Transformation by Graph Transformation: A Comparative Study, ACM/IEEE 8<sup>th</sup> International Conference on Model Driven Engineering Languages and Systems, Montego Bay, Jamaica (2005)
- [30] Rozenberg, G. (ed.): Handbook on Graph Grammars and Computing by Graph Transformation: Foundations, Vol. 1, World Scientific, Singapore (1997)
- [31] Habel, A., Heckel, R., Taentzer, G.: Graph Grammars with Negative Application Conditions, *Fundamenta Informaticae*, Vol. 26, 287-313 (1996)
- [32] Blostein, D., Fahmy, H., Grbavec, A.: Issues in the Practical Use of Graph Rewriting, In proceedings of the 5<sup>th</sup> International Workshop on Graph Grammars and Their App to Computer Science, Williamsburg, USA, LNCS, Vol. 1073, Springer-Verlag, 38-55 (1996)
- [33] Guerra, E., de Lara, J.: Event-driven Grammars: Relating Abstract and Concrete Levels of Visual Languages, *SoSym*, Vol. 6, 317-347 (2007)
- [34] Fujaba Tool Suite website, <http://www.fujaba.de/>
- [35] Kövesdán, G., Asztalos, M. and Lengyel L.: Architectural Design Patterns for Language Parsers, *Acta Polytechnica Hungarica* 11:(5), 39-57 (2014)
- [36] Giese, H., Glesner, S., Leitner, J., Schafer, W., Wagner, R.: Towards Verified Model Transformations, In *ModeV Va06* (2006)
- [37] de Lara, J., Taentzer, G.: Automated Model Transformation and its Validation with ATOM3 and AGG, in *Diagrammatic Representation and Inference*, Lecture Notes in Artificial Intelligence, Vol. 2980, Springer, 182-198 (2004)
- [38] de Lara, J., Vangheluwe, H., Alfonseca, M.: Metamodelling and Graph Grammars for Multi-Paradigm Modelling in ATOM3, *Journal of Software and Systems Modeling*, Vol. 3(3), 194-209 (2004)
- [39] Varró, D.: Automated Formal Verification of Visual Modeling Languages by Model Checking, *Journal on Software and System Modeling*, Vol. 3(2), 85-113 (2004)

# The Analysis and Classification of Birth Data

**Raul Robu**

Department of Automation and Applied Informatics, University Politehnica  
Timișoara, Bulevardul Vasile Pârvan, Nr. 2, 300223, Timișoara, Romania,  
raul.robust@aut.upt.ro

**Ștefan Holban**

Department of Computers, Faculty of Automation and Computers, University  
Politehnica Timișoara, Bulevardul Vasile Pârvan, Nr. 2, 300223, Timișoara,  
Romania, stefan.holban@cs.upt.ro

---

*Abstract: The paper presents a study regarding the births that took place at the Bega Obstetrics and Gynecology Clinique, Timișoara, Romania in 2010. The analysis began from a dataset including 2325 births. The article presents a synthesis of the studies that analyze birth data. The Apgar score is the main subject in many studies. On one hand, researchers investigated the relation between the Apgar score and different factors such as the newborns' cry, the level of glucose in the blood from the umbilical cord, the mother's body mass index before the pregnancy, etc. On the other hand, there are studies that demonstrate that the Apgar score is important for the ulterior evolution of babies. The article presents the attributes from the dataset and how they were preprocessed in order to be analyzed with Weka. The values of each attribute were investigated and the results were presented. The past experience regarding births, expressed through the dataset values, was then used to build classification models. With the help of these models, the Apgar score can be estimated based on the known information regarding the mother, the baby and possible medical interventions. The purpose of these estimations is consultative, to help in identifying which values of the input variables will lead to an optimal Apgar score, in certain circumstances. The classification models were built and tested with the help of ten classification algorithms. After the model that produces the best results of classification was determined, a dedicated application was developed, with the aid of the Weka API which classifies the birth data by using LogitBoost algorithm.*

*Keywords: birth data; classification; data mining; LogitBoost; Weka*

---

# 1 Introduction

Everyday, databases of enormous size are collected. The analysis of these data may help extract interesting and useful information. This analysis could be done individually for each attribute, following the frequency of certain values, of the medium, maximum or minimum values, etc., but also, treating the attributes together, by using data mining techniques, such as classification, clustering, discovering association rules, etc.

The classification techniques apply successfully in the medical field. Building classification models that allow an estimate, with a certain degree of confidence, of a class attribute based on different values of the input variables, may be particularly useful especially if the class attribute is a characteristic that is difficult to obtain through medical methods that they either endanger the patient's life or are prohibitively expensive. Such a classification model was built in [1], to estimate if a pulmonary nodule is cancerous or not, according to different input variables that are the nodule's characteristics and taking into consideration the patient's characteristics, obtained through noninvasive tests (the SPN diameter, border character, presence of calcification, patient's age, smoking history, results of CT densitometry, etc.). The classification model built in [1] can estimate if the nodule is cancerous or not. The medical procedure that determines whether such a nodule is cancerous or not is invasive and implies tissue sampling and analysis, an operation that is not always recommended due to certain patient's health state.

Another example of classification study in the medical field is the classification of patients that develop post operative complications following gastric cancer operations. Two methods were used: logistic regression and neuronal networks. The data from patients in Taiwan that suffered a gastric cancer operation were used. The ANN model had a better performance than logistic regression [2].

The classification models can also be built to simulate the value of the output variable based on different values of the input variables in order to choose values for the input variables so that the output variable has the desired value.

In the dedicated literature there are many studies that investigate birth data. A part of these studies analyze and classify the newborns' cry. Another part of the studies focuses on the Apgar score. In different papers the following are investigated: the relation between the Apgar score and the newborn's cry, between the Apgar score and the glucose level from the blood of the umbilical cord, between the Apgar score and the body mass index of the mother in the pre-pregnancy period, etc. There are several studies that investigate the effects of the Apgar score on the subsequent evolution of the baby. A short synthesis of the studies regarding birth related data is presented in chapter two. These studies demonstrate that it is important for the newborns future evolution that each newborn obtains an Apgar score that is as high as possible. That is why, in this paper, we built classification models which allow an estimation of the interval in which the Apgar score will be

situated considering mother's data, baby's data and data regarding the medical interventions that could possibly help the birth. The purpose of this estimation is to determine, based on the experience of previous birth data, experience represented by the classification model and based on the mother's and baby's data, which medical interventions will lead to the best Apgar score. For example, for a given birth, we can simulate, based on experience (the model) and on the mother's and baby's data, the Apgar score if the mother gives birth naturally as opposed to a caesarean delivery. After testing the models, the result was the model performed with an 80% confidence level. These estimations can be used with a consultative role by doctors and they should help them in the decision making process, so that the newborns have an Apgar score that is as high as possible.

The analyzed data contains information regarding births that took place in 2010, at the *Bega Obstetrics and Gynecology Clinique, Timișoara, România* which we will refer to hereinafter as the *Bega Clinique*. The initial data set contains information for 2325 births and 19 attributes for each birth. The analysis was done with *Weka* [3]. Preprocessing the data was necessary in order to introduce and analyze data in *Weka*. Following this operation, data regarding 2086 births remained as well as 15 attributes for each birth. Data were analyzed both statistically and using some classification techniques. The statistical analysis revealed interesting information, such as the fact that approximately 60% of the women that gave birth in 2010 had a caesarean operation; 62 under aged women became mothers; the medium weight of the newborns was 3194 grams, but two babies that weighed over 5000 grams, etc.

The classification models were built using *Naive Bayes* [4], *J48* [5], *k-Nearest Neighbour* [6], *Random Forest* [7], *Support Vector Machines* [8], *AdaBoost* [9], *LogitBoost* [10], *JRip* [11], *REPTree* [12], *SimpleCart* [13] algorithms. They were tested through cross validation with 10 folds. The best model, from the prediction accuracy point of view, was obtained using the *LogitBoost* algorithm. It allows an estimation with an 80% accuracy of the interval of the newborn's Apgar score based on data regarding the mother, baby and medical interventions. The classification models built in the beginning of the study had little accuracy. In order to increase their accuracy we had to redo the preprocessing phase, as well as the building and testing phases, several times. After we successfully built the models with a satisfactory accuracy, the algorithm that built the best model was chosen and we developed an application dedicated to classifying data regarding births that uses the *Weka's API*. The application allows building classification models with the aid of *LogitBoost* algorithm and has an interface in which data regarding the mother, newborn and medical interventions can be added, these data are used by the classification model to estimate the interval of the *Apgar score*. The reliability of the prediction resulted, after testing the model through cross validation, is approximately 80%.



## 2 Studies on Birth Data

In the current literature, there are different studies that investigate birth data. Some of these studies, analyze and classify the newborns cry. Newborn babies use their cries by instinct, to communicate their needs. The different cries of the infant can indicate different requirements. The paper [14] proposes a method to determine the meanings of infant cries according to a baby expert. It applies the novel *Neuro-fuzzy* techniques for the classification and Perceptual *Linear Prediction* for recognition the infant cries. The results showed that the classification performance obtained by using the *Neuro-fuzzy* techniques yielded the most desirable accuracy over other popular methods. In the paper [15] the authors created a classification model with the aid of the *Support Vector Machines* algorithm which can classify, with 76.23% accuracy, the pathologic cry. Two types of pathologies were analyzed: asthma and ischemic encephalopathy.

There are other studies that analyze the birth data that have in the middle the Apgar score. The Apgar score was introduced by *Virginia Apgar* in 1952 as a means to evaluate the health of the newborns immediately after birth [16]. Apgar score is a clinical test performed on a newborn one and five minutes after birth. It is a composite measure of breathing effort, heart rate, muscle tone, reflexes, and skin color. It is an indicator of the newborn's need for medical attention, shortly after the birth [17] [16]. The Apgar score has survived the test of time as it is still used nowadays in maternities [18] [19].

In [20] the cry characteristics of newborn infants were investigated and correlated with the Apgar scores. The cry of premature and mature infants with low and normal Apgar scores was analyzed using *principle component analysis (PCA)*. The reduced dimension cry signal was investigated to extract features and to correlate them with Apgar scores. The paper [20] proposes the foundation of the design of an automatic algorithm, to replace the manual Apgar scoring system.

In [21] the authors define birth asphyxia based on fetal condition as measured by umbilical artery blood pH, Apgar scores, and neurologic condition of newborns.

A study of the connection between the glucose level from the blood of the umbilical cord and the Apgar score was done in [22]. The study had two major objectives. First, it tried to determine a standard reference level for the glucose in the umbilical cord. The second important objective of the study was to determine if an abnormal level of glucose in the blood of the umbilical cord influences in a negative manner the Apgar score. Following the investigations, no relation between the two factors was determined.

Obesity is a global health problem and maternal obesity may be associated with an increased risk of pregnancy complications and neonatal death. The purpose of the study in [23] was to evaluate the effect of the maternal pre-pregnancy body mass index (*BMI*) on the newborns Apgar score. The study concludes that is

recommended that obese and overweight women should be treated to normalize their *BMI* prior to pregnancy.

There are also a few studies that investigate the effect of the Apgar score to the child evolution in time. In [24] the authors investigate the association of Apgar score at five minutes with long-term neurologic disability and cognitive function. The conclusion was that five-minute Apgar score less than seven has a consistent association with prevalence of neurologic disability and with low cognitive function in early adulthood. In [25] the authors have investigated the relationship between the Apgar Scores at 5 minutes after birth and School Performance at 16 years of age. The study included 877 individuals in the analysis. Newborns with Apgar scores less than 7, at 5 minutes after birth, showed a significant increased risk of not receiving graduation grades, presumably because they went to special schools, due to cognitive impairment or other special educational needs. One out of 44 newborns with an Apgar score of less than 7 at 5 minutes after birth will go to a special school because of the antenatal or perinatal factors that caused the low Apgar score. Nearly all school children who had Apgar scores of less than 7 at 5 minutes after birth showed an increased risk of graduating from compulsory school without graduation grades in a specific subject or receiving the lowest possible grades and were also less likely to receive the highest possible grade.

As the above mentioned studies demonstrate, it is important that each newborn obtains an Apgar score as high as possible.

### 3 The Dataset

The data made available for analysis came from Bega Clinique. The initial data set contained data for 2325 births that took place in this hospital in 2010. Data regarding the births have been stored in a Microsoft EXCEL spreadsheet file.

For each birth, the following data were available:

- ID number
- Month of birth
- Mother's name
- Mother's age
- City of residence
- Location- urban or rural
- Gesta – the number of the pregnancy
- Para – the number of children bore by the mother

- The number of gestation weeks
- Presentation - the position of the fetus when exiting the uterus, more precisely the part of its body that is going to emerge first. The possible values for this field are: cephalic, pelvic, facial and transversal. Cephalic presentation is the normal instance, in which the fetus has the spinal cord parallel to the mother's and the head down with the chin next to the chest. The pelvic presentation means that the fetus emerges with its feet or bottom in front. The facial presentation is when the fetus looks straight ahead and its face will come out first. If the fetus' spine is not parallel (the fetus has an oblique position in the belly) the presentation is transversal.
- The Apgar score - Immediately after birth, even in the first 60 second after expulsion, in the delivery room, an assessment of the newborn's health state is made, evaluating the vital functions and its capacity to adapt to the extra uterine environment. Simultaneously with providing the first nursing measures, the neonatologist will write down the clinical state and behavior of the newborn, quantifying the vital functions with the aid of the Apgar. The Apgar score has values between 0 and 10.
- The baby's gender
- The baby's weight
- If the birth was natural or through a cesarean operation
- Videx – column that indicates if the fetus was extracted using a metal device on its head in order to help the natural birth
- The reason for which a cesarean birth was indicated
- Episiotomy – column that indicates if the perineum was cut or not in order to help the natural birth (0 – it was not, 1 – it was)
- The number of labor hours
- EMP = Manual Extraction of the Placenta, the value 0 indicates that such an intervention was not done and the value 1 indicates that it was done.

## 4 Data Preprocessing

Preprocessing data was done through a series of actions:

- The columns unimportant for the study were eliminated, such as the id number, the name of the mother (also for privacy reasons), the city of

residence. The column month of birth was kept for statistical analysis, but was eliminated from the number of columns used to build the model

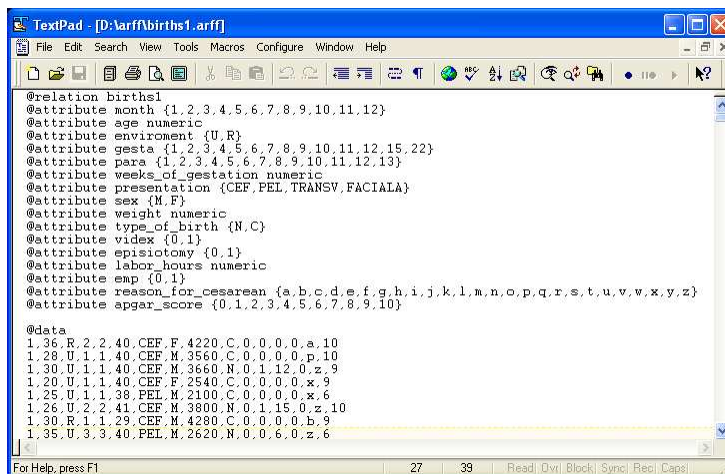
- The errors that could be rectified were corrected. For example:
  - The weight registered for two babies was of 34000g and 34450g. We presumed that a zero was added by mistake at the end of the real weight and we removed this zero.
  - The column hours of labour contained numeric data between 0 and 30, but for two persons 20 minutes had been inserted. The column had to be uniform so it either contains the number of hours of labour or of minutes of labour, so we transformed everything in hours and replaced the two values of 20 min with 0.33 hours
  - For three women that gave birth through caesarean there was no value filled in the episiotomy field. The value 0 (without episiotomy) was filled in because this intervention is specific for the natural birth.
- The instances that contain incorrect values that could not be corrected were eliminated. Those were the following columns:
  - The column number of gestation weeks usually contained the exact number of the gestation week (for example week 40) but for 178 persons, the exact number of the gestation week was not known so the filled in values were 37/38, 38/39, etc. The corresponding instances were deleted.
  - The columns person's age, gesta, para, weeks of gestation have the value zero for ten instances
  - The Apgar score– for other 4 newborns, instead of having a score in this column, the reason for which such a score could not be indicated was filled in –that is because two of them were born at home and the other two were born in the ambulance
- The instances that have missing values in different columns were removed. Columns with missing values were person's age, gesta, para, the Apgar score, sex, weight, type of the birth, videx, reason for caesarean, episiotomy, labour hours. There were 47 instances removed because of missing values
- The attribute recommendation for caesarean was transformed from a String attribute into a nominal attribute. In a first stage, 490 unique values of this attribute were identified with the help of a developed software instrument. Then, we searched for a solution to obtain a smaller number of nominal values. The solution that was found was to group the reasons for the caesarean recommendation. Analyzing data, a number of twenty-six groups of reasons were coded using the letters of the alphabet. Each of the 490 reasons for the caesarean is part of one of the 26 identified groups. For example, the cardio-vascular group contains cerebral aneurism, aneurism, varicose disease,

hypertension, cardiopathy, cardiac insufficiency, maternal cardiac pathology, preeclampsia, mitral valve prolaps, wpw syndrome, thrombophlebitis, bradycardia. In order to fill in the place of each reason, the code corresponding to the group it is part of, a software tool was developed. The identified groups and the letters used for codification are the ones below:

- a- obstetrical antecedents of the mother
- b- cardio-vascular
- c- pelvis conformation
- d- umbilical cord
- e- placenta disattachment
- f- cervix dystocia
- g- endocrine
- h- twin pregnancy
- i- amniotic liquid
- j- broken membranes
- k- neurologic
- l- ophthalmologic
- m- fetus particularities
- n- operatory particularities or accidents
- o- placenta previa
- p- fetus presentation
- q- primipara in age
- r- negative labour trial
- s- renal
- t- uterus rupture
- u- over the term pregnancy
- v- pregnancy with treatment
- w- serological
- x- height of the mother
- y- fetal suffering

- z- unrecommended caesarean (value filled in for the mothers that gave birth naturally)
- The columns were named according to their content. For example, in the file that we received, the letter G was used to name the column that indicates the number of pregnancies (gesta) as well as the column that indicates the weight of the newborn. The newly elected names of the two columns were gesta and weight.

The analysis of the EXCEL file was done using the Filter option. A filter was set for each column and every column was checked individually so that it contained only valid data. After eliminating the instances that contained missing or incorrect data, a number of 2086 valid instances remained (out of 2325). The next phase of preprocessing was to create the *ARFF* file, with the remaining valid data and loading it into *Weka*. The built file *births1.arff* is presented in Figure 1.



```

@relation births1
@attribute month {1,2,3,4,5,6,7,8,9,10,11,12}
@attribute age numeric
@attribute environment {U,R}
@attribute gesta {1,2,3,4,5,6,7,8,9,10,11,12,15,22}
@attribute para {1,2,3,4,5,6,7,8,9,10,11,12,13}
@attribute weeks_of_gestation numeric
@attribute presentation {CEF,PEL,TRANSV,FACIALA}
@attribute sex {M,F}
@attribute weight numeric
@attribute type_of_birth {N,C}
@attribute videx {0,1}
@attribute episiotomy {0,1}
@attribute labor_hours numeric
@attribute emp {0,1}
@attribute reason_for_caesarean {a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z}
@attribute apgar_score {0,1,2,3,4,5,6,7,8,9,10}

@data
1,36,R,2,2,40,CEF,F,4220,C,0,0,0,0,a,10
1,28,U,1,1,40,CEF,M,3560,C,0,0,0,0,p,10
1,30,U,1,1,40,CEF,M,3660,N,0,1,12,0,z,9
1,20,U,1,1,40,CEF,F,2940,C,0,0,0,0,x,6
1,25,U,1,1,38,PEL,M,2100,C,0,0,0,0,x,6
1,26,U,2,2,41,CEF,M,3800,N,0,1,15,0,z,10
1,30,R,1,1,29,CEF,M,4280,C,0,0,0,0,b,9
1,35,U,3,3,40,PEL,M,2620,N,0,0,6,0,z,6

```

Figure 1

The file *births1.arff*

## 5 Statistical Analysis

Next, the preprocessed data were loaded into *Weka* and with its help the values of each attribute were analyzed (see Figure 2). The number of instances is 2086.

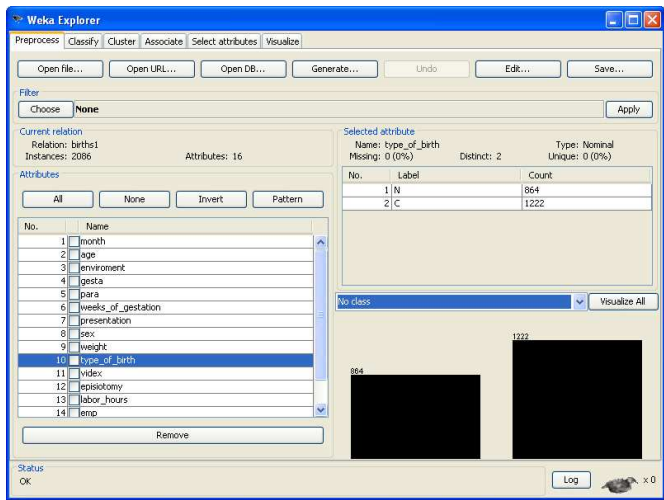


Figure 2  
Statistical analysis for each attribute with *Weka*

As it is shown in Figure 3, most births were registered in May, (217 births) which means that September is most frequent month to plan a child.

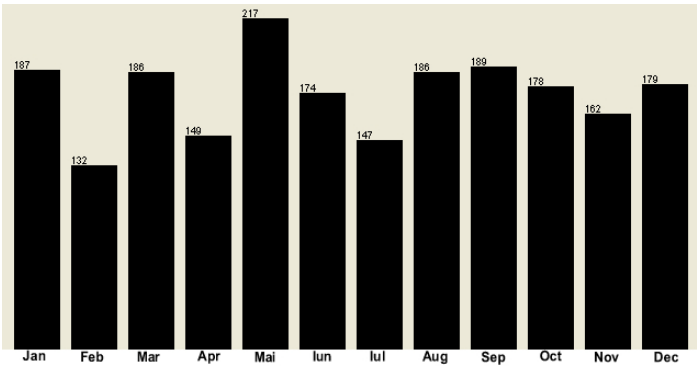


Figure 3  
Number of births for each month of the year

The age of the women who gave birth had been between 12 and 45. Most women become pregnant around the age of 28. 62 legally “minor” women became pregnant and 11 of them had the age between 12 and 15. 43 other women who were over 40 years old became pregnant.

Considering their place of residence, 1363 of the women that gave birth in 2010 at Bega Clinique come from an urban location and 723 come from the rural locations.

As far as the indicator *gesta* is concerned (the number of prior pregnancies) it was observed that the majority of women previously had one pregnancy (849 women) or two (596 women), but there are also extreme situations in which women had over 10 pregnancies (23 women).

The indicator *para* states that the great majority of women had their first (1195 women) or second birth (602 women). As usual, we also encounter extreme situations, 7 women that gave birth to over 10 children.

The number of gestation weeks is a numeric value between 19 and 42, and the average value is 38.467 with a standard deviation of 2.409.

The normal presentation, cephalic is the most common (1936 cases), followed by the pelvic one 144 cases, the transversal presentation with 5 cases, and the rarest is the facial presentation (one single case).

As far as the sex of the babies is concerned we see that more boys (1091) were born than girls (995) in Bega Clinique in 2010.

The average weight of the newborns is of 3194 grams, with a standard deviation of 578 grams, but there were 2 babies born with a weight over 5000 grams (see Figure 4).

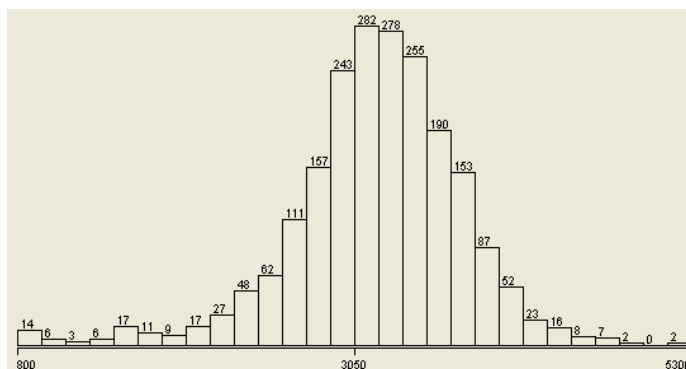


Figure 4

The weight of the newborns

Analyzing the number of women that give birth, we could say that natural birth was less recommended or desired by women than the birth through caesarean because in 2010, 864 women gave birth naturally (approximately 40%) and 1222 women through caesarean (approximately 60%).

The metal cap (*videx*) was used for only 18 of the 864 babies that were born naturally.

Episiotomy was necessary for 620 cases out of 864 (approximately 70%).



The number of labor hours is between zero and thirty. Usually zero hours of labour had been registered for births through caesarean.

The manual extraction of the placenta was done for only five of the women that gave birth naturally.

The most frequent reasons for caesarean recommendation are:

- The obstetric antecedents of the mother (scared uterus, double uterus, agglutinate cervix, uterus fibroma, etc.) 187 cases
- Negative labor trial, 154 cases
- Cardio-vascular reasons (preeclampsia, hypertension, cardiopathy, etc.) 132 cases

The grade given at birth is usually 10 (908 cases) or 9 (746 cases), but as it can be seen in Figure 5, the whole range of grades (from 0 to 10) was given to babies, including the very low ones.

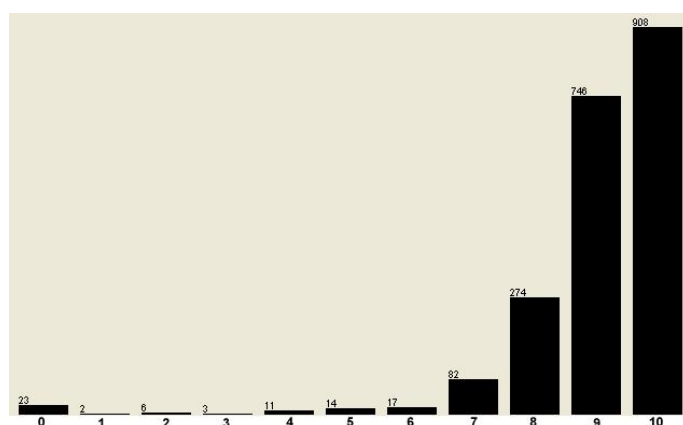


Figure 5  
The Apgar score

## 6 Data Classification

In the next phase, we wanted to build with *Weka*'s help, classification models on birth related data with an accuracy rate as high as possible. The first set of data for which the classification models were built is presented in Figure 1, with the amendment that the attribute month was eliminated among the input attributes, as it was considered irrelevant for the classification. The data set contains 2086 instances and numeric and nominal attributes. The classification models were built

using the *Naive Bayes*, *J48*, *k-Nearest-Neighbour*, *Random forest*, *Support Vector Machines*, *AdaBoost*, *LogitBoost*, *JRipp*, *REPTree* and *SimpleCart* algorithms. These models were tested through cross validation, using 10 folds. Testing with this technique implies dividing the dataset in 10 subsets. Each algorithm is run 10 times. For each run, a different subset is chosen for testing and 9 other subsets are chosen for training. One by one, each of the 10 subsets will be used for testing. The prediction accuracy of the algorithm is in fact the average of the accuracy of predictions obtained by testing the algorithm on each of the 10 subsets. Testing a subset is performed by estimating the Apgar score for all instances from that subset and then comparing the real score which is known, to the estimated one. The accuracy of the prediction represents the percent of the estimations that were correct. Because the classification models initially built had a very poor accuracy, we returned to the preprocessing phase several times, worked on the dataset and rebuilt and tested the classification models. Finally, the best results were obtained for the dataset in Figure 6.

```
@relation births5
@attribute age {'(-inf-18.6]','(18.6-25.2]','(25.2-31.8]','(31.8-38.4]','(38.4-inf)'}
@attribute gesta {'U,R'}
@attribute para {'(-inf-3.8]','(3.8-6.6]','(6.6-9.4]','(9.4-12.2]','(12.2-inf)'}
@attribute weeks_of_gestation {'(-inf-3.4]','(3.4-5.8]','(5.8-8.2]','(8.2-10.6]','(10.6-inf)'}
@attribute weeks_of_gestation {'(-inf-23.6]','(23.6-28.2]','(28.2-32.8]','(32.8-37.4]','(37.4-inf)'}
@attribute presentation {CEF, PEL, TRANSV, FACIALA}
@attribute sex {M, F}
@attribute weight {'(-inf-1700]','(1700-2600]','(2600-3500]','(3500-4400]','(4400-inf)'}
@attribute type_of_birth {N, C}
@attribute videx {0, 1}
@attribute episiokony {0, 1}
@attribute labor_hours {'(-inf-6]','(6-12]','(12-18]','(18-24]','(24-inf)'}
@attribute emp {0, 1}
@attribute reason_for_cesarean {a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z}
@attribute apgar_score {'(-inf-2]','(2-4]','(4-6]','(6-8]','(8-inf)'}

@data
'(31.8-38.4]','R','(-inf-3.8]','(-inf-3.4]','(37.4-inf)','','CEF,F','(3500-4400]','C
'(25.2-31.8]','U','(-inf-3.8]','(-inf-3.4]','(37.4-inf)','','CEF,M','(3500-4400]','C
'(25.2-31.8]','U','(-inf-3.8]','(-inf-3.4]','(37.4-inf)','','CEF,M','(3500-4400]','N
'(18.6-25.2]','U','(-inf-3.8]','(-inf-3.4]','(37.4-inf)','','CEF,F','(1700-2600]','C
```

Figure 6  
The file *births5.arff*

As it can be observed the numeric attributes *age*, *weight*, *weeks\_of\_gestation* and *labour\_hours* were discretized by distributing the values on 5 equal intervals, and the nominal attributes *gesta*, *para*, *weeks\_of\_gestation*, *Apgar\_score* were transformed in numeric attributes and discretized by dividing the values into 5 equal intervals. The new values of these attributes can be viewed in Figure 6.

The purpose for the discretization was to reduce the number of attribute values from the dataset, so that classification algorithms could build models with a higher accuracy.

The discretizations were realized using the *Filter* option from the *Preprocess* tab in *Weka* and the filter *weka.filters.unsupervised.attribute.Discretize*.

The results obtained by the classification algorithms on these data are presented in Table 1.

Table 1  
The accuracy of the prediction on the *births5.arff* data set

Algorithm	Accuracy %
Naive Bayes	78.57
J48	79.09
k-Nearest Neighbour	74.92
Random Forest	76.22
Support Vector Machines	79.62
AdaBoost	79.91
LogitBoost	80.24
JRipp	79.67
REPTree	79.62
SimpleCart	80.00

We can see that the best classification model obtained has a 80.24% accuracy and was built using the *LogitBoost* algorithm. Next, it will be used to make predictions.

We also investigated which is the attribute with the greatest influence on the Apgar score and we determined that it is the weight of the newborn with a correlation coefficient of 38%.

## 7 Predictions

With the aid of the built classification model we can estimate the interval of the newborn's Apgar score based on the input variables. The input variables are data regarding the mother (age, location, gesta, para, number of weeks of gestation, number of hours of labour, recommendation for caesarean), data regarding the baby (sex, weight) which can be determined through an ecography and data regarding possible medical interventions in order to help the birth (natural or through a cesarean operation, presentation, videx, episiotomy, manual extraction of the placenta). Practically, based on the historical data consisting of 2086 recorded births a classification model was built and it can be used to make different simulations of the interval in which the Apgar score of a newborn will be, for different values of the input variables. We can simulate in which interval the Apgar score will be for different medical interventions on the mother. For

example, we can simulate which will be the Apgar score for a mother if she gives birth naturally or through caesarean.

In order to make predictions we can use *Weka*, although the mechanism used to make predictions with *Weka* is not intuitive. It implies creating a new *ARFF* file in which the instance or instances to be predicted will be filled in. This *ARFF* file is loaded in *Weka* using *Supplied test set* command. Then the *Output predictions* option is checked. Finally, we make a right click on the model built with the *J48* algorithm and choose the *Reevaluate model on current test set* option. For each instance from the test set, the class that the user filled in the *ARFF* file will be displayed as well as the class predicted by the model [26].

The second way to make predictions is to use the *Weka* version that has a dynamic interface [27] with which we can easily make predictions (see Figure 7).

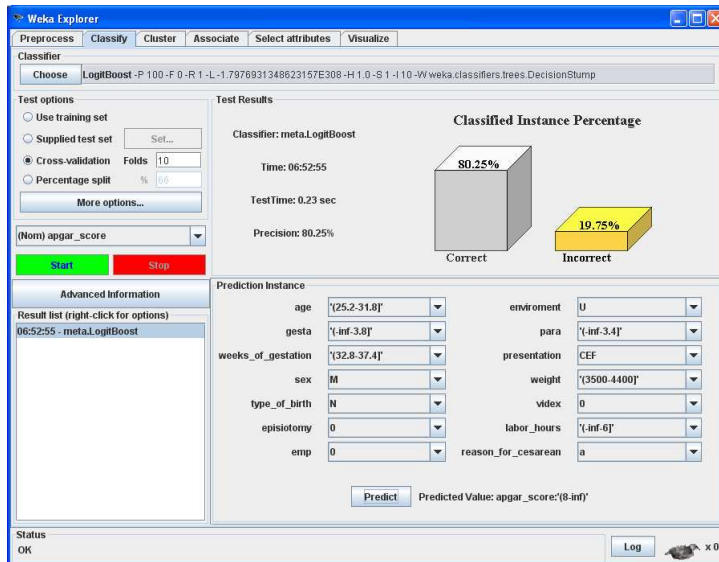


Figure 7

Making predictions with extended Weka

In this *Weka* version, after the classification model is built, the data for the instance that will be predicted can be filled in a dynamic interface, such as the one in the figure and then press the *Predict* button. The result of the prediction is displayed near the button, and the accuracy of the prediction is graphically displayed. The instance that will be predicted is introduced using *JTextField* components for the numeric attributes or *JComboBox* for nominal attributes.

Still, using the classical or extended versions of *Weka* to build classification models in order to make predictions is a little bit much considering that *Weka* offers a lot of facilities that are not used during this process. That is why it is

preferable to develop an application dedicated to classifying data regarding births, that builds the same classification model as *Weka*, using *LogitBoost* algorithm and the data regarding births from the *births5.arff* file. Such an application was developed using *Weka's API* and *Java* and is presented below. The application allows using the created model to make predictions.

## 8 Developing an Application Dedicated to Classifying Data regarding Births

In order to build the classification model using the *LogitBoost* algorithm based on the data set that contains 2086 instances regarding births from 2010 at the Bega Clinique and to make predictions with this model, an application using *Weka's API* was developed. The graphic interface was realized using *Swing* library. In Figure 8 we can see the tab *Model* that allows building the classification model using the *LogitBoost* algorithm.

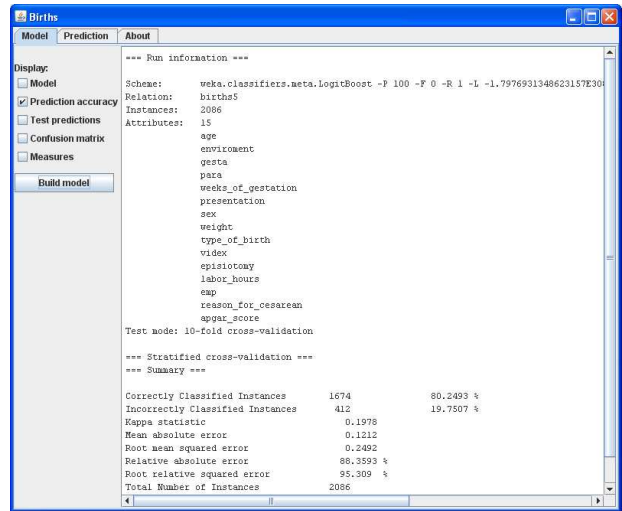


Figure 8  
Births tool - model user interface

The display of created model is a user option. Checking the *Test prediction* determines the display of the actual and predicted class for each of the instances tested through *Cross validation*. This option actually corresponds to the option *Output predictions* from *Weka*. We can also display the confusion matrix and *TP Rate*, *FP Rate*, *Precision*, *Recall*, *F-Measure*, *ROC Area*, *Class* measurements by checking *Confusion Matrix* and *Measures*. Practically this tab *Model* corresponds to the *Classify* tab from *Weka* and displays the same information.

In Figure 9 the Prediction tab can be visualized. All variables from the *births5.arff* file are nominal, that is why, in the interface, there are *JComboBox* type components that permit choosing one of the nominal values for each attribute. We introduce the instance to be predicted in the interface and press the button *Predict* and the prediction made by the classification model is displayed near that *Predict* button. On the right side, there is a graph that shows the accuracy of the Prediction obtained by testing the model through Cross validation.

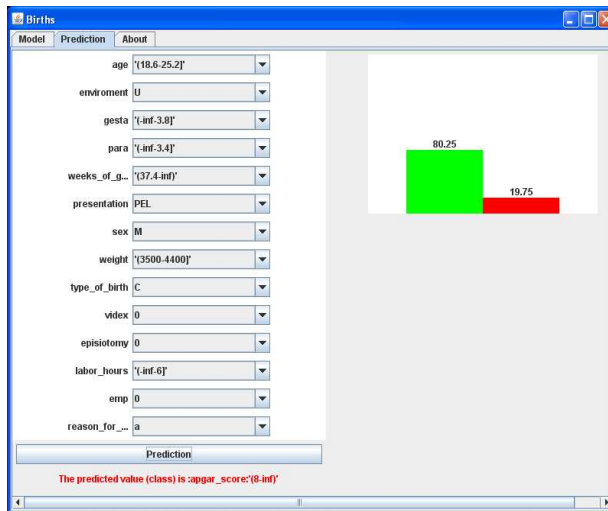


Figure 9

Births tool – prediction user interface

The most important classes from *Weka's API* which have been used while developing this tool are: *Instance*, *Instances*, *Classifier*, *LogitBoost* and *Evaluation*.

The classification model was build using the *buildClassifier()* method that gets as a input parameter the instances that represent the training data set.

```
Classifier classifier = new LogitBoost ();
classifier.buildClassifier(inst);
```

Classifying an instance was done with the help of the *classifyInstance()* method from the *Classifier* class, but it can also be performed with the help of the *distributionForInstance()* method from the same class. The instance to be classified is transmitted as an input parameter to one of the two methods. The *distributionForInstance()* method returns a *double* type vector with a number of elements equal to the number of values of the class attribute. Each element from the vector represents the probability with which the instance belongs to the corresponding class. The instance is classified as belonging to the class that has the highest probability.

The evaluation of the classifier was done with the aid of the *evaluateModelOnceAndRecordPrediction( classifier, test\_instance)* method from the *Evaluation* class.

## Conclusions

The paper presents an analysis on the data regarding the births that took place in the Bega Clinique. The initial data set contained 2325 records and after preprocessing, 2086 instances remained. Preprocessing consisted of eliminating the attributes considered irrelevant for the study, eliminating instances that contained missing or incorrect values, rectifying, if possible, the data that were incorrectly filled in, etc. Next, a statistical analysis for each attribute was made and it revealed some interesting information, such as the fact that approximately 60% of the women that gave birth in this hospital underwent a cesarean operation. The following step was building the model, with the help of ten classification algorithms from *Weka* classifiers that allow an estimation of the interval for the value of the Apgar score based on data from the mother, newborn and medical interventions. Finally, a dedicated tool was developed in *Java*, using *Weka API* which builds a classification model with the help of the *LogitBoost* algorithm for the data set regarding births. The application performs predictions with the aid of the built classification model. This tool helps make different simulations of the Apgar score for different values of the input variables with the purpose to choose the input variables so that the Apgar score is optimal.

## Acknowledgment

This work was partially supported by the strategic grant POSDRU/159/1.5/S/137070 (2014) of the Ministry of National Education, Romania, co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007-2013.

## References

- [1] A. Kusiak, J. Kern, K. Kernstine and B. Tseng, Autonomous Decision-Making: A Data Mining Approach, *IEEE Transactions on Information Technology in Biomedicine*, pp. 274-284, 2000
- [2] L. Yi-Chih, L. Yang-Chu and L. Tian-Shyug, Mining the Complication Pattern of Gastric Cancer Patients by Using Artificial Neural Networks and Logistic Regression, *The Journal of Human Resource and Adult Learning*, pp. 151-155, November 2006
- [3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten, The WEKA Data Mining Software: An Update, *ACM SIGKDD Explorations Newsletter* (2009), Volume 11, Issue 1, pp. 10-18
- [4] G. H. John and P. Langley, Estimating Continuous Distributions in Bayesian Classifiers, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338-345, 1995

- [5] J. R. Quinlan, C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers, Inc.*, pp. 235-240, 1993
- [6] D. Aha and D. Kibler, Instance-based Learning Algorithms, *Machine Learning*, Volume 6, pp. 37-66, 1991
- [7] L. Breiman, Random Forests, *Machine Learning* 45(1), pp. 5-32, 2001
- [8] C. C. Chang and C. J. Lin, LIBSVM: a Library for Support Vector Machines, *ACM Transactions on Intelligent Systems and Technology*, 2, No 3, 2011
- [9] Y. Freund and R. E. Schapire, Experiments with a New Boosting Algorithm, *In: Thirteenth International Conference on Machine Learning*, San Francisco, pp. 148-156, 1996
- [10] J. Friedman, T. Hastie and R. Tibshirani, Additive Logistic Regression: a Statistical View of Boosting (with discussion and a rejoinder by the authors), *The annals of statistics*, 28, No. 2, pp. 337-407, 2000
- [11] W. W. Cohen, Fast Effective Rule Induction, *In: Twelfth International Conference on Machine Learning*, pp. 115-123, 1995
- [12] T. Elomaa and M. Kaariainen, An Analysis of Reduced Error Pruning, *Journal of Artificial Intelligence Research*, Volume 15, pp. 163-187, 2001
- [13] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification and Regression Trees, *Wadsworth International Group*, 1984
- [14] K. Srijiaranon and N. Eiamkanitchat, Application of Neuro-Fuzzy Approaches to Recognition and Classification of Infant Cry, *In TENCON 2014-2014 IEEE Region 10 Conference*, pp. 1-6, 2014
- [15] A. Chittora and H. A. Patil, Classification of Pathological Infant Cries using Modulation Spectrogram Features, *In Chinese Spoken Language Processing (ISCSLP), 2014 9<sup>th</sup> International Symposium on*, pp. 541-545, 2014
- [16] V. Apgar, A Proposal for a New Method of Evaluation of the Newborn, *Curr Res Anaesth*, 32, pp. 260-267, 1953
- [17] V. Naeser, N. Kahr, L. G. Stensballe, K. O. Kyvik, A. Skytthe, V. Backer, C. G. Carson and S. F. Thomsen, Apgar Score Is Related to Development of Atopic Dermatitis: Cotwin Control Study, *Journal of Allergy*, pp. 1-6, 2013
- [18] M. Finster and M. Wood, The Apgar Score has Survived the Test of Time, *Anesthesiology*, 102, No. 4, pp. 855-857, 2005



- [19] B. M. Casey, D. D. McIntire and K. J. Leveno, The Continuing Value of the Apgar Score for the Assessment of Newborn Infants, *New England Journal of Medicine*, 344, No. 7, pp. 467-471, 2001
- [20] R. Sahak, W. Mansor, L.Y. Khuan, A. Zabidi and F. Yasmin, An Investigation into Infant Cry and Apgar Score using Principle Component Analysis, *Signal Processing & Its Applications, CSPA 2009. 5<sup>th</sup> International Colloquium on*, pp. 209-214, 2009
- [21] L. C. Gilstrap, K. J. Leveno, J. Burris, M. L. Williams and B. B. Little, Diagnosis of Birth Asphyxia on the Basis of Fetal pH, Apgar Score, and Newborn Cerebral Dysfunction, *American journal of obstetrics and gynecology*, 161, No. 3, pp. 825-830, 1989
- [22] K. Khan and A. R. Saha, A Study on the Correlation between Cord Blood Glucose Level and the Apgar Score, *Journal of clinical and diagnostic research*, Volume 7, Issue 2, pp. 308-11, 2013
- [23] L. Sekhavat and R. Fallah, Could Maternal Pre-Pregnancy Body Mass Index Affect Apgar Score?, *Archives of gynecology and obstetrics*, 287, No. 1, pp. 15-18, 2013
- [24] V. Ehrenstein, L. Pedersen, M. Grijota, G. L. Nielsen, K. J. Rothman, and H. T. Sørensen, Association of Apgar Score at Five Minutes with Long-Term Neurologic Disability and Cognitive Function in a Prevalence Study of Danish Conscripts, *BMC Pregnancy and childbirth*, 9, No. 1, 2009
- [25] A. Stuart, P. O. Olausson and K. Källen, Apgar Scores at 5 Minutes after Birth in Relation to School Performance at 16 Years of Age, *Obstetrics & Gynecology* 118, No. 2, Part 1, pp. 201-208, 2011
- [26] D. Rodríguez, Making Predictions on New Data using Weka, Available at: <http://www.cc.uah.es/drg/courses/datamining/ClassifyingNewDataWeka.pdf>, Accessed: 2014.09.15
- [27] R. Robu and C. Hora, Medical Data Mining with Extended Weka, *Proceedings of the 16<sup>th</sup> IEEE International Conference on Intelligent Engineering Systems (INES)*, pp. 347-350, 2012

# An Ontology Model-based Minnesota Code

**Norbert Sram, Márta Takács**

Óbuda University

Bécsi út 96/b, H-1034 Budapest, Hungary

E-mail: sramm.norbert@phd.uni-obuda.hu; takacs.marta@nik.uni-obuda.hu

---

*Abstract: In this paper, the authors present an approach towards modeling a classical expert system using an ontology-based solution. The aim was to have an extensible setup, where multiple reasoning methods can be used, to provide the desired outcome. The case study, is a hierarchical rule-based system, for the evaluation of reference ECG signals called the Minnesota Code. This paper describes the practical limitations of the original expert system based definition, of the Minnesota code and describes an approach to represent it as an ontology that provides support for various reasoning methods. The authors present here, a possible solution to use the ontology model and ontology reasoning to provide a diagnostic evaluation of ECG information added to the Minnesota code ontology that corresponds to the rules defined by the expert system based solution.*

*Keywords: ontology; expert system; Minnesota Code*

---

## 1 Introduction

Cardiovascular diseases are some of the most common causes of death. Based on the World Health Organization reports [9], about 30% of deaths are caused by either Ischemic Heart Disease or stroke. The prediction of sudden cardiac deaths, is still a concern and mostly unsolved [6, 7]. It is now well-recognized that classifications based on clinical circumstances can be misleading and often impossible, because 40% of sudden deaths may be unwitnessed [8]. There are examples of systems providing medical assistance to experts regarding diagnostics, using different models, based on expert systems and extended with novel mathematical models [16]. One of the authors has been involved in a research project which aims to shed further light on these cases by introducing telemedicine systems to alleviate the situation [10]. The authors of the paper suggest taking this one step further and introduce a proven diagnostic classification system on top of the monitoring system which could identify possible cardiovascular diseases.

## 2 The Minnesota Code

The diagnostic classification system chosen by the authors is the Minnesota Code algorithm. The Minnesota Code [5] is a classification system for the electrocardiogram that utilizes a defined set of measurement rules to assign specific numerical codes according to severity of the ECG (Electrocardiography) findings. It is the most widely used ECG classification system in the world for clinical trials and epidemiological studies. It incorporates ECG classification criteria that have been validated, widely employed, and accepted by clinicians. From the definition's point of view, the Minnesota Code is a structured list of rules that examines certain characteristics of ECG waveforms. The Minnesota Code combines three major elements: a set of measurement rules, a classification system for reporting ECG findings and a set of exclusion rules. The relationship between the three major sets is vaguely defined.

### 2.1 General Overview of the Diagnostic System

In order to be able to provide a complete diagnostic result with the Minnesota code, it is required to have a 12-lead ECG and the corresponding various parameters of the ECG. This means approximately 55 parameters for each lead. In practice, providing all these parameters can pose significant problems. This is one of the weaknesses of the Minnesota code. Because of the numerous input parameters, the dependencies between various diagnostic rules, the rigid value definition used by the diagnostic rules, the results of the Minnesota code are sensitive to the precision of the input information. By introducing partial processing and measurement error toleration the usability of the diagnostic system could be greatly improved. In the case of an insufficient input dataset, the diagnostic system can still fall back to partial processing. In addition, measurement errors need to be taken into consideration.

The Minnesota code organizes the diagnostic rules based on ECG leads and waveform types. Each diagnostic rule has a unique identification (for example 1.1.1) and in some cases they are referred by multiple ECG lead groups. To resolve a diagnostic rule, it is required to take into consideration all occurrences of that rule. The output of a diagnostic rule equals the aggregation of the results of all evaluated occurrences of the specific rule. The Minnesota code definition states that the aggregation is done with the logical 'and' operator.

### 2.2 Practical Usage of the Diagnostic System

The input of the diagnostic system is an ECG cycle (heartbeat cardiac cycle) and its corresponding waveforms (P, Q, R, S, T) for all the 12 ECG leads (I, II, III, V1, V2, V3, V4, V5, V6, aVR, aVL, aVF). The waveforms are the visually

identifiable parts of the ECG signal, that are used to characterize the ECG signal based on the waveform properties such as duration, amplitude, etc. The process of identifying the waveforms in a cardiac cycle is called annotation. The annotation of the signal can be done by hand or in an automated way using various algorithms [15]. Figure 1 displays a single ECG cycle with its corresponding waveforms annotated. From the point of evaluating the diagnostic rules, this is not relevant. The Minnesota Code deduces the diagnostic output based on annotated ECG signals. In practice this means that the recorded ECG signal is coupled with the required annotation values and this composition is forwarded to the diagnostic system. The output of the system is the state of the Minnesota codes for the examined ECG input. In detail, this means that the diagnostic system examines whether an input ECG cycle and annotations meet the requirement of one or more diagnostic rules (for example the rule 1.1.1). Since the goal is to provide partial diagnostic support besides providing the "true" and "false" states, the missing rule states convey information as well. If a diagnostic rule code is not present in the output, it means that the required information for evaluating that diagnostic rule was not available.

The length of an ECG recording varies in practice, from a few seconds up to several hours. In the case of the ECG databases used herein, the length of the recording varies. In order to avoid the redundant and extreme cases, the ECG recordings are preprocessed before the diagnostic steps. The preprocessing covers the identification of the "typical beats" for ECG records that contain multiple full ECG cycles. The output of the preprocessing for these cases is 1-3 ECG cycles (Figure 1, displays a single cycle) which correspond to the "average" ECG samples of the recording.

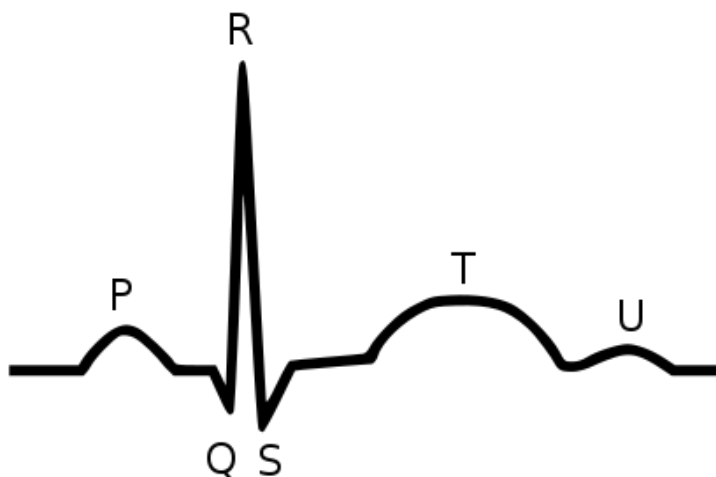


Figure 1

A cardiac cycle (heartbeat) with the corresponding waveforms highlighted on the signal [12]

Measurement rules can have various properties and definitions, but most follow a common format. As an example, the diagnostic rule identified as 1-1-1 has the following definition and is made up of multiple statements (one for each ECG lead group):

*Q/R amplitude ratio  $\geq 1/3$  and Q duration  $\geq 0.03$  sec in lead I or V6*

*Q/R amplitude ratio  $\geq 1/3$  and Q duration  $\geq 0.03$  sec in lead II*

*Q/R amplitude ratio  $\geq 1/3$  and Q duration  $\geq 0.03$  sec in any lead V2, V3, V4, V5*

The rule 1-1-1 is true if one of the statements is fulfilled.

The exclusion rules define which diagnostic rules need to be ignored once a specific exclusion rule's requirements are met. Not all diagnostic rules can be ignored through exclusion rules. The Minnesota Code [5] classification system defines a table that contains the exclusion rules. Table 1, displays a subset of the exclusion rule definitions.

Table 1  
Example of the Minnesota code incompatible rule definitions

<i>Code</i>	<i>Suppress this code(s)</i>
<i>All Q-, QS-codes</i>	<i>7-6</i>
<i>Q &gt; 0.03 in lead I</i>	<i>7-7</i>
<i>3-1</i>	<i>1-3-2</i>
<i>6-8</i>	<i>All other codes</i>

The classification is done through diagnostic rules, by identifying baseline ECGs that are used to categorize a study population into groups based on major and minor abnormalities. The classification definitions are similar to the exclusion rule definitions in the sense that each ECG baseline is associated with a set of diagnostic rules. Table 2 shows an entry of the categorization definitions.

Table 2  
Classification code for prevalent ECG abnormalities

<i>ECG Categories Associated With Myocardial Infarction / Ischemia</i>		
<i>Definition and Description</i>		<i>Minnesota Code</i>
<i>Q wave MI</i>	<i>Q wave MI; major Q waves with or without ST-T abnormalities</i>	<i>1-1-x</i>
	<i>Q wave MI; moderate Q waves with ST-T abnormalities</i>	<i>1-1-1 plus 4.1, 4.2</i>
<i>Isolated minor Q and ST- T abnormalities</i>	<i>Minor Q waves without ST - T abnormalities</i>	<i>1-3-x</i>
	<i>Minor ST-T abnormalities</i>	<i>4-3, 4-4, 5-3, 5-4</i>

The classification table is used to provide the diagnostic output. As Table 2 shows, the classification table is grouped based on ECG categories, where each row is a specific entry. The Minnesota Code column of Table 2F contains the requirements for the entries. Based on this, it is possible to produce a diagnostic result using the diagnostic rules that met their requirements.

Several studies have shown that the effectiveness of the computer-based Minnesota Code evaluation of ECG signal can be as effective as humans are with visual analysis [1]. There were studies that showed that the precision of the Minnesota Code system can be improved through the application of Fuzzy Logic [2]. These facts led to the decision to choose the Minnesota Code as the knowledge base.

Based on previous experiences gained in the modification and improvements of the Minnesota Code, the diagnostic rule set does not address the case of missing or incomplete inputs which, in turn, leads to insufficient diagnostic results for practical cases [4]. In order to incorporate improvements into the whole diagnostic process and not just a specific group [2], the definition and structure of the Minnesota Code needs to be more robust. There are efforts for providing a coherent and robust system for medical applications and providing a semantic representation of the patient data [11]. The Minnesota code-related research has not yet addressed the topic of semantically defining the diagnostic rules and dependencies. In order to establish a robust and extensible representation of the Minnesota Code-based diagnostic process, the authors needed to define a suitable representation. For this, they chose to follow the existing, general principles of semantic medical coding systems [11] and create an Ontology model for the Minnesota Code, which is described in detail in this paper.

### 3 Ontology

In order to restructure the representation of the Minnesota Code, there are multiple requirements that need to be taken into consideration, that are not provided by the current definition and its structured representation:

- The possibility to clearly identify the relationship between different elements
- The grouping of input states (different rules referring to the same crisp values)
- Extensibility without compromising the original definition
- The possibility to provide a partial diagnostic output.

An ontology model-based representation provides the possibility to clearly define the relationships between diagnostic rules and inputs using axioms. Extensions to the diagnostic model are possible by introducing new concepts or axioms between concepts. The evaluation is also feasible by using an ontology reasoner.

In order to provide partial diagnostic support, the usage of an inference system is required. The output of the desired inference system, would be made up of all the Minnesota code rules that have the required input at their disposal. For supporting partial diagnostic, the system needs to be reduced to its building blocks. These building blocks are used to create a customized diagnostic system, which can be evaluated based on the available input data. The relationship between the building blocks, diagnostic rules and inputs is defined by an ontology.

Ontology is a powerful knowledge representation formalism for modeling real-world concepts, basic mechanism and relationships used in different fields from semantic web modeling for the annotation of life events, goals, sub-goals, services, and other specific concepts from the public administration domain [7]. An ontology[2] (O) organizes domain knowledge in terms of concepts (C), properties (P), relations (R) and axioms (A), and can be formally defined as a 4-tuple  $O = (C, P, R, A)$ , where: C is a set of concepts defined for the domain. A concept is often considered as a class in ontology. P is a set of concept properties. A property  $p \in P$  is defined as an instance of a ternary relation of the form  $p(c, v, f)$ , where  $c \in C$  is an ontology concept,  $v$  is a property value associated with  $c$  and  $f$  defines restriction facets on  $v$ . R is a set of is a set of binary semantic relations defined between concepts in C.  $R_t = \{\text{one-to-one, one-to-many, many-to-many}\}$  is e set of relation type. A is a set of axioms. An axiom is a real fact or reasoning rule.

## 4 Ontology-based Approach for the Minnesota Code

### 4.1 Analysis of the Minnesota Code System

The diagnostic rules are the core and defining elements of the Minnesota code. This makes the diagnostic rules the starting point of analysis for constructing an ontology-based model. In order to provide an insight into how the diagnostic rules fit into the Minnesota code and provide the diagnostic process we will provide a detailed explanation for a specific use case. The "Q and QS Patterns" group will be used as a case study. This is the group for the rules that work with the Q and QS waveform patterns. This group can be further divided into three subgroups based on the ECG leads, meaning the Anterolateral site (I, aVL, V6), the Posterior site (II, III, aVF) and the Anterior site (V1, V2, V3, V4, V5). It can be observed that the same rule identification can occur under multiple subgroups. The difference in the rule occurrences is the ECG lead the diagnostic rules process. An example for this scenario is the rule 1-1-1. There are diagnostic rule definitions that are only present in the case of a specific sub-group (for example, the rule 1-1-3 is only present in the Anterolateral site [5]).

Rule Identifier	Group	Condition
Rule 1-1-1	Anterolateral site (leads I, aVL, V6)	Q/R amplitude ratio $\geq 1/3$ , plus Q duration $\geq 0.03$ sec in lead I or V6.
	Posterior (inferior) site (leads II, III, aVF)	Q/R amplitude ratio $\geq 1/3$ , plus Q duration $\geq 0.03$ sec in lead II.
	Anterior site (leads V1, V2, V3, V4, V5)	Q/R amplitude ratio $\geq 1/3$ plus Q duration $\geq 0.03$ sec in any of leads V2, V3, V4, V5.

Figure 2

The definition of the diagnostic rule 1-1-1

As seen in Figure 2, the definition of the diagnostic rule 1-1-1 is split into three parts. From the definition of the diagnostic rule it can be identified that two inputs are examined, the Q/R amplitude ratio and the length of the Q waveform. In the diagnostic rule the input values are compared to predefined crisp values. In the case of the Q/R amplitude ratio this would be greater than  $1/3$ , while in the case of the Q waveform length it would be greater than 0.03s.

The identification of the components that make up the system is the first step towards the construction of the ontology. In order to do this, the first step is to create a structured representation of the diagnostic rules. Representing the diagnostic rules as trees is one possible solution as shown in Figure 3. Figure 3 displays a generic tree representation of a Minnesota code diagnostic rule, where each diagnostic rule is an aggregation of predicate. The predicate branches represent the requirements as defined by a diagnostic rule. Each rule can contain one or more predicates. The predicate branch contains the ECG property that needs to fulfill the requirement as specified by the value node. A concrete example is shown in Figure 4, which is the tree structure based representation of the diagnostic rule 1-1-1 (original definition shown in Figure 2). The goal of this representation is to provide a method for identifying the core components and values of the Minnesota code. This is done through querying all ECG property nodes to identify the input types. Using the identified input nodes, one can query for all criteria values that belong to a specific input node type. The query results of the criteria values for a specific input property led to the conclusion that each input property is compared to a specific entry from a set of waveform states and never compared to a computed or dynamic value. As an example, let us consider



the case of the Q waveform length, where the diagnostic rules examine various states, such as:

- Value that is in the range of 0.02s and 0.03s
- Value that is in the range of 0.03s and 0.04s
- Value that is greater than 0.04s
- Value that is greater than 0.05s

After conducting the same analysis for the other diagnostic rules belonging to the "Q and QS Pattern" group, it can be concluded that the Q waveform length has six possible states in the various rules. As the presented example shows, the studied parameter can be categorized into various states that have a strict definition. It can be concluded that the inputs, waveform states and diagnostic rules are the main concepts that make up the Minnesota code.

The Minnesota code does not define the waveform states. In order to gather all possible waveform states used in the diagnostic rules, an in-depth analysis of all diagnostic rules is needed. The in-depth analysis constitutes of the construction of the tree structure based representation for all diagnostic rules. Figure 3 provides a general overview of a tree structure representing a diagnostic rule which can have one or more predicates. The input types are represented by the ECG property nodes. By filtering out all unique instances we are able to acquire the input types. The possible waveform states are identifiable by paring the ECG property node occurrences (input types) with the value nodes (Figure 3 shows the relationship between the two node types). The diagnostic rules are clearly defined by the Minnesota code, coupled with the tree structure-based representation one has the necessary information for designing the ontology model.

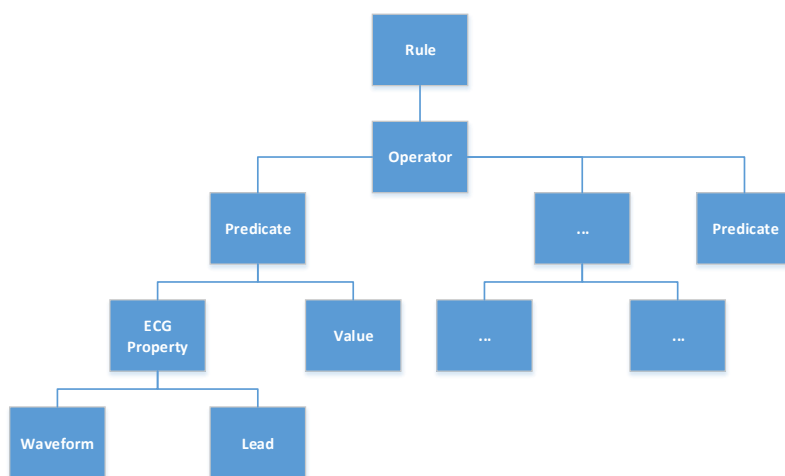


Figure 3

General tree structure based representation of diagnostic rules

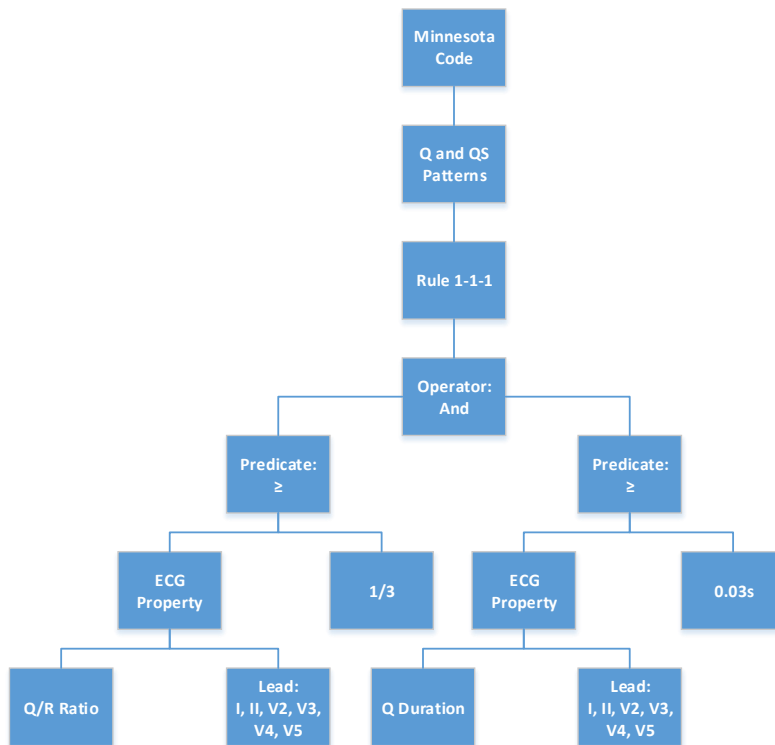


Figure 4

Tree structure-based definition of the Minnesota Code rule 1-1-1

## 4.2 The Ontology Model

The modeling of the ontology is done with the help of the Protégé [13] application. It is used for prototyping the ontology model and reviewing the results of the ontology inference systems. The inference system used by the authors is the Hermit reasoner [14].

The design of the ontology model starts with the identification of the building blocks of the Minnesota code. These building blocks correspond to the ontology concept representation of the elements identified by the analysis of Minnesota rules. In order to achieve a robust solution it is mandatory to have an ontology concept representation for even the smallest elements. The diagnostic inference is made possible by employing the correct relationships between these elements. Based on the original definitions, the elements used for constructing the rules are the waveforms, ECG leads, waveform value states and the grouping concepts. This means that the ontology model of the Minnesota code is divided into 4 main groups: Sample, ECG leads, Waveforms and Rules.

The first group only contains a single concept, the “Sample”. All input samples are an instance of the “Sample” concept. Various axioms are bound to the instances of the “Sample” concept to provide diagnostic inference support.

The second group is made up by the ECG leads and the corresponding instances. The ECG leads have 3 different sub-concepts (subclasses). These are the “AnteriorSite”, “AnterolateralSite” and the “PosteriorSite”. Since the ECG leads are globally used identifiers, each ECG lead group has a set of predefined instances for each ECG lead belonging to the specific sub-concept group. The ECG concepts and instances are used to describe the diagnostic rules in the ontology. Figure 5, shows the ontology model of the ECG leads and Figure 6, shows the “PosteriorSite” sub-concept definition in Protégé.

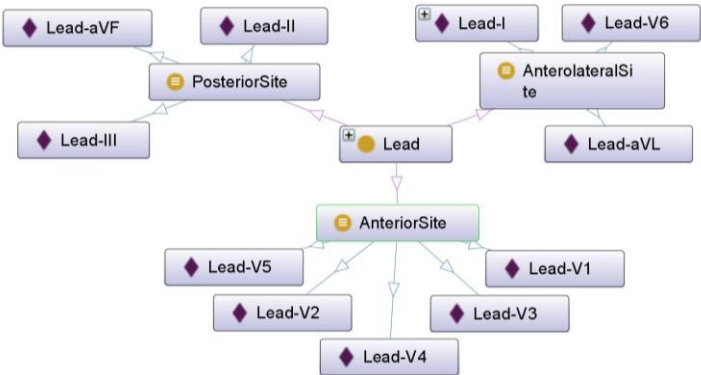


Figure 5  
Graph of the Ontology model for the ECG leads created with Protégé [13]

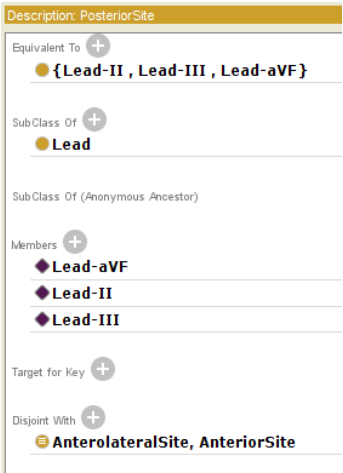


Figure 6  
Ontology model of a specific lead group in Protégé

The third major concept used in the ontology is the “Waveform” concept. As its name implies, this concept is used to represent the waveform characteristics. Every concept in the ontology that is a sub-concept of the “Waveform” concept, marks a parameter of the diagnostic system. The sub-concept tree is made up using the results of the conducted analysis of the Minnesota code rules. This means that all waveforms that were referenced at least once by one of the diagnostic rules. The “Waveform” concept has approximately 17 sub-concepts. The major waveforms concepts are shown in Figure 7. The sub-concepts of these groups contain the state concepts for each waveform. The ontology concepts representing the specific waveform states have been designed based on the analysis results of the rules. Each waveform state concept uniquely identifies a waveform type and crisp value/range found in one of the Minnesota code diagnostic rules. The state concepts are unique to the waveform type that they belong to. A different waveform type can have a state that matches the same crisp state. In the ontology model, it is important to differentiate between different waveform types regardless of their crisp states, in order to support the inference system.

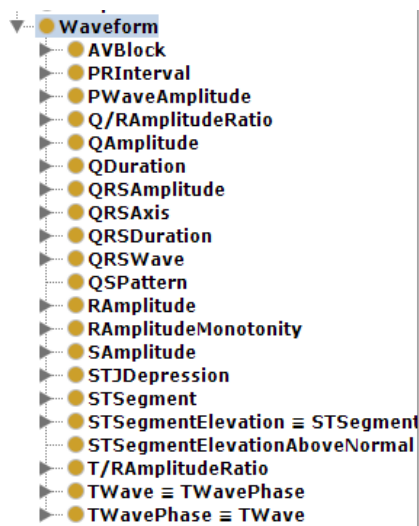


Figure 7  
Waveform groups

The major functionality of the waveform grouping concepts is to identify the waveform type of a specific waveform state provides based on the original Minnesota code definitions. Every waveform grouping concept contains one or more possible state concepts. Figure 8, shows the Q waveform length ontology based state concepts. A distinctive property of the waveform state concepts is a data attribute that captures their state value. For example, the “QDurationLong” concept defines that the value of the instance is greater than or equal to 0.05 ms.



Figure 8

Ontology model of the Q waveform state

The fourth and last major group contains the ontology representation of the diagnostic rules, which are all sub-class (sub-concepts) of the “Rule” concept. The direct descendants correspond to the rule groups defined by the Minnesota code. The diagnostic rules defined by the Minnesota code are modeled as equivalent classes in the ontology (Figure 9). This means that the instances for these concepts are identified based on a specific set of properties/attributes being present for an ontology instance. Diagnostic rule instances are not created or classified directly in the ontology model. This approach is relevant for establishing the diagnostic output. In ontologies, compared to regular classes (concepts) equivalent classes represent a two-way relationship. What this means in practice is that if an ontology instance has the necessary properties, it can be categorized as an instance of an equivalence class, regardless of the existing and established hierarchical relationship. Every diagnostic rule is modeled as a single equivalence class, i.e. a one-to-one mapping compared to the original Minnesota rules. In a scenario, where a set of requirements defined by one of the equivalence classes are met by an instance, the ontology will mark the instance as a type of that specific diagnostic rule class. This method is applied for identifying all the diagnostic rules applicable for a specific ECG sample. Figure 9, shows the structure of diagnostic rule 1-1-1 in the ontology model. The representation is close to the original definitions. The difference is in the application of concepts to represent waveform states. The ontology representation could have applied the same approach, where rules would refer to crisp values. This approach was abandoned in favor of concepts providing the possibility to use various representations for waveform values (fuzzy, crisp, interval-based etc.). In the ontology, every measured value is an instance of a concept. The waveform instances can have various axioms in addition to the measured value. For example, one can extend a waveform instance to have a fuzzy definition.

```

● (hasLead some ({Lead-I , Lead-II , Lead-V2 , Lead-V3 , Lead-V4 , Lead-V5 , Lead-V6}))
  and (hasWaveform some Q/RamplitudeRatioUpperThreshold)
  and (hasWaveform some QDurationShort)
  
```

Figure 9

Ontology representation of Minnesota code diagnostic rule 1-1-1

Inside the ontology the various types, properties and values are connected together with axioms. In the case of the diagnostic system, a predefined set of axioms is

used. For example, an axiom is used to connect the measurement value to the “Waveform” concept instance. This axiom is a function that operates on “Waveform” concepts and floating point values. The described axiom is called “hasCrispValue”. The exact functionality of the “hasCrispValue” is to define the measurement/input value of an ECG waveform. The “hasCrispValue” and similar axioms such as “hasWaveform”, “hasLead” are used by the rule equivalence class definitions. These axioms operate on a specific ontology instances and connect the various ECG properties to that instance. In the case of the “hasWaveform” axiom, the operands always have the type of Waveform and Sample, while the “hasLead” axiom operates on a Waveform instance and a Lead instance.

## 5 Ontology-based Diagnostics

The first step towards providing a diagnostic output is the population of the Minnesota code ontology system. One by one, all the available ECG samples are added to the ontology. For every sample, the same algorithm is executed. The first step of the algorithm is to create the ontology representation for the input, meaning an instance from the “Sample” concept. After creating the instance, the algorithm sets the known properties. The algorithm does this by creating the necessary connections in the ontology, using various axioms (“hasLead”, “hasWaveform”, “hasCrispValue”). ECG leads and the corresponding waveform values are attached to the ontology instance. Since axioms operate on ontology instances, the available waveform properties need to be created. Inputs belong to a specific “Waveform” concept. Waveforms are grouped by type, all groups contain the concepts representing the specific states for a particular waveform. To find the appropriate waveform type concept for an input value, the diagnostic system uses an inference system to acquire all the specific sub-concepts (waveform states) for a waveform type. Figure 10 illustrates a result of the ontology populating step in Protegé. The Figure displays an instance of “Sample” called “testSample” and the corresponding axioms for the “Waveform” properties belonging to the “testSample”.

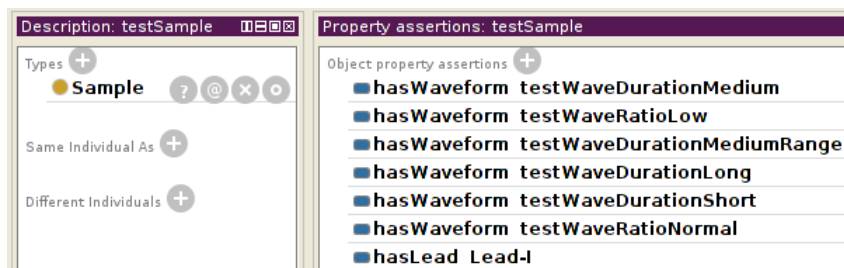


Figure 10

Definition of a “Sample” instance and the corresponding waveforms in Protegé

The result of the inference system is a list of concepts. For all the acquired concepts, the diagnostic algorithm creates an instance with the type and measured value set. The same measured value is set for all the state instances. As an example, a case study of Q waveform length is conducted. The Q waveform length is used by the Minnesota code. In the case of this waveform type (Q) and its parameter (length), the Minnesota code differentiates between 6 states (sets). The ontology system handles the mapping for the specific states. This step is required since in the original Minnesota code rules this waveform property is represented as a single factor, while in the ontology-based solution it is divided into 6 factors because of the fact that the ontology model works with waveform states. For all 6 states the same measured value is set, the difference is in the ontology types of the instances.

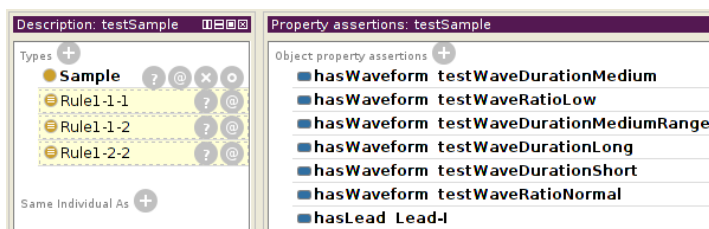


Figure 11

Definition of a “Sample” instance and the corresponding waveforms in Protegé after the rule inference step

After the ontology is populated with the available information, the next step is to identify which diagnostic rules have the required information available for evaluation. In the ontology the definition of diagnostic rules is performed with equivalent classes. This implies that the input samples inferred types contain all the diagnostic rule types that can be evaluated. The type – diagnostic rule – identification is performed by an ontology reasoner [14]. Figure 11 displays the result produced by the Hermit [14] reasoner for the “testSample” shown in Figure 10. As the picture shows, the reasoner identified that “testSample” meets the criteria for 3 possible “Rule” sub-concepts.

After the inference system finished the execution, the types of the sample instance need to be analyzed. All the type associations of the sample instance that have the parent class of 'Rule' mark the diagnostic rule types that have all the required information available for evaluation. The evaluation step starts with acquiring all the types associated with the ECG sample that are the sub-types of the diagnostic rule modeling concept. In the case when the number of diagnostic rule representing types is greater than zero, all the diagnostic rules are evaluated. Since all the waveform concepts have a crisp value associated through the “hasWaveformValue” axiom, the evaluation is a matter of logical comparison. The truthiness values of the waveforms need to be aggregated into a single value thus producing the diagnostic output for a specific rule.

## Conclusions

Our ontology-based Minnesota code model meets the outlined goals. It models the basic concepts of the diagnostic system and the related relationships. The input types and states are explicitly defined. Using the explicit definitions partial diagnosis is also supported using a two-step approach. First, it is determined which diagnostic rules have all the mandatory information available, then the identified rules are evaluated. The compatibility between the original expert system based definitions and the ontology based solution is ensured by crosschecking the results of the ontology model based solution with the results produced by the expert system based solution. By default, both solution use the same methodologies. However, the ontology based definitions are based on state concepts instead of value ranges. The state concepts are the key to the extensibility of the system. Besides the Minnesota code-based value range definitions it is possible to attach various other representations and evaluation algorithms to apply different diagnostic algorithms as well. For example, all waveform states can have fuzzy definitions associated with them. In most fuzzy-ontology solutions the fuzzification of the ontology concepts is carried out using ontology annotations, when a concept has associated multiple fuzzy membership functions. Using this approach, all the ontology concepts represent a single fuzzy variable, where the fuzzy membership functions are stored in an ontology annotation. In the case of the Minnesota code ontology model, the state concepts can be used to store the membership function definitions. For each state concept one membership function definition is assigned. The outcome of this approach is that a fuzzy variable is represented by a grouping concept and the sub-concepts contain the membership function definitions of the variable. The task of the ontology system becomes identification of the branches of the decision tree, which can be evaluated. Usage of the fuzzy characteristics is required for the evaluation of the diagnostic rules. The application of fuzzy logic, in the diagnostic steps, through the usage of fuzzy-ontology, is the next step for improving the current solution presented herein.

## Acknowledgement

The research was supported by the Hungarian OTKA projects 106392 and 105846, and project of the Vojvodina Academy of Sciences and Arts "Mathematical models of intelligent systems and their applications".

## References

- [1] Peter W. M, Shahid L., "Automated Serial ECG Comparison Based on the Minnesota Code", Journal of Electrocardiology, Vol. 29, Sup. 1, pp. 29-34, 1996
- [2] G. Balázs, K. Haraszti, Gy. Kozmann, "Increasing the Efficiency of the "Excluding Rules" of the Minnesota Coding System using the Fuzzy Logic", Measurement Science Review, Volume 5, Section 2, 2005



- [3] Fernando Bobillo, Umberto Straccia, “Fuzzy Ontology Re Presentation using OWL 2”, *International Journal of Approximate Reasoning*, 2011, pp. 1073-1094
- [4] Sram, N., Takacs, M. “Minnesota Code: A Fuzzy Logic-based Approach”, *Proc. of the 11<sup>th</sup> International Symposium on Computational Intelligence and Informatics (CINTI 2010)*, pp. 233-236, Budapest, Hungary, 2010
- [5] Prineas, Ronald J., Crow, Richard S., Zhang, Zhu-ming, “The Minnesota Code Manual of Electrocardiographic Findings”, ISBN 978-1-84882-777-6
- [6] Engelstein ED, Zipes DP., “Sudden Cardiac Death”, *The Heart, Arteries and Veins*. New York, NY: McGraw-Hill; 1998:1081-1112.4
- [7] Myerburg RJ, Castellanos A., “Cardiac Arrest and Sudden Death”, *Heart Disease: A Textbook of Cardiovascular Medicine*. Philadelphia, Pa: WB Saunders; 1997:742-779
- [8] de Vreede Swagemakers JJM, Gorgels APM, Dubois-Arbouw WI, van Ree JW, Daemen MJAP, Houben LGE, Wellens HJJ. “Out-of-Hospital Cardiac Arrest in the 1990’s: a Population-based Study in the Maastricht Area on Incidence, Characteristics and Survival”, *J Am Coll Cardiol*. 1997; 30:1500-1505
- [9] World Health Organization, "Annex Table 2: Deaths by Cause, Sex and Mortality Stratum in WHO Regions, Estimates for 2002", *The world health report*, 2004
- [10] Sándor Khoór, István Kecskés, Ilona Kovács, Dániel Verner, Arnold Remete, Péter Jankovich, Rudolf Bartus, Nándor Stanko, Norbert Schramm, Michael Domijan, Erika Domijan, “Heart Rate Analysis and Telemedicine: New Concepts & Maths”, *Proceedings of the IEEE SISY 2007*, pp. 61-66, Subotica, Serbia, 2007
- [11] Ceusters W Ab, Smith B Bcd, De Moor G A, “Ontology-based Integration of Medical Coding Systems and Electronic Patient Records”, *IFOMIS Reports*, 2004
- [12] John R. Hampton, “The ECG Made Easy 8<sup>th</sup> Edition”, 2013
- [13] <http://protege.stanford.edu>
- [14] Birte Glimm, Ian Horrocks, Boris Motik, Giorgos Stoilos, Zhe Wang, “HermiT: An OWL 2 Reasoner”, *Journal of Automated Reasoning*, October 2014, Volume 53, Issue 3, pp. 245-269
- [15] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101(23):e215-e220 [Circulation Electronic Pages; <http://circ.ahajournals.org/cgi/content/full/101/23/e215>]; 2000 (June 13)
- [16] Doina Drăgulescu, Adriana Albu, “Medical Predictions System”, *Acta Polytechnica Hungarica* Vol. 4, No. 3, 2007, p. 89

# Estimation of Phosphorus Content in Archaeological Iron Objects by Means of Optical Metallography and Hardness Measurements

**Ádám Thiele**

Budapest University of Technology and Economics, Faculty of Mechanical Engineering, Department of Materials Science and Engineering  
Bertalan Lajos u. 7, Bdg. MT, H-1111 Budapest, Hungary, thiele@eik.bme.hu

**Jiří Hošek**

Institute of Archaeology of the ASCR, Prague, v.v.i., Letenská 4, 118 01 Prague 1, Czech Republic, hosek@arup.cas.cz

---

*Abstract: In order to facilitate everyday archaeometallographic research into archaeological and/or historical objects, a method employing results of metallographic examination and hardness measurements to estimate phosphorus content in iron artifacts is introduced in the paper. Furthermore, phosphorus contents encountered in phosphoric iron that was used deliberately as a special material (for pattern-welding etc.) are discussed here. Despite certain limitations, the proposed method can be used for the estimation of the phosphorus content of archaeological iron examined either currently or in the past.*

*Keywords: Phosphoric iron; archaeometallurgy; archaeometallography; Vickers hardness*

---

## 1 Introduction

### 1.1 Archaeological and Archaeometric Background

Iron with enhanced phosphorus content is known in archaeometallurgy, as phosphoric iron, the term being used for iron containing more than 0.1 wt% P [1]. It is commonly encountered in archaeological iron objects regardless of the dating and provenance.

Phosphorus, a natural admixture coming from bog iron ore, gives specific properties to iron, and it is not surprising that this issue has become a subject of interest to many researchers. It is currently well-known, that certain types of phosphoric iron were highly valued in the past, particularly for the possibility to be distinguished with the naked eye (under certain conditions) from non-phosphoric iron and steel.

Phosphorus is an avoided element in modern steel industry. Its detrimental effects include various forms of embrittlement, which reduce the toughness and ductility of steel. Phosphorus, a ferrite-stabilizing element, can be created with up to the maximum of 2.8 wt% in ferrite, as the Fe-P dual phase diagram shows in Fig. 1 [2].

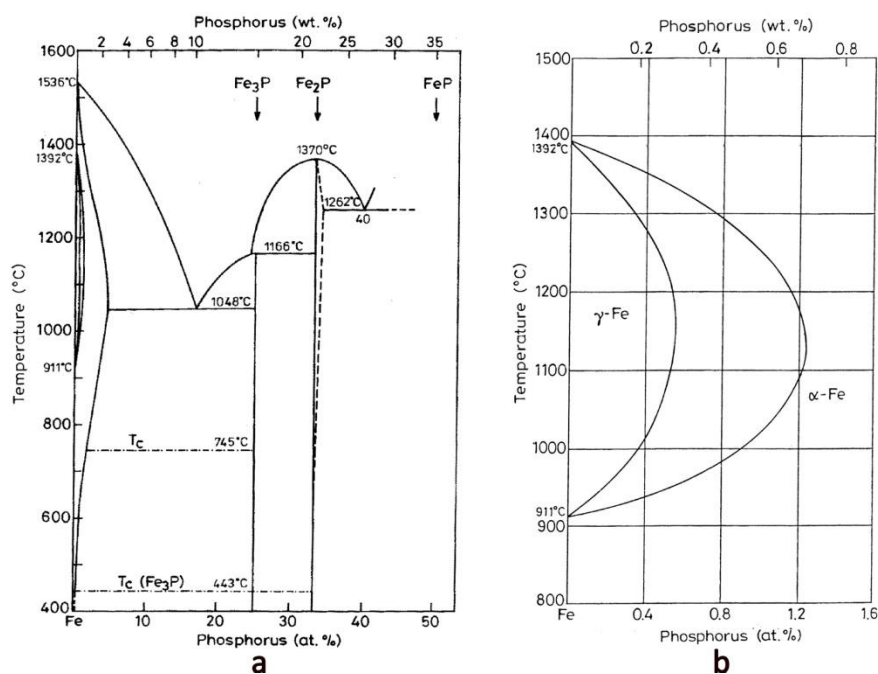


Figure 1

Fe-P dual phase diagram: a – phosphorus content up to 35wt%, b –  $\alpha$ - $\gamma$  the dual phase field

The melting point of the eutectic of iron and phosphorus (Fe-Fe<sub>3</sub>P) is 1048°C. Above a certain temperature and phosphorus content (1048°C and 2.8 wt% P), a molten phase appears on the grain boundaries. Phosphoric iron of high phosphorus content is not forgeable, as the Fe-Fe<sub>3</sub>P eutectic melts on the grain boundaries above the eutectic temperature. On the other hand, phosphoric iron of low phosphorus content cracks during cold-working due to its reduced ductility (cold shortness). If the amount of the Fe-Fe<sub>3</sub>P eutectic is low, it is possible to forge weld the ferrite grains after the molten eutectic effuses just as molten slag does.

Despite its poor mechanical properties, phosphoric iron was deliberately used in medieval metal-working for the manufacture of ostentatious blades of swords, saxes, knives and heads of spears. Such blades and spearheads reveal various forms of pattern-welding and/or strips of phosphoric iron attached to the cutting edges by straight or serrated welds [3-5, etc.]. Phosphoric iron is also a typical feature of Scandinavian variants of sandwiched blades, contrary to the Old-Russian variants employing non-phosphoric iron [6].

On the other hand, phosphoric iron can appear in medieval tools and weapons unintentionally, either as a result of the lack of non-phosphoric iron, or due to the use of unsorted heterogeneous or scrap iron. This is well illustrated by the research conducted by Piaskowski [7], who found high average phosphorus contents (from 0.3 up to 0.85 wt%) in iron implements coming from several archaeological smelting sites in Poland.

With regard to the current research carried out, it seems that iron with only certain amounts of phosphorus was deliberately involved in the forging of ostentatious blades, whose pattern-welded variants are the most widely known and famous ones in [5, 8, 9]. The ability to distinguish this type of phosphoric iron from those which could appear unintentionally while forging is very important in archaeometallurgic practices; particularly when forged semi-products (such as bars uncovered in smithy workshops) or blades of unusual type are the subject of examination and assessment. Establishing the range of the phosphorus content that is most typical for phosphoric iron used deliberately in the past for aesthetic purposes is therefore, the first goal of this study.

Phosphoric iron can be reliably identified in archaeological weapons and tools using the combination of optical microscopy (OM), due to its highly coarse-grained structure having light appearance when etched with Oberhoffer's reagent or ghost structure, and by hardness measurements, because phosphorus increases the hardness of iron alloys. Nevertheless, archaeometallurgists often need to go beyond simple identification. Today it is possible to measure the P content in selected areas of metallographic specimens by SEM-EDS/WDS or LA-ICP-MS analysis. But even the most common SEM-EDS analysis is still a money and time consuming method, which is not regularly available for everyday archaeometallographic research. An easier and less expensive method to employ the results of metallographic examination and hardness values and preliminarily estimate the phosphorus content in iron is needed. Hence, the second and the main aim of this study is to establish a hardness-phosphorus content function valid for archaeological iron objects (primarily for blades of knives and swords) and to determine the accuracy of the model.

## 1.2 Detecting the Presence of Phosphoric Iron in Archaeological Objects

Metallographers can be alerted to the presence of phosphoric iron by the specific appearance of ferritic structure. The so-called ghost structure (Fig. 2a) and coarse-grained ferrite (with grain boundaries often entirely invisible) are the structures typical for phosphoric iron. Ghost structure (GS) can be observed when Oberhoffer's or Klemm's reagents are applied. As Oberhoffer's reagent creates Cu deposits on low-phosphorus areas, these have a darker appearance when observed under OM [10], Klemm's reagent makes the low-phosphorus areas darker as well and, according to Radzikowska [11], it distinguishes the high and low phosphorus areas, even more precisely than Oberhoffer's etchant. Techniques employing phosphoric iron, such as the pattern-welding technique, can be quickly identified in this manner, if the artifact is sufficiently preserved (see Fig. 2b).

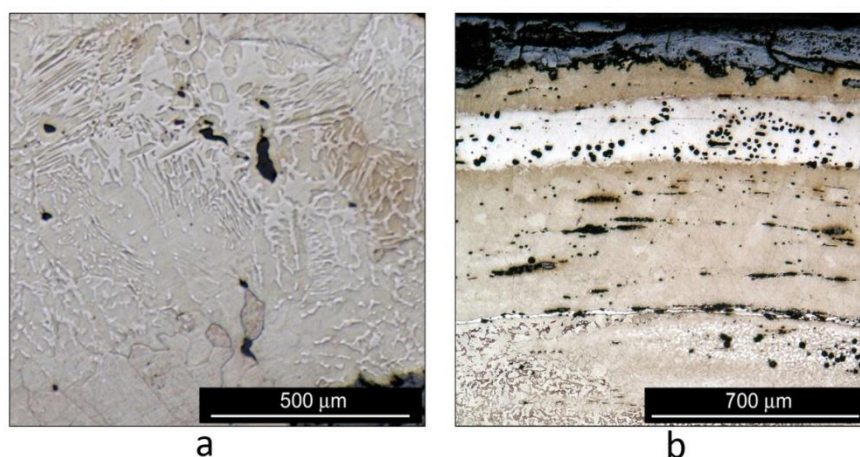


Figure 2

Identification of phosphoric iron when etched: a - ghost structure revealed by etching with nital; b - strip (light) of phosphoric iron of a pattern-welded surface panel when etched by Oberhoffer's reagent

The formation of GS is associated with the domain of the coexistence of austenite and ferrite, which appears in the Fe-P binary diagram between c. 0.1-0.7 wt% and 911-1392°C (cf. Fig. 1b). Since phosphorus has lower solubility in austenite than in ferrite, the austenite, which is formed along the ferrite grain boundaries as allotriomorphs and which additionally grows into the ferrite grains in a needle-shaped morphology, will contain less phosphorus than the untransformed ferrite [12]. When cooling is sufficiently quick (air cooling), the uneven phosphorus distribution will stay unchanged because of the limited diffusion of phosphorus in ferrite. The ghost structure therefore consists of low- (former austenite) and high-phosphorus ferrite, with higher phosphorus content in the grain cores. Literature [1] suggests that structural zones revealing GS can actually contain 0.1-0.7 wt% P.

As the ratio between the low- and high-phosphorus ferrite depends on both the overall phosphorus content in the alloy and the applied heat treatment, this ratio cannot be used to estimate the phosphorus content of the iron. Furthermore, if the cooling is sufficiently slow, GS becomes blurred and can even totally disappear.

## 2 Theory, Calculations - Strengthening Effects in Phosphoric Iron

Phosphorus has the strongest solid solution hardening effect on ferrite among substitutional solid-solution strengtheners. Although the difference in the atomic radius of ferrite and phosphorus is higher than 15% (cf. Hume-Rothery rule), phosphorus still enters the crystal lattice as a substitution solute [13].

In general, the strength of metals is dependent on how easily dislocations in their crystal lattice can be propagated. In substitutional solid solutions, the solute atom replaces the solvent atoms in their lattice positions and dislocations are surrounded by a so called Cottrell-cloud of substitutional solute atoms, so the movement of the dislocation is precluded.

Strength is the mechanical property of metals which can be characterized by the characteristic value of yield strength or proof strength. The increment of proof strength in solid-solution and the specific deformation caused by the solute atom are directly proportional. The specific deformation can be calculated after the equation (1) [14]:

$$\varepsilon = (r_A - r_B) / r_A \quad (1)$$

where:  $\varepsilon$  - specific deformation (-),  $r_A$  - atomic radii of solvent atoms (pm),  $r_B$  - atomic radii of solute atoms (pm).

Empirically measured atomic radii for iron and phosphorus are 140 pm and 100 pm (with an accuracy of about 5 pm) according to Slater [15].

The yield strength increment can be calculated after the equation (2) [14]:

$$\Delta R_{p0.2} = G \cdot \varepsilon \cdot X_c / 100 \quad (2)$$

where:  $\Delta R_{p0.2}$  - proof strength increment (MPa),  $G$  - shear modulus (MPa),  $\varepsilon$  - specific deformation (-),  $X_c$  - atomic percentage of the solute atom (at %).

The shear modulus of iron is 83000 MPa. It can be seen that the increment of proof strength in solid solution and the atomic percentage of the solute atom are directly proportional. Replacing the variables in equation (1) and (2), the calculated proof strength increment is 237 MPa for 1 at % of phosphorus.

According to Cahoon et al. [16], the following equation can be used to determine the relation between Vickers hardness and yield/proof strength (3a):

$$R_{p0.2} = (HV_{0.2}/3) \cdot (0.1)^n \quad (3a)$$

where:  $R_{p0.2}$  - proof strength (kg/mm<sup>2</sup>), HV - Vickers hardness (kg/mm<sup>2</sup>),  $n$  - strain-hardening exponent.

Alternatively, when  $n$  is assumed to be zero, the relation can be expressed in the form of

$$HV \approx 0.3 \cdot R_{p0.2} \quad (3b)$$

where: HV - Vickers hardness number,  $R_{p0.2}$  - proof strength (MPa).

Using the equation (3b), the Vickers hardness increment of phosphorus is 71.1 HV for 1 at %. Considering the molar mass of iron (56 g/mol) and that of phosphorus (31 g/mol), the theoretical Vickers hardness increment is 127 HV for 1 wt% of phosphorus, while literature suggests 123 HV or 125 HV hardness increments for 1 wt% of phosphorus, and a hardness of 60-70 HV for unalloyed ferrite [3, 13].

Carbon is the most common element which can appear in phosphoric iron but arsenic can also be detected in elevated concentrations. Such elements as nickel, cobalt or copper are often present in traces but they can easily be revealed because of their increased concentration in welds [17]. These elements, as well as phosphorus, may cause hardness increments.

Phosphoric iron in archaeological objects (like metals in general) can also be strengthened by strain (work) hardening. The flow curve can be calculated using the Ludwig-Hollomon strain hardening equation.

The grain size of phosphoric iron also has an effect on its strength (grain-boundary strengthening). The connection between yield strength and grain size is defined by the Hall-Petch equation.

Heat treatment can also affect the strength of phosphoric iron, although neither Martensitic transformation, nor precipitation (age) hardening appears. The strengthening effect of heat treatment is low, which can be related to the distribution of solved phosphorus (cf. detailed in 1.2). This can be proved by the fact that neither yield strength nor hardness values differ much in case of water-quenched and furnace-cooled states [18].

### 3 Methods and Results

Previously examined metallographic cross sections of four pattern-welded sword blades and six knife blades were chosen for further technical analysis (Fig. 3).

*Sword No. 54* is the 10<sup>th</sup> Century burial find uncovered in the cemetery of Kanín (Bohemia), which belonged to the early medieval stronghold of Libice nad Cidlinou. The sword belongs to type Y, according to Petersen, and represents a

high quality pattern-welded type sword, albeit the quality of the genuine cutting edges remains unanswered because of the extended corrosion. *Sword No. 120* was lifted from an opulent male tomb No. 120 at the burial ground by Libuše-pond near the stronghold of Stará Kouřim (Bohemia). The sword is an unusually short two-edged sword with a high quality pattern-welded blade having unquenched cutting edges of hypereutectoid steel, and unfortunately, bears no significant typological features. According to the enclosed grave articles, the tomb itself can be dated between the first to the second third of the 9<sup>th</sup> Century. *Sword No. 616* comes from Bešeňov (Slovakia). The weapon was lifted from an opulent princely grave and is dated to the 5<sup>th</sup> Century. The hilt is missing. *Sword No. 715* comes from the stronghold of Mikulčice (Moravia), one of the main power centers of the Great Moravian Empire. Lifted from grave No. 715 and dated to the first half of the 9<sup>th</sup> Century, the sword of type H according to Petersen represents all the 16 Mikulčice swords as a pattern-welded weapon of earlier type, made almost entirely of iron. *Knives No. 249, 251 and 252* come from Sekanka - Hradištko u Davle (Bohemia) - the 13<sup>th</sup> Century urban type trading settlement held by the Ostrov Monastery (the Benedictine Monastery of St. John the Baptist at Ostrov). Based on the various evidence of smithy activity and the high concentration of pattern-welded, striped and serrated knives in the craftsman area, it is believed that the production of these opulent knives took place directly at the site. The knives Nos. 251 and 252 were provided with striped blades, knife No. 249 is a basic type of pattern-welded knives. All the knives were products of excellent quality. *Knife No. 274* comes from Lahovice (Bohemia) - the burial ground in open terrain, which was used from the mid-9<sup>th</sup> to the 11<sup>th</sup> Century. The knife was lifted from grave No. 274, which can, in general, be dated to the first half of the 10<sup>th</sup> Century, and which (regarding the enclosed grave goods) does not belong to wealthy burials in the cemetery. *Knife No. 423* comes from Mutějovice (Bohemia), where traces of both a settlement from the 10<sup>th</sup> to 12<sup>th</sup> Century and two rural smithies (the first of which was in use during the first half, the second in the second half of the 13<sup>th</sup> Century) were uncovered. The pattern-welded knife No. 423 was a product of superior quality and despite the uncertainties in the dating (12<sup>th</sup> or 13<sup>th</sup> Century) it confirms the fact that high quality knives may have appeared even in rural settlements, perhaps as products coming from craft centers. *Knife No. 667* was uncovered at Budeč (Bohemia), which was an important stronghold of the 10<sup>th</sup> and 11<sup>th</sup> Centuries (founded as early as the turn of the 8<sup>th</sup> and 9<sup>th</sup> Centuries) held by the Premyslid family. Knife No. 667 was found at the central part of the stronghold and is interpreted as a 10<sup>th</sup> to 11<sup>th</sup> Century striped blade of good quality.



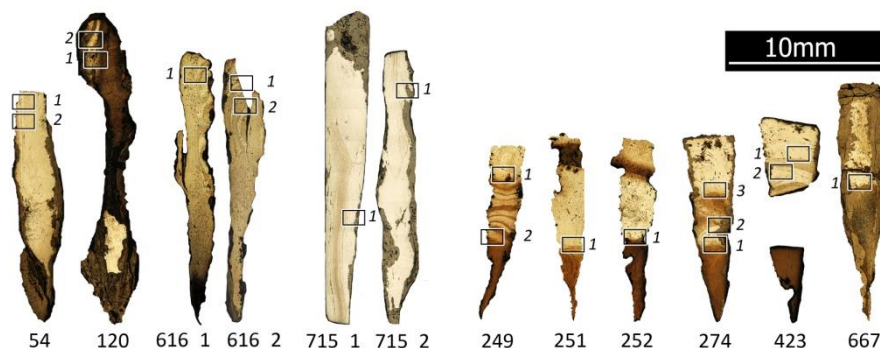


Figure 3

Macro-photographs of investigated metallographic cross sections of medieval sword blades (Nos. 54, 120, 616\_1, 616\_2, 715\_1 and 715\_2) and knife blades (Nos. 249, 251, 252, 274, 423 and 667). Areas of Vickers hardness measurements with phosphoric iron layers are marked with a rectangle.

The metallographic cross sections of the blades were polished and the Nital2%-etched surface was examined under OM and SEM-EDS to identify the phosphoric iron layers. Fig. 3 shows the macro-photographs of the investigated blade cross sections. The identified phosphoric iron layers on which Vickers hardness measurements were carried out are indicated with rectangles.

The Vickers hardness was measured using a Boehler 1105 micro Vickers hardness tester with a load of 0.2 kgf and a loading time of 10 s. Five hardness measurements were performed on individual phosphoric iron layers (cf. Table 1) of each sample (Fig. 4a). The chemical composition of the area of the indentations imprinted in the surface was then measured using a Philips XL30 Scanning Electron Microscope equipped with an Energy Dispersive Spectrometer (Fig. 4b). The detection limit for phosphorus in phosphoric iron was 0.5 at % and ca. 0.3 wt% respectively. This way, a direct relationship between the Vickers hardness and the phosphorus content was found in a total of 90 cases. The results are summarized in Table 1 and Fig. 5.

Table 1

Hardness measurement and EDS analysis results; \*anomalous values revealed by linear regression analysis

Object area tested	Vickers hardness (HV)					Phosphorus content (wt%)					Other element
	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>	5 <sup>th</sup>	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>	5 <sup>th</sup>	
54_1	260*	278	270	231	243	0.89	1.16	1.24	1.09	1.35	
54_2	226	235	214*	234	244	0.99	0.91	1.27	0.93	0.85	
120_1	303	264	275	280	276	1.41	1.50	1.26	1.55	1.37	
120_2	274	241	244	257	272	1.18	0.47	1.18	0.60	1.00	C
616_1_1	299	263	234	221	240	0.51	0.45	0.64	0.55	0.38	As

616_2_1	229	234	220	235	230	0.60	0.70	0.16	0.64	0.60	As
616_2_2	220	214	225	231	230	0.48	0.55	0.58	0.64	0.39	As
715_1_1	159	178	156	149	154	0.40	0.56	0.34	0.29	0.27	
715_2_1	157	151	150	214	146	0.44	0.30	0.32	0.77	0.18	
249_1	263	252	220	254	220	0.60	0.53	0.39	0.37	0.22	C
251_1	185	185	169	174	163	0.66	0.42	0.55	0.61	0.63	
252_1	189	188	171	163	148	0.69	0.61	0.60	0.61	0.43	
274_1	186	220	199	175	166	0.76	0.85	0.73	0.60	0.60	
274_2	188	207	206	199	192	0.41	0.87	0.46	0.56	0.47	C
274_3	160	182	179	187	158	0.50	0.46	0.53	0.59	0.60	
423_1	150	141	147	165	159	0.46	0.26	0.43	0.46	0.63	
423_2	217	229	232	205	180	0.88	0.99	0.76	0.81	0.79	
667_1	165*	205	235	215	225*	0.79	0.81	0.89	0.65	0.45	

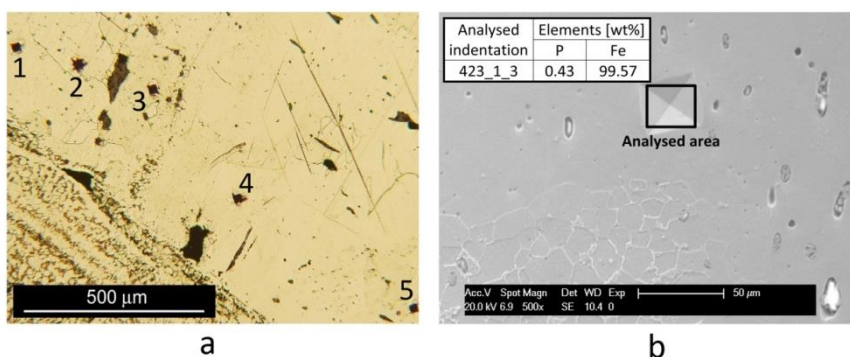


Figure 4

Indentations observed under OM in sample 423 area 1 (a) and EDS analysis on the 3<sup>rd</sup> indentation under SEM (b); (example)

## 4 Discussion

According to our results obtained by means of hardness measurement and EDS analysis (cf. Fig. 5), it seems that phosphoric iron used for the manufacture of ostentatious blades might have contained from 0.4 to 1.4 wt% of phosphorus on average (0.4 to 0.9 wt% in case of the analyzed knives and 0.4 to 1.4 wt% in case of the swords). This is in accordance with the values stated in literature [5, 8, 9]. The difference between the minimum and maximum content of P was 0.35 wt% on average in a single area tested (in one layer of pattern-welding for instance), but in particular cases it might have been as much as twice higher. In some cases, the phosphoric iron also contained arsenic or carbon besides phosphorus.

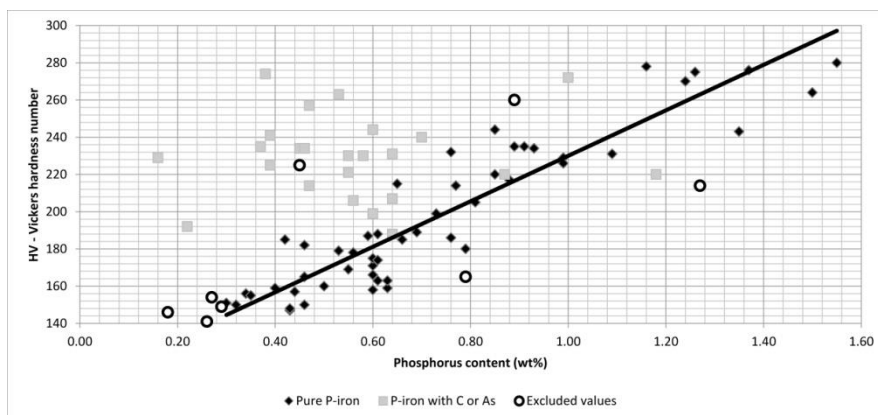


Figure 5

Fitted line plot for the Vickers hardness and the phosphorus content function

In order to determine the relationship between the phosphorus content and Vickers hardness correctly, certain data have to be excluded from further processing. Phosphoric iron enriched in carbon or arsenic has higher hardness than pure phosphoric iron; therefore results related to impure phosphoric iron were excluded (Fig. 5 grey rectangular). Because of the detection limit of the SEM-EDS analysis applied, we excluded the results with less than 0.3wt% of phosphorus as well (Fig. 5 white circle). Finally, four anomalous values were also excluded (cf. Table 1, Fig. 5 white circle).

Using the modified data, the following equations have been derived by means of linear regression:

$$HV = (110.1 + (119.8 \cdot P)) \pm 15.7 \quad (4)$$

$$P = (-0.919 + (0.0083 \cdot HV)) \pm 0.13 \quad (5)$$

where:

P - Phosphorus content (wt%)

HV - Vickers hardness number

According to equation (6), the Vickers hardness increment for 1 wt% of phosphorus in archaeological iron we have analyzed is 120 HV (thus practically making no difference from the values measured and calculated in modern Fe alloys) and non-phosphoric iron would theoretically have a Vickers hardness of about 110 HV, which means that the phosphoric iron we analyzed was somewhat strengthened additionally, regardless of its phosphorus content.

In our samples, additional strengthening effects in ferrite, besides the solid solution strengthening effect of phosphorus do not significantly affect the deviation of the data, which was most likely caused by measurement uncertainties.

The strengthening effect of other elements which were in concentration under the detection limits of SEM-EDS should be low. Grain boundary strengthening could also be negligible as a typical coarse-grained microstructure (with a grain size of 100-500  $\mu\text{m}$ ) was observed in all pure-phosphoric iron layers. The blades analyzed underwent some sort of quenching, but this heat treatment does not provide a significant increase in the hardness of phosphoric iron. The strain hardening effect cannot play an important role because significant cold working of the heat-treated blades is unlikely. In conclusion, these effects all together must have caused the hardness increment, which does not depend on the phosphorus content.

The fact that the established relationship is of general application is supported by the analysis of the variance of hardness residuals, which suggests that there are no significant differences among individual objects analyzed in terms of the means of residual hardness values (in other words, all the analyzed blades follow the established HV-P relationship in a similar manner). The standard deviation  $\pm 16$  HV (the distribution of residuals is reasonably close to a normal distribution) covers both measurement uncertainty and, in general, the insignificant effect of other factors on the hardness of iron alloys.

Within a more complex study of phosphoric iron, performed by Stewart et al. [18], hardness was measured on phosphoric iron containing from 0.1 to 0.38 wt% P, i.e. in a range which was not researched in this study. The published results show lower hardness values of phosphoric iron, in general, in comparison to our data, which seems to have been caused by the different hardness values of the phosphorus-free iron used.

For an easy estimation of the phosphorus content in archaeological iron objects, conversion Table 2 was developed. When using the table, standard deviation should be considered; for instance, for a hardness of 200HV the corresponding phosphorus content is  $0.75 \pm 0.13$  wt%, i.e. there is a 68% probability that the phosphorus content will be in the range of 0.62 to 0.88 wt%.

Table 2

Vickers hardness values (HV) and corresponding phosphorus content P (wt%) according to equation (5). Standard deviation for phosphorus content is 0.13wt%

HV	P (wt%)	HV	P (wt%)	HV	P (wt%)	HV	P (wt%)
140	0.25	185	0.63	230	1.00	275	1.38
145	0.29	190	0.67	235	1.04	280	1.42
150	0.33	195	0.71	240	1.08	285	1.46
155	0.37	200	0.75	245	1.13	290	1.50
160	0.42	205	0.79	250	1.17	295	1.54
165	0.46	210	0.83	255	1.21	300	1.59
170	0.50	215	0.88	260	1.25	305	1.63
175	0.54	220	0.92	265	1.29	310	1.67
180	0.58	225	0.96	270	1.33	315	1.71

The admixture of arsenic in phosphoric iron will lead to the misinterpretation of the phosphorus content when the above method is used. Nevertheless, based on our investigations, it seems that arsenic was not a common admixture of phosphoric iron that we encounter in ostentatious objects we deal with (arsenic was detected only in the object No. 616). It should be noted, however, that in case of arsenic the detection limit and the accuracy of the EDS method are poor. Albeit arsenic makes the determination of phosphorus content by the proposed equation (5) impossible, it is an important admixture of bloomery iron. Arsenic appears only in certain bog-ores and can be removed from iron only with difficulty; therefore its presence in phosphoric iron can serve as a useful guideline in the complex issue of determining provenance [19]. In our case, the sword from Bešeňov is a weapon which differs from the other objects analysed in this study in both dating and provenance.

Unfortunately, there is no way to distinguish pure phosphoric iron from those also containing arsenic by common means of optical metallography [10]. The presence of carbon is suggested by the presence of pearlite (cementite in general) in the structure; therefore, the assessment of such structures should be avoided to estimate the phosphorus content by the proposed method.

## Conclusion

The hardness measurements with detailed chemical SEM-EDS analysis was preformed on archaeologically excavated swords and knives, followed by the statistical treatment of the data obtained. This allowed the following conclusions to be drawn:

- 1 Phosphoric iron with a wide range of average content 0.4-1.4 wt% P (the difference between the minimum and maximum content in a single tested area appears to average 0.35 wt%) was used for aesthetic purposes in the manufacture of ostentatious blades
- 2 When the observed structure of phosphoric iron consists of ferrite without traces of pearlite or ghosting, Vickers hardness (HV) can be used to estimate the phosphorus content P (wt%) using the equation:

$$P = (-0.919 + (0.0083 \text{ HV})) \pm 0.13$$

This is particularly but, not exclusively, suitable for heat-treated blades. The accuracy of the estimation is  $\pm 0.13$  wt%. This equation is not valid when the iron contains arsenic or carbon in addition to the phosphorus. Similarly, when a ghost structure is revealed by etching, the use of the above stated formula may cause misinterpretation.

## Acknowledgements

This 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 New Hungary Development Plan (Project ID:

TÁMOP-4.2.1/B-09/1/KMR-2010-0002). The presented research was conducted with the support Czech Science Foundation (project P405/12/2289).

## References

- [1] Vega, E., Dillmann, P., Lheritier, M., Fluzin, P., Crew, P., Benoit, P. (2003): Forging of Phosphoric Iron. An Analytical and Experimental Approach. In *Archaeometallurgy in Europe*, Vol. II, Milan, 337-346
- [2] Okamoto, H. (1990): The Fe-P (Iron-Phosphorus) System. *Bull. Alloy Phase Diagrams* 11(4), 404-412
- [3] Tylecote, R. F., Gilmour, B. J. J. (1986): *The Metallography of Early Ferrous Edge Tools and Edged Weapons* (BAR British Series 155)
- [4] Pleiner, R. (2006): *Iron in Archaeology. Early European Blacksmiths*. Praha: AU AVČR
- [5] Hošek, J., Malý, K., Zav'álov, V. (2007): Železná houba ze Žďáru nad Sázavou ve světle problematiky fosforového železa ve středověkém nožířství. In *Archaeologia technica* 18. TM Brno. 10-17
- [6] Zav'álov, V., Rozanova, L. S., Terechova, N. N. (2012): *Tradicii i innovacii v proizvodstvennoj kul'ture Severnoj Rusi*. Ankil: Moskva
- [7] Piaskowski, J. (1989): Phosphorus in Iron Ore and Slag, and in Bloomery Iron, *Archaeomaterials* 3, 47-59
- [8] Kinder, J. (2003): Pattern-welded Viking-Age Sword Blades - What can Modern Metallurgical Investigation Contribute to the Interpretation of their Forging Technology? In: *Archaeometallurgy in Europe*, Vol. I. 239-248
- [9] Thålin, L. (1967): Metallografisk undersökning av ett vendeltida praktsvärd, *Fornvännen*. 225-240
- [10] Stewart, J. W., Charles, J. A., Wallach, E. R. (2000): Iron-Phosphorus Carbon System. Part 2: Metallographic Behaviour of Oberhoffer's Reagent. *Material Science Technology* 16, 283-290
- [11] Radzikowska, J. (1998): The Use of Selective Colour Etching to the Metallographic Identification of Phosphorus Segregation and Technology of Early Implenents Made of Bloomery Iron and Steel. In *Metallography'98*, 14-19
- [12] Piccardo, P., Ienco, M. G., Balasubraman, R., Dillmann, P. (2004): Detecting Non-Uniform Phosphorus Distribution in Ancient Indian Iron by Colour Metallography. *Current Science* 87, 650-653
- [13] Key to Metals – [www.keytometals.com](http://www.keytometals.com) (viewed 7 Oct 2013)
- [14] Artinger, I. (1982): *Anyagszerkezzattan*, Tankönyvkiadó, Budapest, 19-27

- [15] Slater, J. C. (1964): Atomic Radii in Crystals. *Journal of Chemical Physics* 41 (10), 3199-3205
- [16] Cahoon, J. R., Broughton, W. H., Kutzak, A. R. (1971): The Determination of Yield Strength from Hardness Measurements. *Metallurgical Transactions* 2(7), 1979-1983
- [17] Tylecote, R. F., Thomsen, R. (1973): The Segregation and Surface Enrichment of Arsenic and Phosphorus in Early Iron Artefacts. *Archaeometry* 15/2, 193-198
- [18] Stewart, J. W., Charles, J. A., Wallach, E. R. (2000): Iron-Phosphorus-Carbon System. Part 1 - Mechanical Properties of Low Carbon Iron-Phosphorus Alloys. *Material Science Technology* 16, 275-282
- [19] Piaskowski, J. (1984): Das Vorkommen von Arsen im antiken und frühmittelalterlichen Gegenständen aus Renneisen, *Archäologie*, Vol. 18, 213-126

# Friction Compensation in TP Model Form - Aeroelastic Wing as an Example System

**Béla Takarics, Péter Baranyi**

Computer and Automation Research Institute  
Hungarian Academy of Sciences  
Kende u. 13-17, H-1111 Budapest, Hungary  
E-mail: takarics@sztaki.hu, baranyi@sztaki.hu

---

*Abstract: The aim of this paper is to fit the friction compensation problem in the field of modern polytopic and Linear Matrix Inequality (LMI) based control design methodologies. The paper proves that the exact Tensor Product (TP) type polytopic representations of most commonly utilized friction models such as Coulomb, Stribeck and LuGre exist. The paper also determines and evaluates these TP models via a TP model transformation. The conceptual use of the TP model of the friction is demonstrated via a complex control design problem of a 2D aeroelastic wing section. The paper shows how the friction model and the model of the aeroelastic wing section can be merged and transformed to a TP type polytopic model - by TP model transformation - whereupon LMI based control performance optimization can immediately be executed to yield an observer based output feedback control solution to given specifications. The example is evaluated via numerical simulations.*

*Keywords: friction compensation; LMI-based multi-objective control; TP model transformation; qLPV systems*

---

## 1 Introduction

Friction is a highly nonlinear phenomenon and may result in steady state errors, limit cycles, poor performance in high precision applications. Linearizing the friction phenomenon in the control design process can also lead to poor control performance ([1, 2]). There are many general, starting from simple to complex and sophisticated friction models presented in the technical literature given as qLPV (quasi Linear Parameter Varying) models, which can be suitable for inclusion in the model for the controlled plant.

Multi-objective LMI (Linear Matrix Inequality)-based control design for qLPV systems has been in the focus of modern control theories, the pioneers Gahinet, Balas, Chilali, Boyd, and Apkarian were responsible for establishing this new concept



[3, 4, 5]. The aim of this paper is to fit friction models, and by this friction compensation of qLPV systems to the LMI-based multi-objective control design framework, both theoretically and in practice.

One direction of LMI-based theories needs the system to be formulated in convex polytopic form. The proposed methodology to obtain the convex polytopic form is to apply Tensor Product (TP) model transformation, which is based on the Higher-Order Singular Value Decomposition (HOSVD) of tensors ([6]). The concept of the HOSVD of continuous functions was given in [7] and [8], TP model transformation was introduced to control in [9] as a methodology for system control design.

Various polytopic forms of the same model affect the performance of LMI-based controllers (see [10, 11, 12] for more information) and TP model transformation is capable of systematic generation of different convex hulls for tensor functions and qLPV models. In this way, the TP model transformation introduces an additional possibility for multi-objective control optimization techniques, namely the convex hull manipulation-based optimization, which is the key property of TP model transformation in control design. The properties of TP model transformation are given below [13, 14, 15]:

- It can be executed uniformly (irrespective of whether the model is given in an analytical form, as an outcome of soft computing-based identification techniques or as a result of a black-box identification) in a routine fashion without analytical interaction, within a reasonable amount of time.
- It generates the HOSVD-based canonical form of qLPV models. This is a unique representation. This form extracts the unique structure of a given qLPV model in the same sense as the HOSVD does for tensors and matrices, in such a way that: the number of LTI components are minimized; the weighting functions are one variable functions of the parameter vector in an orthonormal system for each parameter; the LTI components are also in orthogonal position; the LTI systems and the weighting functions are ordered according to the higher-order singular values of the parameter vector; unique form.
- TP model transformation was extended to generate different types of convex polytopic models. This was motivated by the fact that the type of convex hull of the polytopic models considerably influences the feasibility of the LMI-based design and the resulting performance. This means that instead of developing new LMI equations for feasible controller design, we may rather focus on the systematic modification of the convex hull ([10, 11]).
- Based on the higher-order singular values (that express the rank properties of the given qLPV model for each element of the parameter vector in  $L_2$  norm), the TP model transformation offers a trade-off between the complexity of further LMI design and the accuracy of the resulting TP model.
- TP model transformation is executed before utilizing the LMI design. This means that when we start the LMI design we already have the global weight-

ing functions and during control we do not need to determine a local weighting of the LTI systems for feedback gains.

One can find several applications and related work for TP model transformation in [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. Chumalee et al., Rangajeeva et al., Gai et al., Sun et al. and Qin et al. introduce TP model transformation based novel approaches in avionics related control problems [29, 30, 31, 32], thus leading to pioneering conceptual frameworks.

The first step in fitting friction compensation to LMI-based multi-objective control design methodology is to check whether the most commonly used friction models can be defined by a finite element polytopic form. The second step is to bring friction phenomenon and friction models to the same polytopic structure with the other elements of the control system, which makes LMI-based multi-objective control design methodology available for friction compensation in the same conceptual level.

The paper is organized as follows: Section 2 gives a brief introduction to the commonly used friction models, Section 3 introduces the proposed methodology to transform friction models to polytopic form and LMI-based multi-objective control design. Section 4 lists the polytopic form of the commonly used friction models and Section 5 deals with the friction compensation of an example dynamic system, namely 2 degrees-of-freedom aeroelastic wing section. The conclusions finish the paper.

## 2 Commonly Used Friction Models

Friction is a physical phenomenon and expressed in quantitative terms as a force  $F_f$ , being the force exerted by either of two contacting bodies tending to oppose relative tangential displacement of the other [33]. The following friction models are commonly used in the control of mechatronic systems.

**Definition 1** (*Coulomb friction force*): The Coulomb friction force is a force of constant magnitude, acting in the direction opposite to motion  $v(t)$ . When  $v(t) \neq 0$ :

$$F_f(t) = -F_c \text{sign}(v(t)), \quad F_c(t) = \mu F_n, \quad (1)$$

where  $F_n$  is the normal component of the force pressing surfaces together and  $\mu$  is the frictional coefficient.  $\mu$  is determined by measurements under certain conditions.

**Definition 2** (*Viscous friction*): The viscous friction models the friction force as a force proportional to sliding velocity. When  $v(t) \neq 0$ :

$$F_f(t) = -F_v v(t), \quad (2)$$

where  $F_v$  is the coefficient of viscous friction.

**Definition 3** (Static friction): The static friction is the influence of an external force for the friction at rest, however, this leads to a discontinuous function.

**Definition 4** (The Stribeck model) The Stribeck friction model is defined by the following equation, when  $v(t) \neq 0$  [34]:

$$F_f(t) = - \left( F_c + (F_s - F_c) e^{-\left| \frac{v}{v_s} \right|^\delta} \right) \text{sign}(v(t)) - F_v v, \quad (3)$$

where  $v_s$  is the Stribeck velocity,  $\delta$  is an empirical parameter,  $F_s$  is the static friction.

**Definition 5** (The LuGre model) The LuGre model is a dynamic friction model presented in [35]. Friction is modeled as the average deflection force of elastic springs. When a tangential force is applied the bristles will deflect like springs. If the deflection is sufficiently large enough, the bristles start to slip. The LuGre model has the form

$$\frac{dz}{dt} = v - \delta_0 \frac{|v|}{g(v)} z, \quad F = \delta_0 z + \delta_1(v) \frac{dz}{dt} + f(v), \quad (4)$$

where  $z$  denotes the average bristle deflection. The parameter  $\delta_0$  is the stiffness of the bristles, and  $\delta_1(v)$  the damping. The function  $g(v)$  models the Stribeck effect, and  $f(v)$  the viscous friction. A reasonable choice of  $g(v)$ , which gives a good approximation of the Stribeck effect, is  $g(v) = \alpha_0 + \alpha_1 e^{-(v/v_0)^2}$ . The sum  $\alpha_0 + \alpha_1$  then corresponds to stiction force and  $\alpha_0$  to Coulomb friction force. The parameter  $v_0$  determines how  $g(v)$  varies within its bounds  $\alpha_0 < g(v) \leq \alpha_0 + \alpha_1$ . A common choice of  $f(v)$  is linear viscous friction  $f(v) = \alpha_2 v$ . The main properties of the LuGre friction model are also listed in [36], who also applied the model.

### 3 TP Model Transformation

Assume we have the qLPV model as:

**Definition 6** (qLPV model): Consider the quasi Linear Parameter Varying State Space model:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(\mathbf{p}(t))\mathbf{x}(t) + \mathbf{B}(\mathbf{p}(t))\mathbf{u}(t) \quad (5)$$

$$\mathbf{y}(t) = \mathbf{C}(\mathbf{p}(t))\mathbf{x}(t) + \mathbf{D}(\mathbf{p}(t))\mathbf{u}(t),$$

with input  $\mathbf{u}(t) \in \mathbb{R}^m$ , output  $\mathbf{y}(t) \in \mathbb{R}^l$  and state vector  $\mathbf{x}(t) \in \mathbb{R}^k$ . The system matrix

$$\mathbf{S}(\mathbf{p}(t)) = \begin{pmatrix} \mathbf{A}(\mathbf{p}(t)) & \mathbf{B}(\mathbf{p}(t)) \\ \mathbf{C}(\mathbf{p}(t)) & \mathbf{D}(\mathbf{p}(t)) \end{pmatrix} \quad (6)$$

is a parameter-varying object, where  $\mathbf{p}(t) \in \Omega$  is time varying  $N$ -dimensional parameter vector, where  $\Omega = [a_1, b_1] \times [a_2, b_2] \times \dots \times [a_N, b_N] \in \mathbb{R}^N$  is a closed hypercube.  $\mathbf{p}(t)$  can also include some elements of  $\mathbf{x}(t)$ , in this case (6) is termed as quasi LPV (qLPV) model and as such, it belongs to the class of non-linear models.

**Definition 7** (Finite element TP type polytopic model):  $\mathbf{S}(\mathbf{p}(t))$  in (5) is given for any parameter as the parameter-varying combination of LTI system matrices  $\mathbf{S}_r \in \mathbb{R}^{(m+k) \times (m+k)}$  as:

$$\mathbf{S}(\mathbf{p}(t)) = S \boxtimes_{n=1}^N \mathbf{w}(p_n(t)) \quad (7)$$

by applying the compact notation based on tensor algebra (Lathauwer's work [6]), where the  $(N+2)$  dimensional coefficient tensor  $\mathbf{S} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N \times (m+k) \times (m+k)}$  is constructed from the LTI vertex systems  $\mathbf{S}_{i_1, i_2, \dots, i_N}$  (7) and the row vector  $w_n(p_n(t)) \in [0, 1]$  contains one variable and continuous weighting functions  $w_{n, i_n}(p_n(t))$ , ( $i_n = 1 \dots I_N$ ).

**Definition 8** (Convex type TP model): The TP model is SN (Sum Normalized) if the sum of the weighting functions for all  $p \in \Omega$  is 1 and it is NN (Non-Negative) if the values of the weighting functions for all  $p \in \Omega$  are non-negative. The TP model is convex if it is SNNN:

$$\forall n, i, p_n(t) : w_{n, i}(p_n(t)) \in [0, 1]; \quad \forall n, p_n(t) : \sum_{i=1}^{I_n} w_{n, i}(p_n(t)) = 1 \quad (8)$$

It was shown in [10, 37] that the type of the convex hull has a large influence on the feasibility and control performance of the LMI-based convex optimization, therefore, various types of convex hulls are defined [38]:

**Definition 9** (NO/CNO, Normal type TP model): The convex TP model is a NO (normal) type model, if its  $w(p)$  weighting functions are Normal, that is, if it satisfies (8) and the largest value of all weighting functions is 1. Also, it is CNO (close to normal), if it satisfies (8) and the largest value of all weighting functions is 1 or close to 1.

**Definition 10** (IRNO, Inverted and Relaxed Normal type TP model): The TP model is IRNO type if the smallest values of all weighting functions are 0, and the largest values of all weighting functions are the same.

The basic definitions and the steps of TP model transformation are also given in [37, 13, 39, 40, 41, 42, 43].

### 3.1 Steps of TP Model Transformation

The paper presents a TP model transformation of functions describing nonlinear friction models. Therefore, the steps of TP model transformation of functions are given, which can be generalized for transforming qLPV systems to convex polytopic form.

#### Step 1: Discretization.

The goal of this step is to represent the given function ( $y = f(\mathbf{x})$ ) or qLPV model ( $\mathbf{S}(\mathbf{p}(t))$ ) by its discretized tensor  $\mathcal{F}^{D(\Omega, M)}$ .

We define the transformation space  $\Omega$  in which we expect the TP function to be relevant and the hyper-rectangular grid  $M$ . As a result of discretizing the function  $y = f(\mathbf{x})$  over the grid points we have  $\mathcal{F}^{D(\Omega, M)} \in \mathbb{R}^{M_1 \times \dots \times M_N}$ .

#### Step 2: Extracting the discretized TP function.

The goal of this step is to reveal the TP structure of the given function. We use HOSVD to find the TP structure of the function.

As a result of STEP 1, we have the discretized system tensor  $\mathcal{F}^{D(\Omega, M)}$ . The CHOSVD results in  $\mathcal{F}^{D(\Omega, M)}$ :

$$\mathcal{F}^{D(\Omega, M)} = \mathcal{D} \boxtimes_{n=1}^N \mathbf{U}_n = \mathcal{D} \boxtimes_{n=1}^N \mathbf{w}_n^{D(\Omega, M)}, \quad (9)$$

where the size of  $\mathcal{D}$  is  $R_1 \times R_2 \times \dots \times R_N$ . Since  $R_n = \text{rank}_n(\mathcal{F}^{D(\Omega, M)})$ ,  $R_n \leq M_n$ , for all  $n = 1..N$ . Tensor  $\mathcal{D}$  contains the constant components  $\mathbf{D}_{i_1, i_2, \dots, i_N}$ ,  $i_n = 1..I_n$ . Based on the previous section, we have  $\mathbf{U}_n = \mathbf{w}_n^{D(\Omega, M)}$ .

#### Step 3: Reconstruction of the continuous TP function.

The weighting functions can be determined over any points of intervals  $[a_n, b_n]$  by the help of the given  $y = f(\mathbf{x})$ . In order to determine the weighting functions in vector  $\mathbf{w}_d(x_d)$ , let  $x_n$  be fixed to selected grid-lines as:

$$x_n = g_{n, i_n} \quad n = 1 \dots N, \quad n \neq d, \quad i_n \in \{1 \dots I_n\},$$

where  $i_n$  can be chosen arbitrarily. Then for  $x_d$ :

$$\mathbf{w}_d(x_d) = (\mathbf{y}(\mathbf{x}))_{(3)} \left( \left( \mathcal{B} \boxtimes_{n \neq d}^N \mathbf{u}_{n, i_n} \right)_{(d)} \right)^+, \quad (10)$$

where vector  $\mathbf{x}$  consists of elements  $x_n$  and  $x_d$  as  $\mathbf{x} = (g_{1, i_1} g_{2, i_2} \dots g_d \dots g_{N, i_N})$ , and superscript “+” denotes pseudo inverse and  $\mathbf{u}_{n, i_n}$  is the first row vector of  $\mathbf{U}_n$ . The third-mode matrix  $(\mathbf{y}(\mathbf{x}))_{(3)}$  of function  $\mathbf{y}(\mathbf{x})$  is understood such that function  $\mathbf{y}(\mathbf{x})$  is considered as a three-dimensional tensor, where the length of the third-dimension is one. This in practice means that the function  $\mathbf{y}(\mathbf{x})$  is stored in a one-row vector.

## 4 TP Type Polytopic Form of Some Commonly Used Friction Models

### 4.1 TP Type Polytopic Form of the Coulomb Friction Model

The Coulomb friction in Definition 1 is described by (1).

Due to finite precision of numerical calculations it is necessary to involve an arbitrary parameter to handle the sharp transition between  $v = 0$  and  $v \neq 0$ . A solution is to approximate the signum function in the following way:

$$\text{sign}(v) = \frac{2}{(1 + e^{-1000v})} - 1, \quad (11)$$

where exponent 1000 can be arbitrarily set. The value does not influence the structure of the model, only the slope of the resulting weighting functions.

Based on Definition 1 and the approximation of the signum function as in (11), we can execute the TP model transformation on the Coulomb friction given in the following form:

$$F_{Coulomb} = f(v) = -F_c \left( \frac{2}{(1 + e^{-1000v})} - 1 \right). \quad (12)$$

The aim is to give the TP type polytopic form of the Coulomb friction as:

$$F_{Coulomb}(v) = \mathcal{B} \boxtimes_{n=1}^N \mathbf{w}_n(v), \quad (13)$$

where  $F_{Coulomb}$  is scalar and  $v \in \mathbb{R}$ ,  $\mathcal{B} \in \mathbb{R}^{I_1 \times \dots \times I_N}$  is bounded and the row vector  $\mathbf{w}_n(v) \in [0, 1]$  contains one variable and continuous weighting functions  $w_{n,i_n}(v)$ , ( $i_n = 1 \dots I_N$ ). As a first step of the TP model transformation we have to define the transformation space  $\Omega$ . We expect the model to be valid in speed region from  $a$  m/s to  $b$  m/s. Therefore, we define the transformation space as  $\Omega = [a, b]$  (note that these intervals can be arbitrarily defined). TP model transformation starts with the discretization over a rectangular grid. Let the density of the discretization grid be  $M$ .

#### 4.1.1 Convex Type TP Models of the Coulomb Friction

LMI-based multi objective controller design requires the friction model  $F_f(v)$  to be given in a multi-dimensional convex hull, thus we construct various multi-dimensional convex hulls of  $F_f(v)$  from the HOVSD-based canonical form of the Coulomb friction.

**Definition 11** (Convex model of the Coulomb friction): By applying convexity criteria (SNNN type convex hull), we get that the Coulomb friction  $F_f(v)$  of (12) can be reconstructed from 2 LTI vertex systems as:

$$F_{Coulomb}(v) = \sum_{i=1}^2 w_i^{snnn}(v) B_i^{snnn} = \mathcal{B}_{snnn} \boxtimes_{n=1}^N \mathbf{w}_{n_{snnn}}(v). \quad (14)$$

**Definition 12** (CNO type convex model of the Coulomb friction): The CNO type convex hull represents a tight hull, which is advantageous for controller design.

$$F_{Coulomb}(v) = \sum_{i=1}^2 w_i^{cno}(v) B_i^{cno} = \mathcal{B}_{cno} \boxtimes_{n=1}^N \mathbf{w}_{n_{cno}}(v). \quad (15)$$

**Definition 13** (Convex model of the Coulomb friction): The IRNO type convex hull represents a large convex hull.

$$F_{Coulomb}(v) = \sum_{i=1}^2 w_i^{irno}(v) B_i^{irno} = \mathcal{B}_{irno} \boxtimes_{n=1}^N \mathbf{w}_{n_{irno}}(v). \quad (16)$$

The weighting functions of the canonical form and different convex type TP models of an academic Coulomb friction are given in Figure 1.

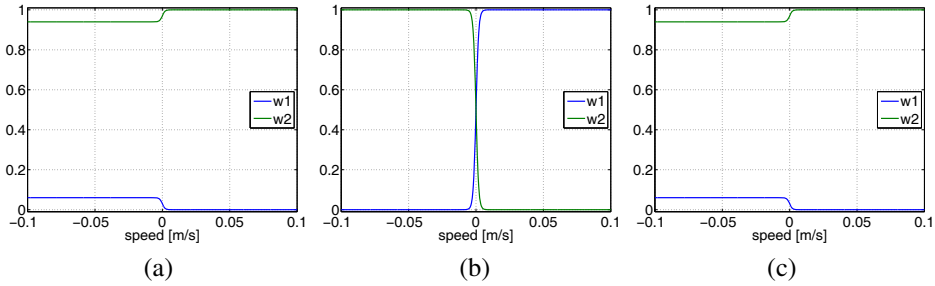


Figure 1

Weighting functions of the Coulomb friction model a) SNNN type convex model; (b) CNO type convex model; (c) IRNO type convex model;

## 4.2 TP Type Polytopic Form of the Coulomb and Viscous Friction Model

Based on Definition 2 and the approximation of the signum function as in (11), we can execute the TP model transformation on the viscous and Coulomb friction model given in the following form:

$$F_f = f(v) = -F_c \left( \frac{2}{(1 + e^{-1000v})} - 1 \right) - F_v v. \quad (17)$$

The aim is to give the TP type polytopic form of the viscous and Coulomb friction as:

$$F_{Coulomb+viscous}(v) = \mathcal{B} \boxtimes_{n=1}^N \mathbf{w}_n(v), \quad (18)$$

where  $F_{Coulomb+viscous}$  is scalar and  $v \in \mathbb{R}$ ,  $\mathcal{B} \in \mathbb{R}^{I_1 \times \dots \times I_N}$  is bounded and the row vector  $\mathbf{w}_n(v) \in [0, 1]$  contains one variable and continuous weighting functions  $w_{n,i_n}(v)$ ,  $(i_n = 1 \dots I_N)$ .

### 4.3 TP Type Polytopic Form of the Stribeck Friction Model

Based on Definition 4 and the approximation of the signum function as in (11), we can execute the TP model transformation on the Stribeck friction model given as:

$$F_f = f(v) = - \left( F_c + \frac{(F_s - F_c)}{\left( 1 + \left( \frac{v}{v_s} \right)^2 \right)} \right) \left( \frac{2}{(1 + e^{-1000v})} - 1 \right) - F_v v. \quad (19)$$

The aim is to give the TP type polytopic form of the Stribeck friction as:

$$F_{Stribeck} = \mathcal{B} \boxtimes_{n=1}^N \mathbf{w}_n(v), \quad (20)$$

where  $F_{Stribeck}$  is scalar and  $v \in \mathbb{R}$ ,  $\mathcal{B} \in \mathbb{R}^{I_1 \times \dots \times I_N}$  is bounded and the row vector  $\mathbf{w}_n(v) \in [0, 1]$  contains one variable and continuous weighting functions  $w_{n,i_n}(v)$ ,  $(i_n = 1 \dots I_N)$ . The weighting functions can be seen in Figure 2.

### 4.4 TP Type Polytopic Form of the LuGre Friction Model

Based on Definition 5, we can give the LuGre friction in a form suitable for TP model transformation as:

$$F_f = f(\mathbf{x}) = \delta_0 z + \delta_1 \left( v - \frac{\delta_0 |v|}{F_c + (F_s - F_c) e^{-|v/v_s| z}} \right) + \delta_2 v. \quad (21)$$

The aim is to give the TP type polytopic form of the LuGre friction as:



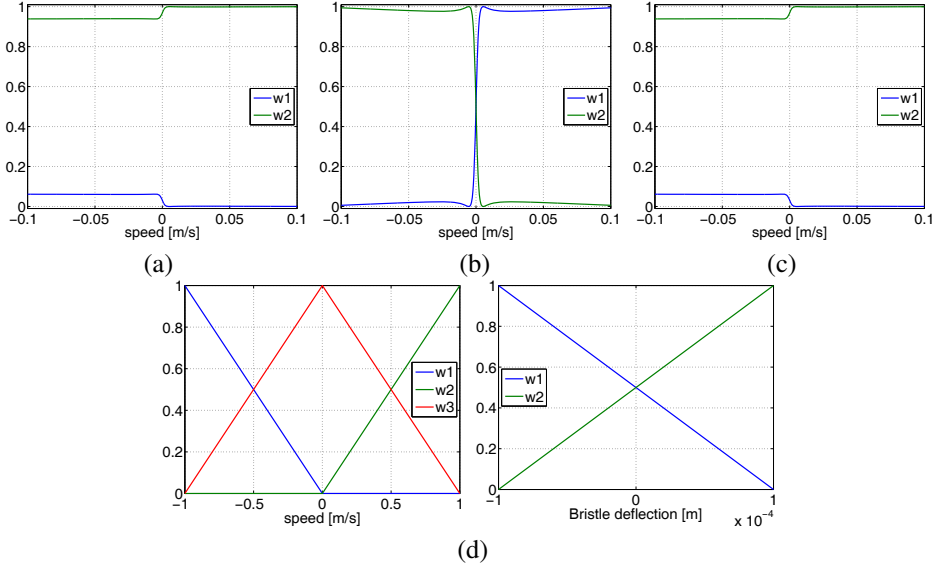


Figure 2

Weighting functions of the Stribeck and LuGre friction models a) SNNN type convex Stribeck model; (b) CNO type convex Stribeck model; (c) IRNO type convex Stribeck model; (d) CNO type convex LuGre model

$$F_{LuGre} = \mathcal{B} \boxtimes_{n=1}^N \mathbf{w}_n(\mathbf{x}), \quad (22)$$

where  $F_{LuGre}$  is scalar and  $\mathbf{x} \in \mathbb{R}^2$ ,  $\mathcal{B} \in \mathbb{R}^{I_1 \times \dots \times I_N}$  is bounded and the row vector  $\mathbf{w}_n(\mathbf{x}) \in [0, 1]$  contains one variable and continuous weighting functions  $w_{n,i_n}(\mathbf{x})$ , ( $i_n = 1 \dots I_N$ ). Vector  $\mathbf{x}$  contains the velocity  $v$  and  $z$ , which denotes the average bristle deflection.  $F_c$  denotes the Coulomb friction,  $F_s$  the static friction,  $v_s$  the Stribeck velocity,  $\delta_0$  the stiffness of the bristles,  $\delta_1$  the microdamping and  $\delta_2$  the viscous friction. We expect the model to be valid in a speed region from  $a_1$  m/s to  $b_1$  m/s and average bristle deflection from  $a_2$  m to  $b_2$  m. After executing the steps of TP model transformation we get the convex polytopic form of the LuGre friction, which can be exactly reconstructed with minimum 6 LTI vertex models as:

$$F_{LuGre}(v, z) = \sum_{i=1}^3 \sum_{j=1}^2 w_{1,i}(v) w_{2,j}(z) B_{i,j} = \mathcal{B} \boxtimes_{n=1}^2 \mathbf{w}_n(p_n(t)), \quad (23)$$

where  $B_{i,j}$  contains the vertex points of tensor  $\mathcal{B} \in \mathbb{R}^{3 \times 2}$ . The same types of convex hull can also be constructed for the LuGre model, Figure 2 shows the weighting functions of the CNO type convex form.

## 5 Example Friction Compensation: 2 DoF Aeroelastic Wing Section

### 5.1 qLPV Model of the Aeroelastic Wing with Nonlinear Friction

The aeroelastic wing model with nonlinear friction was validated by experiments in [44]. Similar models with linear friction were used in [45] [46], [47] and [48]. The parameters and the equations of motion with qLPV representation of the 2 DoF prototypical aeroelastic wing section are given in [47]. The difference between the present qLPV model and the qLPV model given in [47] lies in the friction. Authors have reported discrepancies between the experimental measurements and the analytical analysis [49]. These discrepancies can most likely be accounted for in the Coulomb damping forces that occur within the pitch bearing and plunge slider motion that is not taken into account in many of the models [50]. The equations for the plunge and pitch damping forces are as follows:

$$F_h = \mu_h mg |\dot{h}| \dot{h}^{-1}, \quad F_\alpha = \mu_\alpha M_f |\dot{\alpha}| \dot{\alpha}^{-1}. \quad (24)$$

$M_f$  is the frictional moment due to the nonlinear cam and  $\mu_h, \mu_\alpha$  are the frictional coefficients. Equation (24) is identical to (1), thus we apply the Coulomb friction from Definition 1. The parameters of the friction models are given in [50].

### 5.2 Execution of TP Model Transformation

First of all we define the transformation space  $\Omega : [14, 25] \times [-0.1, 0.1] \times [-2.5, 2.5] \times [-2.5, 2.5]$  in which we expect the TP model be relevant, then we discretize the qLPV model in  $M_1 \times M_2 \times M_3 \times M_4$  points, where  $M_1 = M_2 = M_3 = M_4 = 102$ .

In this paper we generate the exact minimized form, this means that we eliminate only the zero singular values. The number of non-zero singular values on the first dimension is 3 and on the second, third and fourth dimension is 2. The numerical values are the following respectively: 519297, 196869, 401.154, 469459, 296706, 555361, 1009.52, 555361 and 51.7953. We found that the model can be exactly given in TP type polytopic form with the combination of 24 LTI vertex systems. The CNO type weighting functions can be seen in Figure 3.

### 5.3 Controller and Observer Design for the Aeroelastic Wing with Nonlinear Friction

LMI based multi-objective, robust control design can be executed immediately on convex TP type polytopic models [9]. Therefore, the aim is to transform the friction models to convex polytopic form via TP model transformation. An LTI feedback

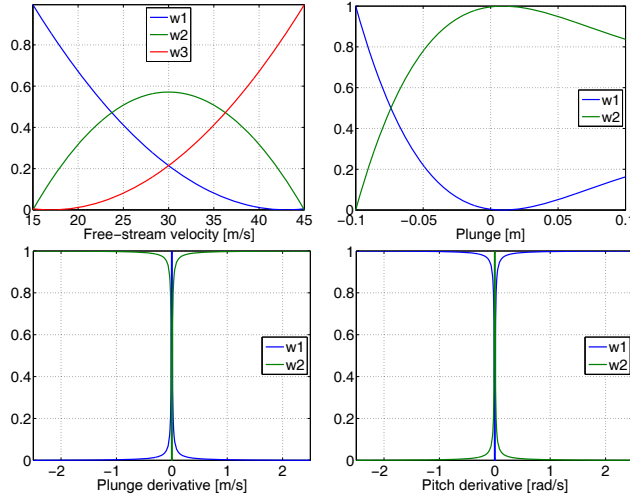


Figure 3

CNO type weighting functions of the aeroelastic model with nonlinear friction

gain is determined for each LTI vertex system of a given convex TP model based on feasibility test of the selected LMIs. In practical applications the state of the system is often not readily available. In this case one has to define the polytopic observer structure. There are various alternative ways to achieve output feedback and observer design (in this regard it is referred to [51, 52]). The observers are required to satisfy the following:

$$\mathbf{x}(t) - \hat{\mathbf{x}}(t) \rightarrow 0 \quad \text{as } t \rightarrow \infty,$$

where  $\hat{\mathbf{x}}(t)$  denotes the state-vector estimated by the observer. This condition guarantees that the steady-state error between  $\mathbf{x}(t)$  and  $\hat{\mathbf{x}}(t)$  converges to 0. The present case is restricted to systems in which  $\mathbf{p}(t)$  does not contain values from the estimated state-vector  $\hat{\mathbf{x}}(t)$ . The following strategy for controller and observed design can be used in this case:

$$\begin{aligned} \hat{\dot{\mathbf{x}}}(t) &= \mathbf{A}(\mathbf{p}(t))\hat{\mathbf{x}}(t) + \mathbf{B}(\mathbf{p}(t))\mathbf{u}(t) + \mathbf{K}(\mathbf{p}(t))(\mathbf{y}(t) - \hat{\mathbf{y}}(t)) \\ \hat{\mathbf{y}}(t) &= \mathbf{C}(\mathbf{p}(t))\hat{\mathbf{x}}(t). \end{aligned}$$

The polytopic model form of this structure is:

$$\begin{aligned}
\hat{\mathbf{x}}(t) &= \mathcal{A} \bigboxtimes_{n=1}^N \mathbf{w}_n(p_n(t)) \hat{\mathbf{x}}(t) + \mathcal{B} \bigboxtimes_{n=1}^N \mathbf{w}_n(p_n(t)) \mathbf{u}(t) + \\
&\quad + \mathcal{K} \bigboxtimes_{n=1}^N \mathbf{w}_n(p_n(t))_r (\mathbf{y}(t) - \hat{\mathbf{y}}(t)) \\
\hat{\mathbf{y}}(t) &= \mathcal{C} \bigboxtimes_{n=1}^N \mathbf{w}_n(p_n(t)) \hat{\mathbf{x}}(t).
\end{aligned} \tag{25}$$

The goal of the controller and observer design is to determine gains  $\mathcal{F}$  and  $\mathcal{K}$  in such a way that the stability of the observer and the controller is simultaneously guaranteed. The control signal in output feedback control gets the following form:

$$\mathbf{u}(t) = -\mathcal{F} \bigboxtimes_{n=1}^N \mathbf{w}_n(p_n(t)) \hat{\mathbf{x}}(t). \tag{26}$$

There are several LMI theorems available for observer and controller design to derive the vertex gains  $\mathcal{K}$  of the observer and the feedback gains  $\mathcal{F}$  of the controller. The feasibility test of the following LMI was executed in the example case:

**Theorem 1** (*Globally and asymptotically stable observer and controller with decay rate*) Based on LMIs presented in [52] one can derive the following LMI:

$$\begin{aligned}
\mathbf{P}_1 \mathbf{A}_r^T - \mathbf{M}_{1,r}^T \mathbf{B}_r^T + \mathbf{A}_r \mathbf{P}_1 - \mathbf{B}_r \mathbf{M}_{1,r} + 2\alpha \mathbf{P}_1 &< \mathbf{0}, \\
\mathbf{A}_r^T \mathbf{P}_2 - \mathbf{C}_r^T \mathbf{N}_{2,r}^T + \mathbf{P}_2 \mathbf{A}_r - \mathbf{N}_{2,r} \mathbf{C}_r + 2\alpha \mathbf{P}_2 &< \mathbf{0}, \\
\mathbf{P}_1 \mathbf{A}_r^T - \mathbf{B}_s \mathbf{M}_{1,s} - \mathbf{M}_{1,s}^T \mathbf{B}_r^T + \mathbf{A}_s \mathbf{P}_1 - \mathbf{B}_r \mathbf{M}_{1,s} + \mathbf{P}_1 \mathbf{A}_s^T - \mathbf{M}_{1,r}^T \mathbf{B}_s^T + \mathbf{A}_s \mathbf{P}_1 + 4\alpha & \\
\mathbf{P}_1 &< \mathbf{0}, \\
\mathbf{A}_r^T \mathbf{P}_2 - \mathbf{C}_s^T \mathbf{N}_{2,r}^T + \mathbf{P}_2 \mathbf{A}_r - \mathbf{N}_{2,r} \mathbf{C}_s + \mathbf{A}_s^T \mathbf{P}_2 - \mathbf{C}_r^T \mathbf{N}_{2,s}^T + \mathbf{P}_2 \mathbf{A}_s - \mathbf{N}_{2,s} \mathbf{C}_r + 4\alpha \mathbf{P}_2 &< \mathbf{0},
\end{aligned}$$

which leads to an asymptotically stable observer and controller decay rate.

One can apply further LMIs in order to guarantee various additional constraints.

**Remark 1** *Convex hull manipulation is also an important step of the optimization, but it is out of the scope of this paper. More information about convex hull manipulation in general and with a special focus on the aeroelastic wing section model can be found in [10].*

The aim of the controller is to stabilize the 2 DoF prototypical aeroelastic wing section in the flutter velocity range. The controller has to satisfy the following criteria: The closed-loop system is stable (in this case we aim at achieving asymptotic stability); The closed-loop system exhibits good settling behavior for a class of initial conditions. The aeroelastic model was transformed to finite element convex TP

model form in the previous section, upon which LMI-based control design can immediately be executed. In the present case  $U$ , the plunge deflection ( $x_1(t)$ ) and pitch angle ( $x_2(t)$ ) are measurable, but the unavailable state values  $x_3(t)$  and  $x_4(t)$  have to be estimated. Solving LMIs of Theorem 1 simultaneously leads to the control and observer gains in the form of  $\mathcal{F}$  and  $\mathcal{K}$  respectively. As mentioned earlier, convex hull manipulation is an important optimization technique (see [10] for more details). Since it is out of scope of the paper, based on the results of [10] we conclude that the best hull type for our case is the CNO type (see Definition 9).

## 5.4 Validation of the Controller Design

The numerical examples are performed with  $a = -0.4$  and with free stream velocity  $U = 20m/s$ , a velocity which exceeds the linear flutter velocity  $U = 15.5m/s$ , and for initials  $h = 0.01m$  and  $\alpha = 0.1rad$ . Figure 4 shows the time response of the controlled system. An important issue should be addressed here. Theorem 1 claims that the resulting controller is globally stable. However, the TP-model transformation is a numerical method that can be performed over an arbitrary, but bounded, domain  $\Omega$ . Therefore, the global stability, ensured by Theorem 1, is restricted to  $\Omega$  and as such it is termed quasi global stability. This, however, has practical significance because the accuracy of a given model is also bounded in reality. In the present example the prototypical aeroelastic model is accurate only for low speeds, and we have defined  $\Omega$  accordingly in the design process. The performance of the controller is evaluated based on the following aspects: settling time of the pitch - 0.2s; number of oscillations of the pitch - no oscillation; maximal overshoot of the pitch - no overshoot; settling time of the plunge - 0.75s; number of oscillations of the plunge - 4; maximal overshoot of the plunge - 110 % of the original error; maximal torque of the controller - 52 Nm. Comparing the control results one can observe that the controllers developed in this paper are considerable faster than in [50]. We can notice that the friction compensation is well managed, the controller results are similar as in [47, 53], where the same controller design method was utilized but the aeroelastic model lacks nonlinear friction.

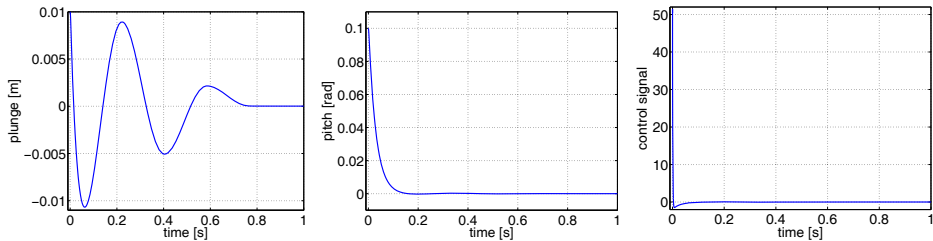


Figure 4

Time response of the designed controller for  $U = 20m/s$  and  $a = -0.4$

## Conclusions

The most commonly used friction models were transformed to a convex polytopic form via TP model transformation. In this way, friction compensation can be fit to a LMI-based, modern and multi-objective control design. All of the presented friction models can be decomposed to a finite element polytopic form.

The extension of qLPV systems with nonlinear friction models increases the number of the dimension of the system. However, extending the qLPV system with the presented friction model does not lead to explosion in the number of necessary vertex systems in the polytopic form and preserves the exact control design, which is the main theoretical result of the paper.

The effectiveness of the proposed friction compensation methodology was presented through the qLPV model of an aeroelastic wing section with Coulomb friction. The qLPV model was transformed to TP type polytopic model on which LMI based controllers can be designed. The polytopic decomposition of such a model is possible and the model can be exactly described by 24 LTI vertex points. The designed controller and observer was validated by simulation and it was compared to the controllers of other authors, dealing with the same problem. We can say that the designed controller performance is on par with the controller designed for the system with linear damping and it is faster than the controllers for nonlinear damping found in the current literature.

## Acknowledgement

The research was supported by the National Research and Technology Agency, (ERC\_09) (OMFB-01677/2009) (ERC-HU-09-1-2009-0004 MTA SZTAKI) and Control Research Group of Hungarian Academy of Science.

## References

- [1] C. C. de Wit, H. Olsson, K. J. Astrom, and P. Lischinsky, "Dynamic friction models and control design," in *In Proceedings of the American Control Conference*, 1993, pp. 1920–1926.
- [2] H. Olsson, *Control Systems with Friction. PhD thesis.* Lund Institute of Technology, University of Lund, 1996.
- [3] P. Apkarian and P. Gahinet, "A convex characterization of gain-scheduled  $H_\infty$  controllers," *IEEE Trans. Aut. Contr.*, 1995.
- [4] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory.* Philadelphia: SIAM books, 1994.
- [5] J. Bokor and G. Balas, "Linear matrix inequalities in systems and control theory," in *Perprints of 16th IFAC Word Congress*, 2005, p. Keynote lecture.

- [6] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, "A multi linear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.
- [7] P. Baranyi, L. Szeidl, P. Várlaki, and Y. Yam, "Definition of the HOSVD-based canonical form of polytopic dynamic models," in *3rd International Conference on Mechatronics (ICM 2006)*, Budapest, Hungary, July 3-5 2006, pp. 660–665.
- [8] —, "Numerical reconstruction of the HOSVD-based canonical form of polytopic dynamic models," in *10th International Conference on Intelligent Engineering Systems*, London, UK, June 26-28 2006, pp. 196–201.
- [9] P. Baranyi, "TP model transformation as a way to LMI based controller design," *IEEE Transaction on Industrial Electronics*, vol. 51, no. 2, April 2004.
- [10] P. K. P. Grof and P. Baranyi, "Different determination of the stability parameter space of a two dimensional aeroelastic system, a tp model based approach," *IEEE 14th International Conference on Intelligent Engineering Systems*, pp. 0–6, 2010.
- [11] P. Baranyi, "Output feedback control of two-dimensional aeroelastic system," *Journal of Guidance, Control, and Dynamics*, vol. 29, no. 3, pp. 762–767, May-June 2005.
- [12] P. Grof, P. Galambos, and P. Baranyi, "Convex hull manipulation based control performance optimization: Case study of impedance model with feedback delay," *IEEE*, Jan. 2012, pp. 495–499.
- [13] Z. Petres, "Polytopic decomposition of linear parameter-varying models by tensor-product model transformation," Ph.D. dissertation, Budapest, Hungary, November 2006.
- [14] S. Nagy, P. Baranyi, and Z. Petres, "Centralized tensor product model form," *IEEE*, Jan. 2008, pp. 189–193.
- [15] L. Szeidl and P. Várlaki, "Hosvd based canonical form for polytopic models of dynamic systems," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 13, no. 1, pp. 52–60, Jan. 2009.
- [16] R. Precup, L. Dioanca, E. M. Petriu, M. Radac, S. Preitl, and C. Dragos, "Tensor product-based real-time control of the liquid levels in a three tank system," in *2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, Montreal, QC, Canada, jul 2010, pp. 768–773.
- [17] Z. Szabó, P. Gáspár, and J. Bokor, "A novel control-oriented multi-affine qLPV modeling framework," in *Control Automation (MED), 2010 18th Mediterranean Conference on*, Jun. 2010, pp. 1019–1024.

- [18] C. Sun, Y. Huang, C. Qian, and L. Wang, "On modeling and control of a flexible air-breathing hypersonic vehicle based on LPV method," *Frontiers of Electrical and Electronic Engineering*, vol. 7, no. 1, pp. 56–68, 2012.
- [19] T. Luspay, T. Péni, and B. Kulcsar, "Constrained freeway traffic control via linear parameter varying paradigms," *Control of Linear Parameter Varying Systems With Applications*, p. 461, 2012.
- [20] P. Korondi, "Sector sliding mode design based on tensor product model transformation," in *Intelligent Engineering Systems, 2007. INES 2007. 11th International Conference on*, Jul. 2007, pp. 253–258.
- [21] R. Precup, C. Dragos, S. Preitl, M. Radac, and E. M. Petriu, "Novel tensor product models for automatic transmission system control," *IEEE Systems Journal*, p. In print., 2012.
- [22] S. Ilea, J. Matusko, and F. Kolonic, "Tensor product transformation based speed control of permanent magnet synchronous motor drives," in *17th International Conference on Electrical Drives and Power Electronics, EDPE 2011 (5th Joint Slovak-Croatian Conference)*, 2011.
- [23] B. Takarics and P. Baranyi, "TP Model-based Robust Stabilization of the 3 Degrees-of-Freedom Aeroelastic Wing Section," *Acta Polytechnica Hungarica*, vol. 12, no. 1, Feb. 2014.
- [24] P. Baranyi, "TP model transformation as a manipulation tool for qLPV analysis and design," *Asian Journal of Control*, vol. 17, no. 2, pp. 497–507, Mar. 2015.
- [25] J. Kuti, P. Galambos, and Á. Miklós, "Output feedback control of a dual-excenter vibration actuator via qLPV model and TP model transformation," *Asian Journal of Control*, vol. 17, no. 2, pp. 432–442, Mar. 2015.
- [26] S. Chumalee and J. F. Whidborne, "Gain-scheduled  $H_\infty$  control for tensor product type polytopic plants," *Asian Journal of Control*, vol. 17, no. 2, pp. 417–431, Mar. 2015.
- [27] R.-E. Precup, E. M. Petriu, M.-B. R?dac, S. Preitl, L.-O. Fedorovici, and C.-A. Drago?, "Cascade control system-based cost effective combination of tensor product model transformation and fuzzy control," *Asian Journal of Control*, vol. 17, no. 2, pp. 381–391, Mar. 2015.
- [28] T. T. Wang, W. F. Xie, G. D. Liu, and Y. M. Zhao, "Quasi-Min-Max model predictive control for image-based visual servoing with tensor product model transformation," *Asian Journal of Control*, vol. 17, no. 2, pp. 402–416, Mar. 2015.
- [29] S. Chumalee and J. Whidborne, *LPV Autopilot Design of a Jindivik UAV*. American Institute of Aeronautics and Astronautics, 1801 Alexander Bell Dr., Suite 500 Reston VA 20191-4344 USA,, 2009.



- [30] W. Qin, Z. Zheng, G. Liu, J. Ma, and W. Li, "Robust variable gain control for hypersonic vehicles based on LPV," *Systems Engineering and Electronics*, vol. 33, no. 6, pp. 1327–1331, 2011.
- [31] S. Rangajeeva and J. Whidborne, "Linear parameter varying control of a quadrotor," in *Industrial and Information Systems (ICIIS), 2011 6th IEEE International Conference on*, Aug. 2011, pp. 483–488.
- [32] W. Gai, H. Wang, T. Guo, and D. Li, "Modeling and LPV flight control of the canard rotor/ wing unmanned aerial vehicle," in *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on*, Aug. 2011, pp. 2187–2191.
- [33] P. Korondi, P. Bartal, and F. Kolonic, "Friction model based on tensor product transformation," in *7th International Symposium of Hungarian Researchers on Computational Intelligence, Budapest*, Budapest, 2006, pp. 83–94.
- [34] D. P. Hess and A. Soom, "Friction at a lubricated line contact operating at oscillating sliding velocities," *Journal of Tribology*, vol. 112, no. 1, pp. 147–152, 1990.
- [35] C. C. de Wit, H. Olsson, K. J. Astrom, and P. Lischinsky, "A new model for control of systems with friction," *IEEE Transactions on Automatic Control*, vol. 40, no. 3, pp. 419–425, 1995.
- [36] P. J. Dolcini, "Contribution to the clutch comfort," Ph.D. dissertation, Grenoble, France, May 2007.
- [37] P. Gróf, P. Baranyi, and P. Korondi, "Convex hull manipulation based control performance optimisation," *WSEAS Trans. Sys. Ctrl.*, vol. 5, no. 8, pp. 691–700, Aug. 2010.
- [38] P. Baranyi, "Convex hull generation methods for polytopic representations of LPV models." *IEEE*, Jan. 2009, pp. 69–74.
- [39] P. Z. Baranyi and Y. Yam, "Uniform observer design for linear parameter varying systems," in *Computational Intelligence. Proceedings of the 5th International Symposium of Hungarian Researchers, Budapest*. Budapest: BME, 2004, pp. 371–381.
- [40] B. Takarics, "Parallel distributed compensation based sector sliding mode control of takagi-sugeno type polytopic models." *IEEE*, Jan. 2012, pp. 501–506.
- [41] Z. Petres, B. Resko, and P. Baranyi, "Reference signal control of the TORA system: a TP model transformation based approach," vol. 2. *IEEE*, Jul. 2004, pp. 1081–1086.
- [42] Z. Petres and P. Z. Baranyi, "Approximation and complexity trade-off by tp model transformation in controller design: a case study of the tora system,"

- in *8th international symposium of Hungarian researchers on computational intelligence and informatics. Magyar kutatók 8. nemzetközi szimpóziuma. Budapest, 2007.*, Budapest, 2007, pp. 441–455.
- [43] P. Galambos, P. Z. Baranyi, and P. Korondi, “Extended tp model transformation for polytopic representation of impedance model with feedback delay,” *WSEAS Transactions on Systems and Control*, vol. 5, no. 9, pp. 701–710, September 2010.
- [44] J. J. Block and T. W. Strganac, “Applied active control for nonlinear aeroelastic structure,” *Journal of Guidance, Control, and Dynamics*, vol. 21, no. 6, pp. 838–845, November–December 1998.
- [45] Y. C. Fung, *An Introduction to the Theory of Aeroelasticity*. John Wiley and Sons, New York, 1955.
- [46] E. H. D. (Editor), H. C. J. Curtiss, R. H. Scanlan, and F. Sisto, *A Modern Course in Aeroelasticity*. Stifthoff and Noordhoff, Alpen aan den Rijn, The Netherlands, 1978.
- [47] P. Baranyi, “Tensor product model based control of 2-d aeroelastic system,” *Journal of Guidance, Control, and Dynamics*, vol. 29, no. 2, March–April 2006.
- [48] P. Marzocca and L. Librescu, “Aeroelastic response of nonlinear wing sections using a functional series technique,” *AIAA Journal*, vol. 40, no. 5, May 2002.
- [49] R. Duvigneau and M. Visonneau, “Optimization of a synthetic jet actuator for aerodynamic stall control,” *Laboratoire de Mécanique des Fluides CNRS UMR 6598*, 2005.
- [50] J. Block and H. Gilliatt, “Active control of an aeroelastic structure,” *AIAA Meeting Papers on Disc*, 1997.
- [51] C. W. Scherer and S. Weiland, *Linear Matrix Inequalities in Control*, ser. DISC course lecture notes, DOWNLOAD: <http://www.cs.ele.tue.nl/SWeiland/lmid.pdf>, 2000.
- [52] K. Tanaka and H. O. Wang, *Fuzzy Control Systems Design and Analysis - A Linear Matrix Inequality Approach*. John Wiley and Sons, Inc., 2001, 2001.
- [53] P. Baranyi, P. Korondi, and H. Hashimoto, “Global asymptotic stabilization of the prototypical aeroelastic wing section via TP model transformation,” *Asian Journal of Control*, vol. 7, no. 2, pp. 99–111, Oct. 2008.

# Performance Issues in Cloud Computing: KVM Hypervisor's Cache Modes Evaluation

**Borislav Đorđević<sup>1</sup>, Nemanja Maček<sup>2</sup>, Valentina Timčenko<sup>3</sup>**

<sup>1,3</sup> Mihailo Pupin Institute, University of Belgrade, 15 Volgina Street, 11060 Belgrade, Serbia; borislav.djordjevic@pupin.rs, valentina.timcenko@pupin.rs

<sup>2</sup> Department of Computer Technologies, The School of Electrical and Computer Engineering of Applied Studies, Vojvode Stepe 283, 11000 Belgrade, Serbia  
e-mail: nmacek@viser.edu.rs

---

*Abstract: This paper examines the performance of bare-metal hypervisors within the context of Quality of Service evaluation for Cloud Computing. Special attention is paid to the Linux KVM hypervisors' different cache modes. The main goal was to define analytical models of all caching modes provided by hypervisor according to the general service time equation. Postmark benchmark software is used for random performance testing with two different workloads, consisting of relatively small objects that simulate a typical mail server. Sequential performance is evaluated with Bonnie++ benchmark software. The experiments were conducted separately with a single virtual machine, and, two and three virtual machines running on the hypervisor. The interpretation of obtained benchmark results according to the proposed models is the main contribution of this research.*

---

*Keywords: cloud computing; virtualization; hypervisor; KVM; cache modes*

---

## 1 Introduction

Cloud computing (CC) is a concept of sharing hardware and software resources that are available on request. CC is based on virtualization, which allows hardware consolidation, provides resource isolation, leads to a higher level of security and reliability of the available IT infrastructure and significantly reduces maintenance costs [1-4]. The CC is closely related to the Quality of Service (QoS), which is a guaranteed level of performance and availability of service provided to users [5].

The hypervisor (virtual machine manager) is a software layer that creates and manages virtual machines. Guest operating systems (OS) are executed on virtual machines. There are the two types of hypervisors. Type-1 (also known as bare-metal or native) hypervisor is executed directly on the physical hardware, while type-2 (also known as hosted) hypervisor runs on the host OS, thus providing the guest OS with lower performance due to overhead produced by the host OS.

Linux Kernel-based Virtual Machine (KVM) [6] is type-1 hypervisor integrated into the Linux OS as a kernel module. This type of implementation allows KVM to follow modern kernel improvements. KVM uses a modified QEMU emulator for the block and network devices [7]. KVM provides a large number of tuneable parameters to the administrators, including three different caching modes, which differ in performance and achievable reliability.

In this paper, KVM hypervisor block device performance and reliability are examined. The obtained results are interpreted according to analytical models of workloads, reading, normal write cycles and flushing or direct writing in different caching modes. We have set up several preliminary research hypotheses on I/O performance, which were validated, along with the model, with the synthetic benchmarking. The main contribution of this paper is the proposed analytical modelling of workload and read/write (R/W) operations in the specific caching virtual environment, which allows us to make recommendations for the optimal KVM cache mode selection in specific situations.

## 2 Related Work

QoS in CC is a complex concept which includes many factors such as, performance, reliability, availability and security. These factors are mutually dependent, thus making the QoS evaluation a hard task. For example, security in CC is a research area that includes, but is not limited to the following issues: privileged user access, regulatory compliance, data location, data segregation, recovery, defence against the attacks and long-term viability. Availability and recovery are mutually dependent, as well as performance and reliability. Hence, analyzing all the factors would outreach the scope of this research.

The scope of this research is hypervisor performance evaluation as one of fundamental factors of the QoS. There are several different performance evaluation approaches reported in the literature that differ in methodology. For instance, the most common is the comparative performance analysis of Xen, KVM, VMware and OpenVZ hypervisors using different benchmark tools such as Bonnie++, IOzone, LINPACK and LMBench [8-12].

Some recent studies have focused on the emerging problem of fast input/output (I/O) support for a growing number of applications in the CC, thus targeting block device and general I/O performance analysis in virtual environment [9, 13-15]. Thus, it is often seen that the experimental results draw the attention mainly to the impact of performance overheads on the adoption of CC technology. Consequently, it is important to pay attention when making decisions or choices of virtual infrastructure management solutions, as there can be a rather limited capacity to react to changes on demand in stochastic and dynamic environments where not all the choices would be appropriate [9].

Another class of topics arise from the need for resource management optimization, virtual machine migration and resource replicas in CC [16-18]. These issues are highly correlated to the dynamic fluctuation of the system workload, further producing load imbalance, lower utilization and workload hotspots. Some of these approaches enforce the introduction of automatic management of virtualized resources in cloud environments, and the particular control system that would compute the necessary resource allocations for each used application, provide dynamic adjustment and virtual machine rearrangement in the cloud, mainly based on statistical machine learning techniques [17].

Studies that target a highly pervasive need for energy efficiency in the context of performance in CC are presented in [19-20]. These studies provide some fundamental insights on the impact of virtualization on energy usage, consider the energy overhead increase correlation with the increase of physical resources utilization, and also suggest some possibilities of CC server consolidation in data centers for reducing energy cost [19]. Alternatively, an approach for applying appropriate allocation schemes of dynamic requests for virtual servers is proposed for server farms applications [20], and efficient and secure use of system resources [21]. Additionally, CC technology can highly improve different business processes, e.g. in the context of universities where CC can provide a more intense data processing environment and enhance specific I/O quality dimensions [22].

The research presented in this paper belongs to the comparative performance analysis approach group. Although it is partially similar with the research presented in [9], which analyzes caching modes without any in-depth interpretation, all results presented here are interpreted according to the proposed analytical model. Best practices for block I/O performance and recommendations for the selection of appropriate KVM cache mode according to the type of storage that is used are discussed in [23]. For example, writethrough mode is recommended for local or direct-attached storage, as it ensures data integrity and provides acceptable I/O performance, while "none" mode is recommended for remote NFS storage, as it effectively turns all guest I/O operations into direct I/O operations on the host.

### 3 KVM Cache Modes

The operating system's page cache improves the disk read/write operation performance. Within the Kernel-based Virtual Machine (KVM) environment, both the host and the guest OS maintain their own page caches. The page cache is copied to a permanent storage using flushing (fsync), while direct I/O requests bypass the page cache. There is also a disk R/W cache, resulting in three independent caches. There are three caching modes available for KVM guest

operating systems – writethrough, write-back and none, resulting in three different write operation options:

- data will be cached in the host-side page cache if the cache mode is set to writeback;
- data will be immediately flushed to the physical disk cache if the cache mode is set to none;
- data will be immediately flushed to the physical disk platters if the cache mode is set to writethrough.

If KVM caching mode is set to write-back, both the host OS page cache and the disk write cache are enabled for the guest. QEMU-KVM interacts with the disk image file with writeback semantics for guest's flushing and direct write cycles: write operations are reported to the guest as completed when the data is placed in the host page cache. The guest's virtual storage controller is expected to send down flush commands. Guest OS application's I/O performance is good, but the data is not protected from power failures. As a result, writeback caching mode is recommended only if the potential data loss is not a major concern.

If KVM caching mode is set to none, the host OS page cache is disabled. QEMU-KVM interacts with the disk image file with writethrough semantics for guest's flushing and direct write cycles; the host page cache is bypassed and I/O is performed directly between the QEMU-KVM buffers and the storage device. Write operations are reported to the guest as completed when the data is placed in the disk R/W cache. The guest's virtual storage controller is expected to send down flush commands. Guest's write performance in this mode is expected to be optimal because the write operations bypass the host OS page cache and go directly to disk R/W cache. However, due to host OS page cache being disabled, the guest's read performance is not as good as in writeback and writethrough modes. This cache mode is suitable for guests with large I/O requirements, and is generally the best choice, as it is the only mode that supports migration.

KVM's writethrough mode enables different caches for reading and writing. Host OS page cache and the disk cache are enabled for the guest's reading operations. QEMU-KVM interacts with the disk image file with write-through semantics for guest's flushing and direct write cycles, and write operations are reported as completed only when the data has been fully committed to the storage device. The guest's virtual storage controller does not need to send down flush commands. Guest's application read performance is good as in the write-back mode, as the host OS page cache is enabled for reading. However, this mode provides lowest writing performance and is prone to scaling problems, because data is written through to the physical storage medium. This cache mode is suitable for systems with small number of guests that have low I/O requirements.

## 4 Modelling the Workload and Cache Modes

Performance characteristics of each workload are based on times required to complete read and write operations. Both reading and writing can be either random or sequential. Thus, the total workload processing time  $T_W$  is given by:

$$T_W = T_{RR} + T_{SR} + T_{RW} + T_{SW}, \quad (1)$$

where  $T_{RR}$  denotes random read time,  $T_{SR}$  sequential read time,  $T_{RW}$  random write time and  $T_{SW}$  sequential write time. For the specified workload, expected access time for the file system includes five components given by the following equation:

$$T_W = T_{DIR} + T_{META} + T_{FL} + T_{FB} + T_J, \quad (2)$$

where  $T_W$  is the total time needed to complete all operations on the workload,  $T_{DIR}$  the time needed to complete all directory related operations,  $T_{META}$  the time needed to complete all metadata operations,  $T_{FL}$  the time needed to complete all free lists operations,  $T_{FB}$  the time needed to complete direct file blocks operations and  $T_J$  the time needed to complete journaling operations.

General service time equation that can be applied to any caching system is:

$$T_{srv} = P_{HIT} \cdot T_{CACHEsrv} + P_{MISS} \cdot T_{CACHEsrv}, \quad (3)$$

where  $T_{CACHEsrv}$  denotes an effective disk access time with caching functionality (cache hit/miss service time), and  $P_{HIT}$  and  $P_{MISS}$  denote probabilities of hits and misses in the cache, respectively.

Cache service time is calculated as ratio of request size and cache transfer rate (the throughput of the cache).

### 4.1 Reading Operations in Different Cache Modes

Most of the timing equations presented here are based on the proper application of general service time equation (3).

Reading cycles in writeback and writethrough, as shown on Figure 1 (left), use all three caches. Application's reading service time is a function of guest OS page cache for the hit cycles, while host OS page cache and disk R/W cache are included in the miss cycles:

$$T_{srvR} \approx P_{HIT\_R\_guest} \cdot T_{srvR\_GuestPageCache} + P_{MISS\_R\_guest} \cdot T_{srvR\_VirtDisk}. \quad (4)$$

Virtual disk read service time is the function of host OS page cache for hit cycles, while disk R/W cache is included in the miss cycles:

$$T_{srvR\_VirtDisk} \approx P_{HIT\_R\_host} \cdot T_{srvR\_HostPageCache} + P_{MISS\_R\_host} \cdot T_{srvR\_PhysDisk}. \quad (5)$$

Physical disk service time is the function of disk R/W cache for hit cycles, while disk platters are included in the miss cycles:

$$T_{\text{srvR\_PhysDisk}} \approx P_{\text{HIT\_R\_disk}} \cdot T_{\text{srvR\_DiskCache}} + P_{\text{MISS\_R\_host}} \cdot T_{\text{srvR\_DiskPlatters}} \quad (6)$$

Throughputs of guest and host OS page cache depend on the operating memory, while disk R/W cache throughput is considerably smaller and depends on the disk interface speed. Disk platters' reading service time is given as a sum of consumed seek time  $T_{\text{seek}}$ , generated rotational latency  $T_{\text{latency}}$  and media transfer time  $T_{\text{media}}$ :

$$T_{\text{srvR\_DiskPlatters}} = T_{\text{seek}} + T_{\text{latency}} + T_{\text{media}} \quad (7)$$

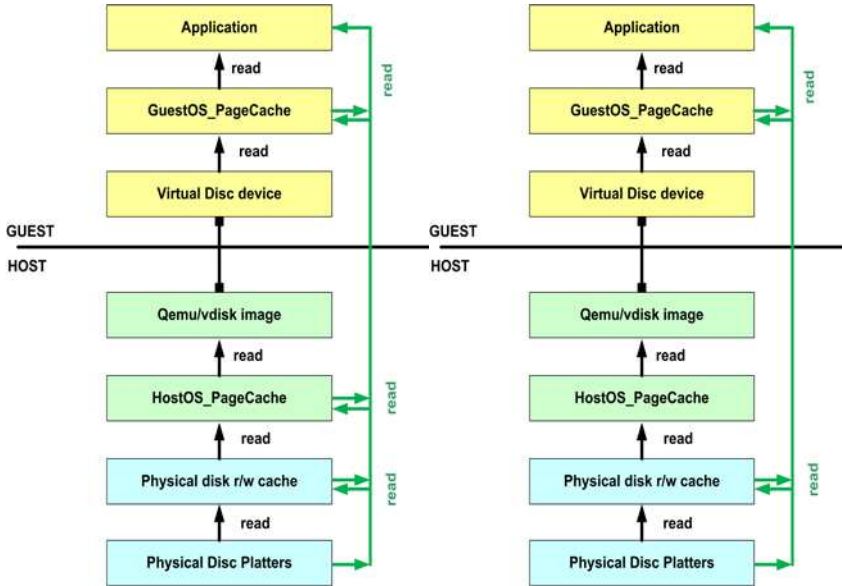


Figure 1

Reading when cache mode is set to writeback or writethrough (left) and set to none (right)

If the KVM cache mode is set to none, only guest OS page cache and disk R/W cache are active for reading operations, while host OS page cache is disabled, as shown in Figure 1 (right). Disabling the host OS page cache degrades reading performance, compared to writeback or writethrough mode.

Application's reading service time depends on guest OS page cache, while disk R/W cache is included into miss cycles, as given in (4). The virtual disk read service time, which is now an equivalent of physical disk service time given by equation (6), is the function of disk R/W for hit cycles, while disk platters are included into the miss cycles.



## 4.2 Normal Write Operations in all Cache Modes

Normal writing cycles use only guest OS page cache in all cache modes, as shown in Figure 2 (left). Host OS page cache and disk R/W cache are used for miss cycles as the disk block allocation function.

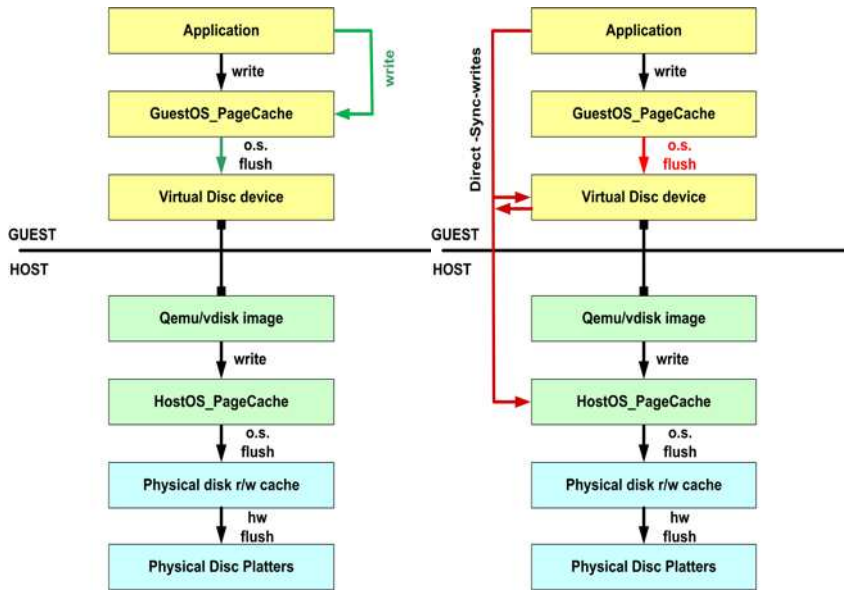


Figure 2

Normal writing operations in all cache modes (left) and flushing or direct-sync writing when cache mode is set to writeback (right)

Application's writing service time is a function of guest flushing time, guest OS page cache for hit cycles, while host OS page cache and disk R/W cache are included into the miss cycles as guest block allocation service time:

$$T_{srvW} \approx P_{HIT\_W\_host} \cdot T_{srvW\_GuestPageCache} + P_{MISS\_W\_guest} \cdot T_{GuestBlockAllocate} + T_{GuestFlush} \quad (8)$$

If we exclude guest flushing, guest application's writing service time is almost identical to guest OS page cache service time; normal write performance without guest flushing time is almost identical for all three cache modes. However, guest flushing time is different for these modes: in writeback, flushing goes into the host OS page cache, in none caching mode flushing goes into the disk R/W cache, and in the writethrough mode it goes into the disk platters.

### 4.3 Flushing or Direct-sync Writing in Different Cache Modes

Let the direct write service time denote guest flushing and direct sync write time.

In writeback cache mode, only host OS page cache is active for flushing – direct writing, while disk R/W cache is active for miss cycles as disk block allocation function. Flushing and direct-sync write cycles are shown in the Figure 2 (right). Direct write service time is a function of host OS page cache for hit cycles, while disk R/W cache is defined as host block allocation service time in the miss cycles:

$$T_{srvDW} \approx P_{HIT\_W\_host} \cdot T_{srvW\_HostPageCache} + P_{MISS\_W\_host} \cdot T_{HostBlockAllocate} + T_{HostFlush} . \quad (9)$$

If host flushing is excluded, the guest's direct writing service time is almost identical to host OS page cache service time (very fast).

If cache mode is set to none, only disk R/W cache is active for flushing – direct writing, while disk platters in miss cycles are considered for disk block allocation function. Flushing and direct-sync write cycles are shown in the Figure 3 (left).

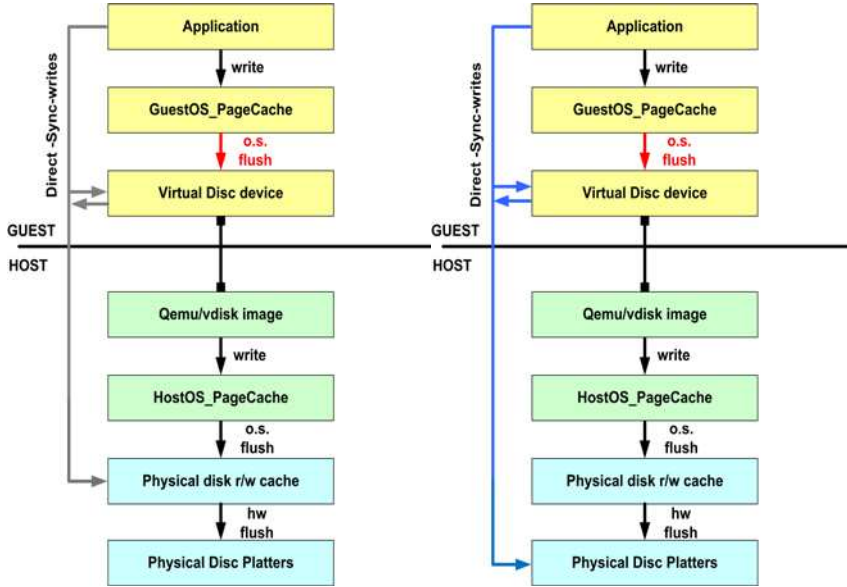


Figure 3

Flushing or direct-sync writing when cache mode is set to none (left) and set to writethrough (right)

Direct write service time is a function of disk R/W cache for hits, while disk platters are included in the miss cycles as disk block allocate service time:

$$T_{srvDW} \approx P_{HIT\_W\_disk} \cdot T_{srvW\_DiskCache} + P_{MISS\_W\_disk} \cdot T_{DiskBlockAllocate} + T_{DiskFlush} . \quad (10)$$

If disk flushing is excluded, the guest's direct writing service time is almost identical to disk R/W cache service time.

If writethrough cache mode is used, each guest flushing or direct write cycle finishes directly to disk platters, as shown in Figure 3 (right). Direct write service time is a function of disk platter service time, and host page cache and disk R/W cache for hits:

$$T_{srvDW} \approx P_{HIT\_W\_host} \cdot T_{srvW\_HostPageCache} + P_{HIT\_W\_disk} \cdot T_{srvW\_DiskCache} + T_{DiskPlatters} . \quad (11)$$

#### 4.4 Hypotheses on KVM I/O Performance

KVM virtual environment provides three levels of caching (guest OS cache, host OS cache and disk R/W cache) and three different cache modes that mainly differ in write semantics and interaction with the host OS cache. Writeback and writethrough modes employ all three levels of caching in writeback and writethrough semantics, respectively, while "none" cache mode bypasses the host OS cache and employs disk R/W cache in writeback semantics. Single or multiple virtual machines running different applications can be executed on a hypervisor.

The following research hypotheses on overall I/O performance, based on the proposed model, are set:

- H1: I/O performance depends on the type of applications running on VMs.
- H2: I/O performance depends on number VMs running.
- H3: Number of VMs running on the hypervisor affects the amount of RAM available to the host OS cache, resulting in overall I/O performance degradation; this is more evident if the hypervisor runs on the system with smaller amount of RAM and large number of VMs running.

According to the presented model, writeback and writethrough cache modes are expected to provide a distinctive reading operations advantage over the "none" cache mode on the basis of equation (5). The preliminary hypotheses on reading operations performance are set as follows:

- H4: Writeback and writethrough are expected to provide remarkably better throughput for workloads with dominant random read components.
- H5: If the dominant component of the workload is the sequential reading, the writethrough and writeback should also provide much better performance under the following conditions: the host page cache is large enough and it relies on applied read ahead technology.
- H6: The impact of equation (5) directly depends on the size of available host page cache and on the usage of the read ahead technologies.

Write operations are correlated to flushing and direct sync cycles and the writeback mode should provide the best results according to equation (9). The preliminary hypotheses on writing operations performance are set as follows:

- H7: If the workloads' dominant components create a large number of random and sequential flushing and direct sync cycles, the usage of writeback caching should provide the system with the best performance.
- H8: Under the same circumstances, the writethrough mode is expected to provide the system with the worst performance, according to equation (11).
- H9: The impact of the equation (9) directly depends on the size of the available host page cache and on the usage of the block-allocation technologies. Also, as with the reading operations, the larger cache is expected to provide the system with better performance.

These hypotheses and the presented model are validated by a set of performance measurements (synthetic benchmarking) and result interpretation presented in next section of the paper.

## 5 Experiments

We have used Postmark benchmark [24] for hypervisor's random performance testing and Bonnie++ benchmark [25] for hypervisor's sequential performance testing. Postmark simulates Internet mail server workload. It creates a large pool of randomly generated files, performs a set of operations, such as creation, reading, and deletion, and measures the time required to perform these operations. Bonnie++ is a benchmark with the ability to perform several performance tests on the file system, including sequential throughput and CPU overhead monitoring during the test.

Experiments were performed on the dual core Intel Xeon CPU E3110 @ 3.00GHz server with 4GB RAM, 1TB hard disk (7200 rpm, 6Gb/s). Centos\_OS\_6.5\_final is the underlying operating system with the ext4 as the test file system. Centos based Linux Kernel 2.6.32-358.18.1.el6.x86\_64 is chosen as a native host for KVM hypervisors. Centos 6.5\_final is also the guest operating system.

The test environment with only one virtual machine running is not completely valid representation of a cloud. Thus, we have performed three sets of experiments: one with the single VM running, and two additional tests with two and three VMs running. All the results obtained with one VM running are interpreted according to the analytical model presented in section 4 of this paper. The model is also validated with the results obtained from the experiments with multiple VMs, which are more representative of the CC environment.

5.1 Random and Sequential Performance Testing with Single VM Running

Random performance is measured with two different test sets. Obtained experimental results are given in Table 1 and graphically presented in Figure 4.

The workload specifications for the first test set are: small number of files (4,000) ranging in size from 1 KB to 100 KB, moderate number of create/delete operations, and smaller amount of read/write operations (1.6 GB for reading, 1.8 GB for writing). The workload specifications for the second test set differ in file size ranging in size from 100 KB to 300 KB, and a larger amount of read/write operations (4.6 GB for reading, 5.4 GB for writing).

Table 1  
Postmark random performance testing results

Workload	Operation	Cache mode		
		Writeback	Writethrough	None
Test Set 1	Random read	59.77 MB/s	3.98 MB/s	5.41 MB/s
	Random write	69.89 MB/s	4.66 MB/s	6.33 MB/s
Test Set 2	Random read	26.83 MB/s	7.36 MB/s	8.57 MB/s
	Random write	31.15 MB/s	8.55 MB/s	9.94 MB/s

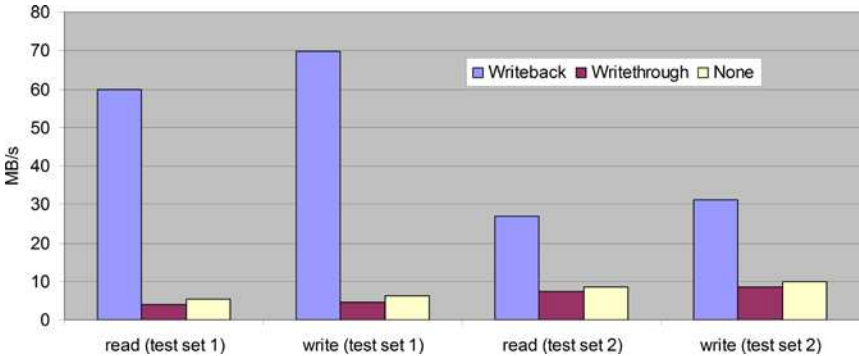


Figure 4  
Random performance testing results

The writeback mode provides the best performance on both workloads. The results obtained from this test set indicate that random read and write components are dominant in (1), while the direct file block component is dominant in (2). The second test set enforces the workaround of sequential components. Writeback mode remarkably outperforms the writethrough mode. Although read (4-7) and normal write operation performances (8) are almost identical, as both modes employ host and guest OS page cache, there is a huge throughput difference due to the flushing and direct write cycles. Writeback mode uses the host OS page cache

with the writeback semantics given by (9) which is remarkably faster than writethrough synchronous writing mode (11). Big throughput differences indicate that the flushing and direct write cycles are very intensive, with the dominant random components of the workload; throughput difference decreases with the second test set, which increases the sequential component of workload.

Writeback mode remarkably outperforms the hypervisor with cache mode set to none as well. Although normal write operation performances (8) are almost identical, read cycles and flushing and direct write cycles provide huge throughput difference. Writeback mode's usage of host OS page cache with writeback semantics (9) provides faster write cycles than disk R/W cache with writeback semantics used if cache mode is set to none (10). Sequential components of workload are more pronounced in the second test set, resulting in smaller differences in read performance and flushing and direct write performance related to first test set.

The performance of hypervisor with the cache mode set to none is slightly better than the writethrough cache. Although normal write operation performances (8) are almost identical, read cycles and flushing and direct write cycles provide minor throughput difference. There are two reasons for the throughput differences. Writethrough read cycles through host OS page cache (5-8) are much faster than reading cycles without any cache. Flushing and direct write cycles using disk R/W cache with writeback semantics (10) are remarkably faster than writethrough synchronous writing mode (11). To conclude, writethrough mode reads data faster, data is written faster if cache mode is set to none. Differences in performance decrease if the workload with stronger sequential components is used.

Bonnie++ sequential throughput test is used to measure sequential writing, reading and rewriting performance of different KVM cache modes. Obtained experimental results are given in Table 2 and graphically presented in Figure 5.

The throughput differences for sequential writes providing writeback's superior performance appear as the result from formulas (9-11) for flushing and direct write cycles.

Rewrite operation reads the contents of the file, deletes the contents and writes new data into the file. Writeback's top performance results from flushing and direct write cycles. Throughput differences between writeback and writethrough modes in rewriting operations are decreased due to read cycles, as both cache modes use host OS page cache (4-7). Writeback mode outperforms the hypervisor's none cache mode, due to the flushing and direct write cycles (10-11) and due to the read cycles, by using host OS page cache.

The writeback slightly outperforms writethrough mode in sequential reading operations, as they both use host OS page cache. If caching mode is set to none, host OS page cache is not used resulting in 6 times lower throughput.

Table 2  
Bonnie++ sequential performance testing results

Operation	Cache mode		
	Writeback	Writethrough	None
Sequential write	97.585 MB/s	37.797 MB/s	85.681 MB/s
Sequential rewrite	42.722 MB/s	26.006 MB/s	36.366 MB/s
Sequential read	618.881 MB/s	584.503 MB/s	98.082 MB/s

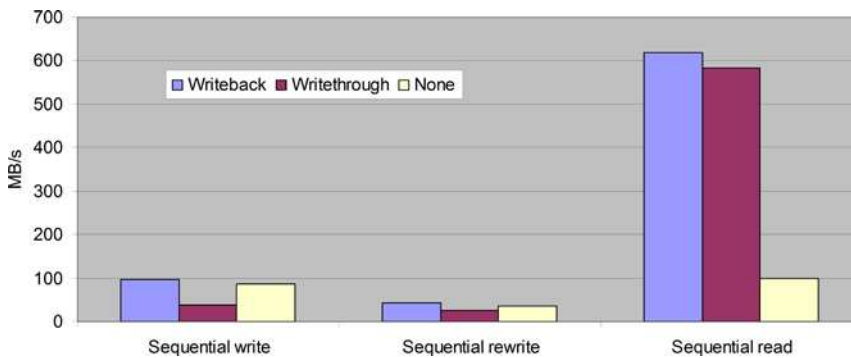


Figure 5  
Sequential performance testing results

## 5.2 Validation of the Model and Hypotheses with Single VM Running

Regarding validation of the presented model for different caching modes, the experiments with one VM running have provided the results that were expected.

It is evident that with a single virtual machine running, the host operating system has more RAM than when multiple virtual machines are running. This results in the largest host page cache, i.e. the greatest impact to read and write operation performances, given by equations (5) and (9), respectively.

For the group of random tests, the performance of the writeback mode is, according to semantics given by equation (9), much better than the performance of writethrough mode, given by equation (11). The reading in writeback and writethrough, performed according to equation (5), is faster than reading when caching mode set to none. As the host page cache is larger than disk R/W cache, writeback also writes faster – which validates equations (9) and (10). Writethrough also provides better read performance than cache mode set to none according to (5), but significantly lower write performances, which validates equations (10) and (11).

According to the group of sequential tests:

- The results of the sequential write tests have shown that writeback caching mode is superior, resulting from (9), while according to (11) the writethrough is the worst option.
- Sequential rewrite operation combines sequential read and write operations on the same blocks. The performances of systems with the caching modes set to none and writeback mode are similar. This results from a big positive impact provided by disk R/W cache (10). Writethrough is the worst option again (11).
- With the largest possible host page cache size, the writeback and writethrough modes provide much better read performances (5) when compared to caching mode set to none.

### 5.3 Experiments and Validation with Multiple VMs Running

The next set of the experiments is carried out with multiple VMs running on the same hypervisor. The same Postmark workload is used for the two and three VMs running scenarios, and average values were used for result interpretation.

The results of random read and random write experiments with two and three VMs are given in Tables 3 and 4, respectively.

Table 3  
Postmark random performance testing results (two VMs running)

Workload	Operation	Cache mode		
		Writeback	Writethrough	None
Test Set 1	Random read	21.15 MB/s	1.76 MB/s	2.28 MB/s
	Random write	24.73 MB/s	2.06 MB/s	2.67 MB/s
Test Set 2	Random read	4.75 MB/s	2.60 MB/s	3.26 MB/s
	Random write	5.51 MB/s	3.02 MB/s	3.78 MB/s

Table 4  
Postmark random performance testing results (three VMs running)

Workload	Operation	Cache mode		
		Writeback	Writethrough	None
Test Set 1	Random read	8.20 MB/s	1.08 MB/s	1.89 MB/s
	Random write	9.59 MB/s	1.26 MB/s	2.21 MB/s
Test Set 2	Random read	3.09 MB/s	1.63 MB/s	2.57 MB/s
	Random write	3.59 MB/s	1.89 MB/s	2.99 MB/s

According to random test results, the performance of the virtual machine used for measurement decreases with each new VM added to the hypervisor, as each VM added to the system consumes a part of the available RAM (1 GB in our case).



Thus, there is less memory available to host OS, resulting in smaller host page cache. This has a direct impact on the equations (5) and (9): reading performances are decreased according to equation (5), and writing performances are decreased according to (9).

The results obtained from the first test set indicate that writeback caching mode slightly decreases its superior performance with the increased number of VMs running. However, it is still the very best option for the test set 1. Results of this test are in accordance with the expected ones, thus validating the analytical model properly.

Writeback caching mode outperforms the writethrough mode, mostly due to writeback semantics given with equation (9) that depends directly on the host page cache size, unlike writethrough semantic given with the equation (11). This difference decreases with each virtual machine added to the system, as each addition provides a negative impact on writeback performance, which is, as expected, in accordance with the equation (9).

Writeback provides faster reading operations if compared to the caching mode set to none, which is in accordance with (5) and mostly depends on the host page cache size. Writeback also provides better write operations performance, as given with equation (9) and (10). The increased number of VMs running decreases the performance difference between aforementioned modes, as equation (5) has weaker impact on writeback performance.

Writethrough provides the system with better reading operation performance than the system with caching mode set to none according to equation (5), which highly depends on host page cache size. But, when the write operations are analyzed, the so called "none" caching mode is outperforming writethrough, and this originates from differences expressed by equations (10) and (11). With the increasing the number of VMs running, this difference becomes even more evident. The explanation relies on application of the equation (5), as it has weaker effects with the writethrough, and in those circumstances the "none" mode overtakes the precedence in performance analysis.

According to the results obtained from the test set 2, the writeback mode loses some of its superiority with the increased number of VMs running, while system with caching mode set to none still outperforms writethrough mode. The explanation of this behaviour is similar to the one provided for the test set 1 – smaller amount of host page cache used for reading and writing.

The results of sequential read, write and rewrite operations obtained from Bonnie++ with two and three VMs are given in Tables 5 and 6, respectively.

According to the obtained test results, the performance decreases with each new VM added to the hypervisor.

Table 5

Bonnie++ sequential performance testing results (two VMs running)

Operation	Cache mode		
	Writeback	Writethrough	None
Sequential write	37.98 MB/s	22.58 MB/s	32.21 MB/s
Sequential rewrite	10.62 MB/s	6.77 MB/s	13.96 MB/s
Sequential read	36.45 MB/s	32.35 MB/s	28.08 MB/s

Table 6

Bonnie++ sequential performance testing results (three VMs running)

Operation	Cache mode		
	Writeback	Writethrough	None
Sequential write	31.19 MB/s	14.64 MB/s	21.09 MB/s
Sequential rewrite	6.79 MB/s	4.39 MB/s	7.24 MB/s
Sequential read	23.90 MB/s	21.48 MB/s	20.40 MB/s

The results of the sequential write tests have shown that writeback caching mode dominates over the other two modes, for both scenarios with two and three VMs running. The explanation of such behaviour relies on equation (9), and it's comparison to equations (10) and (11), which is in accordance with the presented analytical model.

The best results in sequential rewrite operations testing are achieved when caching mode is set to none. When testing the rewriting operation, the overall effect of reduced host page cache from equations (5) and (9) influences in such a way that the "none" mode (which does not utilize the host page caching, but still utilizes disk R/W cache) provides better results than writeback mode. That behaviour was detected for the sequence that covers: sequential block reading, editing, and writing. Writethrough mode provides remarkable lower performance when compared to other caching modes.

The results of sequential read operations test indicate that, for both two and three VMs running, writeback and writethrough cache modes still provide higher I/O throughput. Thus, the presented model is completely validated for sequential read operations. When testing is performed with only one virtual machine, the available host page cache is large, thus writeback and writethrough modes significantly outperform the system with caching mode set to none, according to equation (5). This difference is decreased with each virtual machine added into the testing system, and in those circumstances the equation (5) has a weaker influence on the writeback and writethrough modes performances. With sufficient number of virtual machines, host page cache available to each virtual machine would be reduced and the cache mode performances would probably not differ that much.

## Conclusions

While comparing the results obtained from testing the system with the one, two and three virtual machines running, we have detected the biggest performance difference between different cache modes employed in the single virtual machine scenario. This difference results from the hosts OS provided with largest amount of RAM. With the introduction of two or three VM, the performances of all cache modes were reduced, as well as the difference between them. The writeback mode employs three different caches in the writeback semantics resulting in superior performance. The best performances are achieved for random read/write workloads, as for sequential workloads when the amount of host OS page cache is large and the read ahead technique is dominating. However, writeback can endanger data integrity in case of power surges. Writethrough mode also employs three cache types with all caching outside of the virtual machines working with the writethrough semantics. Although data integrity is ensured, writethrough performs poorly, especially with the random workload when flushing and direct write cycles are dominant. Hypervisor's none cache mode, which employs disk R/W cache only outside of virtual machine, is faster than writethrough and slower than writeback mode, but does not provide total data integrity.

The results indicate that the amount of RAM has a huge impact on the performance, as it directly affects the host OS page cache size. According to the results, the writeback cache mode provides the system with the best performance if a small number of VMs is running. Each VM added to the system with fixed RAM size results in overall performance degradation, regardless of the employed cache mode, and decreased performance difference between the systems that employ "none" and writeback cache modes. Due to decreased impact of host OS cache it is expected that this performance difference between "none" and writeback modes will become almost negligible when a large number of VMs (10 or more VMs) is running. According to that, we strongly recommend the usage of writeback mode on hypervisors running a small number of VMs and "none" cache mode on hypervisors running a large number of VMs. It should be noted that the amount of RAM available to the host directly affects the performance differences between systems that run large numbers of VMs – the larger amount of RAM available to the host results in bigger performance difference between systems that employ different cache modes.

Experimental results are not surprising and they are not contradictory to our preliminary hypotheses. The exception that was not expected is the rewrite operation results. This indicates that in some rare situations I/O performance does not benefit that much from the cache.

Further research will include KVM cache mode examination while running large number of virtual machines, as well as the impact of disk schedulers, I/O modes and virtual disk image formats towards virtualization performance.

## Acknowledgement

This paper has been partially financed by Serbian Ministry of Education, Science and Technical Development (Development Projects III 43002, TR 32037 and TR 32025).

## References

- [1] K. Xiong, H. Perros: Service Performance and Analysis in Cloud Computing. In SERVICES 2009, 5<sup>th</sup> 2009 World Congress on Services, Bangalore, India, 2009, pp. 693-700
- [2] J. G. Hansen, E. Jul, Lithium: Virtual Machine Storage for the Cloud. In 2010 SoCC'10: Proc. 1<sup>st</sup> ACM Symp. Cloud Comput., ACM Press, New York, USA, 2010, pp. 15-26
- [3] D.-J. Kang, C.-Y. Kim, K.-H. Kim, S.-I. Jung: Proportional Disk I/O Bandwidth Management for Server Virtualization Environment. In 2008 Int. Conf. Comput. Sci. Inf. Technol., Piscataway, NJ, USA, 2008, pp. 647-653
- [4] J. Nakajima, K. M. Asit: Hybrid Virtualization—Enhanced Virtualization for Linux. In 2007 Proc. Linux Symp, 2007
- [5] T. Imada, M. Sato, and H. Kimura: Power and QoS Performance Characteristics of Virtualized Servers. In Proc. 2009 10th IEEE/ACM Int. Conf. Grid Computing (GRID), Piscataway, NJ, USA, 2009, pp. 232-240
- [6] KVM, Kernel-based Virtual Machine. <http://www.linuxkvm.org>
- [7] QEMU, Open Source Processor Emulation. <http://www.qemu.org>
- [8] T. Deshane, Z. Shepherd, J. Matthews, M. BenYehuda, A. Shah, B. Rao: Quantitative Comparison of Xen and KVM. 2008 Xen Summit, Berkeley, CA, USA, USENIX Association, 2008
- [9] D. Armstrong, K. Djemame: Performance Issues in Clouds: An Evaluation of Virtual Image Propagation and I/O Paravirtualization. The Computer Journal, Vol. 54, No. 6, 2011, pp. 836-849
- [10] P. Padala, X. Zhu, Z. Wang, S. Singhal, K. G. Shin: Performance Evaluation of Virtualization Technologies for Server Consolidation. Tech. Report, HP Labs, USA, 2008
- [11] X. Xu, F. Zhou, J. Wan and Y. Jiang: Quantifying Performance Properties of Virtual Machine. In 2008 Linux; Program Testing; Software Performance Evaluation; Systems Analysis; Virtual Machines, Vol. 1, Piscataway, NJ, USA, 2008, pp. 24-28
- [12] J. Che, Q. He, Q. Gao, D. Huang: Performance Measuring and Comparing of Virtual Machine Monitors. In 2008 IEEE/IFIP 5<sup>th</sup> Int. Conf. Embedded and Ubiquitous Computing. EUC2008, Vol. 2, Piscataway, NJ, USA, 2008, pp. 381-386

- [13] S. Y. Liang, X. Lu: An Efficient Disk I/O Characteristics Collection Method Based on Virtual Machine Technology, In 2008 Proc. 10<sup>th</sup> IEEE Int. Conf. High Performance Computing and Commun., HPCC2008, Dalian, China, 2008, pp. 943-949
- [14] Y. Dong, J. Dai, Z. Huang, H. Guan, K. Tian, Y. Jiang: Towards High-quality I/O Virtualization. In 2009 ACM Int. Conf. Proc. Series, Haifa, Israel, 2009, pp. 12-22
- [15] X. Liao, H. Jin, J. Yu, D. Li: A Performance Optimization Mechanism for SSD in Virtualized Environment. The Computer Journal, Vol. 56, No. 8, 2013, pp. 992-1000
- [16] A. Sallam, K. Li: A Multi-objective Virtual Machine Migration Policy in Cloud Systems. The Computer Journal, Vol. 57, No. 2, 2014, pp. 195-204
- [17] Q. Li, Q. Hao, L. Xiao, Z. Li: An Integrated Approach to Automatic Management of Virtualized Resources in Cloud Environments. The Computer Journal, Vol. 54, No. 6, 2011, pp. 905-919
- [18] W. Zhao, P. M. Melliar-Smith, and L. E. Moser: Low Latency Fault Tolerance System. The Computer Journal, Vol. 56, No. 6, 2013, pp. 716-740
- [19] Y. Jin, Y. Wen, Q. Chen and Z. Zhu: An Empirical Investigation of the Impact of Server Virtualization on Energy Efficiency for Green Data Center. The Computer Journal, Vol. 56, No. 8, 2013, pp. 977-990
- [20] T. V. Do: Comparison of Allocation Schemes for Virtual Machines in Energy-Aware Server Farms. The Computer Journal, Vol. 54, No. 11, 2011, pp. 1790-1797
- [21] L. Vokorokos, A. Baláz, N. Ádám: Secure Web Server System Resources Utilization. Acta Polytechnica Hungarica, Vol. 12, No. 2, 2015, pp. 5-19
- [22] I. Petkovics, P. Tumbas, P. Matković, Z. Baracska: Cloud Computing Support to University Business Processes in External Collaboration. Acta Polytechnica Hungarica, Vol. 11, No. 3, 2014, pp. 181-200
- [23] Kernel Virtual Machines (KVM): Best Practices for KVM (second edition). IBM Corporation, 2012
- [24] J. Katcher: PostMark: A New File System Benchmark, Technical Report TR3022. Network Appliance Inc, 1997
- [25] Bonnie++ Benchmark Suite. <http://www.coker.com.au/bonnie++/>

# Supply Chain Risk Management Using Software Tool

**Matotek Marija, Barać Ivan, Regodić Dušan**

Singidunum University, 32 Danijelova St., 11000 Belgrade, Republic of Serbia,  
matotek@vts-zr.edu.rs; dregodic@singidunum.ac.rs

**Grubor Gojko**

Sinergija University, Raje Banjičića St., 76300 Bijeljina, Bosnia & Herzegovina,  
ggrubor@sinergija.edu.ba

---

*Abstract: Risk management is an integrating part of every supply chain aspect. Unsuccessful risk management can have negative economic and ecologic consequences and cause total or partial lost in the business of a company. Supply Chain Risk Management (SCRM) technology helps managers plan for and handle disruptions in the supply chain. Companies that invest in the emerging field of SCRM technology are less likely to sustain costly supply disruptions or negative press because of the actions of their suppliers. In this paper a model of risk management in supply chain is suggested, according to the international standard ISO 31000:2009 recommendation. An automatic risk evaluation has been performed by software application and as a result 2393 individual risk assessments have been obtained. In this case study the levels of risk factors obtained by risk treatment have been accepted. In order to lessen the subjectivity of the analysts it is necessary to use the same methodology for risk assessment by more than one risk analysts and then to compare their results to provide an objective risk assessment, a key phase in risk management.*

*Keywords: risk; risk management; software tool; supply chain*

---

## 1 Introduction

As a general answer to the question “What is management risk?” the theory offers [1] the following answer: “To many analysts, politicians and academics it is the management of the natural environment and of technology-generated macro-risks that appear to threaten our existence! In general, risk management can be defined as a life discipline which considers the likelihood of future events causing unwanted effects. However, risk may not be considered as an avoidable category.

According to the ISO 31000 standard risk management consists of risk identification, risk assessment, risk prioritization, and effectively allocate and using resources for risk treatment in order to minimize, monitor and control likelihood or impact of the unwanted events, and to maximize realization of the expected successes.

In theory, supply chains work as a cohesive, singularly competitive unit, accomplishing what many large, vertically integrated firms tried and failed to accomplish in years past. The difference is that independent companies in a supply chain are relatively free to enter and leave a supply chain relationship if these relationships are no longer proving beneficial; it is this free market alliance – building that allows supply chains to operate more effectively than vertically integrated conglomerates [2]. A supply chain risk appraisal process can help make strategic decisions and operational plans to reduce the quantity of supply chain defects [3]. The identified compliance issues require response strategies, in the form of developing policies and procedures, to address compliance risks effectively. Infrastructure and resources (including material, human, IT, financial, technical) have to be provided and assigned to enforce the program. To illustrate the high-complexity of compliance management, consider that, to manage for example “product liability compliance” the following business functions could be related: engineering, procurement, manufacturing, quality control, sales and distribution, and more. To ensure product compliance the organization needs a certain level of control across the entire supply chain [4].

Economic globalization and the resultant complexity of the supply chain network plus the uncertainty of the environment makes risk and vulnerability a major challenge to related firms [5]. The risk assessment segment in the supply chain management is a rather devalued research field. In this paper the authors will present a risk management model in supply chains. The preliminary results of the risk assessment shown in this paper, suggest that the software tool application in risk assessment can be useful, mainly for decreasing overall total risk. Thus, risk management is getting simpler and more effective requiring less bureaucracy.

## **2 Historical Perspective of Risk Management**

Risk management has been familiar to humanity since ancient times. The authors of the article [6] reported that the first risk assessment was recorded with the Asipu group who lived in the Tigris-Euphrates valley in 3200 BC. They tried to make contributions to risk assessment in construction businesses, business start – ups, and even concerning marriage arrangements. The Asipu analyzed available risk factors suggesting alternatives and proposing the best solution for quantitative risk mitigation [6]. In 792 BC bottomry contracts were used in order to mitigate risk. This form consisted of three elements: loan, interest rate and risk premium. In

this period, the concept of general average was also developed, only to evolve into insurance concept. Apparently, insurance is one of the oldest forms of the risk mitigation strategy. The advance in probability theory development in the 17<sup>th</sup> and 18<sup>th</sup> Centuries made it an important tool which can be used for risk management. Since the nineteenth century, probability theory continues in many disciplines, using specific tools and methods from finance to system engineering disciplines.

### 3 Basic Elements of Supply Chains Management

According to [7], the supply chain is a set of physical elements, in which their activities and processes are related to their mutual interactions. The executive part of this chain consists of appropriate ways and rules of realization related to logistic activities and processes, and it is an operative part by its nature. The management of the executive part and determination of the fixed part of the supply chain both constitute the supply chain management used to define its performance (for example, costs and service levels). In [8], the authors developed the framework for supply chains management including three basic elements (Figure 1):

- supply chain structure (defining the key elements of supply chain which will be connected through business processes)
- business processes (defining the business processes used to connect certain elements of supply chain) and
- managing components (defining the level of integrated management for each business process).

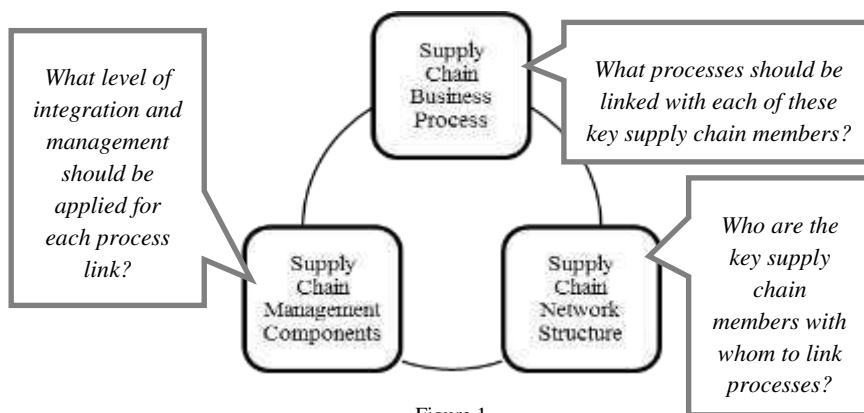


Figure 1

Supply chain management framework: Elements and key decisions [8]



Each of these elements is directly related to the degree of fulfillment of end-user demands, taking into account the key performance indicators. They refer to a relatively small number of critical dimensions of business which have a huge impact on the success in the market. Key performance indicators compare the efficiency and / or effectiveness of a system with standardized or target (desired) values. A well-defined set of performance indicators of the supply chain can help identify technological reserves in key logistics processes and their later utilization.

Companies have reacted to the apparent opportunities and threats of globalization through various global production practices that have increased supply chain complexity and various forms of risk. Through increasing supply chain integration, companies have attempted to manage this increased level of complexity. Supply chain integration has been identified as a key practice to manage supply chains and achieve superior performance [9].

## **4 Risk Taxonomy in The Supply Chain**

Risk taxonomy in the supply chain is possible according to several criteria. According to the exposure area, the following assets are exposed to the risk:

- people,
- technologies,
- environment and
- business processes.

Risk management concept in the supply chain can be defined as exposure to the risky events that have negative impact to the supply chain operability and performances such as service level, costs or possibility of the fast response. The spectrum of the risky events that can affect supply chain operability is very large going from external risk factors (e.g. supply chain environment) to inter-organization and intra-organization risk factors. The consequences of these risk factors can be categorized according to duration, intensity and likelihood of occurrence as follows:

- (1) from short-term to long-term;
- (2) from small intensity ones to high intensity ones, and
- (3) from very rare to very common ones.

From the logistic point of view, the interaction among supply chain members becomes more and more complex due to the growing uncertainty as a result of the new business models applied to increase logistic efficacy and competitiveness.

Therefore, with the main risk sources in the supply chains the following types of interactions can be identified:

- (1) occurring between supply chain members and the environment, and
- (2) occurring between individual members inside the supply chain.

The traditional method of management of these risks requires the engagement of additional material and time. However, what the companies really need, in the supply chain, is proactive risk management.

When, considering risk control tools, making decisions on appropriate measures for risk treatment and performing their implementation among partners, they become obvious complex tasks, even though within the supply chain a consistent risk management policy is present. In a supply chain, an internal audit entity can develop an appropriate risk management program providing risk assessment by continual monitoring and auditing. However, before a company can conceive a methodology for risk mitigation, the managers have to understand the risk categories, events and conditions that generate risks (Tab. 1). Therefore, using the knowledge about these key risk factors, the companies can choose the most efficient risk mitigation strategy. To mitigate risks by intelligent positioning and dimensioning without profit reduction is a great challenge for management.

Table 1  
Risk categories and its expression in the supply chain

<b>Risk category</b>	<b>Risk expression</b>
Disorder/ interruption	Natural disaster. Business dispute. Supplier bankruptcy. War and terrorism. Dependency on one supply source, and alternative supplier response ability.
Delay	High utilization of the capacity on the chain source Supply source inflexibility. Bad quality and contribution on the chain source Border crossing or mode of transport change.
Systems	ICT infrastructure breakdown. ICT system integration or excessive networking. E-business.
Prediction	Incorrect prediction due to long terms, short life cycle, and small clients data base. “Bullwhip effect” or information distortion due to sale promotion, incentive, lack of visibility, and excessive demand during shortages of products.
Intellectual property	Vertical integration of the supply chain. Global market and outsourcing.
Procurement	Exchange rate risk. Percentage of the key components or raw materials that are supplied from one source. Large utilization of the industrial capacity. Long terms vs. short terms contracts.

Demand	Number of consumers. Finance strength of consumers.
Inventory	Products rate of obsolescence. Holding costs. Value of the products. Uncertainty of the supply and demand.
Capacity	Capacity costs. Capacity flexibility.

## 5 Supply Chain Risk Management – Case Study

This chapter summarizes the results of the case studies in which the process of risk management in supply chains is implemented in accordance with the recommendations of international, globally accepted, standards ISO 31000 : 2009, Risk management - Principles and guidelines.

Risk management process is a systematic application of management policies, procedures and practices of communication activities, consulting, establishing the context, and identifying, analyzing, evaluating, treating, monitoring and reviewing risk [10].

Figure 2 shows the block diagram of the process of risk management. The case study did not discuss the following blocks: "Establishing the context", "Communication and consultation" and "Monitoring and review" because it was assumed that the risk management framework had already been established, including the specified sub-processes. Within the study, emphasis is given to the assessment of risk (comprising the steps of risk identification, risk analysis and risk assessment) and treatment of risk. For the automation purposes of the abovementioned sub – processes and the qualitative analysis the application, whose results were used solely for scientific purposes, was used.

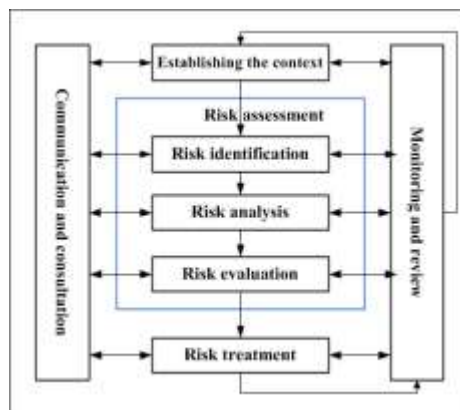


Figure 2  
Risk management process (Adapted from [10])

Table 2 shows the basic concepts and definitions needed for a description of the risk management process. Most of the concepts and definitions have been taken from ISO 31000:2009 [10] standard, and the rest from ISO/IEC 27005 [11] which gives the recommendations for risk assessment in ISMS.

Risk assessment is the overall process of risk identification, risk analysis and risk evaluation [10].

The risk assessment used the statistics and numerical approximations or qualitative methods, defining the values on the basis of quality. As the output from the phase of the risk assessment, the obtained parameters were used in the treatment of stage risk (reducing the risk to an acceptable level). The control measures to be implemented in order to improve supply chain risk through treatment were identified.

Table 2  
The basic terms and definition [10], [11]

Terms	Definition
risk	effect of uncertainty on objectives
risk management	coordinated activities to direct and control an organization with regard to risk
assets/ resource	the organization should allocate appropriate resources for risk management
threat/ risk source	element which alone or in combination has the intrinsic potential to give rise to risk
vulnerability	intrinsic properties of something resulting in susceptibility to a risk source that can lead to an event with a consequence
risk criteria	terms of reference against which the significance of a risk is evaluated
level of risk	magnitude of a risk or combination of risks, expressed in terms of the combination of consequences and their likelihood
residual risk	risk remaining after risk treatment
control	measure that is modifying risk
risk assessment	overall process of risk identification, analysis and risk evaluation
risk identification	process of finding, recognizing and describing risks
risk analysis	process to comprehend the nature of risk and to determine the level of risk
risk evaluation	risk evaluation process of comparing the results of risk analysis with risk criteria to determine whether the risk and/or its magnitude is acceptable or tolerable
risk treatment	process to modify risk

## 5.1 Risk Management Methodologies

The organization should define criteria to be used to evaluate the significance of risk. The criteria should reflect the organization's values, objectives and resources.

Some criteria can be imposed by, or derived from, legal and regulatory requirements and other requirements to which the organization subscribes. Risk criteria should be consistent with the organization's risk management policy, be defined at the beginning of any risk management process and be continually reviewed [10].

Table 3 presents the basic concepts and criteria that are defined in the methodology for risk management, which is defined before the phase of risk assessment and cannot be altered during the phase of risk assessment and risk treatment. The validity of the methodology can be analyzed during and upon completion of the above two stages in risk management, and it can be changed for future risk assessment (the block diagram in Figure 1 "Monitoring and review").

Table 3 shows that the risk is calculated through the influence of the three characteristics of the supply chain (impact on confidentiality (e.g., information), integrity (of the supply chain as a whole) and availability (e.g., the services offered through the supply chain). In the considered case study, the risk actually represents the probability that a particular threat risk source will exploit the vulnerability of the assets / resource, which is demonstrated through the loss of confidentiality, integrity or availability of the supply chain.

Table 3  
Risk management methodologies terms and criteria [10]

Term	Description of criteria		
$T_v$ -threat value	(1-low; 2-medium and 3-high)		
$V_v$ -vulnerability value			
$P = T_v \times V_v$ – probability $P_l$ – level of risk	<del><math>R = \{1, 2, 3, 4, 5\}</math></del> <del><math>R = \{1, 2, 3, 4, 5\}</math></del>		
$I_{vx}$ - impact value (x= confidentiality, integrity or availability)	(1- very low; 2- low; 3- medium; 4- high; 5- very high)		
$R_x = I_{vx} \times P_l$ - Risk (x= confidentiality, integrity or availability)	risk level	action	risk treatment
$1 \leq R_x \leq 2$	very low	not require any action	risk retention
$3 \leq R_x \leq 4$	low	required monitoring of risk	risk retention
$5 \leq R_x \leq 10$	medium	need some action and monitoring risk	risk retention
$12 \leq R_x \leq 16$	high	necessary actions	risk reduction
$20 \leq R_x \leq 25$	very high	necessary actions	risk reduction

The case study used a modified version of the software tools Hestia Risk (Croatian company that distributes Bluefield Ltd.), integrated software solution for managing risks by using a qualitative methodology, which is based on the application of international standards ISO 31000, and other related standards such as ISO / IEC 27005, ISO 22301 and others. The authors have taken advantage of the flexibility and parametrization software to adjust for risk assessment in SCM. The authors define a list of vulnerabilities, threats list, their relations, the key processes in the SCM, as well as measures to be implemented to mitigate the risk. In addition, this defines the method of risk assessment which is later used in the case study.

## 5.2 Risk Identification

The organization should identify sources of risk, areas of impacts, events (including changes in circumstances) and their causes and their potential consequences. The aim of this step is to generate a comprehensive list of risks based on those events that might create, enhance, prevent, degrade, accelerate or delay the achievement of objectives. It is important to identify the risks associated with not pursuing an opportunity. Comprehensive identification is critical, because a risk that is not identified at this stage will not be included in further analysis [10].

Identification should include risks whether or not their source is under the control of the organization, even though the risk source or cause may not be evident. Risk identification should include examination of the knock-on effects of particular consequences, including cascade and cumulative effects. It should also consider a wide range of consequences even if the risk source or cause may not be evident. As well as identifying what might happen, it is necessary to consider possible causes and scenarios that show what consequences can occur. All significant causes and consequences should be considered [10].

### 5.2.1 Assets

Consideration should be given to the following: people, skills, experience and competence; resources needed for each step of the risk management process; the organization's processes, methods and tools to be used for managing risk; documented processes and procedures; information and knowledge management systems; and training programmes [10].

Figure 3 shows the review of 27 processes in supply chains, which are grouped into nine specific groups. Each process was analyzed individually in order to investigate the effect of the potential risks on confidentiality, integrity and availability. After analysis, each group processes were associated with certain risk groups.

### 5.2.2 Threats/Risk Sources

A risk source can be tangible or intangible [10]. Threats have the power to damage the information, assets, process or system. Threats may arise from natural or human origin and can be accidental or intentional. It should identify sources of accidental and intentional threats, as well as to estimate how their occurrence probability is. It is necessary to expect threats, because failure to do so may cause a risk to the organization. The level of threat is defining for each of the identified threats. This study deals with 73 different kinds of threats and risk sources according to confidentiality, integrity and availability in supply chains.

	IMPACTS		
	Confidentiality	Integrity	Availability
COMMUNICATION WITH THE REGION AND THE LEGAL PROCEDURES AND COMMUNICATION	2	2	3
COMPLIANCE WITH LEGAL PROCEDURES AND STANDARDS	3	2	3
DESIGN PROCESS			
DELEGATION OF AUTHORITY	2	1	4
DEPARTMENTALIZATION	1	1	2
ESTABLISHING A RANGE OF CONTROLS	2	2	4
IDENTIFICATION AND ARRANGEMENT OF JOBS	1	2	3
FINANCIAL MANAGEMENT			
JOBS OF ACQUISITION AND PLACEMENT OF FUNDS	3	5	3
JOBS OF FINANCIAL MANAGEMENT	5	4	3
JOBS OF PAYMENT AND LIQUIDATION	5	4	3
INFORMATION MANAGEMENT			
COLLECTION AND PROCESSING OF INFORMATION	4	3	4
INFORMATION USAGE	5	4	5
MATERIAL MANAGEMENT			
PLANNING OF MATERIALS	1	4	4
PRODUCTION CONTROLS	3	4	5
PURCHASE OF MATERIALS	1	2	3
STORAGE	1	3	4
PLANNING AND DEMANDS ANTICIPATION			
DISTRIBUTION PLANNING	4	5	3
PLANNING AND DEMANDS ANTICIPATION	4	4	2
PLANNING SECURITY STOCKS	3	2	1
SUPPLY NETWORK PLANNING	5	5	3
PRODUCTION PROCESS			
EXECUTION OF PRODUCTION	3	4	5
FINALISATION	3	4	4
PREPARATION OF PRODUCTION	3	4	5
RETURN PROCESSES			
REPLACEMENT	1	2	1
RETURN	1	1	1
THE TRANSPORT PROCESS			
LOADING	2	2	4
TRANSPORT	2	2	4
UNLOAD	2	1	4

Figure 3

List of supply chain processes

### 5.2.3 Vulnerabilities

Vulnerability is a weakness in an asset or group of assets. An asset's weakness could allow it to be exploited and harmed by one or more threats [11]. The vulnerability factors of supply chains may be classified into five groups [12], which may be complemented by two further risk sources [13]:

- disturbances in the value-added process (manufacturing, purchasing, storage, delivery, scheduling),
- control (non-existence or failure),

- demand (lack of information, unpredictability, unexpected events),
- supply (unreliability, lack of capacity, vis major),
- environmental (economic and political events, accidents, natural disasters),
- enterprise structure (if non-conformance with enterprise processes) and
- 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 [14].

The study deals with 67 different types of vulnerabilities in supply chain processes.

#### **5.2.4 Group of Risks**

After the analysis, the threats were divided into four risk groups (human factor, processes, system and external factors). Each group process is associated with one or more risk groups. At the stage of risk assessment, risks for each process are analyzed by each individual threat that is associated with the risk groups.

#### **5.2.5 Controls**

Controls include any process, policy, device, practice, or other actions which modify risk. Controls may not always exert the intended or assumed modifying effect [11].

The study defined 76 – control measures to be applied in order to reduce the level of risk to an acceptable one after the risk treatment phase. Control-measures can be organizational, procedural and technical.

### **5.3 Risk Analysis**

Risk analysis involves developing an understanding of the risk. Risk analysis provides an input to risk evaluation and to decisions on whether risks need to be treated, and on the most appropriate risk treatment strategies and methods. Risk analysis can also provide an input into making decisions where choices must be made and the options involve different types and levels of risk [8].

Risk analysis involves consideration of the causes and sources of risk, their positive and negative consequences, and the likelihood that those consequences can occur. Factors that affect consequences and likelihood should be identified.



Risk is analyzed by determining consequences and their likelihood, and other attributes of the risk. An event can have multiple consequences and can affect multiple objectives. Existing controls and their effectiveness and efficiency should also be taken into account [10].

5.4 Risk Evaluation

The purpose of risk evaluation is to assist in making decisions, based on the outcomes of risk analysis, about which risks need treatment and the priority for treatment implementation. Risk evaluation involves comparing the level of risk found during the analysis process with risk criteria established when the context was considered. Based on this comparison, the need for treatment can be considered. Decisions should take account the wider context of the risk and include consideration of the tolerance of the risks borne by parties other than the organization that benefits from the risk. Decisions should be made in accordance with legal, regulatory and other requirements [10].

This application was used for automatic evaluation of risk, and as a result 2393 individual risk assessments were obtained. Risk assessments are divided according to levels of risk (risk level "5" - 1084 risks; level "4" - 709; the level of "3" - 537; the level of "2" – 50; level "1" - 13 risk).

The methodology for risk management in the supply chain was used to define that only "4" and "5" risk levels would be treated, which makes the total of 1793 different risks in this study with all controls - measures for reducing risk used up in the risk treatment phase. Figure 4 shows the list of risk levels.






Number Of Asset Types		10
Number Of Assets		26
Number of risk assessments		2,394
Number of risk assessments5		1,084
Number of risk assessments4		709
Number of risk assessments3		537
Number of risk assessments2		50
Number of risk assessments1		13
Number of risk assessments0		1
Number of available controls		209
Number of used controls		23
Document made by	Document controlled by	Document approved by

Figure 4  
List of risk levels

## 5.5 Risk Treatment

Risk treatment involves selecting one or more options for modifying risks, and implementing those options. Once implemented, treatments provide or modify the controls [10].

In the risk evaluation phase, it was determined that all the available control – measures would be used in order to reduce the risks levels "4" and "5". The principle that the application of each control – measure reduces the level of individual threats for value 1 was used.

Figure 5 shows part of the risk treatment plan. By applying the defined control / measures of risk treatment plan the risk levels "4" and "5" are reduced.

Clause	ISO 27001:2005 Controls		Treatment ID	Budget	Person responsible	Implementation dates		Finished
	Section	Control				Start	End	
WS 1.3.	EDUCATION OF EMPLOYEES		2416	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.4.	ESTABLISHMENT OF THE PRICE FOR CERTAIN PRODUCTS THAT ENABLE THE IMPLEMENTATION OF BUSINESS PLANS		178	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.5.	FORMATION OF A UNIFIED INTERNAL REGISTER TO STORE THE ORIGINAL COPY OF THE CONTRACT, ANNEXES, DECISIONS, STATEMENTS		194	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.6.	FORMING A WORKING UNIT		38	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.7.	DO NOT USE INFORMATION ON BUSINESS SECRETS FOR THE PURPOSE OF ADVERTISING THROUGH PUBLIC SOCIAL NETWORKS		82	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.8.	INNOVATION IN P / U AND MONITORING TRENDS		30	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.9.	INTERNAL CONTROL		30	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.10.	EXPLORE THE MARKET AND CONSUMER NEEDS		4	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.11.	RESEARCH AND ADAPTING TO MARKET NEEDS		85	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.12.	MARKET RESEARCH		91	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.13.	EXPLORE THE MARKET AND CONSUMER NEEDS		75	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.14.	CREATING A STRATEGIC PLAN FOR INTER-INSTITUTIONAL COOPERATION		103	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.15.	DEVELOPMENT, ADOPTION AND COMPLIANCE WITH THE FINANCIAL PLAN FOR EACH YEAR / QUARTER		886	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.17.	CLEARLY DEFINED DESCRIPTION OF THE POSITION, RIGHTS AND OBLIGATIONS OF EMPLOYEES		137	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.19.	CLEARLY DEFINE PRIORITIES		535	100,000.00 €	István Szencs	4.10.2014	31.12.2014	
WS 1.20.	CONTROL OF CONFORMANCE OF THE GOODS AND DISPATCH NOTES		87	100,000.00 €	István Szencs	4.10.2014	31.12.2014	

Figure 5

Risk treatment plan

Figure 6 shows part of the risk list before risk treatment, and risk – treated values after applying controls – measures, i.e., the reduction of threat levels for value 1.

After the risk treatment phase had been completed, 85 risk levels "5" remained (reduced by 999 risks) and 203 risk levels "4" (reduced by 506 risks).

By analyzing the results of the risk level "5", and after completion of the risk evaluation phase the following processes were identified as the most critical in the supply chain: "distribution planning", "production control" and "supply networking planning" - 15 risks; "information usage" - 13 risks; "execution of production", "jobs of acquisition and placement of funds" and "preparation of production" - 9 risks each at level "5".

Rec #	Risk ID	Asset name	Threats	Vulnerability	Before risk treatment						After risk treatment								
					T	V	P	C	I	A	Treatment	T	V	P	C	I	A	Risk	
13	8.347	COLLECTION AND PROCESSING OF INFORMATION	CS-COMPETENCE OF THE WORKFORCE	CS-INCREASING COSTS	3	3	3	4	3	4	5	3	2	2	4	4	3	4	5
14	8.348	COLLECTION AND PROCESSING OF INFORMATION	CS-COMPETENCE OF THE WORKFORCE	CS-REDUCED PRODUCT QUALITY	3	3	3	4	3	4	5	3	2	2	4	4	3	4	5
15	8.370	COLLECTION AND PROCESSING OF INFORMATION	CS-CONFUSING PLANS	CS-OWNS IN BUSINESS	1	2	2	4	3	4	3	1	1	2	2	4	3	4	3
16	8.346	COLLECTION AND PROCESSING OF INFORMATION	CS-CONFUSION DURING UNLOADING PROCESS	CS-EMPLOYEE INJURY	3	3	3	4	3	4	5	3	2	2	4	4	3	4	5
17	8.415	COLLECTION AND PROCESSING OF INFORMATION	CS-CREDIT INSOLVENCY	CS-SUBSIDIARY OF EQUIPMENT ACQUISITION	3	3	3	4	3	4	5	3	2	2	4	4	3	4	5
18	8.359	COLLECTION AND PROCESSING OF INFORMATION	CS-DELIVERED GOODS DO NOT CONFORM TO THE CONTRACT	CS-TERMINATION OF THE CONTRACT	2	2	4	4	3	4	4	3	1	2	2	4	3	4	3
19	8.411	COLLECTION AND PROCESSING OF INFORMATION	CS-DELIVERED GOODS DO NOT CONFORM TO THE CONTRACT	CS-TERMINATION OF THE CONTRACT	1	2	2	4	3	4	3	1	1	2	2	4	3	4	3
20	8.306	COLLECTION AND PROCESSING OF INFORMATION	CS-DISSATISFACTION OF WORKERS	CS-INEFFECTIVE AND INEFFICIENT PERFORMANCE OF EACH INDIVIDUAL TASK	3	3	3	4	3	4	5	3	2	2	4	4	3	4	5
21	8.361	COLLECTION AND PROCESSING OF INFORMATION	CS-DISSATISFACTION OF WORKERS	CS-INTENTIONAL WRONGFUL LABELING	3	2	4	4	3	4	5	3	2	1	2	4	3	4	3
22	8.396	COLLECTION AND PROCESSING OF INFORMATION	CS-EQUIPMENT NOT REQUIRED	CS-ADDITIONAL COSTS FOR THE RESTORING OF THE EQUIPMENT TO OPERATING STATE	2	3	4	4	3	4	5	3	1	2	2	4	3	4	3
23	8.308	COLLECTION AND PROCESSING OF INFORMATION	CS-FAILURE TO BREAK DOWN ALL THE TASKS OF THE COMPANY	CS-DUPLICATING WORK / EXTENSION OF THE PRODUCTION PROCESS	2	2	4	4	3	4	4	3	1	2	2	4	3	4	3
24	8.434	COLLECTION AND PROCESSING OF INFORMATION	CS-FAILURE TO BREAK DOWN ALL THE TASKS OF THE COMPANY	CS-DUPLICATING WORK / EXTENSION OF THE PRODUCTION PROCESS	3	2	4	4	3	4	5	3	2	1	2	4	3	4	3

File name: 000

Figure 6

List of risk, before and after risk treatment

The next step includes deciding as to whether to accept the estimated levels of risk or to implement the new controls / measures; next, it is necessary to re – execute the risk evaluation phase. In this study, the levels obtained by risk evaluation were accepted.

## Conclusions

Nowadays, decreasing the frequency and probability of risks occurring – from internal and environmental processes – can greatly reduce tensions within companies. Those reductions prepare conditions for formulating proper and operational strategies, thus, securing the continuation of those companies within the global markets.

Handling or managing risk is becoming a necessary and an objective assumption of business success and successful, risk management should provide continuous existence of the system. The documentation resulting from the risk management process enables the process of accreditation and authorization of the risk management process. Risk management should not be linked only to organizations, but it should apply to all individual activities of short-term or long-term nature.

The concept of supply chain management brings additional risks, which in accordance with the objective of management increases the level of integration and coordination among supply chain partners, and should be dealt with concurrently. These risks include the internal risks in the supply chain, such as co-operative and operational risks.

Risk management should be extended from the financial and corporate perspective, to the domain of logistics and inter-organizational cooperation. However, although it sounds simple, the problems to be solved are:

- the challenge of transforming risk management from legal obligations into a tool for planning;
- identification of hidden risks and their dissolution;
- quantification and the probability of the level of damage;
- application of bottom - up risk management in supply chains in order to avoid overloading the top management, as is the case in the top - down management.

Segments of the supply chain simulation by stochastic modeling in order to support mapping of risk impact assessment of different risk parameters variation may represent an elegant but a complicated task. Therefore, supply chain partners need to develop an understanding of the importance of the process of identification of the structure of key risks. Efficient security and the protection of the supply chain include basic standards for physical security, access control, personnel security, education and training, procedural security, IT security, business partners, as well as the safe transfer from point of origin to the final destination within the supply chain [15].

This paper describes the process of risk management in the supply chain by application of software tools. The disadvantage of the methodology used in this case study is the use of qualitative methods for risk assessment, which entails a certain degree of subjectivity of risk analysts and greatly depends on their experience. To minimize this subjectivity, it is necessary to use the same methodology, to have the risk assessed by several analysts and to compare their results to ensure a more objective assessment which is the key phase in the supply chain risk management.

This is an assignment project that was conducted at the University Singidunum (Belgrade, Serbia) by PhD students and their mentors. Contribution of authors can be seen in the detailed description of the risk assessment process in SCM, an adaptation of a software tool Hestia Risk, enabling the assessment of the risks as much as possible automate, accelerate and make more objective. Outputs from these case study can be used for educational purposes, as well as real-SCM system. Further research projects will be expanded by creating new Web applications to risk assessment in SCM, based on experiences and results of conducted case studies.

## References

- [1] F. H. Kloman (1999) Risk Management Agonistes, *Risk Analysis Journal*, 10(2): 201

- [2] J. Wisner, K. C. Tan, G. Leong (2015) *Principles of Supply Chain Management: A Balanced Approach*, Cengage Learning
- [3] Zurich Insurance Company, *Supply Chain Risk Assessment*, Zurich, Switzerland (2010)
- [4] P. Benedek (2012) *Compliance Management – a New Response to Legal and Business Challenges*, *Acta Polytechnica Hungarica*, 9(3): 135
- [5] Y. Yu, W. Xiong, Y. Cao (2015) A Conceptual Model of Supply Chain Risk Mitigation: The Role of Supply Chain Integration and Organizational Risk Propensity. *Journal of Coastal Research: Special Issue 73 - Recent Developments of Port and Ocean Engineering*: 95
- [6] V. T. Covello, J. Mumpower (1985) Risk Analysis and Risk Management: An Historical Perspective, *Risk Analysis*, 5(2): 103
- [7] M. Maslarić (2014) *Development of Model for Logistics Risk Management in Supply Chains*, PhD Thesis, Faculty of Technical Sciences, Novi Sad, Serbia
- [8] M. Cooper, D. Lambert, J. Pagh (1997) Supply Chain Management: More Than a New Name for Logistics. *The International Journal of Logistics Management*, 8(1): 1
- [9] F. Wiengarten, P. Humphreys, C. Gimenez, R. McIvor (2015) Risk, Risk Management Practices, and the Success of Supply Chain Integration, *International Journal of Production Economics*
- [10] International Organization for Standardization, *ISO 31000 – Risk management – Principles and guidelines* (2009)
- [11] International Organization for Standardization, *ISO/IEC 27005-Information technology – Security techniques – Information security risk management* (2011)
- [12] M. Christopher, H. Peck (2004) Building the Resilient Supply Chain, *International Journal of Logistics Management*, 15(2): 1
- [13] G. Smith, K. Watson, W. Baker, J. Pokorski (2007) A Critical Balance: Collaboration and Security in the IT-enabled Supply Chain, *International Journal of Production Research*, 45(11): 2595
- [14] P. Michelberger Jr., C. Lábodi (2012) After Information Security – Before a Paradigm Change (A Complex Enterprise Security Model), *Acta Polytechnica Hungarica*, 9(4): 101
- [15] M. Matotek, D. Regodić (2015) Human Resource Risk Management in Supply Chain, *Dyna Management*, 3(1): 1

# Factors Affecting the Selection and Implementation of a Customer Relationship Management (CRM) Process

**Regina Reicher, Ágnes Szeghegyi**

Óbuda University Keleti Faculty of Business and Management, Institute of Enterprise Management Tavaszmező u. 17, 1084 Budapest, Hungary  
reicher.regina@kgk.uni-obuda.hu; szeghegyi.agnes@kgk.uni-obuda.hu

---

*Abstract: The situation of Hungarian micro, small and medium-sized enterprises has become very difficult as the result of the economic changes of the recent years. The open European market, the economic crisis and the ever increasing competition demand fast reaction, creating a serious challenge for these enterprises coping with a lack of capital and other resources. Therefore, companies in the SME sector pay an increasing attention to serve their customers at a high level. In order to achieve this, they often seek an IT solution. Owing to these circumstances it is extremely important for the enterprises coping with harsh economic conditions to choose and implement the most suitable CRM IT solution quickly, efficiently and at the lowest risk of possible failure. This requires, however, the thorough knowledge of competencies of the organization, the wide range of solutions available on the market and adequate methodology which offers success for them. The present research aims to explore the factors affecting the decision-making process in the course of which the SME heads select and implement a CRM system. As the result of our research, the specification defined by experts and heads of user companies could be determined and classified.*

*Keywords: CRM implementation; CRM selection; factors affecting successful implementation; SME*

---

## 1 Introduction

The complexity of internal procedures within enterprises has reached a level to where the use of IT management tools has become a necessity. Fast development of business IT has helped large enterprises to develop strong capacities within this area. In addition, the introduction of IT systems prompted the rationalization of internal procedures.

Small and Medium-sized Enterprises (SMEs) took an interest in using IT assets as well. The market offered them a solution in the form of business management

software – even though these did not cover all operations of the company, just some fields, still, they have made some procedures more transparent and economic.

The primary goal of a CRM strategy is to develop and retain customer loyalty. Customer loyalty means a dedicated activity of the customer, which presupposes a positive customer attitude and a consequential customer behaviour.

The aim of using a CRM system is to organize the information arriving from various channels and to display it in a unified manner, sorting it by customers. Such cumulated customer history may improve the effectiveness of the company in the field of customer management and product development. It may define customer value and therefore it could further improve the effectiveness of handling important customers.

With this research we aim to concentrate exclusively on the Hungarian micro- and SME sector. The reason for this is the fact that large enterprises operate under different environmental conditions (capital strength and creditworthiness). We are not researching public service providers, Telecom service providers or companies in the banking sector, which are typical CRM users.

It is also relevant to mention that the number of CRM suppliers is rather large, and the competition in this sector undoubtedly strong.

In the light of the importance and relevance of the topic, we have set out the following aims.

- Analysis of the procedure and general issues of the implementation of a CRM IT system from the viewpoint of the supplier in case of SME users;
- Classification of the phases of the selection and implementation process, searching for common and different characteristics with the help of principal component analysis.

We expect that the results of the aimed research would help in identifying and organizing the factors within and surrounding IT suppliers and CRM buyer SMEs that influence the implementation and operation of CRM IT systems.

## **2 The CRM System**

According to Rust [14] strategic actions aiming at the construction of personalized, long-term relations were observed on the B2B market for many years. Examples include, a global accounts director appearing in an independent position, client managers receiving managerial positions and products that are developed in order to meet the demand of a single client. Similar examples can be seen on B2C markets, too. There are also, some initiatives observed in

multinational companies to improve the level of customer services, to map consumer needs and to develop the products or services according to the consumers' requirements.

Mester says that two important trends were involved in the development of CRM. One of them is the decreasing differences between products - this forces the market actors to distinguish themselves from their competitors by tailoring their products. The other one is the development of information technology which enabled the companies to collect and analyze the data about the clients with the help of various software products. [11]

Thus the operational strategy and the marketing concepts of companies changed fundamentally. When the product is not in the focus of marketing any more, the aim is not the sale of a certain product in any quantity to anyone with the maximum profit attainable, but the consumer is being put in the crosshairs of marketing. The common objective of marketing experts, salespeople and the company is to sell the most different products to the consumer at the highest possible profit, to the satisfaction of the customer. [7]

This change of attitudes describes the strategy of the companies fighting and surviving in the current economic crisis. This attitude affects the business activities and all the organizational units of the company, the areas of finance, marketing, logistics, quality assurance, information technology and sales, as well as all the levels of staff from the senior executives to the direct customer service representatives in the front office.

One of the main possibilities of CRM systems is the preparation of customer analyses. By using the database, the system is able to create client groups on the basis of given parameters, to make data analyses and to calculate client value. [3]

If we intend to group the functions, three areas can be distinguished. There are internal procedures where the aim is to support work-flow and to track staff performance. There are customer processes where the help-desk function and customer history is managed. And, finally, there are the statistics which include: customer segmentation, market basket analysis, attrition analysis and risk analysis.

The collected data can be analyzed, reports and statement can be compiled on the basis of different filters and complex conditions. That data is a great assistance in drafting forecasts, in the planning process, during product development and in the overall development of the whole organization. It may also help find and mark the path to be followed as management is not left to groping around in the dark.

There are several definitions of CRM in the literature. Approaching it by focusing on its content, in a functional manner, or from an IT viewpoint, company experts defined the features of CRM strategy and software differently.



According to Mátyás Gritsch CRM is „A customer-oriented philosophy which – based on modern information and communication techniques – aims to form profitable customer-relations in the long run”. [4]

Adrian Payne says that „CRM is a strategic method, the aim of which is to create higher shareholder value by developing proper relations with customers and customer groups. The CRM combines the possibilities of information technology and relationship marketing strategy in order to form profitable and long-term relations. It is important to note that CRM ensures wider possibilities for the use of data and information in order to reach better understanding of customers and to realize relationship marketing strategies at a higher level. This requires the integration of people, actions, processes and marketing possibilities across several functions which are enabled by information, technology and applications.” [12]

There are a number of other definitions, however, we can conclude that some key concepts emerge in all of them. All the definitions discuss the concepts of strategy, technology, client, organization and cooperation. It is obvious that the CRM definition can be drafted in several ways, these will differ only in regards to the approach but the above mentioned five key concepts are the same. It is also clear that the role of the technological framework, the information technology solution is the support of a CRM strategy with a modern tool. Furthermore, availability and implementation of a CRM philosophy is not required for to ensure a high-level of operation of customer services within a given company. The operation of a CRM IT solution is based on the customer strategy of the company and should support it.”The CRM is a company-wide, customer-oriented strategy which – in order to meet customer needs as effectively as possible – integrates the business processes into an IT solution representing the latest technology built on a database which contains all the data of customers. The IT support and the unified database enable the automatization of customer procedures and the processing of customer data in order to maximise profit. The human factor is a particularly important element of customer relations treatment.” [10]

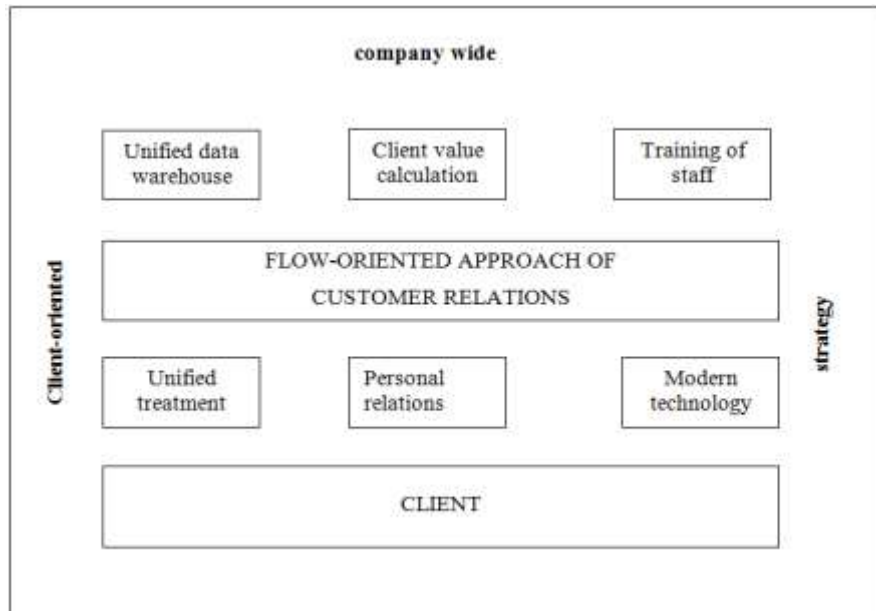


Figure 1

Comprehensive approach of customer relations [10]

## 2.1 CRM Types

From an information technology viewpoint, a CRM is a customer-oriented e-business solution. The CRM is basically built on an information database or data warehouse to which internet applications are connected. Its aim is to collect and analyse available data about customers in order to obtain a considerable volume of very valuable information about their purchasing habits, buying preferences and thus to determine the group of the most important clients.

On the basis of the usual classification offered by references, the CRM systems can be organised into three large groups on the basis of their functions and tasks within the company. These typizations are the most frequent on the distributors' market, too.

- Operative CRM is the part of the system which „deals with the automatization of business processes including front office costumer contact points. These areas apply the automatization of sales, marketing and help-desk services. Traditionally the operative CRM has a substantial share within corporate costs because a lot of companies do develop and utilize call centers and sales automatization systems. Consequently, the suppliers of CRM systems try to offer more and more types of alternative CRM-solutions”. [12]

Hetyei adds that this system controls and synchronises the client interactions in the field of services in addition to marketing and sales. He underlines the importance of recording all the customer data at the place where it arises and to make it available where and when it is needed. Front office staff and the staff of customer contact points belong to this function. [8]

MESTER, however, also draws attention to the fact that the operative CRM systems can display a wide range of client data but are not suitable to analyse them.” [11]

- Thus, in addition to the software solution providing operative functions, it can be necessary to use analytical CRM, which „includes the recording, storage, organisation, analysis and use data arising during the daily operation of an enterprise. The integration of analytical CRM solutions with operative CRM solutions is an important aspect.” [12]

Hetyei says that – in order to enhance the loyalty of profitable customers – analytics helps the optimisation of information sources by organising customers according to different aspects, customer segmentation, analysing client migration or by defining the target group of marketing campaigns. It enables the company to carry out analysis with marketing purposes. Moreover, it provides a way to make analysis concerning cross-sales, analysis regarding product affinity, to use customer value calculation models, to apply attrition models and to operate campaign management solutions. [8]

According to Mester, who examines the analytical tool with an IT approach „the data warehouse and all the related analysis tools can be actually defined as analytic CRM. What makes the totality of tools become a solution, is the repeatedly mentioned integration among them which – in an ideal case – hides the limits of the individual tools from the user and the preliminary, broadly tailored business intelligence – at least at a sectoral level – which provides solid basis for the solution of problems related with the information needs of the company.” [11]

- In addition to these, there is the collaborative CRM which “means the use of services and infrastructure and creates relations between the company and the multichannel sales system. This system connects the customers, the company and the employees.” [12]

Hetyei adds that the collaborative solution enhances cooperation with suppliers, partners and customers in order to refine processes and to serve the needs of customers. The collaborative CRM extends communication with clients to all the marketing channels, for example – in case of a bank – to bank branches, mobile bank services, internet and others. [8]

According to Mester, “the aim of collaborative CRM is to support the interactions between customers, suppliers, partners and company; as well as to distribute the customer information among the members of the corporation in order to create efficient communication and coordination and to fulfil the needs arising”. [11]

These boundaries blur in the practice. The individual functional types in themselves can operate less efficiently than when operating together. The usefulness of data collection by an operative CRM can be questioned if the user is not able to make analytics, reports, analyses and forecasts utilizing them. As the philosophy of a CRM application permeates the entire company, it becomes clear that the company should extend it not only to the customers, but to the partners and staff as well.

## **2.2 Implementation and Operation of a CRM**

If the company management makes a decision to implement a CRM software, the aspects of choosing the software do not always correspond to the real needs. The decision-making may be influenced by the size of the company, the special features of the industrial branch, financial possibilities and a lot of other factors. Thus the true needs are not necessarily reflected in the aspects of selection.

The implementation of CRM is a complex, long project task concerning multiple areas, in which the company staff and IT suppliers should cooperate. If the implementation of CRM is successful, it causes basic changes in the life of companies. Révész says that these changes mean increasing sales, growing satisfaction of consumers, specific decline of general and marketing costs, improvement and increasing efficiency of coordination among internal processes. [13]

Gritsch says that the starting point of CRM implementation must always be the description of strategic objectives of the company and the recording of customer-oriented corporate processes. [4]

Payne is slightly more sophisticated when giving advice concerning the implementation of CRM. He gives high priority to making a survey at the company before the implementation to see how much they are prepared for CRM activities. In this phase it is determined at what level the company is in, concerning, the field of customer treatment and what are the basic requirements for implementing the CRM. The factors that could hamper the successful implementation project are mapped. For example, the lack of skills, inadequate data quality, lack of participation on behalf of the management or the insufficiency of performance measuring systems.

Alshawi et al. performed research concerning the factors which affect SME actors concerning the acceptance of CRM. The research has revealed that the data quality, the organisational changes and the technical questions basically determine the relationship to the system. [1]

## 2.3 Small and Medium-Size Enterprises in Hungary

“The SME sector is one of the basic factors of economic growth in Hungary, and similarly to the developed countries of the world and Europe. The SMEs make up more than 99% of the total number of Hungarian enterprises.” [2]

The economic austerity, the heavy fiscal adjustments and the impacts of the global economic crisis have affected the results of companies very strongly. The global economic crisis has had a determinant impact on the Hungarian companies as well. Therefore, the small and medium-size enterprises also face the problem of declining domestic consumption. In order to maintain the level of their revenues, they have to implement new tools and methods in order to remain competitive on the narrowing market.

The SMEs react to every economic event very sensitively. All the cyclical fluctuations can be felt significantly and can be measured immediately in revenue and profit fluctuations. Since the economic environment is not permanent, the responsibility and the role of management becomes of exponential importance in this sector. The permanently changing environment, however, has made this sector rather flexible. The SMEs specialise, change their profile and corporate form more easily and they have a simpler organisational structure than large-scale companies.

The income generation of small and medium-size enterprises is more balanced than that of the large-scale companies. This leads to the conclusion that large-scale companies have a primary role in the emergence and persistence of economic disparities. The employment structure in case of small and medium-size enterprises is much more balanced in regional terms than in the case of large-scale companies (2013). [21]

On the basis of the flash report on competitiveness, made by Chikán *et al.* in 2010, it can be concluded that the ratio of external funds within the sources of the companies was 39% in 2004. This ratio declined to 27.7% by 2009. [5]

As a summary: the significance of small- and medium-sized enterprises cannot be neglected in the economic sector of Hungary. The Economic Development Operative Program (GOP), launched in the frames of New Hungary Development Plan and New Széchenyi Plan, was aimed primarily at the improvement of competitiveness of micro, small and medium-sized enterprises. Between 2007 and 2013, the SME sector could apply for IT development funds within the program. The aim of the program was to implement information and communication technologies in the business processes among companies and in services provided by the enterprises. Thus the enterprises could purchase and implement CRM systems as well.

According to the info-communication report of 2011, only 10% of the surveyed companies applied for funding towards their planned investments. [16]

Consequently, the preconditions of the appearance of CRM include information technology development as well. By analysing consumer databases it can be revealed which consumers bring profit and which do not, which services are high quality and which are not, and which marketing tools are efficient. The development of the internet was also a very strong influencing factor in the creation of CRM. The evolving e-trade has imposed new challenges for supplying and servicing companies.

### **2.3.1 CRM as Competitive Advantage**

The representative research of Deák and Mester in Hungary warns that the Hungarian companies have not been mature enough for the implementation of CRM systems. Almost 45% of CEOs think that conscious customer management means registration. Two-third of companies do not examine the reasons for losing customers and 25% of companies do not follow the needs of customers - by their own admission. It is clear that the companies are not aware of their possibilities. [6]

However, by 2012 the situation was completely different. The developed Western societies were quick to realise the business opportunities in customer loyalty therefore they soon started to use CRM applications. Loyalty 360 and SAS carried out a joint research program and examined 150 partner companies in the B2B and B2C market. They surveyed trends in the field of customer loyalty programs. The results were not very promising. Hardly a fourth of respondents regarded their program to be very good. Only 19% considered the reduction of client migration a strategic task. In the communication with the customers, mostly email and social media were emphasised. Only 36% of respondents integrated the loyalty data with other client data. "Companies which determine the customer life-cycle and adjust their loyalty building and customer-retention activities to this are usually more successful - said Wilson Raj, the Global Director of Customer Intelligence at SAS. [20]

## **2.4 Situation and Opportunities regarding the Competitiveness of SMEs in Hungary**

In Hungary, the small and medium-size enterprises are disadvantaged in regard to competitiveness compared to large-scale companies. Among other reasons, this is due to the lack of experts, lack of information and lack of corporate relations.

According to the competitiveness report of 2009, which is a research made among enterprises with at least 10 employees, it is common in many European countries that small and medium-size enterprises are not informed in the matters of the globalised world economy, they do not operate on international markets and are not confident enough to utilise the advantages offered by information and

communication technologies (ICT). Among the examined countries in the region, internet use connected with internal business processes, the use of an ERP system, CRM, e-invoicing or digital signature is the least widespread in Hungary and is far below the EU average. (Figure 2)

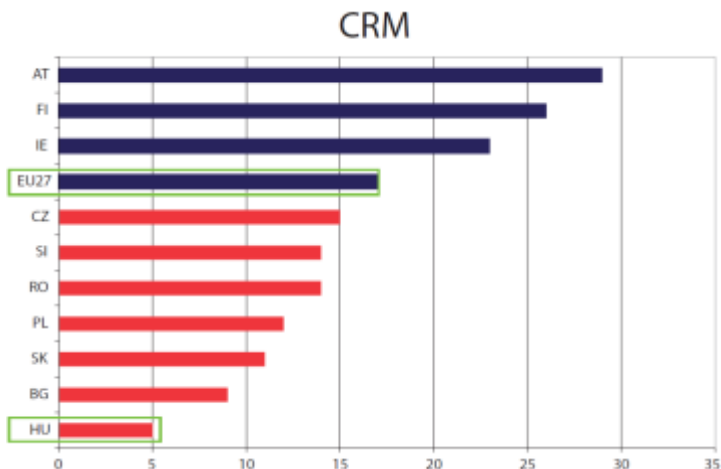


Figure 2

Expansion of CRM application in some EU member countries (%) [15]

The info-communication report of 2011 states that there is a huge gap between large-scale companies and small, medium-size enterprises in many areas, including investments. At present, SMEs concentrate primarily on infrastructural elements in their procurements, like PCs, hardware equipment, network development, etc. Investments supporting business objectives do emerge among large-scale companies and, the only exception is perhaps the solutions ensuring internet presence. Still, there are only a few of these in the SME sector. [16]

The research made by the Association of Chartered Certified Accountants (ACCA) in the third quarter of 2011 has shown that the 661 small and medium-sized enterprises in Europe do not utilise the advantages of the digital world although they are aware of the cost-reducing impact of technological applications. [17]

One of the most important opportunities for recovery, from the crisis, is the search for new markets, especially towards the former Comecon countries and Visegrád countries. The reduction of product quality was regarded as the least appropriate alternative; this may be explained due to the strong pressure of market competition. The increase of demand and the reduction of costs also received high scores from the leaders participating in the research. These two areas could be supported properly by the implementation of a CRM system.

According to a survey made by K&H Bank in 2012, staff increase of SMEs stagnated. The heads of the SMEs said that they did not plan to hire new employees but did not want to fire anyone, either. 70% of respondents did not think of any expansion in regards to staff. The research showed different ratios in case of each size unit. 14% of micro-enterprises plan to hire new staff, this ratio is 22% in the case of small-scale companies and more than double, 30% in case of medium-sized enterprises. [18]

„The SME trust index research made by K&H among the chief executives of Hungarian SMEs in the second quarter of 2013 confirms that the investment willingness of enterprises exceeded its lowest point. The ratio of SMEs considering investment increased again in the first quarter of this year, at present 58% of them plan investments. As regards to the subject of investments, the IT development projects are still given priority”. [19]

### **3 The Situation in Hungary**

#### **3.1 Expert Opinions**

Since there is only a limited amount of literature on the implementation and operation of CRM systems, and these focus mainly on large companies, we performed a primary exploratory qualitative assessment. For the purpose of analysing the Hungarian CRM suppliers and users, we organised our research around the following two primary quantitative assessments.

The interviewed companies were characterised by two types of data. One was the number of clients, ranging from 3 to 30. The other was the number of employees using the system. Here, the answers ranged from tens to 2000; these are approximations as no interview subjects had exact data. The answers were influenced by the fact that several companies could indicate only those clients in their answers who agreed to the use of their name as reference. Additionally, we observed another tendency: where the distributor has a foreign parent company, where that company offered remote access to its servers. This affects distributors active on the international market. For them it is hard to assess the number of domestic clients, as the parent company of the distributor may be offering remote access to a Hungarian subsidiary of a foreign client. The Hungarian CRM distributor will not know about this, as the license agreement was signed in the country of the parent company and access to the CRM system is provided from there.

Based on the literature we collected the critical points of CRM implementations, the general pitfalls, and then formulated the questions around these.



By characterising the companies which implement a CRM, the respondents could not indicate a typical industry. They said to have clients both in the manufacturing and the services sector. They mentioned several *motivation factors*. It is primarily the market stress, the serious competition for clients that motivates clients to implement a CRM system as a need for marketing support arises. Another motivating factor was state aid offered to help IT projects, as well as the cases where sales experts with experience working in multinational companies appeared as new employees at SMEs.

There were great differences in product *selection criteria* in the case of various CRM purchasers. Often the client has no expertise in CRM systems and the connected IT solutions, furthermore they negotiate without bringing an advisor or expert along, and therefore their decisions on matters beyond the price will be subjective, superficial and often unfounded. The *economic crisis* had *no positive results*. Hungarian companies did not recognize the importance of customer retention in this situation. It is mainly companies with foreign parents who have their own *implementation strategy*. With these companies the implementation is performed following a central strategy developed based on international experience. US, German and Swedish parent companies require a high level of support in this area as well from the representatives in every country. Two of the respondents mentioned an *implementation failure rate* higher than 50%. They concluded that more than 50% of their clients are not using the system after its implementation, regardless of education, good support and proper preparation. Everyday tasks and user resistance often raise serious issues in the company implementing a CRM. Another serious issue is raised by previously unplanned tasks becoming part of the everyday routine. Concerning the *duration of the implementation*, answers were spread out over on a broader scale, from the current week to one and a half years. Among the auxiliary services offered to the system, education and support were mentioned uniformly. Some mentioned advisory services at the development of the system plan, the assessment of existing hardware, the purchase of new hardware at a good price, periodical updates, patching and remote support.

Based on the experience of the IT suppliers it can be said that there are frequent cases where a CRM is implemented but not used; the reasons for this may be found in using a wrong procurement model. Usually the problem is caused by the non-optimal size of the procurement unit and its members not having the proper professional qualities. Choosing a supplier without an objective set of criteria is a very risky endeavour. The draft contract of the IT companies always foresees potential mistakes to be made by the purchaser; however, the purchaser often fails to contemplate its own potential and its limits.

### 3.2 Circumstances of Quantitative Research, Expert and User Sample Characteristics

We contacted 57 distributors asking them participate with our expert questionnaire and to forward our user questionnaire to their clients. We received 31 expert answers and 104 user answers. This means that 54% of the distributors (experts) have participated.

However, only estimations can be given on the percentage of user<sup>1</sup> answers, as no exact data is available on the number of users that have received the questionnaire. The estimate value was calculated based on the total number of clients of the software distributors contacted; this is about 1300-1400. Hence, the user response ratio is about 12-13%, which corresponds to the typical answer ratio of research conducted in the corporate sector.

The expert questionnaire was sent to 57 CRM system providers who are involved in the implementation, development and support of CRM systems. We paid attention to contacting a variety of experts which concerned the type of CRM systems implemented. Only three of them were of the previous interview subjects. Of the 31 experts, some were implementing installed software while others offered a cloud-based service. There was also a variety in the size of their clients, ranging from micro-enterprises to medium-sized or large companies. All software companies performed continuous development, and some of them offered individual development as well.

The 31 experts had a total experience of 170 years. The most experienced worked for 13 years as an expert and six of them had more than 10 years of experience; the average work experience is 5 and a half years. 11 experts worked on the implementation of 3 or more types of CRM systems, while 10 of them had experience with only one type.

The questions concerning CRM implementation can be organised in 3 clearly distinguishable groups on the basis of our previous experiences, the references and the interviews with experts. It is important to mention that human factors, the questions of IT support and corporate strategy play an important role. We tried to form the principal components in conjunction with these.

The opinion of respondents was measured in the following two questions, in a 10-grade high measuring level interval scale

---

<sup>1</sup> 'Users' are here companies who have bought a CRM system from the software distributors. We asked each company to fill out exactly one questionnaire.

16/22 Please evaluate on a scale of 1-10 how important you regard the following conditions for the SUCCESSFUL IMPLEMENTATION

	<i>1 not important at all, 10 very important</i>									
	1	2	3	4	5	6	7	8	9	10
mapping corporate maturity										
rationalisation of processes										
involving an independent consulting firm										
development of customer management strategy										
permanent management support										
exact definition of needs										
informing staff										
exact mapping of financial sources										

17/22 Please evaluate on a scale of 1-10 how important you regard the following conditions for the SUCCESSFUL OPERATION!

	<i>1 not important at all, 10 very important</i>									
	1	2	3	4	5	6	7	8	9	10
proper training during the implementation										
proper test running during the implementation										
permanent feedback to the management										
permanent system supervision by supplier										
defining an authorisation system										
using a correct database										
client meetings organised by supplier, dissemination of information										

The factor or principal component analysis is important because it reveals the hidden interrelations within the answers of the respondents. The characteristics belonging to one factor are linked in the mind of respondents as well, therefore they cannot be treated separately in the product policy. The high-weight variable is an important element of the factor or principal component, but to decide what is regarded as a high value the factor weight table is being used. [9]

Based on the main component analysis we could define three well separable components that are of great importance for our respondents and belong together according to their evaluation (Figure 3).

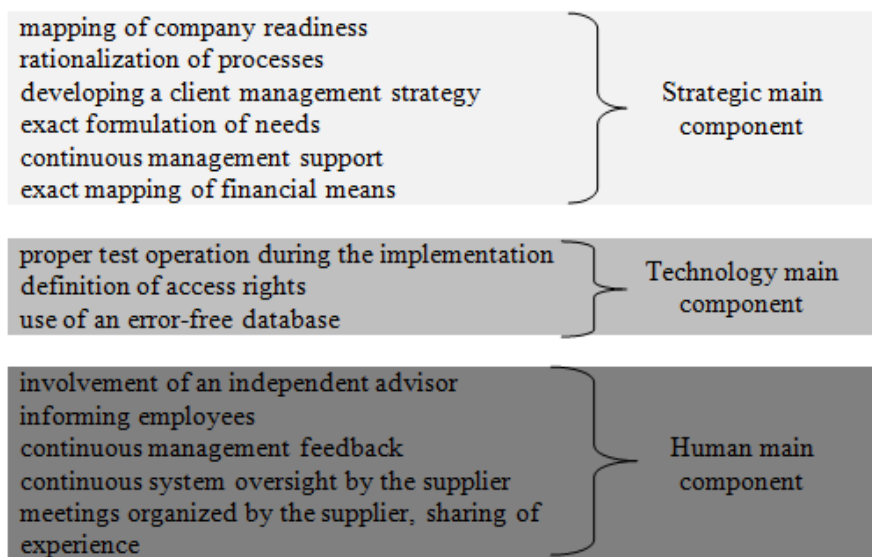


Figure 3

Results of the main component analysis

The examination and analysis of the final model seems thus to verify our original hypothesis that based on the answers of the supplier and the purchaser side, three sets of variables can be separated that are not correlating to each other and keep the information content of the different variables to the greatest extent. These three sets of variables represent the groups of human, strategic and technology variables. Thereafter, the attitude of the respondents can be examined alongside these components.

During our research we looked for the answer to the question what similarities there are between the views of suppliers and purchasers on the decisive reason for choosing a specific CRM and the motivation factors for the implementation.

Based on the answers of the supplier side and the purchaser side, three sets of variables can be separated: human, strategic and technological factors. The results of our questionnaires show that the (human, strategic and technological) factors defining the preparation and the implementation can be identified among the Hungarian suppliers and purchasers.

Identifying the (human, strategic and technological) elements of preferences in the preparation and implementation habits, effecting the decisions of Hungarian CRM suppliers and purchasers, give basis for efficient planning and implementing related to the preparation and the implementation.

Based on the results of the research we have proved that these factors play an important role in the success of the implementation for the respondents. The

examination was performed using quantitative methods: main component analysis and component testing.

### **Conclusions**

Having mapped the IT implementations it can be said that CRM implementations are mostly characterised by the feeling of failure from the side of both the supplier and the purchaser. The number of implementations does not decrease despite these feelings of failure and experts hope for a slow increase as a result of the improvement of the conditions after the waning of the economic crisis. During our research we could not aim to have a representative sample as no data is available on the statistical population. Based on the estimation of the experts, we can estimate that the yearly number of CRM implementations is 10 to 20. This is, however, the number of systems actually needing an implementation. Presently, the spreading of cloud-based systems increases and even multiplies this number. However, these software solutions often do not provide proper CRM background service, only sales support, which is only a fraction of the tasks of a CRM.

No reliable data is available either on the supplier or the purchaser market. Therefore the views of the two sides of each other are based only on experience. Even if the suppliers (understandably) have more experience, this experience cannot be generalised to cover the entire purchaser side. One of the aims of our studies was to reveal the real motivation factors behind the choice and implementation of a CRM system and to compare this to the picture developed by the supplier side.

In the next part of our research principal component analysis was used on the views of the respondents on the preparation and the implementation, where they assigned a value of importance to the various tasks on a scale of 1 to 10, 1 being the least important, while 10 being the most important.

The explained variance of the various principal components is split roughly and proportionately. Of course the results of the survey cannot be considered as a basis for conclusions for the entire domestic market given the deficiencies of the sampling; however, the general trends can be shown based on the answers of the suppliers and the purchasers.

It can be stated that the respondents divide the procedure of the preparation and the implementation into 3 main parts. The first contains steps containing strategic issues, involving the rationalisation of processes, the exact definition of demand, the development of a client management strategy, the mapping of company readiness and the exact mapping of financial potential. Although the weight of the various factors does not show large differences, the sequence here is not chronological, but that of importance based on the factor weights.

The second main component contains technology issues, involving the use of an error-free database, the definition of access rights and the management of the test

operation during the implementation. The order here again shows the order of the score values.

The third main component contains the human factors. In the order of factorial weight these are: client meetings organised by the supplier, experience sharing, continuous management feedback, continuous system overview by the supplier, continuous management support and the involvement of an independent advisor company in the process of the preparation and the implementation.

These interrelated elements extend beyond the examined stages. The groups formed of the various variables are independent of the stage of the implementation in which the given task arises. However, the technology component is an exception to this, as its appearance only makes sense after the selection.

Our research has proven that the implementing side suffers a serious lack of professional expertise. The methodical deficiencies of the supplier side are not helping in the solution of this issue. Furthermore, the supplier side has a superficial knowledge about the motivation and demands of the implementing side; therefore even if they have a methodology it may not put the emphasis on the proper points.

The results of principal component analysis have highlighted that the experts have different attitudes, they attach importance to different areas. Therefore later we would give priority to performing segmentation with the help of cluster analysis. It may help - after the survey made among existing staff - to form expert pairs to support the implementation project who can facilitate the success of the project by complementing each other. Thus none of the areas would prevail and proper emphasis can be given to each part of the task.

## References

- [1] Alshawi, S., Missi, F., Irani, Z. (2011): Organisational, Technical and Data Quality Factors in CRM Adoption – SMEs Perspective, *Industrial Marketing Management*, Vol. 40, Issue 3, April 2011, pp. 376-383
- [2] Bíró, P. (ed.) (2011): *Company Management and Marketing in E-Ages*, Tbalint Publisher, Törökbálint
- [3] Bohnné, K. K. (2005): *The Customer is Satisfied?* Public Press Ltd., Budapest
- [4] Chikán, A., Wimmer, Á. (2004): *Business Glossary*, Alinea Publisher, Budapest
- [5] Chikán, A., Czakó, E., Zoltayné, P. Z. (ed.) (2010): *Corporate Competitiveness in Times of Crisis*, Competitiveness Research Center, Budapest Corvinus University
- [6] Deák, Cs., Mester, Cs. (2005): *Change Management in the Backstage of CRM Projects*, *Business Studies*, Volume 3, Number 1, pp. 101-112

- [7] Erdélyi, E., Kovács, B., Merényi, A., Számely, É. (2006): Experience of the Best Practice in CRM Implementation in Industry, T-Mobile, Telecom Hungary XVII. évf. 2006/3 pp. 16-23
- [8] Heteyi, J. (2004): ERP systems in Hungary XXI. century, Computerbooks, Budapest
- [9] Hoffmann, M., Kozák, Á., Veres, Z. (2000): Market research, Technical publisher, Budapest
- [10] Mester, Cs. (2006): How to become a determining element of competitiveness of companies in the CRM? Management sciences XXXVII. year. 2006/ special issue pp. 87-97
- [11] Mester, Cs. (2007): The power of CRM, or customer relationship management in the Hungarian general corporate practice, PhD. Dissertation, University of Miskolc
- [12] Payne, A. (2007): CRM handbook, customer relations on high level, HVG publisher Ltd., Budapest
- [13] Révész, B. (2004): The Influence of CRM and E-CRM Systems on Customers' Opinion of a Company, 3<sup>rd</sup> International Conference for Young Researches pp. 244-251
- [14] Rust, R. T., Thompson, D. V., Hamilton, R. W. (2006): Do not over complicate this product! Harward Business manager, VIII. évf. 2006/9 pp. 50-59
- [15] Competitiveness Yearbook, GKI, 2009
- [16] Hungarian Information and Communications Report 2011
- [17] [http://www.vallalkozas-online.hu/index.php?option=com\\_content&task=view&id=515&Itemid=105](http://www.vallalkozas-online.hu/index.php?option=com_content&task=view&id=515&Itemid=105)
- [18] [http://www.marketinginfo.hu/hirek/article.php?id=24112&referer\\_id=newsletter](http://www.marketinginfo.hu/hirek/article.php?id=24112&referer_id=newsletter)
- [19] [https://www.kh.hu/publish/kh/hu/khcsoport/sajtokozlmeny/2013/2013\\_\\_II\\_\\_negyedev/mersekeltten\\_noevkv\\_\\_beruhazasi\\_kedv.html](https://www.kh.hu/publish/kh/hu/khcsoport/sajtokozlmeny/2013/2013__II__negyedev/mersekeltten_noevkv__beruhazasi_kedv.html)
- [20] [http://www.sas.com/resources/asset/Strategic\\_Imperative.pdf](http://www.sas.com/resources/asset/Strategic_Imperative.pdf)
- [21] <http://www.kaleidoszkop.nih.gov.hu/documents/15428/123426/kkv12>

# An Empirical Examination of Investment Risk Management Models for Serbia, Hungary, Croatia and Slovenia

**Vladimir Djakovic**

University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia, v\_djakovic@uns.ac.rs

**Igor Mladenovic**

University of Nis, Faculty of Economics, Trg Kralja Aleksandra 11, 18000 Nis, Serbia, igor.mladenovic@eknfak.ni.ac.rs

**Goran Andjelic**

Educons University, Faculty of Business Economy, Vojvode Putnika 87, 21208 Sremska Kamenica, Serbia, goran.andjelic@educons.edu.rs

---

*Abstract: The research presented in the study is the analysis and implementation of parametric and non-parametric Value at Risk (VaR) calculation models for predicting risk and determining the maximum potential loss from investment activities. The study sample includes stock indices of Serbian (BELEX15), Hungarian (BUX), Croatian (CROBEX) and Slovenian (SBITOP) markets, from 1<sup>st</sup> January 2006 to 31<sup>st</sup> December 2012. The methodology connotes the use of analysis and synthesis, as well as relevant statistical and mathematical methods. The study is based on the assumption that there is no statistically significant difference among the different models of risk management, in relation to the performance of investment risk prediction in the markets of the observed transition economies. The main aim of the study is to assess the performances of risk management models in practice, in order to operationally optimize investment decisions. The research results indicate the implementation adequacy of the tested models in the observed transitional markets, with full consideration of their specifics.*

*Keywords: risk management; value at risk; extreme value theory; historical simulation; delta normal VaR*

---



# 1 Introduction

Contemporary business conditions, characterized by dynamic changes both at micro and macro levels and frequent crises, challenge investing activities with constantly new and increasing risks. In order to successfully overcome the investment risk, in practice, numerous different models of risk management are developed to answer the challenges of ever changing markets and encounter the holders of investment decisions. This study analyzes and tests the implementation of various parametric and non-parametric Value at Risk (VaR) calculation models in order to predict the risk, focusing on the specific transitional markets, namely: Serbia, Hungary, Croatia and Slovenia. The lack of adequate data about the implementation of the chosen empirical models in transitional markets is particularly challenging. Such markets are interesting for the study especially because of their transitional features, and this fact takes on a whole new dimension with the beginning of the global economic crisis.

In accordance with the above mentioned, this study analyzes and tests the implementation of the following models: Extreme Value Theory (EVT), Historical Simulations (HS VaR) and Delta Normal VaR (D VaR), with the confidence level of 97.5% for 100 and 300 days (rolling windows), in the period from 2006 to 2012, for the observed markets. The real origin of the study objective is the researchers' ambition to affect significantly the risk level included in the investing activities by testing the models in observed transitional markets. Thus, the process of investing in these markets will be more acceptable to potential investors, their propulsive and liquidity will be increased, and real preconditions created for their further development and stabilization. In this context, the main objective of the study is to determine the performances of risk management models in practice to optimize investment decisions. The additional targets are set as follows: Ascertain whether the applied risk management models properly evaluate and predict the distribution tails of index daily returns in the target markets; Determine risks and analyze environmental factors important for successful investing; Analyze the implementation possibilities of the selected models in turbulent business conditions, and establish the fundamentals and guidelines for further development of the models in practice.

The basic hypothesis ( $H_0$ ) is that there is no statistically significant difference among the different models of risk management in relation to the performance of investment risk prediction in transition economy markets.

The derived hypotheses are the following:

- $H_1$ : There is no statistically significant difference of the risk management models' effectiveness in risk prediction for the indices in the period 2007-2012.

H<sub>2</sub>: There is no clearly defined boundary among the risk management models, in relation to the performance of risk prediction for the indices in the period 2007-2012.

H<sub>3</sub>: There is no statistically significant difference of the risk management models' effectiveness in risk prediction for the indices in particular years.

The research results are derived from the author's doctoral thesis [11] and will be useful both to academic and investment communities (public and regulatory). In fact, the results of the research should define the challenges and implementation opportunities of the selected investment risk management models, consistent with current trends in this area. In this way, the relevant guidance and possibilities will be provided to actively monitor the performance of the given risk management models' implementation in the observed transitional markets (Serbia, Hungary, Croatia and Slovenia), to optimize the investment process.

## 2 Theoretical Background

Numerous studies in the field of risk management have been conducted in developed markets, with special focus on the investment process and optimization of investment decisions. There are also studies in this area that indicate the specifics of transitional markets in the context of adequate risk prediction. This study is determined to analyze and test the performance of the risk management models' implementation in transitional markets because of their similar characteristics, i.e. the significant level of similarity in the markets of Serbia, Hungary, Croatia and Slovenia [1].

Basak and Shapiro [8] analyze the factors that affect the design of optimal, dynamic portfolio, with special emphasis on the market risk management using the Value at Risk models. In this context, they specifically investigate the impact of risky assets on the extreme portfolio losses. In order to invest efficiently, the authors propose an alternative model of risk management (LEL-RM) in order to eliminate VaR deficiencies, i.e. reduce the potential loss from investing activities. The authors particularly emphasize the necessity of incorporating the stock market volatility in the concerned model and the importance of permanent credit risk, which affects the expected return from the investment process.

Ait-Sahalia and Lo [12] suggest the implementation of nonparametric Value at Risk model of risk management to evaluate adequately extreme losses in investments. The authors imply how important it is to adjust the given model to risk aversion, investment time horizon and other important factors that influence the assessment and validation of investment risk. They also conclude that the statistical risk measures do not fully reflect the level of portfolio uncertainty, and that definition of maximum potential loss involves the analysis of environmental

conditions and economic validation of investment activities. Consequently, the authors present an alternative to statistical VaR and Arrow-Debreu framework in evaluating the economic VaR.

Rosenberg and Schuermann [4] indicate the importance of integrated risk management as a basis for measuring and managing risk and capital, with particular emphasis on the market and investment credit risk. They analyze measures of skewness and kurtosis and the thickness of the tails, i.e. the investment return distribution. This research is significant because it indicates the correlation between business risk and total investment risk, and correlative links between different types of risk. The above mentioned points to the complexity of investment risk prediction using a Value at Risk model. Consequently, the authors suggest a hybrid approach to enable adequate risk predictions, and particularly emphasize the importance of information, correlations, and unique risk characteristics.

Zikovic and Pecaric [9] analyze the implementation of Value at Risk investment model in distinctly transitional Croatian market. The focus of the research is the index of the Zagreb Stock Exchange (CROBEX). The research is particularly important because it points to the starting points and guidelines in the application of VaR models in transitional markets, especially in time of global economic crisis. The testing results of the implementation of VaR models indicate the discrepancy between actual and expected levels of risk resulting from extreme events in the market or extreme movements in stock returns. The authors conclude that the risk management models based on Extreme Value Theory (generalized Pareto distribution) have better performance in the prediction of risk/return of CROBEX stock index.

Campbell et al. [7] develop a portfolio selection model which is focused on maximizing the return from invested assets. In addition, the authors apply the Value at Risk model, i.e. the limits that are established by risk managers in determining the maximum potential loss from investing activities. The research sample consists of risky assets (U.S. stocks and bonds). The authors conclude that there are some similarities when it is assumed that investment returns are normally distributed compared with the mean-variance approach. They also highlight the significance of the impact of alternative time horizons and specific risks in the adequate portfolio selection, i.e. appropriate risk management in investments.

Bucevska [10] analyzes the importance of an adequate risk measure in the business environment which is characterized by negative consequences of global economic crisis. The author tests the Value at Risk performance models on the MBI10 stock exchange index in order to measure transitory market volatility, i.e. risk prediction, in the investment in market of the Republic of Macedonia. The research results of the GARCH model in turbulent business conditions point to the implications of econometric estimate of the given VaR models and show that EGARCH provides the best VaR estimation in the tested transitional market.

Uppal and Mangla [5] test the extreme losses, i.e. risks because of financial turbulence as a result of global financial crisis. The authors include ten countries in the research, in the period before and during the crisis. Investment risk prediction involves the analysis of the stock returns in the countries. The performances of the Extreme Value Theory model indicate different parameters of the estimated return distribution in the context of adequate investment risk prediction. Uncertainty is a fundamental problem in the implementation of the tested risk management model, especially in the period of frequent extreme events in the monitored markets. The research is important because it identifies key problems of creating extreme risks in investments, i.e. the maximum potential loss from investing activities.

The aforementioned studies provide valuable information for the investment risk prediction, both in developed and transitional markets, particularly in the context of optimal allocation of capital. The actuality of the studies is particularly evident because of existing and new challenges consequent of turbulent business conditions and global economic crisis. The lack of adequate empirical results from this area in transitional markets confirms the necessity of continuous tests of implementation performance of different risk management models, with the special emphasis on prediction of maximum potential loss from investing activities.

### 3 Methodology

The research methodology in the study involves the use of appropriate statistical and mathematical methods to analyze and test performances of different models of investment risk management, i.e. parametric and non-parametric models. Testing of implementation performances of Extreme Value Theory (EVT), Historical Simulation (HS VaR) and Delta Normal VaR (D VaR) models, with the confidence level of 97.5% for 100 and 300 days (rolling windows), is conducted from 2006 to 2012, in the markets of Serbia, Hungary, Croatia and Slovenia. The research focus is on adequate risk prediction or maximum potential risk from investing activities. The rolling windows of 100 and 300 days are used for the robust risk prediction of different models of risk management. The study sample includes values (returns) of BELEX15, BUX, CROBEX and SBITOP stock indices in the observed markets. The year 2006 is the period over which is calculated the initial value of VaR. The investment risk prediction is calculated on a daily basis depending on VaR rolling windows (100 and 300 days), and at the end of one year period, the number of days with unsuccessful prediction is compared with stock indices in the observed markets. The calculation results for a rolling window of 300 days shown in the subsequent tables cover the period from 2008 to 2012, because of the calculation characteristics, i.e. specifics, for a given

window. The sample characteristics have non-parametric attributes, so they are therefore analyzed with nonparametric procedures by frequency modalities. The multivariate procedures - MANOVA and discriminant analysis, can be applied to the scaled data. By calculating the discrimination coefficient (d. coeff), the features determining the specifics of subsamples (EVT, HS VaR and D VaR) are insulated, together with the features (years) that should be excluded from further processing, i.e. the observed research space is reduced. The applied univariate procedures are the following: Roy's Test, Pearson's Coefficient of Contingency ( $\chi$ ) and Multiple Correlation Coefficient (R). The purpose of conducting these procedures is to determine the characteristics of each subsample, as the basis of reliable and accurate prediction, with a certain confidence. MANOVA analysis (Multivariate Analysis of Variance) is used to establish whether there is a difference in the effectiveness of the various risk management models per year during the study period (2007-2012), while Roy's test determines the exact feature (year) of the difference (each year is tested separately). MANOVA analysis tests the hypothesis  $H_1$ , the discriminant analysis tests the hypothesis  $H_2$ , while Roy's test is used for the hypothesis  $H_3$ .

Mathematically, the risk management model can be calculated as follows: The  $\Delta p$  is the change in portfolio value in the next N days. VaR then corresponds to the loss (100-a) percentile of the  $\Delta p$  distribution. Then:

$$P(\Delta P < VaR) = 1 - a \quad (1)$$

VaR is the (100-a) percentile of the distribution value and it is generally calculated on a daily basis, with different levels of confidence. Delta normal VaR is calculated as follows:

$$VaR_{\alpha}^{\%} = Z_{\alpha} \sigma \quad (2)$$

where:  $Z_{\alpha}$  – the value of theoretical distribution,  $\sigma$  – standard deviation.

$Z_{\alpha}$  depends on the confidence level of VaR calculation. [6] in [2]

Extreme Value Theory studies the asymptotic properties of random values, in the forms:

$$M_n = \max\{X_1, X_2, \dots, X_n\}; m_n = \min\{X_1, X_2, \dots, X_n\}$$

when  $n \rightarrow \infty$ , whereat  $X_1, X_2, \dots, X_n$  are random values with given probability distributions. If for the random value  $M_n$  is valid:

$$P\left\{\frac{M_n - b_n}{a_n} \leq x\right\} \rightarrow G(x), n \rightarrow \infty \quad (3)$$

where  $G(x)$  is the non-degenerate distribution function, and  $a_n > 0$  and  $b_n$  ( $n \in \mathbb{N}$ ) being real numbers, then it is said that  $G(x)$  is determined by the marginal distribution of linearly normalized maxima  $M_n$ , with  $a_n$  and  $b_n$  being stipulating constants. [3]

## 4 Preliminary Data Analysis

Owing to the available historical data and the possibility of an adequate performance analysis of the implementation of Extreme Value Theory (EVT), Historical Simulation (HS VaR) and Delta Normal VaR (D VaR) models for investment risk predicting in the markets of Serbia, Croatia, Slovenia and Hungary, the research sample includes the daily returns of the market indices during the period from 1<sup>st</sup> January 2006 to 31<sup>st</sup> December 2012, i.e. a total of 1764 observation days. The year 2006 is used as the initial period for calculating VaR.

Table 1  
Kolmogorov-Smirnov test of the normality of the distribution sample in 2006

		CROBEX	BUX	SBITOP	BELEX15
N		252	252	252	252
Normal Parameters <sup>ab</sup>	Mean	2714.8106	22522.3334	1190.7842	1246.1715
	Std. Deviation	426.87002	1274.48005	163.50757	191.01019
Most Extreme Differences	Absolute	0.152	0.098	0.146	0.222
	Positive	0.120	0.058	0.146	0.222
	Negative	-0.152	-0.098	-0.063	-0.111
Kolmogorov-Smirnov Z		2.419	1.553	2.318	3.522
Asymp. Sig. (2-tailed)	(p)	0.000	0.016	0.000	0.000

Tables 1-7: <sup>a</sup> Test distribution is Normal; <sup>b</sup> Calculated from data

Source: the author's calculations [11]

Based on the 2006 Kolmogorov-Smirnov test results, it can be concluded that the values of the observed indices are not normally distributed, i.e. there are significant differences between the sampling distribution and the normal distribution, as indicated by the values of (p) for CROBEX (0.000), BUX (0.016), SBITOP (0.000) and BELEX15 (0.000) (Table 1).

Table 2  
Kolmogorov-Smirnov test of the normality of the distribution sample in 2007

		CROBEX	BUX	SBITOP	BELEX15
N		252	252	252	252
Normal Parameters <sup>ab</sup>	Mean	4583.3379	26099.6362	2159.6792	2615.7545
	Std. Deviation	558.84969	1881.77510	397.25782	378.11747
Most Extreme Differences	Absolute	0.213	0.101	0.177	0.203
	Positive	0.094	0.101	0.138	0.106
	Negative	-0.213	-0.034	-0.177	-0.203
Kolmogorov-Smirnov Z		3.377	1.599	2.814	3.221
Asymp. Sig. (2-tailed)	(p)	0.000	0.012	0.000	0.000

Based on the 2007 Kolmogorov-Smirnov test results, it can be concluded that the values of the observed indices are not normally distributed, i.e. there are significant differences between the sampling distribution and the normal distribution, as indicated by the values of (p) for CROBEX (0.000), BUX (0.012), SBITOP (0.000 ) and BELEX15 (0.000) (Table 2).

Table 3

Kolmogorov-Smirnov test of the normality of the distribution sample in 2008

		CROBEX	BUX	SBITOP	BELEX15
N		252	252	252	252
Normal Parameters <sup>ab</sup>	Mean	3414.7449	19800.2232	1673.6265	1436.8296
	Std. Deviation	914.35999	4141.47022	433.49060	543.20959
Most Extreme Differences	Absolute	0.196	0.211	0.126	0.118
	Positive	0.084	0.124	0.071	0.108
	Negative	-0.196	-0.211	-0.126	-0.118
Kolmogorov-Smirnov Z		3.117	3.343	2.005	1.877
Asymp. Sig. (2-tailed)	(p)	0.000	0.000	0.001	0.002

Based on the 2008 Kolmogorov-Smirnov test results, it can be concluded that the values of the observed indices are not normally distributed, i.e. there are significant differences between the sampling distribution and the normal distribution, as indicated by the values of (p) for CROBEX (0.000), BUX (0.000), SBITOP (0.001 ) and BELEX15 (0.002) (Table 3).

Table 4

Kolmogorov-Smirnov test of the normality of the distribution sample in 2009

		CROBEX	BUX	SBITOP	BELEX15
N		252	252	252	252
Normal Parameters <sup>ab</sup>	Mean	1855.5760	16047.3393	977.6396	597.7590
	Std. Deviation	267.78054	3876.14243	79.86016	136.76785
Most Extreme Differences	Absolute	0.084	0.130	0.164	0.073
	Positive	0.067	0.115	0.094	0.073
	Negative	-0.084	-0.130	-0.164	-0.057
Kolmogorov-Smirnov Z		1.328	2.057	2.604	1.160
Asymp. Sig. (2-tailed)	(p)	0.059	0.000	0.000	0.136

Based on the 2009 Kolmogorov-Smirnov test results, it can be concluded that the values of BUX and SBITOP indices are not normally distributed, i.e. there are significant differences between the sampling distribution and the normal distribution, as indicated by the values of (p) for BUX (0.000) and SBITOP (0.000). The values of CROBEX index are normally distributed, with the increased conclusion risk (0.059). BELEX15 index values are normally distributed (0.136) (Table 4).

Table 5  
Kolmogorov-Smirnov test of the normality of the distribution sample in 2010

		CROBEX	BUX	SBITOP	BELEX15
N		252	252	252	252
Normal Parameters <sup>ab</sup>	Mean	1990.4215	22482.2929	888.5902	659.3855
	Std. Deviation	136.02370	1161.77717	66.44689	39.43656
Most Extreme Differences	Absolute	0.173	0.072	0.158	0.147
	Positive	0.173	0.072	0.158	0.147
	Negative	-0.146	-0.043	-0.110	-0.083
Kolmogorov-Smirnov Z		2.739	1.150	2.514	2.338
Asymp. Sig. (2-tailed)	(p)	0.000	0.142	0.000	0.000

Based on the 2010 Kolmogorov-Smirnov test results, it can be concluded that the values of CROBEX, SBITOP and BELEX15 indices are not normally distributed, i.e. there are significant differences between the sampling distribution and the normal distribution, as indicated by the values of (p) for CROBEX (0.000), SBITOP (0.000) and BELEX15 (0.000). The values of BUX index are normally distributed (0.142) (Table 5).

Table 6  
Kolmogorov-Smirnov test of the normality of the distribution sample in 2011

		CROBEX	BUX	SBITOP	BELEX15
N		252	252	252	252
Normal Parameters <sup>ab</sup>	Mean	2083.0105	20468.0758	724.8801	670.1108
	Std. Deviation	200.92035	2879.32410	88.81687	103.59536
Most Extreme Differences	Absolute	0.233	0.225	0.132	0.215
	Positive	0.138	0.157	0.132	0.123
	Negative	-0.233	-0.225	-0.113	-0.215
Kolmogorov-Smirnov Z		3.693	3.565	2.093	3.416
Asymp. Sig. (2-tailed)	(p)	0.000	0.000	0.000	0.000

Based on the 2011 Kolmogorov-Smirnov test results, it can be concluded that the values of the observed indices are not normally distributed, i.e. there are significant differences between the sampling distribution and the normal distribution, as indicated by the values of (p) for CROBEX (0.000), BUX (0.000) SBITOP (0.000 ) and BELEX15 (0.000) (Table 6).

Based on the 2012 Kolmogorov-Smirnov test results, it can be concluded that the values of BUX, SBITOP and BELEX15 indices are not normally distributed, i.e. there are significant differences between the sampling distribution and the normal distribution, as indicated by the values of (p) for BUX (0.017), SBITOP (0.001) and BELEX15 (0.000). The values of CROBEX index are normally distributed (0.678) (Table 7).



Table 7  
Kolmogorov-Smirnov test of the normality of the distribution sample in 2012

		CROBEX	BUX	SBITOP	BELEX15
N		252	252	252	252
Normal Parameters <sup>a,b</sup>	Mean	1735.9712	18078.4705	566.7558	473.9669
	Std. Deviation	55.11718	886.48860	34.73749	39.56733
Most Extreme Differences	Absolute	0.045	0.098	0.125	0.174
	Positive	0.045	0.062	0.121	0.174
	Negative	-0.036	-0.098	-0.125	-0.121
Kolmogorov-Smirnov Z		0.720	1.548	1.979	2.768
Asymp. Sig. (2-tailed)	(p)	0.678	0.017	0.001	0.000

Providing the above mentioned, based on the 2012 Kolmogorov-Smirnov test results, it can be concluded that the values of CROBEX, BUX, SBITOP and BELEX15 indices are different from a normal distribution in 2006, 2007, 2008 and 2011, while there is the normal distribution for the following index values: for CROBEX and BELEX15 in 2009, for BUX in 2010 and for CROBEX in 2012.

## 5 Results and Discussion

In this section, the hypotheses  $H_1$ ,  $H_2$  and  $H_3$  will be either proved or rejected for CROBEX, with the confidence level of 97.5%, for 100 days, in the period from 2007 to 2012.

Table 8

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for CROBEX (97.5%; 100 days) in the period 2007-2012

Analysis	n	F	p
MANOVA	6	4.088	0.000
Discriminant	6	4.086	0.000

Tables 8, 10, 12, 14, 16, 18, 20 and 22:

Legend:  $n$  – years (features),  $F$  – the values of Fisher distribution,  $p$  – significance level

Source: the author's calculations [11]

Based on the value of  $p=0.000$  (MANOVA analysis) and  $p=0.000$  (discriminant analysis), the hypotheses  $H_1$  and  $H_2$  are rejected; the alternative hypotheses  $A_1$  and  $A_2$  are accepted for CROBEX (97.5%; 100 days). Consequently, there is the difference and the clearly defined boundary among VaR calculation models for CROBEX (97.5%; 100 days).

Table 9

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for CROBEX (97.5%; 100 days) in particular years

Year	$\chi$	R	F	p	d. coeff
2007	0.087	0.087	2.884	0.057	0.020
2008	0.020	0.020	0.150	0.861	0.009
2009	0.023	0.023	0.201	0.818	0.000
2010	0.100	0.100	3.813	0.023	0.020
2011	0.062	0.062	1.476	0.230	0.023
2012	0.026	0.026	0.255	0.775	0.009

Tables 9, 11, 13, 15, 17, 19, 21 and 23:

Legend:  $\chi$  – Pearson's contingency coefficient, R – multiple correlation coefficient, F – Fischer value distribution, p – significance level, d. coeff – discrimination coefficient

Source: the author's calculations [11]

Since  $p < 0.1$  (Roy's test), the alternative hypothesis  $A_3$  is accepted, which means that there is a significant difference among the VaR calculation models for CROBEX (97.5%; 100 days) in the performance of risk prediction, observed by years, such as: in 2007 (0.057) and in 2010 (0.023). Since  $p > 0.1$  (Roy's test), there is no reason not to accept the hypothesis  $H_3$ , which means that no significant difference among the VaR calculation models for CROBEX is observed (97.5%; 100 days) in the performances of risk prediction, by years, such as: in 2008 (0.861), in 2009 (0.818), in 2011 (0.230) and in 2012 (0.775). The discrimination coefficient indicates that the greatest contribution to discrimination, i.e. the biggest difference, is among the VaR calculation models for CROBEX (97.5%; 100 days) in performances of risk prediction, in the following years, respectively: in 2011 (0.023) in 2007 (0.020), in 2010 (0.020), in 2008 (0.009) and in 2012 (0.009). It should be noted that the latent feature is the one in which there is no difference among the VaR calculation models, and discriminant analysis includes the same in the structure in which there is a significant difference among the VaR calculation models. The latent feature models are the years 2008 (0.861), 2011 (0.230) and 2012 (0.775).

In this section, the hypotheses  $H_1$ ,  $H_2$  and  $H_3$  for CROBEX will be either proved or rejected, with a confidence level of 97.5%, for 300 days, in the period from 2008 to 2012.

Table 10

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for CROBEX (97.5%; 300 days) in the period 2008-2012

Analysis	n	F	p
MANOVA	5	0.007	1.000
Discriminant	4	1.728	0.089

Based on the values of  $p=1.000$  (MANOVA analysis) and  $p=0.089$  (discriminant analysis), there is no reason not to accept the hypothesis  $H_1$ , reject the hypothesis  $H_2$  and accept the alternative hypothesis  $A_2$  for CROBEX (97.5%; 300 days). This means that there is no difference among the VaR calculation models, but nevertheless, there is a clearly defined boundary among them. This fact suggests that there probably exist latent characteristics that in conjunction with other features (synthesized) contribute to discrimination of VaR calculation models. The starting unit, i.e. system, is reduced to the system of 4 features instead of 5, with a difference and boundary existing among the VaR calculation models for CROBEX (97.5%; 300 days).

Table 11

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for CROBEX (97.5%; 300 days) in particular years

Year	$\chi^2$	R	F	p	d. coeff
2008	0.040	0.040	0.600	0.549	0.004
2009	0.036	0.037	0.503	0.605	0.002
2010	0.000	0.000	0.000	1.000	0.000
2011	0.071	0.071	1.915	0.148	0.010
2012	0.073	0.073	2.021	0.134	0.006

Since  $p>0.1$  (Roy's test), there is no reason not to accept the hypothesis  $H_3$ , which means that there is no significant difference among the VaR calculation models for CROBEX (97.5%; 300 days) in the performances of risk prediction, in particular years, such as: in 2008 (0.549), in 2009 (0.605), in 2010 (1.000), in 2011 (0.148), and in 2012 (0.134). The discrimination coefficient indicates that the greatest contribution to discrimination, i.e. the most important difference, is among the VaR calculation models for CROBEX (97.5%; 300 days) in performances of risk prediction, in the following years, respectively: in 2011 (0.010), in 2012 (0.006), in 2008 (0.004), and in 2009 (0.002). It should be noted that the latent feature is the one in which there is no difference among the VaR calculation models, and discriminant analysis includes the same in the structure in which there is a significant difference among the VaR calculation models. The latent feature models are the years 2008 (0.549), 2009 (0.605), 2011 (0.148) and 2012 (0.134).

In this section, the hypotheses  $H_1$ ,  $H_2$  and  $H_3$  for BUX will be either proved or rejected, with a confidence level of 97.5% for 100 days, in the period from 2007 to 2012.

Table 12

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for BUX (97.5%; 100 days) in the period 2007-2012

Analysis	n	F	p
MANOVA	6	3.135	0.000
Discriminant	6	3.132	0.000

Based on the values of  $p=0.000$  (MANOVA analysis) and  $p=0.000$  (discriminant analysis), the hypotheses  $H_1$  and  $H_2$  are rejected; the alternative hypotheses  $A_1$  and  $A_2$  are accepted for BUX (97.5%; 100 days). Consequently, there is the difference and the clearly defined boundary among the VaR calculation models for BUX (97.5%; 100 days).

Table 13

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for BUX (97.5%; 100 days) in particular years

Year	$\chi$	R	F	p	d. coeff
2007	0.080	0.081	2.462	0.086	0.013
2008	0.034	0.034	0.441	0.644	0.008
2009	0.078	0.078	2.315	0.100	0.008
2010	0.056	0.056	1.171	0.311	0.014
2011	0.040	0.040	0.619	0.539	0.024
2012	0.030	0.030	0.337	0.714	0.003

Since  $p < 0.1$  (Roy's test), the alternative hypothesis  $A_3$  is accepted, which means that there is a significant difference among some VaR calculation models for BUX (97.5%; 100 days) in the performances of risk prediction, observed by years, such as in 2007 (0.086). Since  $p > 0.1$  (Roy's test), there is no reason not to accept the hypothesis  $H_3$ , which means that no significant difference among the VaR calculation models for BUX is observed (97.5%; 100 days) in the performances of risk prediction, by years, such as: in 2008 (0.644), in 2009 (0.100), in 2010 (0.311), in 2011 (0.539) and in 2012 (0.714). The discrimination coefficient indicates that the greatest contribution to discrimination, i.e. the biggest difference, is among the VaR calculation models for BUX (97.5%; 100 days) in performances of risk prediction, in the following years, respectively: in 2011 (0.024), in 2010 (0.014), in 2007 (0.013), in 2008 (0.008), in 2009 (0.008) and in 2012 (0.003). It should be noted that the latent feature is the one in which there is no difference among the VaR calculation models, and discriminant analysis includes the same in the structure in which there is a significant difference among the VaR calculation models. The latent feature models are the years 2008 (0.664), 2009 (0.100), 2010 (0.311), 2011 (0.539) and 2012 (0.714).

In this section, the hypotheses  $H_1$ ,  $H_2$  and  $H_3$  for BUX will be either proved or rejected, with a confidence level of 97.5% for 300 days, in the period from 2008 to 2012.

Table 14

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for BUX (97.5%; 300 days) in the period 2008-2012

Analysis	n	F	p
MANOVA	5	2.537	0.005
Discriminant	5	2.538	0.005

Based on the value of  $p=0.005$  (MANOVA analysis) and  $p=0.005$  (discriminant analysis), the hypotheses  $H_1$  and  $H_2$  are rejected; the alternative hypotheses  $A_1$  and  $A_2$  are accepted for BUX (97.5%; 300 days). Consequently, there is the difference and the clearly defined boundary among the VaR calculation models for BUX (97.5%; 300 days).

Table 15

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for BUX (97.5%; 300 days) in particular years

Year	$\chi$	R	F	p	d. coeff
2008	0.007	0.007	0.017	0.983	0.003
2009	0.066	0.066	1.645	0.194	0.018
2010	0.048	0.048	0.884	0.414	0.002
2011	0.074	0.074	2.077	0.126	0.012
2012	0.073	0.073	2.011	0.135	0.010

Since  $p>0.1$  (Roy's test), there is no reason not to accept the hypothesis  $H_3$ , which means that there is no significant difference among the VaR calculation models for BUX (97.5%; 300 days) in the performances of risk prediction, in particular years, such as: in 2008 (0.983), in 2009 (0.194), in 2010 (0.414), in 2011 (0.126) and in 2012 (0.135). The discrimination coefficient indicates that the greatest contribution to discrimination, i.e. the biggest difference, is among the VaR calculation models for BUX (97.5%; 300 days) in performances of risk prediction, in the following years, respectively: in 2009 (0.018), in 2011 (0.012), in 2012 (0.010), in 2008 (0.003) and in 2010 (0.002). It should be noted that the latent feature is the one in which there is no difference among the VaR calculation models, and discriminant analysis includes the same in the structure in which there is a significant difference among the VaR calculation models. The latent feature models are the years 2008 (0.983), 2009 (0.194), 2010 (0.414), 2011 (0.126) and 2012 (0.135).

In this section, the hypotheses  $H_1$ ,  $H_2$  and  $H_3$  for SBITOP will be either proved or rejected, with a confidence level of 97.5%, for 100 days, in the period from 2007 to 2012.

Table 16

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for SBITOP (97.5%; 100 days) in the period 2007-2012

Analysis	n	F	p
MANOVA	6	2.336	0.006
Discriminant	6	2.332	0.007

Based on the value of  $p=0.006$  (MANOVA analysis) and  $p=0.007$  (discriminant analysis), the hypotheses  $H_1$  and  $H_2$  are rejected; the alternative hypotheses  $A_1$  and  $A_2$  are accepted for SBITOP (97.5%; 100 days). Consequently, there is the difference and the clearly defined boundary among the VaR calculation models for SBITOP (97.5%; 100 days).

Table 17

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for SBITOP (97.5%; 100 days) in particular years

Year	$\chi$	R	F	p	d. coeff
2007	0.096	0.096	3.509	0.031	0.017
2008	0.025	0.025	0.239	0.788	0.002
2009	0.036	0.037	0.503	0.605	0.000
2010	0.052	0.052	1.026	0.359	0.006
2011	0.043	0.043	0.704	0.495	0.007
2012	0.056	0.056	1.196	0.303	0.013

Since  $p < 0.1$  (Roy's test), the alternative hypothesis  $A_3$  is accepted, which means that there is a significant difference among some VaR calculation models for SBITOP (97.5%; 100 days) in the performances of risk prediction, observed by years, such as in 2007 (0.031). Since  $p > 0.1$  (Roy's test), there is no reason not to accept the hypothesis  $H_3$ , which means that no significant difference among the VaR calculation models for SBITOP is observed (97.5%; 100 days) in the performances of risk prediction, by years, such as: in 2008 (0.788), in 2009 (0.605), in 2010 (0.359), in 2011 (0.495) and in 2012 (0.303). The discrimination coefficient indicates that the greatest contribution to discrimination, i.e. the biggest difference, is among the VaR calculation models for SBITOP (97.5%; 100 days) in the performance of risk prediction, in the following years, respectively: in 2007 (0.017), in 2012 (0.013), in 2011 (0.007), in 2010 (0.006) and in 2008 (0.002). It should be noted that the latent feature is the one in which there is no difference among the VaR calculation models, and discriminant analysis includes the same in the structure in which there is a significant difference among the VaR calculation models. The latent feature models are the years 2008 (0.788), 2010 (0.359), 2011 (0.495) and 2012 (0.303).

In this section, the hypotheses  $H_1$ ,  $H_2$  and  $H_3$  for SBITOP will be either proved or rejected, with a confidence level of 97.5% for 300 days, in the period from 2008 to 2012.

Table 18

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for SBITOP (97.5%; 300 days) in the period 2008-2012

Analysis	n	F	p
MANOVA	5	1.375	0.188
Discriminant	2	2.022	0.090

Based on the value of  $p=0.188$  (MANOVA analysis) and  $p=0.090$  (discriminant analysis), there is no reason not to accept the hypothesis  $H_1$  and reject the hypothesis  $H_2$ , and accept the alternative hypothesis  $A_2$  for SBITOP (97.5%; 300 days). It means that there is no significant difference among VaR calculation models, yet there exists a clearly defined boundary among the VaR calculation models. This fact indicates that probably there exist latent features that in conjunction with other features (synthesized) contribute to discrimination of VaR calculation models. The starting unit, i.e. system, is reduced to the system of 2 features instead of 5, with a difference and boundary existing among VaR calculation models for SBITOP (97.5%; 300 days).

Table 19

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for SBITOP (97.5%; 300 days) in particular years

Year	$\gamma$	R	F	p	d. coeff
2008	0.052	0.052	1.041	0.354	0.000
2009	0.036	0.037	0.503	0.605	0.005
2010	0.052	0.052	1.012	0.364	0.000
2011	0.076	0.076	2.194	0.113	0.000
2012	0.078	0.078	2.294	0.102	0.009

Since  $p>0.1$  (Roy's test), there is no reason not to accept the hypothesis  $H_3$ , which means that there is no significant difference among the VaR calculation models for SBITOP (97.5%; 300 days) in the performances of risk prediction, in particular years, such as: in 2008 (0.354), in 2009 (0.605), in 2010 (0.364), in 2011 (0.113) and in 2012 (0.102). The discrimination coefficient indicates that the greatest contribution to discrimination, i.e. the biggest difference, is among VaR calculation models for SBITOP (97.5%; 300 days) in performances of risk prediction, in the following years, respectively: in 2012 (0.009) and in 2009 (0.005). It should be noted that the latent feature is the one in which there is no difference among the VaR calculation models, and discriminant analysis includes the same in the structure in which there is a significant difference among the VaR calculation models. The latent feature models are the years 2009 (0.605) and 2012 (0.102).

In this section, the hypotheses  $H_1$ ,  $H_2$  and  $H_3$  for BELEX15 will be either proved or rejected, with a confidence level of 97.5%, for 100 days, in the period from 2007 to 2012.

Table 20

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for BELEX15 (97.5%; 100 days) in the period 2007-2012

Analysis	n	F	p
MANOVA	6	3.681	0.000
Discriminant	6	3.688	0.000

Based on the values of  $p=0.000$  (MANOVA analysis) and  $p=0.000$  (discriminant analysis), the hypotheses  $H_1$  and  $H_2$  are rejected; the alternative hypotheses  $A_1$  and  $A_2$  are accepted for BELEX15 (97.5%; 100 days). Consequently, there is the difference and the clearly defined boundary among the VaR calculation models for BELEX15 (97.5%; 100 days).

Table 21

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for BELEX15 (97.5%; 100 days) in particular years

Year	$\chi$	R	F	p	d. coeff
2007	0.039	0.039	0.567	0.568	0.024
2008	0.027	0.027	0.266	0.766	0.013
2009	0.030	0.030	0.337	0.714	0.002
2010	0.068	0.069	1.782	0.169	0.012
2011	0.037	0.037	0.517	0.597	0.020
2012	0.091	0.091	3.160	0.043	0.016

Since  $p < 0.1$  (Roy's test), the alternative hypothesis  $A_3$  is accepted, which means that there is a significant difference among some VaR calculation models for BELEX15 (97.5%; 100 days) in the performances of risk prediction, observed by years, such as in 2012 (0.043). Since  $p > 0.1$  (Roy's test), there is no reason not to accept the hypothesis  $H_3$ , which means that no significant difference among the VaR calculation models for BELEX15 is observed (97.5%; 100 days) in the performances of risk prediction, by years, such as: in 2007 (0.568), in 2008 (0.766), in 2009 (0.714), in 2010 (0.169) and in 2011 (0.597). The discrimination coefficient indicates that the greatest contribution to discrimination, i.e. the biggest difference, is among the VaR calculation models for BELEX15 (97.5%; 100 days) in the performances of risk prediction, in the following years, respectively: in 2007 (0.024), in 2011 (0.020), in 2012 (0.016), in 2008 (0.013), in 2010 (0.012) and in 2009 (0.002). It should be noted that the latent feature is the one in which there is no difference among the VaR calculation models, and discriminant analysis includes the same in the structure in which there is a significant difference among the VaR calculation models. The latent feature models are the years 2007 (0.568), 2008 (0.766), 2009 (0.714), 2010 (0.169) and 2011 (0.597).



In this section, the hypotheses  $H_1$ ,  $H_2$  and  $H_3$  for BELEX15 will be either proved or rejected, with a confidence level of 97.5%, for 300 days, in the period from 2008 to 2012.

Table 22

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for BELEX15 (97.5%; 300 days) in the period 2008-2012

Analysis	n	F	p
MANOVA	5	1.001	0.442
Discriminant	2	1.465	0.212

Based on the values of  $p=0.442$  (MANOVA analysis) and  $p=0.212$  (discriminant analysis), there is no reason not to accept the hypotheses  $H_1$  and  $H_2$ . It means that there is no significant difference or a clearly defined boundary among the VaR calculation models. Not even after the reduction of the starting unit, i.e. system, from 5 to 2 features, there is no difference or boundary existing among the VaR calculation models for BELEX15 (97.5%; 300 days).

Table 23

The significance of the difference among the VaR calculation models in relation to performances of risk prediction for BELEX15 (97.5%; 300 days) in particular years

Year	$\chi^2$	R	F	p
2008	0.025	0.025	0.239	0.788
2009	0.052	0.052	1.005	0.367
2010	0.051	0.051	1.001	0.368
2011	0.045	0.045	0.756	0.470
2012	0.051	0.051	1.001	0.368

Since  $p>0.1$  (Roy's test), there is no reason not to accept the hypothesis  $H_3$ , which means that there is no significant difference among the VaR calculation models for BELEX15 (97.5%; 300 days) in the performances of risk prediction, in particular years, such as: in 2008 (0.788), in 2009 (0.367), in 2010 (0.368), in 2011 (0.470) and in 2012 (0.368).

## Conclusions

When analyzing the results obtained in the research, it can be concluded that the tested risk management models used in investment risk prediction, enable the determination of the maximum possible loss from investment activities, in the observed markets. The results indicate that the rolling window, with fewer days, is more sensitive to changes in the daily index values. With the rolling window of 300 days, there are less mutual variations in effectiveness of the tested models.

Exploring the characteristics of the application of the models D VaR and EVT, based on the results obtained, it can be concluded that the differences in performance of these models are not significant, while with the HS VaR model

these differences are much more considerable. Of course, practice confirms the exceptions, as well. In the case of the year 2011, the HS VaR and EVT models show greater similarity, so this fact can be a good basis for a subsequent study, which will be focused on even more analytical approach to the implementation of these models.

With the analysis of the research results, it can be concluded that the basic hypothesis  $H_0$  is rejected, i.e. there is a statistically significant difference among the Extreme Value Theory (EVT), Delta Normal VaR (D VaR) and Historical Simulation (HS VaR) models, according to successful investment risk prediction in the markets of the transition economies. The hypotheses  $H_1$ ,  $H_2$  and  $H_3$  are also rejected, which only confirms the need for the specific research in terms of testing implementation performances of the observed models in the analyzed markets.

Volatility of business, as a consequence of a global economic crisis, significantly affects the adequate implementation of the tested models of investment risk management, i.e. adequate predictions of the maximum possible losses from investing activities. Such conditions are characterized by extreme events (extreme movements in the observed markets), low liquidity of the observed markets, low level of transparent business operations, their substantial inefficiency, incomplete institutional frameworks, and so on.

The importance of the research in this study has its academic, professional and practical dimensions. The academic dimension stems from the fact that until now very little research has been conducted in this area, focused on the markets of transition economies, so, in this respect, the conclusions obtained through the specific research open the "gate" for further steps in the analysis of the implementation of the tested risk management models for these markets. Professional-practical dimension of the research is the importance of the obtained information on the implementation specifics and modalities of the tested models in the observed transitional markets.

Considering all the above, further research will primarily focus on analysis of the specific implementations of the observed models in transitional markets, in terms of testing their adaptabilities, to the ever-changing conditions and turbulent environment, that is inherent in these markets. In this way, all the interested investment parties (academic and professional communities, policy makers, etc.) will have access to reliable information, that is, in practice, quantitatively tested concerning the possibilities of model implementation in transitional markets.

### **Acknowledgement**

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, within the Project No. III47028.

## References

- [1] G. Andjelic, V. Djakovic, Financial Market Co-Movement between Transition Economies: A Case Study of Serbia, Hungary, Croatia and Slovenia, *Acta Polytechnica Hungarica, Journal of Applied Sciences*, Budapest, Hungary, Vol. 9, No. 3, pp. 115-134, 2012
- [2] G. B. Andjelic, V. Dj. Djakovic, M. M. Sujic, An Empirical Evaluation of Value-At-Risk: The Case of the Belgrade Stock Exchange Index - BELEX15, *Industrija*, Vol. 40, No. 1, pp. 39-60, 2012
- [3] J. Jockovic, Generalized Pareto Distributions in Extreme Value Theory and their Implementations, M.Sc. Thesis, University of Belgrade, Faculty of Mathematics, pp. 1-56, 2009
- [4] J. V. Rosenberg, T. Schuermann, A General Approach to Integrated Risk Management with Skewed, Fat-tailed Risks, Staff Report, Federal Reserve Bank of New York, No. 185, pp. 1-59, 2004
- [5] J. Y. Uppal, I. U. Mangla, Extreme Loss Risk in Financial Turbulence- Evidence from the Global Financial Crisis, *Managerial Finance*, Vol. 39, No. 7, pp. 653-666, 2013
- [6] M. Ottenwaelter, Value-at-Risk for Commodity Portfolios, INP Grenoble – ENSIMAG, pp. 1-43, 2008
- [7] R. Campbell, R. Huisman, K. Koedijk, Optimal Portfolio Selection in a Value-at-Risk Framework, *Journal of Banking and Finance*, Vol. 25, No. 9, pp. 1789-1804, 2001
- [8] S. Basak, A. Shapiro, Value-at-Risk-based Risk Management: Optimal Policies and Asset Prices, *The Review of Financial Studies*, Vol. 14, No. 2, pp. 371-405, 2001
- [9] S. Zikovic, M. Pecaric, Modelling Extreme Events: Application to Zagreb Stock Exchange, *Ekonomski pregled*, Vol. 61, No. 1-2, pp. 19-37, 2010
- [10] V. Bucevska, An Empirical Evaluation of GARCH Models in Value-at-Risk Estimation: Evidence from the Macedonian Stock Exchange, *Business Systems Research*, Vol. 4, No. 1, pp. 49-64, 2013
- [11] V. Djakovic, The Application of the Extreme Value Theory Model in Investments, Ph.D. Thesis, University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia, 227 p., 2013
- [12] Y. Ait-Sahalia, A. W. Lo, Nonparametric Risk Management and Implied Risk Aversion, *Journal of Econometrics*, Vol. 94, No. 1-2, pp. 9-51, 2000

# A Theoretical Game Approach for Two-Echelon Stochastic Inventory System

Saeed Alaei<sup>1</sup>, Alireza Hajji<sup>2</sup>, Reza Alaei<sup>2</sup>, Masoud Behraves<sup>3\*</sup>

<sup>1</sup>Department of Industrial Engineering, K. N. Toosi University of Technology, Pardis St., Molla-Sadra Ave., Vanak Sq., P.O. Box: 19991-43344, Tehran, Iran. s.alaei@kntu.ac.ir, r.alaei@kntu.ac.ir

<sup>2</sup>Department of Industrial Engineering, Sharif University of Technology, Azadi Ave, P.O. Box: 11365-8639, Tehran, Iran, ahaji@sharif.edu

<sup>3</sup>Young Researchers and Elite Club, Marand Branch, Islamic Azad University, Kasrayi St., Daneshgah Sq., P.O. Box: 461/15655, Marand, Iran, behraves@marandiau.ac.ir (corresponding Author)

---

*Abstract: In this paper, we study the differences between the Centralized and Decentralized approaches in a two-echelon stochastic inventory system under the lost sale policy. We formulate the condition in which the retailer applies  $(r, Q)$  inventory policy, and his relation with the upper echelon who acts as a manufacturer. This situation has not been considered in the literature before. The Centralized approach results in optimal solution of the system and the Decentralized one is based on Stackelberg game in which the manufacturer is the leader. The demand arrives according to the stationary Poisson process. We drive the long-run average cost functions, then a set of computational steps are developed to obtain the solutions. Furthermore, we provide a numerical study to compare the two approaches. Here are some conclusions: (a) the Decentralized approach reduces the system's cost efficiency; (b) moreover, the Decentralized approach raises the lead time, the order quantity, and the supply chain inventory relative to the Centralized approach.*

*Keywords: Continuous Review Policy; Two-Echelon Inventory System; Inventory Management; Stackelberg Game*

---

## 1 Introduction

Essentially, a supply chain is composed of independent members, each with its own objectives and individual costs. It is important how the members behave, to manage their inventory. If overall system performance is the objective, then choosing policies to minimize total costs, i.e., the optimal solution. Although this approach is appealing it has an important flaw. Each member may incur only a

fraction of the supply chain cost. So, each member's costs may not be minimized in the optimal solution. For example, a supplier may care more than a retailer about consumer backorders for the supplier's product, or the retailer's cost to hold inventory may be higher than the supplier's [1].

While the firms may agree to cooperate in order to reduce overall system costs, each may face the temptation to deviate from any agreements, as to minimize their own costs. So, when each firm is interested in minimizing their own costs independently, it chooses policies, in which, the overall system performance, cannot necessarily be optimized. In this paper, we study the differences between the Centralized and Decentralized approaches in a two-stage serial supply chain including a manufacturer and a retailer. The Decentralized approach is based on a game theory model called the Stackelberg game.

There are two types of games considered in current literature: cooperative and non-cooperative games. In cooperative games, none of the players dominate the others, and the firms simultaneously choose their policies, While, in non-cooperative games, this is not true.

Stackelberg game is a non-cooperative game. The research conducted in this paper presents a model of (1) the Centralized model in which, the goal is minimizing the overall system costs results in the optimal solution, and (2) the Decentralized or Stackelberg model, where individual firms in the supply chain have their own objectives and decisions to optimize. In this competitive approach, two firms play a game to achieve Stackelberg equilibrium. Stackelberg equilibrium is a pair of policies in which each firm minimizes its own cost assuming the other player chooses its equilibrium policy. Thus, each firm makes an optimal decision given the behaviour of the other firm, and therefore, none has an incentive to deviate unilaterally from the equilibrium.

We consider a supply chain with one manufacturer and a single retailer in our inventory system. An outside supplier supplies raw material to the manufacturer with zero lead time and the manufacturer produces a product at the rate of  $\mu$  and supplies it to a retailer who in turn supplies it to the consumers. Furthermore, assume that the demands arrive according to a stationary Poisson process with the rate of  $\lambda$  to the retailer. The retailer uses the continuous review  $(r, Q)$  inventory policy for controlling costs. The manufacturer operates on a make-to-order basis and uses a lot-for-lot policy, and then he begins to produce a batch of  $Q$ , as soon as he receives an order from the retailer and delivers it to him after the lead time  $(l)$ . Moreover, the manufacturer holds a monopolistic status and has an opportunity to obtain some inventory and demand information of the retailer, so he can gain advantage of this information for decreasing his own costs.

Hsiao et al. (2005) investigate the situation in which, the supplier and retailer choose the lead time and the cycle time respectively. They use Stackelberg game to analyse the problem [2].

In the Decentralized/Stackelberg approach, the manufacturer, as a leader, who is aware of the reaction of the retailer, optimizes his lead time, and the production rate. And as a follower, the retailer takes the manufacturer's optimal decisions as input parameters to determine order quantity and reorder points. This paper is organized as follows. Section 2 reviews the related literature with studies focused on non-cooperative and Stackelberg games. Section 3 describes the notations, and the long-run average cost structure describing the inventory patterns of the manufacturer and the retailer. Then we develop the Centralized and Stackelberg policies for the problem. Section 4 presents a set of computational algorithms in order to search Stackelberg equilibrium and the Centralized solution. Section 5 presents a numerical study and the corresponding sensitivity analysis for some parameters with the purpose of evaluating the influence of these parameters on costs, the manufacturer and the retailer decisions, and the competition penalty. Section 6 summarizes the results and provides concluding remarks.

## 2 Literature Review

In recent years, many papers have been published in the field of multi-echelon inventory management in which the game theory approach is adopted. We focus on studies in which non-cooperative and Stackelberg games have been investigated. There are some papers in the literature that study a vendor-buyer or seller-buyer supply chain, and they analyse the problem as a Stackelberg game in which the vendor/seller is the leader, and the buyer is the follower. Monahan (1984) and Chiang et al. (1994) consider the quantity discount model, in which, the vendor is willing to optimize his discount schedule [3, 4]. Eliashberg and Steinberg (1987) consider production activities such as product delivery and inventory policy, and their relation to marketing strategies such as pricing policies [5]. Hsiao and Lin (2005) discuss an EOQ model and investigate the optimal lead time and cycle time of the supplier and the retailer [2]. Qin et al. (2007) study a system in which demand is price-sensitive [6].

They compare the discounts that the supplier gives to the buyer in the centralized approach with the decentralized one. Liou et al. (2006) study multi-period inventory models in which the economic order quantity is integrated with the economic production quantity (EOQ-EPQ). They investigate the problem under the Stackelberg game approach and obtain the optimality conditions and the optimal replenishment policy [7]. Lal and Staelin (1984) investigate why and how a vendor should develop a pricing scheme, even if such a pricing structure does not change the consumer demand. They show that for any given pricing scheme, there exists, a unified pricing policy, which motivates the buyer to increase their order quantity per order [8]. Moreover, they show that the seller can reduce their costs while leaving the buyer in a stable condition. There are some papers in the literature that handle the problem studied by Lal and Staelin (1984) with the

Stackelberg game approach [8]. For example, see (Rosenblatt and Lee, 1985; Weng, 1995; Munson and Rosenblatt, 2001; Wang, 2002) [9, 10, 11, 12]. Parlar and Wang (1994) investigate discounting decisions for a supplier with a group of homogeneous buyers. They use Stackelberg game and show that the seller set up his quantity discount schedule so that the buyer will order more than his Economic Order Quantity (EOQ) [13]. Moreover, both the seller and the buyer can gain considerably from quantity discount. Wang and Wu (2000) expand the work of Parlar and Wang (1994) to multiple heterogeneous buyers. They propose a discount pricing policy based on the percentage increase from a buyers' order quantity before discount [14, 13].

Li et al. (1995) investigate the advantages of cooperation in a seller-buyer inventory control system [15]. First, the relationship between the seller and the buyer modelled as a non-cooperative game. Then, they develop interactive game theory to address the system cooperation problem. Furthermore, the optimal system order quantity-pricing strategies are determined. Li et al. (1996) focus on advantages of improving buyer-seller cooperation in an inventory system [16]. Both cooperative and non-cooperative games are considered. They show that the order quantity and the total system profit are higher at cooperation than at non-cooperation. Furthermore, the wholesale price is higher at non-cooperation than at cooperation. Cachon and Zipkin (1999) investigate a two-stage serial supply chain in which the two sides use Base Stock policy for managing their inventory [1]. They consider two games: In one game, the firms are committed to tracking echelon inventory and in the other game, they track local inventory. They also, discuss two Stackelberg games that the supplier is the leader in one game and is the follower in another game. They show that competition reduces efficiency, but raises the supply chain inventory. Viswanathan and Piplani (2001) study a one-vendor, multi-buyer supply chain in which the vendor specifies common replenishment periods and offers a price discount to entice the buyers to replenish only at those time periods [17]. They investigate the problem under the Stackelberg game concept. So, the optimal replenishment period and the price discount are determined. Axsater (2001) studies a one-warehouse, multi-retailer supply chain in which the warehouse, as well as its local costs, pays a penalty cost for a delay at the warehouse to the retailer facing the delay. He uses Stackelberg game and shows that if the game is played constantly, the system will approach game equilibrium, but not necessarily, the systems optimal solution [18]. Bylka (2003) considers a vendor-buyer system and defines some conditional games. He shows that competitive approach does not necessarily reduce system efficiency. Some authors investigate a one-manufacturer, multi-retailer in Vendor Managed Inventory system by using a Stackelberg game [19]. Yu et al. (2009.a), discuss how the vendor can benefit in the system for increasing his own profit. They investigate that the Stackelberg equilibrium can be improved by using cooperative contracts [20]. Yu et al. (2009.b) investigate how a manufacturer and his retailers interact with each other with the purpose of optimizing their individual net profits. They consider advertising, pricing, and inventory decisions in their model.

Furthermore, there are some other papers that applied Stackelberg game to analyse multi-echelon inventory systems [21]. For example, see (Gal-or, 1985; Dowrick, 1986; Lau and Lau, 2004; 2005; Yang and Zhou, 2006; Chu et al. 2006) [22, 23, 24, 25, 26, 27]. Summarizing the brief review above, the supplier in a two-echelon supply chain, is often considered as a seller/vendor and there are only a few papers in which the supplier acts as a manufacturer. Moreover, in later systems, in the literature, the demand sometimes varies with price or is deterministic. In this sense, we formulate the condition in which the retailer applies  $(r, Q)$  inventory policy, and his relation with the upper echelon who acts as a manufacturer. This situation has not been considered in the literature before.

### 3 Model Formulation

#### 3.1 Notation

In this paper, we use a notation for representing the parameters and the decision variables to model the inventory management problem in a two-echelon supply chain with the Centralized and Decentralized approaches. We denote the Poisson probability with  $p(j; \lambda) = (\lambda t)^j \cdot e^{-\lambda t} / j!$  and its tail probability with  $P(r; \lambda) = \sum_{j=r}^{\infty} p(j; \lambda)$  subject to  $(j=0, 1, 2 \dots)$ .

##### Parameters:

A	Retailer fixed ordering cost
A'	Manufacturer fixed setup cost
h	Retailer unit holding cost per unit time
H	Manufacturer unit holding cost per unit time
$\pi$	Retailer unit shortage cost per unit of lost sales
$\lambda$	Poisson demand rate
I	The accumulated inventory in a Cycle
L	The number of lost sales in a Cycle
T	Cycle Time
TL	The Portion of the Cycle with lost sales
k1	Time-dependent Production cost per unit time
k2	Technology development cost, per one unit increasing on the production rate
CR	Retailer Cost Function
CM	Manufacturer Cost Function
CT	Total cost(sum of the retailer's and the manufacturer's costs)

##### Decision Variables:

L	Lead time
$\mu$	Manufacturer Production rate
Q	Retailer order quantity
r	Retailer Reorder point



### 3.2 Retailer Cost Model

The retailer uses the continuous review  $(r, Q)$  inventory policy for controlling his costs in which  $r$  and  $Q$  are the reorder point and order quantity, respectively. We assume that the demand is according to stationary Poisson process, and unsatisfied demands are completely lost. The retailer costs include the holding cost, the fixed ordering cost, and the shortage cost. Figure 1 shows the retailer's inventory level. As shown in the figure, the cycle time ( $T$ ) is the time between two successive reorder times. We will assume that at most one order outstanding is allowed. In other words, when the retailer orders a batch of  $Q$ , he is not allowed to place another order, unless he received the previous deliverables. Under these conditions, the cycle time is always greater than or equal to the lead time.

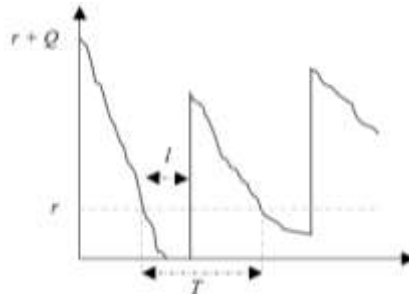


Figure 1

Retailer's Inventory Level

Rabinowitz et al. (1995), consider the partial backorder policy used in conjunction with the traditional  $(r, Q)$  inventory system [28]. Their policy is modelled using a control variable,  $b$ , which limits the maximum number of backorders allowed to be accumulated in any cycle. By setting  $b$  to zero in their model, we have the lost sale policy. So, the expected cost for the retailer is given by:

$$C_R = \frac{E(C)}{E(T)} = \frac{A + h.E(I) + \pi.E(L)}{E(T)} \quad (1)$$

In which  $T$  is the cycle time,  $C$  is the costs accumulated within a cycle,  $I$  is the accumulated inventory in a cycle, and  $L$  is the number of lost sales in a cycle, and we have:

$$E(L) = \lambda l P(r; \lambda l) - r P(r + 1; \lambda l) \quad (2)$$

$$E(T) = \frac{Q}{\lambda} + l P(r; \lambda l) - \frac{r}{\lambda} P(r + 1; \lambda l) \quad (3)$$

$$E(T) = \frac{Q}{\lambda} + l P(r; \lambda l) - \frac{r}{\lambda} \quad (4)$$

Where  $i$  is given by:

$$i = r l [P(r - 1, \lambda l) - P(r, \lambda l)] + \lambda l^2 [P(r - 1, \lambda l) - P(r - 2, \lambda l)] / 2 + r(r + 1) [P(r + 1, \lambda l) - P(r, \lambda l)] / 2 \lambda \quad (5)$$

As long as  $Q^* \geq r + 1$ , otherwise there is no feasible solution (Rabinowitz et al., 1995). According to the definition of Poisson's tail probability function and some of the properties of the Poisson distribution, it can be shown that  $i$  is equal to zero.

### 3.3 Manufacturer Cost Model

The manufacturer Orders from an outside supplier with zero lead time. We consider a cost function for the manufacturer that consists of the holding cost, the setup costs, and time-dependent production cost. Holding cost is incurred only for finished products. We also split the setup costs into two parts: one part is fixed for every production period, and another one is an increasing function of the production rate. For example, assume an assembly line that has the technology for assembling a set of parts that are supplied by a supplier, and there is no cost for raw materials. In this assembly line, time-dependent production cost coincides with the daily production cost. If the production rate exceeds a specific limit, it is necessary to enhance the technology. For simplifying the problem, we assume that for every increasing unit on the production rate, the manufacturer incurred a cost called technology development cost. For instance, if the manufacturer incurred 200\$ for increasing 100 units on the production rate, then the technology development cost will be 2\$. Figure 2 shows the Manufacturer Inventory Level.

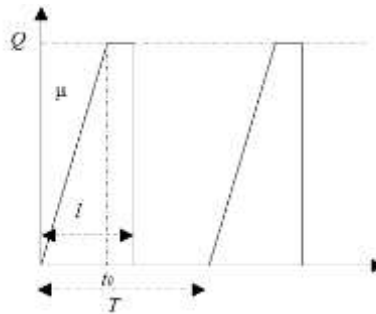


Figure 2

Manufacturer's Inventory Level

In this particular setting, the manufacturer's production cycle was equal to the retailer's replenishment Cycle. The manufacturer operates on a make-to-order basis using a lot-for-lot policy. So, begins to produce a batch of  $Q$  at the rate of  $\mu$ , as soon as he receives an order from the retailer and delivers it to him after the lead time. Let  $T_1, T_2, \dots$  be the sequence of reorder times of retailer,  $C_n$  be the Reward at the time of the  $n$ th renewal, and  $C(t) = \sum_{n=1}^{N(t)} C_n$  be the amount of cost incurred by  $t$ , then,  $\{C(t), t \geq 0\}$  forms a Renewal Reward Process where renewals occur at reorder times. Hence, the long-run average cost is given by:

$$C_M = E(C)/E(T) = [A' + HQ(2l - Q/\mu)/2 + k_1Q/\mu + k_2\mu]/E(T) \quad (6)$$

Remember that the amount of inventory held in a cycle is  $Q(2l - Q/\mu)/2$ ; and  $Q/\mu$  is the portion of the lead time that spent for production. The manufacturer makes a trade-off among inventory, setups, and production costs by producing at an optimal production rate. So, he spends a portion of lead time for production. Using a simple calculus, we can calculate the optimal amount of  $\mu$  (given by the following equation) in order to minimize manufacturer cost.

$$\mu^* = [Q(2k_1 - HQ)/2k_2]^{1/2} \quad (7)$$

Consider that the production time ( $t_0$ ) has to be lower than or equal to the lead time (See Figure 2). This results  $\mu^*$  to be greater than or equal to  $Q/l$ . Furthermore, it is necessary that the right-hand side of the equation (7) to be a real number. Therefore, under these conditions, the equation (7) adjusted to the following equation:

$$\mu^* = \begin{cases} \left[ \frac{Q(2k_1 - HQ)}{2k_2} \right]^{1/2} ; & \text{if } Q \leq \frac{2k_1 l^2}{2k_2 + Hl^2} \\ Q/l; & \text{otherwise} \end{cases} \quad (8)$$

### 3.4 Centralized Approach

In the Centralized approach, the goal is minimizing the sum of manufacturer and retailer costs. Decision variables are reorder point ( $r$ ), order quantity ( $Q$ ), lead time ( $l$ ), and production rate ( $\mu$ ). In this approach, we denote the Centralized solution with  $(r^*, Q^*, l^*, \mu^*)$ .

### 3.5 Decentralized Approach (Stackelberg Game Approach)

In a Stackelberg approach, players are classified as leader and follower. The leader chooses a strategy first, and then the follower observes this decision and makes his own strategy. It is necessary to assume that each enterprise is not willing to deviate from minimizing its own cost. In other words, each player chooses his best strategy. Here, the manufacturer is the leader, and the retailer is the follower. The manufacturer, as the Stackelberg leader, induces/encourages the retailer to choose his strategy by changing the value of lead time. So the manufacturer wants to find a set of  $(r, Q, l)$  that minimizes his costs. In other words, the manufacturer determines his lead time, and acts as a leader by announcing lead time to the retailer in advance, and the retailer acts as a follower by choosing his reorder point and order quantity based on manufacturer strategy. We denote Stackelberg Equilibrium point with  $(r^s, Q^s, l^s, \mu^s)$ . The manufacturer and the retailer play a Stackelberg game to determine game Equilibrium. The two players cost functions are  $C_R$  and  $C_M$  (Equation 1 and 6). The manufacturer knows that if he sets his lead time value to  $l$ , then the retailer will set  $r$  and  $Q$  to values determined by his reaction function:

$$[r(l), Q(l)] = \arg \min_{r, Q} C_R(r, Q, l) \quad (9)$$

In other words, the retailer chooses  $r$  and  $Q$  that minimize his costs, given the manufacturer's strategy ( $l$ ). Then the Manufacturer chooses  $l^*$ , its Cost-Minimizing lead time, given the reaction function of the retailer:

$$[l^s, \mu^s] = \arg \min_{l, \mu} C_M(r(l), Q(l), l, \mu) \quad (10)$$

## 4 Search Algorithms

In order to obtain the Centralized solution and Stackelberg equilibrium, we have four decision variables:  $r$ ,  $Q$ ,  $l$ , and  $\mu$ . As stated in previous sections,  $\mu$  is an explicit function of  $Q$  and  $l$  (equation 9). In spite of the four-dimensional structure of the model, a three-dimensional search should be used to find the solutions. We assume that all variables but  $\mu$  are integer numbers. Algorithms 1 and 2 summarize the computational steps for obtaining the centralized and Stackelberg solutions, respectively. Consider that the step 3 of both algorithms is associated to sub-algorithms 1 and 2, respectively.

---

### Algorithm 1: Steps for obtaining Centralized solution

---

- Step 1. Initialization: Set  $l = 0$
  - Step 2. Set  $l = l + 1$
  - Step 3. Calculate the optimal  $r$ ,  $Q$  and  $\mu$  for current lead time ( $l$ ). And set them  $r(l)$ ,  $Q(l)$  and  $\mu(l)$  respectively (Sub-Algorithm 1)
  - Step 4. Calculate  $C_T(Q(l), r(l), l, \mu(l))$  for current lead time ( $l$ ),  $r(l)$ ,  $Q(l)$  and  $\mu(l)$
  - Step 5. If  $l = 1$ , go to Step 2; otherwise go to the next step
  - Step 6. If  $C_T(Q(l-1), r(l-1), l-1, \mu(l-1)) \geq C_T(Q(l), r(l), l, \mu(l))$ , go to Step 2; otherwise stop. The Centralized solution will be  

$$(Q^*, r^*, l^*, \mu^*) = (Q(l-1), r(l-1), l-1, \mu(l-1))$$
- 

### Algorithm 2: Steps for obtaining Stackelberg equilibrium (Decentralized solution)

---

- Step 1. Initialization: Set  $l = 0$
  - Step 2. Set  $l = l + 1$
  - Step 3. Calculate the retailer's optimal reaction ( $r$ , and  $Q$ ) for current lead time ( $l$ ). And set them  $r(l)$ , and  $Q(l)$  respectively (sub-algorithm 2)
  - Step 4. Calculate the production rate and  $C_M(Q(l), r(l), l, \mu(l))$  that is the manufacturer cost for current lead time ( $l$ ), and retailer's optimal reaction
  - Step 5. If  $l = 1$ , go to Step 2; otherwise go to next Step
  - Step 6. If  $C_M(Q(l-1), r(l-1), l-1, \mu(l-1)) \geq C_M(Q(l), r(l), l, \mu(l))$ , go to Step 2; otherwise stop. Stackelberg equilibrium will be  $(Q^s, r^s, l^s, \mu^s) = (Q(l-1), r(l-1), l-1, \mu(l-1))$
- 

### Sub-Algorithm 1: Algorithm for step 3 of Algorithm 1

---

- Step 1. Initialization: Set  $r = 0$
  - Step 2. Set  $r = r + 1$
  - Step 3. Calculate  $Q^*(r)$  that is the optimal order quantity for current reorder point
    - Step 3.1. Set  $Q = 0$
    - Step 3.2. Set  $Q = Q + 1$
    - Step 3.3. Calculate  $\mu(Q)$ , and  $C_T(Q, r, \mu(Q))$  for current  $r$ ,  $Q$ , and  $\mu(Q)$ .
    - Step 3.4. If  $Q = 1$ , go to Step 3.2; otherwise go to next step
-

---

Step 3.5.	If $C_T(Q-1, r, \mu(Q-1)) \geq C_T(Q, r, \mu(Q))$ , go to Step 3.2; otherwise $Q^*(r) = Q-1$ , so, go to Step 4
Step 4.	If $r = 1$ , go to Step 2; otherwise go to Step 5
Step 5.	If $C_T(Q^*(r-1), r-1, \mu(Q^*(r-1))) \geq C_T(Q^*(r), r, \mu(Q^*(r)))$ , go to Step 2; otherwise $r(l) = r-1$ , and $Q(l) = Q^*(r-1)$ , so, stop

---

Sub-Algorithm 2: Algorithm for step 3 of Algorithm 2

---

Step 1.	Initialization: Set $r = 0$
Step 2.	Set $r = r + 1$
Step 3.	Calculate $Q^*(r)$ that is the optimal order quantity for current reorder point
Step 3.1.	Set $Q = 0$
Step 3.2.	Set $Q = Q + 1$
Step 3.3.	Calculate $C_R(Q, r)$ that is the retailer's cost for current $r$ and $Q$ .
Step 3.4.	if $Q = 1$ , go to Step 3.2; otherwise go to next step
Step 3.5.	If $C_R(Q-1, r) \geq C_R(Q, r)$ , go to Step 3.2; otherwise $Q^*(r) = Q-1$ , so, go to Step 4
Step 4.	if $r = 1$ , go to Step 2; otherwise go to Step 5
Step 5.	if $C_R(Q^*(r-1), r-1) \geq C_R(Q^*(r), r)$ , go to Step 2; otherwise $r(l) = r-1$ , and $Q(l) = Q^*(r-1)$ , so, stop.

---

## 5 Numerical Study

In this section, we present detailed numerical examples to:

- Compare the results of the Centralized approach with the results of Stackelberg game approach
- Evaluate the benefits of Stackelberg game approach to the manufacturer
- Illustrate Stackelberg game equilibrium graphically
- Analyse the sensitivity of the solutions, costs, and the competition penalty with respect to the variation of system parameters

### 5.1 Base Case

In our numerical study, we consider a base-case for implementing our algorithms. The parameter values for the base-case are presented in Table 1.

Table 1  
Parameter values of the base-case

Parameter	$A$	$A'$	$h$	$H$	$\pi$	$\lambda$	$kl$	$k2$
Value	100	100	1	0.5	50	10	50	16

Figure 3 shows the variation of manufacturer cost with respect to the lead time in the game approach. As shown in the figure, when the lead time is equal to 8, the manufacturer cost is minimized. Furthermore, Figure 4 illustrates the variation of

the total cost with respect to the lead time in the Centralized approach. When the lead time is equal to 2, the total cost is minimized.

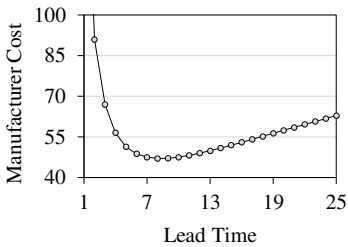


Figure 3  
Manufacturer Cost in the Game Approach

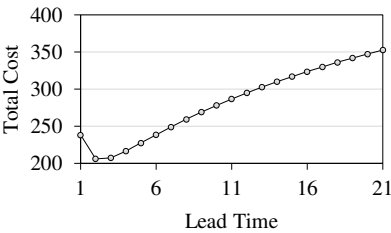


Figure 4  
Centralized Total Cost

Table 2 summarizes the Centralized and Stackelberg solutions for the base-case values. From table 2, the lead time and order quantity have increased in the game approach, but the production rate has decreased, and the reorder point has not changed. The manufacturer cost is decreased by 48%, and the retailer cost is increased by 84% in the game approach. So, the manufacturer gains the advantages of the game approach for decreasing his costs. However, the total cost in the game approach is increased by 26% subject to the centralized approach. This is known as the competition penalty.

Table 2  
Solutions of two approaches for the Base-case Example

	$l^*$	$Q^*$	$\mu^*$	$r^*$	$CR$	$CM$	$CT$
Centralized solution	2	121	60.5	7	115.46	90.64	206.1
Stackelberg Equilibrium	8	212	26.5	7	212.4	47.04	259.4

## 5.2 Stackelberg Equilibrium

In this section, we illustrate the Stackelberg equilibrium point graphically for the base-case values. To simplify the problem, we will assume that the reorder point is constant and equals to 7. Therefore, the manufacturer and the retailer play a Stackelberg game to determine  $Q$  and  $l$ . As noted before, when the manufacturer sets his lead time value to  $l$ , then the retailer's reaction function will be as follows:

$$Q(l) = \arg \min_Q C_R(Q, l) \quad (11)$$

The manufacturer chooses  $l^*$ , given the reaction function of the retailer:

$$l^s = \arg \min_l C_M(Q(l), l) \quad (12)$$

In other words, the manufacturer wants to find a combination of the lead time and the order quantity that minimizes the costs in which the order quantity would be the reaction value of the retailer to the manufacturer strategy ( $l$ ). To analyse the problem graphically, remember that equilibrium will be where the retailer's reaction functions or  $Q(l)$  is tangential to the manufacturer's ISO-cost curve. Consider that the ISO-cost curve consists of all combinations of  $l$  and  $Q$  that yield the same cost for the manufacturer. Figure 5 shows the Stackelberg equilibrium point for the base-case data. There are two ISO-cost curves in the figure. Remember that, the lower ISO-cost curve represents the higher levels of Cost. The Retailer's reaction function is tangential to the Manufacturer's ISO-cost curve with the value of 47.04 where  $l$  and  $Q$  are equal to 8 and 212, respectively. Then Stackelberg equilibrium point will be  $(l^s, Q^s) = (8, 212)$ .

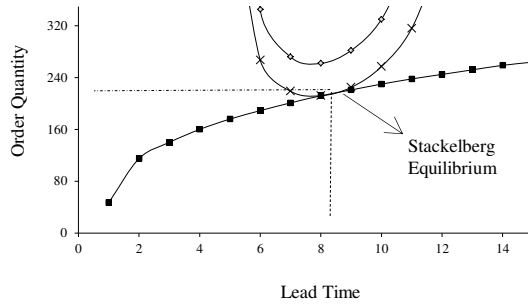


Figure 5  
Stackelberg equilibrium and ISO-Cost Curves

## 5.3 Sensitivity Analysis

In this section, we perform a sensitivity analysis by changing the values of system parameters in the base-case. We solve a sample of 500 problems and then we draw the relevant conclusions in each sub section.

### 5.3.1 Variations of Optimal Solution and Game Equilibrium

The Centralized solution and Stackelberg Equilibrium of the system along with the sensitivity analysis for all parameters are presented in Table 3. We vary the parameters one at a time by doubling and halving the base-case values.

Table 3  
Two approach solutions under variation of Base-Case Parameters

Solutions → Parameters ↓		<i>Centralized Solution</i>				<i>Stackelberg Equilibrium</i>			
		$l^*$	$Q^*$	$r^*$	$\mu^*$	$l^s$	$Q^s$	$r^s$	$\mu^s$
Base-case		2	121	7	60.5	8	212	7	26.5
A	50	2	117	7	58.5	8	210	7	26.25
	20								
	0	3	152	7	50.7	8	215	7	26.9
A ′	50	2	117	7	58.5	8	212	7	26.5
	20								
	0	3	152	7	50.7	9	221	7	24.6
	0.5	3	216	7	72	8	324	7	40.5
	2	2	82	7	41	9	140	7	15.6
	0.2								
H	5	2	121	7	60.5	11	238	7	21.6
	1	2	120	7	60	6	189	7	31.5
$\pi$	25	3	105	7	35	9	142	7	15.8
	10								
	0	1	73	14	73	8	321	7	40.1
$\lambda$	5	2	71	4	35.5	7	103	4	14.7
	20	2	214	15	107	9	438	15	48.7
k1	25	2	117	7	58.5	9	221	7	24.6
	10								
	0	2	128	7	64	7	201	7	28.7
k2	8	1	72	13	72	6	189	7	31.5
	32	4	159	7	39.75	12	245	7	20.4

Here are some representative observations:

- (1) The decentralized approach raises the lead time and order quantity values, but lowers the values of the reorder point and production rate. In other words, we always have:

$$l^* < l^s, \quad \mu^* > \mu^s, \quad Q^* < Q^s, \quad r^* \geq r^s$$

This is because, in the Decentralized approach, the manufacturer wants to increase the cycle time in order to minimize related costs (see equation 6). They could only increase the lead time to achieve this goal. It can be proven that  $dE(T)/dl = P(r, \lambda l) \geq 0$ . As a consequence, in each cycle, he has more time for



production, and then they could reduce the production rate. In the other hand, regarding to the increase on the cycle time, the retailer has to increase his order quantity. This reaction is in the same direction of the manufacturer goal, because  $dE(T)/dQ = 1/\lambda \geq 0$ . It is reasonable for the retailer to increase the reorder point regarding to increase on the lead time in order to reduce his shortage cost, but he acts vice versa. Because, increasing the reorder point is in opposition to the manufacturers' goal. It can be proved that the increasing one unit on the reorder point results in decreasing the cycle time by  $P(r + 1, \lambda l)/\lambda$ . So if the retailer wants to increase the reorder point, the manufacturer increases the lead time to compensate the reduction on the cycle time.

- (2) Furthermore, the decentralized approach raises the supply chain inventory relative to the centralized solution. In other words, if firms choose the optimal solution, they will tend to *decrease* inventory.
- (3) The Manufacturer only spends a portion of lead time for production. This time is equal to  $Q/\mu$ . The remaining time  $(l - t_0)$  is spent to holding inventory without production. The manufacturer delivers a batch as soon as possible after the production period in the Centralized approach. However, in the game approach, he prefers to deliver the batch very late in some cases. This result in increasing the retailer's cost.

### 5.3.2 Variations of the Manufacturer and the Retailer Costs

In the centralized approach, there is no inventory system if the lead time value is greater than 54. In other words, the total average cost of the retailer exceeds  $\pi\lambda$ , then it is reasonable that every order has been lost. In this approach, the maximum value of the order quantity is 320. However, in the decentralized approach, for any value of the lead time, there exists an inventory system. The maximum value of the order quantity is 499 (also for infinite values of the lead time). This is because the holding cost for the 500<sup>th</sup> unit and above values of the order quantity is greater than their shortage cost. Figure 6 shows the variations of the manufacturer and the retailer costs with respect to some selected parameters for two policies. CM and CR represent the manufacturer and the retailer cost, respectively. We select several representative observations summarized as follows:

- (1) The manufacturer gains the advantages of the game approach for decreasing costs. However, the retailer cost in the game approach is always greater than its value in the centralized approach.
- (2) In the centralized approach, the effects of increment in the shortage cost are tolerated by the manufacturer. However, in the game approach, the manufacturer chooses his strategy in a manner that his cost experiences no rise.
- (3) The retailer cost increment is always greater than the manufacturer cost

decrement in the game approach. For example, in the base-case, the game approach results in an increase of 97 units in the retailer cost, however, the manufacturer cost decrement is 43 units.

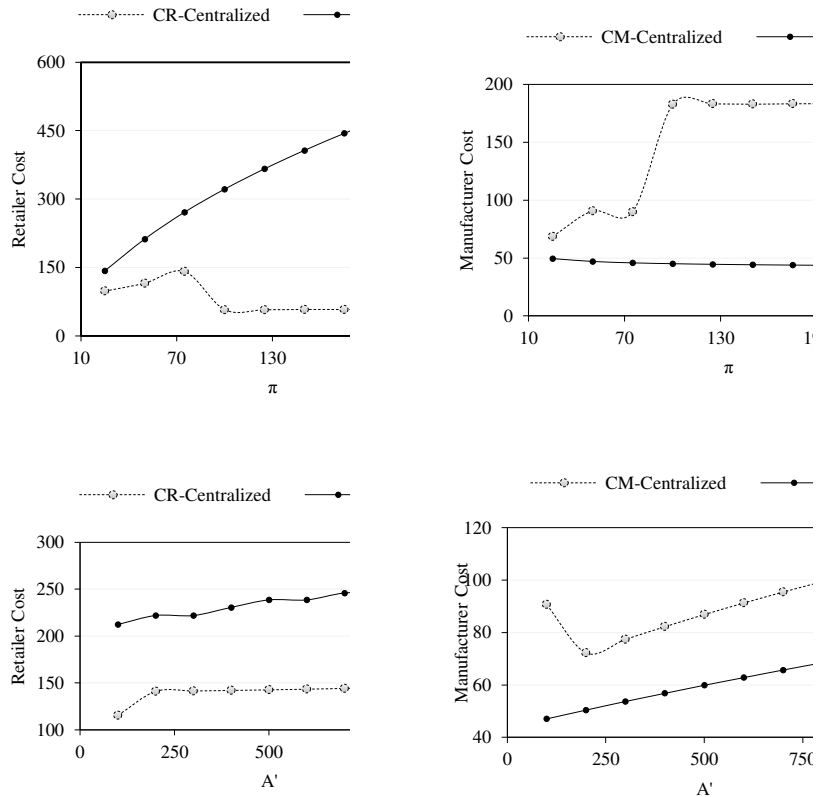


Figure 6

Variations of the manufacturer and the retailer costs in two approaches

### 5.3.3 Variations of the Competition Penalty

The competition penalty is the difference between the Centralized total cost and Stackelberg total cost measured as a fraction of the Centralized total cost. Figure 7 shows the variations of the competition penalty ( $\rho$ ) with respect to some selected parameters of the system. We select several representative observations summarized as follows:

- (1) The value of  $\rho$  is always positive. So, the total cost in the decentralized approach is always greater than that in the Centralized approach.
- (2) The competition penalty is less sensitive to variations of  $A'$ .

- (3) Shortage cost has the most significant influence on the competition penalty.

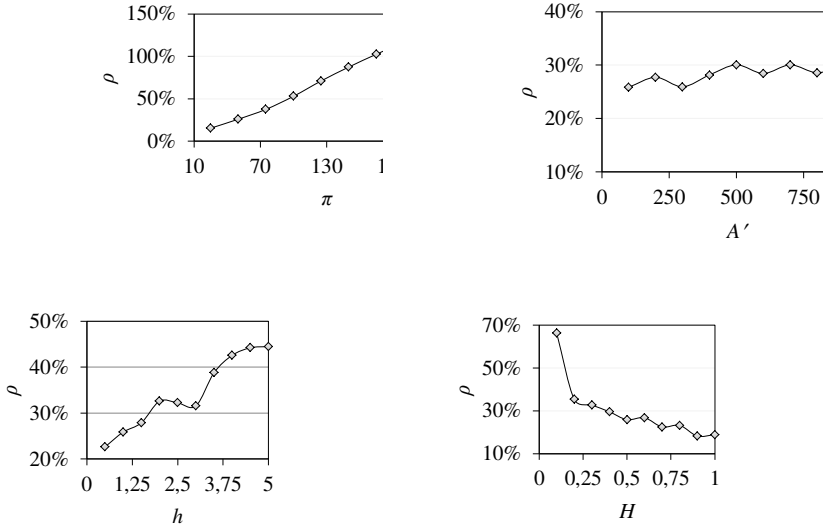


Figure 7

Variations of competition penalty with respect to various parameters

## Conclusions

Whereas, the firms in a supply chain may agree to cooperate in order to minimize overall system costs, each firm may face the temptation to deviate from any agreement to reduce its own costs. This paper has studied the difference between the Centralized and decentralized approaches in a two-stage serial supply chain in which the manufacturer has the opportunity to obtain some inventory and market-related information of the retailer and then they can take advantage of this information for decreasing their own costs. We provided a numerical example and the corresponding sensitivity analysis for evaluating the costs, the competition penalty, and the manufacturer's and retailer's decisions in the two approaches. Here are some important results:

- (1) The Decentralized approach reduces the systems cost efficiency.
- (2) Shortage cost has the most significant influence on the competition penalty. For example, with the increase of the shortage cost from 50 in the base-case to 250, the competition penalty increases from 26% to 150%.
- (3) The manufacturer gains the advantages of the game approach for decreasing their costs. However, the retailer cost increased in the game approach. Additionally, the retailer cost increment is always greater than

the manufacturer cost decrement in the game approach. For example, in the base-case, the game approach results in an increase of 97 units in the retailer cost, but the manufacturer cost decrement is 43 units.

- (4) In the Decentralized approach, manufacturer tends to increase the lead time and decrease the production rate. On the other hand, the retailer wants to increase the order quantity and doesn't want to increase the reorder point.
- (5) The decentralized approach raises the supply chain inventory relative to the centralized approach. In other words, if firms choose the optimal solution, they will tend to *decrease* inventory. However, Cachon and Zipkin (1999) [1] suggest the opposite in a condition that two firms use base stock policy.

Our research could be extended in several possible directions. Because the decentralized approach reduces the system's cost efficiency, there is an opportunity for the players to coordination in order to align their incentives to reduce the supply chain's costs. Numerical studies indicate that the optimal solution is never a game equilibrium, so the firms can sign a contract with the purpose of achieving the system optimal solution, as in a game equilibrium. Moreover, we consider the lost sales policy in our inventory system, however, in some real inventory systems, it is more reasonable to assume that some of the excess demands are backordered, and the rest is lost. Then, we can extend the model in which a partial backorder policy is allowed.

## References

- [1] Cachon, G. P., Zipkin, P. H. (1999) Competitive and Cooperative Inventory Policies in a Two-Stage Supply Chain. *Management Science*, 45 (7): 936-953
- [2] Hsiao, J. M., Lin, C. (2005) A Buyer-Vendor EOQ Model with Changeable Lead-Time in Supply Chain. *International Journal of Advanced Manufacturing Technology*, 26 (7-8): 917-921
- [3] Monahan, J. P. (1984) A Quality Discount Pricing Model to Increase Vendor Profits. *Management Science*, 30 (6): 720-726
- [4] Chiang, C. W., Fitzsimmons, J., Huang, Z., Li, Susan (1994) A Game Theoretic Approach to Quantity Discount Problem. *Decision Sciences*. 25 (1): 153-168
- [5] Eliashberg, J., Steinberg, R. (1987) Marketing-Production Decisions in an Industrial Channel of Distribution. *Management Science*, 33 (8): 981-1000
- [6] Qin, Y., Tang, H., Guo, C. (2007) Channel Coordination and Volume Discounts with Price-Sensitive Demand. *International Journal of Production Economics*, 105 (1): 43-53
- [7] Liou, Y. C., Schaible, S., Yao, J. C. (2006) Supply Chain Inventory Management via a Stackelberg Equilibrium. *Journal of Industrial and*

*Management Optimization*, 2 (1): 81-94

- [8] Lal, R., Staelin, R. (1984) An Approach for Developing an Optimal Discount Pricing Policy. *Management Science*, 30 (12): 1524-1539
- [9] Rosenblatt, M. J., Lee, H. L. (1985) Improving Profitability with Quantity Discounts under Fixed Demand. *IIE Transactions*, 17 (4): 388-395
- [10] Weng, Z. K. (1995) Channel Coordination and Quantity Discounts. *Management Science*, 41 (9): 1509-1522
- [11] Munson, C. L., Rosenblatt, M. J. (2001) Coordinating a Three-Level Supply Chain with Quantity Discounts. *IIE Transactions*, 33 (5): 371-384
- [12] Wang, Q. N. (2002) Determination of Suppliers' Optimal Quantity Discount Schedules with Heterogeneous Buyers. *Naval Research Logistics*, 49 (1): 46-59
- [13] Parlar, M., Wang, Q. (1994) Discounting Decisions in a Supplier–Buyer Relationship with a Linear Buyer's Demand. *IIE Transactions*, 26 (2): 34-41
- [14] Wang, Q., Wu, Z. (2000) Improving a Supplier's Quantity Discount Gain from Many Different Buyers. *IIE Transactions*, 32 (11): 1071-1079
- [15] Li, S., Huang, Z., Ashley, A. (1995) Seller Buyer System Cooperation in a Monopolistic Market. *Journal of the Operational Research Society*, 46 (12): 1456-1470
- [16] Li, S., Huang, Z., Ashley, A. (1996) Improving Buyer Seller System Cooperation through Inventory Control. *International Journal of Production Economics*, 43 (1): 37-46
- [17] Viswanathan, S., Piplani, R. (2001) Coordinating Supply Chain Inventories through Common Replenishment Epochs. *European Journal of Operational Research*, 129 (2): 277-286
- [18] Axsater, S. (2001) A Framework for Decentralized Multi-Echelon Inventory Control. *IIE Transactions*, 33 (2): 91-97
- [19] Bylka, S. (2003) Competitive and Cooperative Policies for the Vendor–Buyer System. *International Journal of Production Economics*, 81-82 (11): 533-544
- [20] Yu, Y., Chu F., Chen H. (2009a) A Stackelberg Game and its Improvement in VMI System with a Manufacturing Vendor. *European Journal of Operational Research*, 192 (3): 929-948
- [21] Yu, Y., Huang G. Q., Liang L., (2009b) Stackelberg Game-Theoretic Model for Optimizing Advertising, Pricing and Inventory Policies in Vendor Managed Inventory (VMI) Production Supply Chains. *Computers & Industrial Engineering*, 57 (1): 368-382

- [22] Gal-Or, E., (1985) First Mover and Second Mover Advantages. *International Economic Review*, 26 (3): 649-653
- [23] Dowrick, S., (1986) Von Stackelberg and Cournot Duopoly: Choosing Roles. *The Rand Journal of Economics*, 17 (2): 251-260
- [24] Lau, A. H. L., Lau, H. S. (2004) Some Two-Echelon Supply-Chain Games: Improving from Deterministic-Symmetric-Information to Stochastic-Asymmetric Information Models. *European Journal of Operational Research*, 161 (1): 203-223
- [25] Lau, A. H. L., Lau, H. S. (2005) A Critical Comparison of the Various Plausible Inter-Echelon Gaming Processes in Supply Chain Models. *The Journal of the Operational Research Society*, 56 (11): 1273-1286
- [26] Yang, S. L., Zhou, Y. W. (2006) Two-Echelon Supply Chain Models: Considering Duopolistic Retailers' Different Competitive behaviours. *International Journal of Production Economics*, 103: 104-116
- [27] Chu, H., Wang, J., Jin, Y., Suo, H. (2006) Decentralized Inventory Control in a Two-Component Assembly System. *International Journal of Production Economics*, 102 (2): 255-264
- [28] Rabinowitz, G., Mehrez, A., Chu, C. W., Patuwo, B. E. (1995) A Partial Backorder Control for Continuous Review of (r, Q) Inventory System with Poisson Demand and Constant Lead Time. *Computers and Operations Research*, 22 (7): 689-700