COMMUNICATIONS

COMMUNICATIONS

Sergey Subbotin - Andrii Oliinyk - Vitaly Levashenko - Elena Zaitseva*

DIAGNOSTIC RULE MINING BASED ON ARTIFICIAL IMMUNE SYSTEM FOR A CASE OF UNEVEN DISTRIBUTION OF CLASSES IN SAMPLE

The problem of development automation of classification rules synthesis on the basis of negative selection in the case of uneven distribution of classes in the sample is solved. The method for the synthesis of classification rules on the basis of negative selection in the case of uneven distribution of class instances of sample is proposed. This method uses a priori information about instances of all classes of the sample. The software implementing the proposed method is developed. Some experiments on the solution of practical problem of gas turbine air-engine blade diagnosis are conducted.

Keywords: *Artificial immune system, instance, negative selection, classification, recognition error, sample.*

1. Introduction

Process of building decision models for non-destructive testing, for technical or medical diagnostics, and for pattern recognition is a topical task [1-4]. The situation when most data of a training set belong to one class is typical for such a process [5 and 6]. We have to develop new models for object formalization or process descriptions. One of the perspective approaches for developing such models is based on conception of artificial immune systems [7-9]. This model can be created based on one class. The difference between numbers of instances belonging to different classes is significant in this case. Then the usage of artificial immune systems with negative selection is proposed in [10-13]. These systems involve the construction of a set of detectors (computational elements) that are capable of recognizing unknown instances [14-16]. This approach allows to detect anomalies or random variations in diagnosed objects [7 and 10], and to recognize instances of "non-self" classes (classes of objects which are not represented in the training set) [8, 12 and 15]. There are well known methods for the synthesis of artificial immune systems based on the negative selection [8-16]. These methods generate an exhaustive number of detectors (the possible solutions) and employ instances with one class only. Instances with other classes are not taken into account. Moreover, these methods have got high requirements for computing resources.

Consequently, the development of methods for the synthesis of artificial immune systems on the basis of negative selection, which are free from these disadvantages, is a topical problem. In addition, the diagnostic models based on artificial immune systems have a low level of generalization. The detectors (rules) of the immune system are easy to understand. However, because of the low level of generalization, a detector system has a large dimension. It is difficult to understand and analyze by human, which generally leads to reduction of interoperability of the diagnostic model.

So, the purpose of this paper is to develop a method of classification rules synthesis on the basis of a set of detectors. These rules handle data of a training set with a significant difference in the number of instances which belong to different classes.

2. Problem statement

Let us assume that there is a training set $S = < P, T >$, where P is a set of input parameters (features) of an objects and set T is a set of values of the output parameter. Set P is represented as a matrix $P = \left(p_{qm} \right)_{QM}$, where p_{qm} is a value of the m-th feature of q-th instance in the set S. Variable m is a feature of the object (m = 1, 2, ..., M). Variable q is a number of instance (object) in the sample S(q = 1, 2, ..., Q). Value M is measure of cardinality of features of set S; Q is a cardinality of instances on set S. Set of values of the output parameter is represented as a vector $T = \left(t_q \right)_Q$, where $t_q \in T'$ is a value of the output

* [1]Sergey Subbotin, [1]Andrii Oliinyk, [2]Vitaly Levashenko, [2]Elena Zaitseva
[1]Department of Program Tools, Faculty of Computer Sciences, Zaporizhzhya National Technical University, Ukraine
[2]Department of Informatics, Faculty of Management Science and Informatics, University of Zilina, Slovakia
E-mail: subbotin@zntu.edu.ua

parameter of q-th instance; T' is a set of possible values of the output parameter (usually in problems of non-destructive quality control and pattern recognition a set T' consists of two elements $T' = \{t'_0, t'_1\}$, determining class of suitability of object, such if $t_q = t'_0$ then q-th object is considered unusable, if $t_q = t'_1$ it is usable, suitable etc.).

The number of instances of the sample of one class (for example, instances of the class $t_q = t'_1$) is significantly different from the number of instances of another class, which is defined by (1):

$$0 \le N_{t_q=t'_0} << N_{t_q=t'_1}, \quad (1)$$

where $N_{t_q=t'_0}$ and $N_{t_q=t'_1}$ are the numbers of instances of the sample $S =< P, T >$, the value of the output parameter t_q of which are equal to t'_0 and t'_1, respectively, $N_{t_q=t'_0} + N_{t_q=t'_1} = Q$.

Then, on the basis of a training set $S =< P, T >$ it is necessary to generate a set $RB = \{rule_1, rule_2, ..., rule_{NR}\}$ of productions $P_r \rightarrow T_r$, that allows to provide an acceptable level of recognition error E. This error E is defined as the ratio of number of incorrectly recognized instances N_{er} to the total number of instances Q (2):

$$E = \frac{N_{er}}{Q}. \quad (2)$$

3. The method of classification rule mining based on negative selection

As noted above, the known methods of negative selection [8-16] have got such disadvantages as the generation of the exhaustive number of detectors, the usage of information of one class instances only, low interoperability of synthesized set of solutions of detectors etc. In addition, most methods based on the principles of negative selection as detectors used hypersphere with a fixed radius. This radius determines the area of feature space covered by the detector. The choice of the radius of the hypersphere-detector is a very complex task. It can be explained by the fact that for large values of the radius recognition accuracy is reduced and for low values the number of generated detectors increases. It lowers the generalization properties of the synthesized model in a form of set of detectors of artificial immune network.

These disadvantages necessitate the extension of essential requirements to the computer resources. It decreases the speed of solutions search and in some cases does not allow to find an acceptable solution. To eliminate these drawbacks it is advisable to use the method of classification rules synthesis on the basis of negative selection in the case of uneven distribution of instances of the sample classes. In this method is used:

- known information about instances of both classes $T' = \{t'_0, t'_1\}$ in generating the set of detectors $AB = \{Ab_1, Ab_2, \ldots, Ab_{N_{Ab}}\}$. It is forming a set of detectors with high approximation and generalizing properties;
- information about individual significance V_m of features p_m. It is eliminating irrelevant and redundant features of the sample $S =< P, T >$;
- a hypercube of maximum possible volume as a form of detector. It is a contrast to known methods of negative selection, in which a hypersphere is used as a form of detector. This hypercube allows to eliminate the necessity of solving a resource intensive problem of search of optimal radius of hyperspheres of detectors.

Evaluation of the significance of features p_m with respect to the output parameter T is the initial stage of the proposed method. It allows to identify and to exclude irrelevant features from further consideration, thereby reducing the search space and time of the method.

As noted above, in this paper we consider the problem in which the initial sample $S =< P, T >$ is characterized by a discrete number of classes $T' = \{t'_0, t'_1\}$. Therefore, to estimate the significance V_m of features p_m it is advisable to apply different criteria. These criteria allow to carry out an assessment of significance of features with respect to a discrete output parameter T [2, 4 and 17-22]. We propose to use entropy as an essential criterion [4 and 17]. The entropy reflects the degree of uncertainty of the state of the object. This criterion is calculated as:

$$V_m = -\sum_{n=1}^{N_{\text{int}}(p_m)} \left(\rho(p_{mn}) \sum_{n=1}^{N_{\text{int}}(T)} \rho(p_{mn}, t_l) \log_2 \rho(p_{mn}, t_l) \right), \quad (3)$$

where $\rho(p_{mn}) = \frac{N(p_{mn})}{Q}$ is a probability of that the feature value p_m of instances in the sample S gets to the m-th range of values; $N(p_{mn})$ is a number of instances in the sample S, the values of the m-th feature belong to the n-th interval of its range; $N_{\text{int}(p_m)}$ is a number of intervals into which the range of the m-th feature is divided; $N_{\text{int}}(T)$ is a number of possible values (intervals into which the range of values is divided) of the output parameter T; $\rho(p_{mn}, t_l) = \frac{N(p_{mn}, t_l)}{N(p_{mn})}$ is a conditional probability that the value of the output parameter T is equal to t_l (will be in l-th interval t_l) provided that the m-th feature p_m gets to the n-th interval p_{mn}; $N(p_{mn}, t_l)$ is a number of instances of the sample S, whose value of output parameter T is equal to t_l (belong

to *l*-th interval of the range of change t_l) with the proviso that the value of *m*-th feature belongs to *n*-th interval p_m.

Features p_m with values of the individual significance below minimum $(V_m < V_{\min})$, are considered uninformative and are excluded from the sample $S =< P, T >$.

We propose to estimate the relationship of features as the significance of one of them in relation to another. It allows to identify groups of interdependent features. Only one highly informative feature is saved in each of these groups. Other features can be excluded from further consideration. Because these features are redundant, they complicate the synthesis of diagnostic models and reduce their interoperability. The evaluation of the significance V_{md} is provided by the use of the entropy in (3). We believe that one of the features p_d is output parameter T (interval of feature values, which is considered as an output parameter p_d. This interval is split into $N_{\text{int}}(T)$ discrete intervals). After that the analogous features are excluded from the sample S (if the value of the mutual significance V_{md} is more than the maximum permissible $V_{md} > V_{\max}$).

Next, a set of detectors (structures that can determine whether the estimated instance belongs to a particular class) is built. The use of the principles of negative selection detectors for the class $T' = t_1'$ can detect unknown instances. Some of these instances do not belong to the relevant class t_1' [9, 11 and 13].

So, we form a set of detectors, which takes two output parameter values t_1' (class "self") and t_0' (class "non-self"). It is necessary for making samples S_0 and S_1 from $S =< P, T >$. We are using instances belonging to classes t_1' and t_0'. $S_1 =< P, T = t_1' >$ (sample of "self" instances) and $S_1 =< P, T = t_0' >$ (sample of "non-self" instances) is true in this case.

Thereafter, the first candidate in detectors $Ab_1 =< Ab_{1\min}, Ab_{1\max} > \in AB_1$ is created. In this case $Ab_{1\min} = \{Ab_{11\min}, Ab_{12\min}, \ldots, Ab_{1M\min}\}$ and $Ab_{1\max} = \{Ab_{11\max}, Ab_{12\max}, \ldots, Ab_{1M\max}\}$ are the sets of minimum and maximum values of the *m*-th features in the candidate detector Ab_1. We would like to note that $Ab_{1m\min} = \min_{q=1,2,\ldots,Q_1}(p_{qm})$, $Ab_{1m\max} = \max_{q=1,2,\ldots,Q_1}(p_{qm})$, m = 1,2,...,M. Q_1 is the number of instances in the sample S_1. This first candidate in detectors Ab_1 is presented in the form of a hypercube. Set AB_1 of detectors AB_k is formed based on the set of "self" instances S_1 and it allows to detect "non-self" instances, i.e. those instances that do not belong to the class t_1'.

We detect correspondence of every *n*-th instance s_q in the sample $S_1 =< P, T = t_1 >$ to candidate detector Ab_k based on (4):

$$eq(Ab_k, s_q) = \begin{cases} 1, \left(\sum_{m=1}^{M} \{1 \mid (Ab_{km\min} < p_{qm}) \wedge (Ab_{kmmax} < p_{qm})\}\right) = M; \\ 0, \left(\sum_{m=1}^{M} \{1 \mid (Ab_{km\min} < p_{qm}) \wedge (Ab_{km\max} < p_{qm})\}\right) \neq M; \end{cases} \quad (4)$$

where sum $\sum_{m=1}^{M} \{1 \mid (Ab_{km\min} < p_{qm}) \wedge (Ab_{km\max} < p_{qm})\}$ determines the number of corresponding feature values p_m of *q*-th instance to candidate A_{bk}. If $1, \left(\sum_{m=1}^{M} \{1 \mid (Ab_{km\min} < p_{qm}) \wedge (Ab_{km\max} < p_{qm})\}\right) = M,$ then it is considered that an instance $s_q =< p_{qm}, t_q >$ corresponds to the candidate detector Ab_k. In other words, this instance is located within the space of a hypercube with coordinates $Ab_{1\min} = \{Ab_{11\min}, Ab_{12\min}, \ldots, Ab_{1M\min}\}$ and $Ab_{1\max} = \{Ab_{11\max}, Ab_{12\max}, \ldots, Ab_{1M\max}\}$.

If there is at least one instance $s_q =< p_{qm}, t_q = t_1' > \in S_1$ with $eq(Ab_k, s_q) = 1$, then the candidate A_{bk} is activated by comparison with the instance S_1. This candidate cannot be a detector. Therefore, when the condition (5) is true, then the stage of additional training of the candidate Ab_k is performed.

$$\exists s_q \in S : eq(Ab_k, s_q) = 1 \quad (5)$$

The purpose of this stage is the conversion of the candidate detector Ab_k. This conversion needs to eliminate instances S_1, which activate the candidate Ab_k. One of the features $Ab_{km} =< Ab_{km\min}, Ab_{km\max} >$ is selected for this purpose. Candidate detector Ab_k should coincide with an instance s_q according to this feature. Next, one of the boundary values of the *m*-th feature of candidate A_{bk} is transformed. New boundary values equal: $Ab_{km\min} = p_{qm} + \eta_n (Ab_{km\max} - Ab_{km\min})$ if value *rnd* > 0.5 or $Ab_{km\max} = p_{qm} - \eta_n (Ab_{km\max} - Ab_{km\min})$ if value *rnd* ≤ 0.5). The value *rnd* is calculated as a randomly generated number between [0;1): *rnd* = *rand*[0;1). The volume of the hypercube Ab_k decreases based on this transformation. It is explained that an instance s_q is located outside the space described by the candidate for detector Ab_k. Ratio η_n is defined by user as a parameter of the method. We define this ratio in the range of $\eta_n \in (0;1]$. This ratio influences the distance between instances of the sample S_1 and hypercube detector Ab_k directly. The higher value of the ratio corresponds to the greater distance.

Candidate detector Ab_k, is re-checked with every instance in the sample S_1 by the condition (5) after conversion of the boundary values of one of the features. When the condition (5) is true then we have to re-transform the boundary values of one feature of the candidate Ab_k. This process will be repeated until the condition (5) is fulfilled.

We search examples s_q in set $S_1 =< P, T = t_1 >$ which activate candidate Ab_k. If these examples are absent then the step of fitness assessment of the candidate Ab_k for generalization is started. The set of detectors $AB_1 = \{Ab_1, Ab_2, \ldots Ab_{N_{Ab}}\}$ is created by using the principles of negative selection. This set allows determining the identity of instances s_q to the same class with high accuracy [10-16]. So, we use expressions (6) and (7) as

evaluation criteria of ability of the detector Ab_k for generalization of the data:

$$G_1(Ab_k) = \frac{1}{M}\sum_{m=1}^{M}\frac{Ab_{km\max} - Ab_{km\min}}{p_{m\max} - p_{m\min}}, \quad (6)$$

$$G_2(Ab_k) = \frac{\prod_{m=1}^{M}(Ab_{km\max} - Ab_{km\min})}{\prod_{m=1}^{M}(p_{m\max} - p_{m\min})}, \quad (7)$$

where $p_{m\min} = \min_{q=1,2,\ldots,Q_1}(p_{qm})$ and $p_{m\min} = \max_{q=1,2,\ldots,Q_1}(p_{qm})$ are minimum and maximum values of the *m*-th feature in the sample S_1.

Criteria (6) and (7) represent the part of search space which is covered by the detector. Criterion $G_1(Ab_k)$ shows the average fraction of coverage by the detector space in each of *M* dimensions of the feature space. Criterion $G_2(Ab_k)$ shows the volume part of the covered space. The values of these criteria influence the size of the search space which covers this detector. The greater values correspond to a greater part of the search space. Consequently, if the criterion for evaluating the quality of generalization $G_1(Ab_k)$ is above the threshold $G_{\min}$, then the candidate Ab_k has a high generalization capability. This candidate can be added to the set of detectors $AB_1 = AB_1 \cup \{Ab_k\}$.

The creation of new candidates Ab_k is carried out until the termination criteria are satisfied. These criteria can be used as follows: recognition accuracy *E*(*S*), the achievement of the maximum number of detectors $(N_{Ab} = |AB| > N_{Ab\max})$, the excess of the limit of search time $(t > t_{\max})$, etc.

The set of detectors $AB_1 = \{Ab_1, Ab_2, \ldots, Ab_{N_{Ab}}\}$ was generated by using negative selection. This set describes the region of the search space $\overline{S_1}$. The space $\overline{S_1}$ is complementary to the region of space in which there is a set of "self" instances S_1. The set $AB_1 = \{Ab_1, Ab_2, \ldots, Ab_{N_{Ab}}\}$ is characterized by high approximation and generalization capability.

Similarly, a set of detectors AB_0 for the set S_0 can be created. However, there are tasks with the uneven distribution of class instances in set $S =< P, T >$. Problems with the creation of detectors, which adequately reflect the space of instances S_0, may occur in these tasks. In particular, the detectors Ab_k in a form of hypercube with too large volume can be generated. These detectors cannot be able to summarize the data adequately. This is due to the insufficient number of instances in the sample $S_0(Q_0 << Q_1)$.

Therefore, we propose to calculate information on the size of the detectors which built on the basis of sample S_1. This information will also be used for generation of detectors for instances S_0.

These detectors will show information about the presence of instances of the sample S_0 in the hypercube. Information about their absence is not shown in this case. It is an essential difference of the proposed method from the approach based on classic negative selection. These detectors will be fully consistent with the detectors built previously for the sample S_1 based on negative selection. These detectors will contain information about areas of the search space in which items S_1 are not arranged.

Detectors $Ab_k^{(0)}$ of sample S_0 are generated so that their centers correspond to the coordinates of instances $s_k =< p_{km}, t_k = t'_0 >\in S_0$ of sample S_0. As the metric for estimation of the coordinates of detectors $Ab_k^{(0)}$ we used average detector length that can be considered as normalized Manhattan distance. The size of their hypercube edges correspond to the same sizes of detectors created on the basis of the sample S_1.

Consequently, the coordinates of the detector $Ab_{km}^{(0)} =< Ab_{km\min}^{(0)}, Ab_{km\max}^{(0)} >$ are determined using (8) and (9):

$$Ab_{km\min}^{(0)} = p_{km} - \frac{1}{2}\Delta Ab_m, \quad (8)$$

$$Ab_{km\max}^{(0)} = p_{km} - \frac{1}{2}\Delta Ab_m, \quad (9)$$

where ΔAb_m is an average length of the edges of detectors $AB_1 = \{Ab_1, Ab_2, \ldots, Ab_{N_{Ab}}\}$ which were created on the basis of the set S_1. Value ΔAb_m was calculated on the basis of the information about detectors $AB_1 = \{Ab_1, Ab_2, \ldots, Ab_{N_{Ab}}\}$ by (10):

$$\Delta Ab = \frac{1}{N_{AB}M}\left(\sum_{k=1}^{N_{Ab}}\sum_{m=1}^{M}(Ab_{km\max} - Ab_{km\min})\right). \quad (10)$$

Then, the comparison of the generated detectors $Ab_k^{(0)}$ to the instances of sample S_1 is performed by (4). If the condition (5) is true then the detectors $Ab_k^{(0)}$ are converted similarly to the above stage of additional training. Then the value of one of criteria $G(Ab_k^{(0)})$ is calculated. This criteria estimates the ability of detector to generalize the data. If its value is above the threshold, the detector $Ab_k^{(0)}$ is added to the set $AB_0 = AB_0 \cup \{Ab_k^{(0)}\}$.

Thus the set of detectors AB_0 is generated. Both set AB_0 and set AB_1 describe the area of the search space $\overline{S_1}$. This space is complementary to the region of arranging instances S_1. Therefore, a recognizing model can be represented as a set of detectors $AB = AB_0 \cup AB_1$. These detectors permit to recognize membership of unknown instances $s'_q =< p'_{qm}, t'_q >\notin S$ to the class of "non-self", i.e. to refer them to the class $t'_0 : t'_q = t'_0$.

Lets us improve the level of interoperability of the synthesized recognition model which was presented as a set of detectors $AB = \{Ab_1, Ab_2, \ldots, Ab_{N_{Ab}}\}$. A set of classification rules $PR_r : P_r \to T_r$ based on a set *AB* is formed for this goal. Note that the left part P_r of the implication is a set of conditions (11):

$$\text{If } \begin{array}{l}(p_1 \in [Ab_{k1\min}; Ab_{k1\max}]) \wedge (p_2 \in [Ab_{k2\min}; Ab_{k2\max}]) \\ \wedge \ldots \wedge (p_M \in [Ab_{kM\min}; Ab_{kM\max}])\end{array} \quad (11)$$

The right part T_r contains the value of the output parameter *T* when performing an *r*-th set of conditions P_r (11).

When generating a set of rules *PR* to the rules P_r antecedently, we will include only the boundaries of

features for which they are not the limit values, i.e., $Ab_{km\min} \neq \min_{q=1,2,\ldots,Q}(p_{qm})$ and $Ab_{km\max} \neq \max_{q=1,2,\ldots,Q}(p_{qm})$. For example, the rule PR_k is created for detector Ab_k=[<5,7>,<8,$p_{2\max}$>,<$p_{3\min}$, $p_{3\max}$>,<4,6>]. This rule has the following structure:
If $(p_1 > 5 \wedge p_1 < 7) \wedge (p_2 > 8) \wedge (p_4 > 4 \wedge p_4 < 6)$ then the instance corresponds to the class of "non-self" $(T \neq t_1)$.

The upper limit of feature p_2 and feature p_3 has not been included in the rule explicitly. Because the corresponding values of the detector do not affect the quality of recognition. In addition, the exclusion of such values from the rule PR_k reduces its complexity. Interoperability and comprehensibility of the rule will be increased from the other side. So, classification rules based on each detector Ab_k are constructed. So, a set PR of N_{Ab} classification rules $PR_r : P_r \rightarrow T_r$ has been created with the using of proposed approach.

The proposed method for the classification rules synthesis is based on negative selection approach. This method is oriented to the case of uneven distribution of class instances of sample in generating a set of detectors. The proposed method uses known information about instances of all classes of the sample. It also takes into account information about the individual significance of features. A hypercube of maximum possible volume is used as a form of detector. It allows to exclude irrelevant and redundant features from the sample, thereby reducing the search space and time of the method implementation. As a result, a set of detectors with high approximation and generalization capability is formed.

The proposed method increases the generalizing properties of the synthesized model by reducing the number of detectors and conditions of antecedents. This method improves interoperability of the model, reduces its dimension (structural and parametric complexity) and volume of the used memory. All of these improvements increase the model performance with sequential computation.

4. Experiments and results

A computer program has been developed for implementation of the proposed method of classification rule synthesis based on negative selection. This software is oriented to the verification and analysis of different characteristics of this method. This software deals with a blade diagnosis of an aircraft engine gas turbine [23]. The blades of gas turbine were characterized by the values of the power spectra of damped oscillations after impact excitation. These values of the the power spectra are used as input features. Classes of blade quality were defined with the help of experts: undamaged and defective (potentially dangerous). Each blade was described by 10240 characteristics of the power spectrum of damped oscillations. Artificial features were constructed to reduce the search space based on these characteristics. A set consisting of 80 artificial features was obtained based on this reducing.

The resulting sample $S =< P, T >$ does not have statistical representativeness, because it does not display the actual frequency distribution of classes. Really, the number of undamaged blades is substantially greater in the general population than the number of defective blades. These defective blades $(t_q = t'_1)$ in the sample represent typical cases of nonconformity, which provides a topological representation of defective blades in the sample. All the possible cases of the class of undamaged blades $(t_q = t'_0)$ cannot be present in a sample from a practical point of view. Therefore, it is necessary to build a diagnostic model for aircraft engine blade class recognition based on the available sample $S =< P, T >$ with uneven distribution of instances of classes.

The sample $S =< P, T >$ contains 42 instances characterizing defective (potentially dangerous) blades and 72 instances representing undamaged blades. The proposed method for the synthesis of classification rules was compared with the existing methods of negative selection. These methods synthesized a set of detectors based on "self" instances $S_1 \subseteq S$ of the sample only. Therefore, the problem of blade diagnosis of gas turbine of aircraft engines was solved with the proposed method two times:

- using a sub-sample $S_1 \subseteq S$, which contains information about defective (potentially dangerous) instances ("self") only;
- using all the original samples $S =< P, T >$.

Experimental investigation of characteristics of the proposed method has been compared with other methods of negative selection.

The first part of experimental results is given in Table 1. This table contains next values. Column N_{it} describes a number of iterations of the method. Column t contains calculation time. Misclassification error on the training data $S =< P, T >$ indicated at column E. Column E_t contains misclassification error of the test data. Columns $P_{t,t_q=t'_1/t_q=t'_0}$ and $P_{t,t_q=t'_0/t_q=t'_1}$ indicated probability of misclassification. So, value $P_{t,t_q=t'_1/t_q=t'_0}$ is error probability of assignment to class "self" $(t_q = t'_1)$ when the instance actually belongs to a class of «non-self» $(t_q = t'_0)$. Similarly, value $P_{t,t_q=t'_0/t_q=t'_1}$ is error probability of assignment to class "non-self" $(t_q = t'_0)$ when an instance actually belongs to a class of «self» $(t_q = t'_1)$. These probabilities are calculated using (12) and (13) respectively:

$$P_{t,t_q=t'_1/t_q=t'_0} = \frac{N_{t,t_q=t'_1/t_q=t'_0}}{N_{t,t_q=t'_0}}, \quad (12)$$

$$P_{t,t_q=t'_0/t_q=t'_1} = \frac{N_{t,t_q=t'_0/t_q=t'_1}}{N_{t,t_q=t'_1}}. \quad (13)$$

These expressions contain next variables. Variable $N_{t,t_q=t'_1/t_q=t'_0}$ is a number of instances of the test sample recognized as "self" $(t_q = t'_1)$, but actually belonging to the class of «non-self»

$(t_q = t'_0)$. Variable $N_{t,t_q=t'_0}$ is a number of instances of the test sample belonging to the class of "non-self" $(t_q = t'_0)$. Variable $N_{t,t_q=t'_0/t_q=t'_1}$ is a number of instances of the test sample recognized as "non-self" $(t_q = t'_0)$, but actually belonging to the class of «self» $(t_q = t'_1)$. Variable $N_{t,t_q=t'_1}$ is a number of instances of the test sample belonging to the class of "self" $(t_q = t'_1)$. Test sample contains 273 instances. The subset of 261 instances from them belongs to the class $t_q = t'_0$ (undamaged products) and the subset of next 12 instances belongs to the class $t_q = t'_1$ (defective instances).

The second part of experimental investigation compared the proposed method with other classification methods. The above-described problem of blade diagnosis of a gas turbine of aircraft engines [23] was solved by these methods. Several different models of this problem were obtained based on these methods. We analyzed the next classification models:

- a model in form of classification rules synthesized by the proposed method MPRSBNS;
- a feed-forward neural network which has been trained by using error back-propagation. This network consists of three layers of neurons. The first layer of the neural network has five neurons, the second has three neurons and the third layer has one neuron only. Neurons in the first and the second layers have got a logistic sigmoid activation function. In the third layer a single neuron has got a threshold activation function;
- a model in form a set of detectors constructed by the method MMD [16].

The whole training set of 114 instances was used for experiment investigation of the first and the second models. This training set included (a) a subset of 42 instances characterizing defective blades and (b) a subset of 72 instances representing undamaged blades. Part of the sample $(S_1 \subseteq S)$ was used for constructing the third model based on a set of detectors using the method MMD. We had to do it because this method deals with instances of one class only.

Some criteria of these models have been calculated for analysis. The results of calculation are given in Table 2. We analyzed the next criteria of such classification models:

- N_{param} is a criterion determining the parametric complexity of the model. This criterion is calculated as the number of model parameters. In particular, the total number of parameters $Ab_{km\,\min}$ and $Ab_{km\,\max}$ for the first and the third models; and the total number of adjustable parameters (weight coefficients) for the second model;
- criteria E, E_t, $P_{t,t_q=t'_1/t_q=t'_0}$ and $P_{t,t_q=t'_0/t_q=t'_1}$ were described in Table 1.

5. Discussion

Table 1 shows that the misclassification error values E produced by the method MMD [16] (E=0.018) and by the proposed method MPRSBNS (E = 0.026 and E = 0.009) are acceptable. The low recognition errors of these methods were provided by the wide coverage of field of "self" instances $S_1 \subseteq S$ by synthesized detectors. The proposed method MPRSBNS synthesized a set of detectors based on instances of all classes of the sample $S =< P, T >$. This method provided more acceptable results (E = 0.009) compared to a set of detectors synthesized using "self" instances $S_1 \subseteq S$ (E = 0.026) only. Method RNS [13] and model V-Detector [14 and 15] had less acceptable misclassification error value E (E = 0.070 and E = 0.035, respectively). This fact indicates that synthesized detectors do not cover the area of "self" instances

Results of the first part of experiments — Table 1

Method	N_{Ab}	N_{ll}	t, ms	E	E_t	$P_{t,t_q=t'_1/t_q=t'_0}$	$P_{t,t_q=t'_0/t_q=t'_1}$
Real-Valued Negative Selection (RNS) [13]	207	50	27.3	0.070	0.136	0.126	0.333
Model V-Detector (MVD) [14, 15]	41	50	24.1	0.035	0.077	0.069	0.250
Method with masking of detectors (MMD) [16]	19	14	13.2	0.018	0.055	0.054	0.083
Method of classification rules synthesis on the basis of negative selection MPRSBNS (using sample $S_1 \subseteq S$)	20	12	12.1	0.026	0.037	0.038	0
Method of classification rules synthesis on the basis of negative selection MPRSBNS (using sample $S =< P, T >$)	31	19	13.7	0.009	0.011	0.011	0

Results of the second experimental part. Comparison of different classification models — Table 2

Model	N_{param}	E	E_t	$P_{t,t_q=t'_1/t_q=t'_0}$	$P_{t,t_q=t'_0/t_q=t'_1}$
Model as a set of classification rules synthesized by the proposed method MPRSBNS	652	0.009	0.011	0.011	0
Feed-forward neural network	427	0.018	0.070	0.065	0.167
Model of a set of detectors constructed by the method with masking of detectors MMD [16]	804	0.018	0.055	0.054	0.083

$S_1 \subseteq S$. The experimental results show that the method RNS [13] and the model V-Detector [14 and 15] generate the largest number of detectors (N_{Ab}= 207 and N_{Ab}= 41, respectively). It increases the learning time *t* and computational costs of computer resources. The time of calculations is not so important criterion in the experiment. Accuracy of prediction (criteria E_t, $P_{t,t_q=t'_1/t_q=t'_0}$ and $P_{t,t_q=t'_0/t_q=t'_1}$) is more important. Method MMD [16] and the proposed method MPRSBNS (using sample $S_1 \subseteq S$) generated significantly fewer number of detectors (N_{Ab}= 19 and N_{Ab}= 20, respectively). It indicates a more efficient operation of these methods. In particular, the method MPRSBNS uses a priori information about the significance of features at the initial stage. This method eliminates from further consideration irrelevant and redundant features that can reduce the search space and create a set of a small number of detectors based on highly informative features with a high approximation and generalization capability.

Criteria E_t, $P_{t,t_q=t'_1/t_q=t'_0}$ and $P_{t,t_q=t'_0/t_q=t'_1}$ were used for the analyzing of properties and characteristics of the investigated methods. These criteria describe misclassification error and the probability of making a wrong decision based on test data. Misclassification errors of models synthesized by the proposed method MPRSBNS and methods [13-16] are shown in Table 1. These errors have been calculated on test data E_t. Misclassification error of the proposed method MPRSBNS is significantly lower than the error of other known methods (E_t = 0.136, E_t = 0.077 and E_t = 0.055 for the methods [13, 15 and 16], respectively). It can be explained by using the characteristics $G(Ab_k)$. These characteristics allow to evaluate the ability of the detector for the generalization of data. The proposed method MPRSBNS allowed to reach misclassification error E_t = 0.037 (using a part of the sample $S_1 \subseteq S$) and E_t = 0.011 (using the full sample $S =< P, T >$).

It is important to note the specificity of the solved problem of blade diagnosis. An error of assignment to "non-self" class $(t_q = t'_0)$ has a very high cost when the instance actually belongs to a «self» class $(t_q = t'_1)$. This error has been evaluated by criterion $P_{t,t_q=t'_0/t_q=t'_1}$. This is due to the fact that the classification of defective blades to the class of undamaged can cost human lives. The test data has zero error probability $P_{t,t_q=t'_0/t_q=t'_1}$ for the proposed method MPRSBNS (see Table 1). This fact indicates high efficiency of the proposed method for solving such problems. The zero level of error probability $P_{t,t_q=t'_0/t_q=t'_1}$ by using the proposed method is explained by:

- a high level of coverage of typical instances of class $t_q = t'_1$. This coverage was made by generated set of detectors $AB = \{Ab_1, Ab_2, \ldots, Ab_{N_{Ab}}\}$. Note that this set of detectors was obtained using a priori information about the importance of features;
- a high generalizing ability of synthesized set of detectors. That is caused by the use of the criteria (6) and (7). These criteria allow to estimate ability of the detector to the data generalization.

The number parameters N_{param} of the model synthesized by the proposed method (N_{param}= 652) is less than a similar model constructed by the method MMD [16] (N_{param}= 804) (see Table 2). This is due to the fact that when using the proposed method, the average size of the generated detector is less. Such reduction can be explained that a priori information of feature significance is used during the process of negative selection. It allows to exclude from further consideration irrelevant and redundant features that complicate the process of synthesis of diagnostic models and reduce their interpretability. Thus, the model synthesized by the proposed method MPRSBNS is more simple and straightforward compared to the model created by the method of [16]. Approximation and generalization capabilities of the model synthesized by the method MPRSBNS are also higher. This fact is confirmed by the values of the criteria *E*, E_t, $P_{t,t_q=t'_1/t_q=t'_0}$ and $P_{t,t_q=t'_0/t_q=t'_1}$.

The comparison of the model synthesized on the basis of MPRSBNS method and the neural network model allows next to conclude. The model constructed by the proposed method has higher generalizing and approximation abilities (criteria *E*, E_t, $P_{t,t_q=t'_1/t_q=t'_0}$ and $P_{t,t_q=t'_0/t_q=t'_1}$). However, the number of model parameters constructed by the method MPRSBNS (N_{param}= 652) is greater than in the neural network model (N_{param}= 427). It can be explained by representation of neural network as a set of neurons interconnected in a certain way and characterized by weighting coefficients as adjustable parameters. And each neuron corresponds to a function of many arguments. At the same time, this neural network model is difficult enough for human perception. The model as a set of classification rules synthesized by the proposed method is more intuitive in comparison with the neural network model. Really, classification rules of the form "if condition, then action" are much more understandable and human-readable than a set of coefficients that reflect the degree of neuronal connections in the neural network model.

Thus, the results of experiments showed that the proposed method due to the usage of a priori information and exclusion of irrelevant and redundant features of the sample makes it possible to reduce the search space and time of execution. Proposed method allows to synthesize classification models in a form of a set of detectors with high approximation and generalization capabilities. Also by reducing the number of detectors and the conditions in antecedents it increases interpretability of the model, reduces its dimension and, therefore, the size of the used memory.

6. Conclusions

In this paper we solve the problem of automation of classification rule synthesis based on negative selection for the case of uneven class distribution in the sample.

The developed method of classification rule synthesis based on negative selection uses a priori information about instances of all classes in the sample at detector set generation. It also takes into account information about the individual feature significance. A hypercube of maximum possible volume is used as a form of detector. It allows to exclude irrelevant and redundant features from the sample, thereby reducing the search space and time of execution of the method, as well as to generate a set of detectors with high approximation and generalization capability. The proposed method improves interoperability of the model, reduces its dimension, size of used memory and improves the model performance for the sequential computation. It is obtained by increasing the generalizing properties of synthesized model by reducing the number of detectors and conditions of antecedents.

An experimental study of the proposed method and its comparison with the known analogues is performed. A practical task of diagnosing the vanes of gas turbine of aircraft engines has been solved. The mathematical approach proposed at [24 and 25] can be used for reliability analysis of the proposed solution.

References

[1] SOBHANI-TEHRANI, E., KHORASANI, K.: *Fault Diagnosis of Nonlinear Systems using a Hybrid Approach.* Springer, 2009.

[2] DENTON, T.: *Advanced Automotive Fault Diagnosis.* Elsevier, 2006.

[3] GERTLER, J.: *Fault Detection and Diagnosis in Engineering Systems.* Marcel Dekker Inc., 1998.

[4] *Encyclopedia of Artificial Intelligence*. DOPICO, J. R., CALLE, J. D., SIERRA, A. P. (Eds). Information Science Reference, 2009.

[5] *Encyclopedia of Machine Learning*. SAMMUT, C., WEBB, G.I. (Eds). Springer, 2011.

[6] MUMFORD, C. L., JAIN, L. C.: *Fault Detection and Diagnosis in Engineering Systems.* Springer-Verlag, 2009.

[7] MO, C. L.: *Handbook of Research on Artificial Immune Systems and Natural Computing: Applying Complex Adaptive Technologies.* Information Science Reference, 2009.

[8] TARAKANOV, A., SKORMIN, V., SOKOLOVA, S.: *Immunocomputing: Principles and Applications.* Springer, 2013.

[9] DASGUPTA D., NINO, F.: *Immunological Computation: Theory and Applications.* Auerbach Publications, 2008.

[10] JI, Z., DASGUPTA, D.: Revisiting Negative Selection Algorithms. *Evolutionary Computation,* 2007, 15: 223-251.

[11] ZENG, J., LI, T., LIU, X.: *A Feedback Negative Selection Algorithm to Anomaly Detection.* Proc. of Natural Computation, Haikou, 2007: 604-608.

[12] ZHANG, P. T., WANG, W., TAN, Y.: A Malware Detection Model Based on a Negative Selection Algorithm with Penalty Factor. *Science China. Information sciences,* 2010, 53(12), 2461-2471.

[13] OLIVEIRA, L. O., DRUMMOND, I. N.: Real-valued Negative Selection (RNS) for Classification Rask. *Recognizing Patterns in Signals, Speech, Images and Videos,* 2010, 6388, 66-74.

[14] JI, Z.: *A Boundary-aware Negative Selection Algorithm.* Proc. of Intern. Conference on Intelligence and Soft Computing, Benidorm, 2005, 49-55.

[15] JI, Z.: *Negative Selection Algorithms: From the Thymus to V-detector*: doctor of philosophy in "Computer Science" thesis, Memphis, 2006.

[16] ZAITSEV, S., SUBBOTIN, S.: *Negative Selection using Masked Detectors.* Proc. of seminar on Modelling of non-equilibrium systems, Krasnoyarsk, 2011, 95-98.

[17] SUBBOTIN, S., OLEYNIK, A.: *Entropy Based Evolutionary Search for Feature Selection.* Proc. of Intern. Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, Lviv, 2007, 442-443.

[18] OLIINYK, A., SUBBOTIN, S.: The Decision Tree Construction Based on a Stochastic Search for the Neuro-fuzzy Network Synthesis. *Optical Memory and Neural Networks (Information Optics),* 2015, 24(1): 18-27.

[19] OLIINYK, A., ZAIKO, T., SUBBOTIN, S.: Training Sample Reduction Based on Association Rules for Neuro-Fuzzy Networks Synthesis. *Optical Memory and Neural Networks (Information Optics),* 2014, 23(2): 89-95.

[20] OLIINYK, A., SKRUPSKY, S., SUBBOTIN, S.: Using Parallel Random Search to Train Fuzzy Neural Networks. *Automatic Control and Computer Sciences,* 2014, 48(6): 313-323.

[21] OLIINYK, A., SKRUPSKY, S., SUBBOTIN, S.: Experimental Investigation with Analyzing the Training Method Complexity of Neuro-Fuzzy Networks Based on Parallel Random Search. *Automatic Control and Computer Sciences,* 2015, 49(1): 11-20.

[22] BODYANSKIY, YE., VYNOKUROVA, O.: Hybrid Adaptive Wavelet-neuro-fuzzy System for Chaotic Time Series Identification. *Information Sciences,* 2013, 220: 170–179.

[23] DUBROVIN, V., SUBBOTIN, S., BOGUSLAYEV, A., YATSENKO, V.: *Intelligent Diagnosis and Prediction of Reliability of Aircraft Engines (in Russian).* Motor-Sich, 2003.

[24] ZAITSEVA, E., KVASSAY, M., LEVASHENKO, V., KOSTOLNY, J.: Reliability Analysis of Logic Network with Multiple Outputs, *Communications - Scientific Letters of the University of Zilina*, 2015, vol. 1A: 44-50.

[25] KVASSAY, M., KOSTOLNY, J., Evaluation of Algorithms for Identification of Minimal Cut Vectors and Minimal Path Vectors in Multi-State Systems, *Communications - Scientific Letters of the University of Zilina*, 2015, vol. 17, No. 4: 8-14.

Jaroslav Janacek - Marek Kvet*

MIN-MAX ROBUST EMERGENCY SERVICE SYSTEM DESIGN

This paper deals with the emergency service system design using the weighted p-median problem formulation. In such systems, not only the disutility of an average user is minimized, but also the disutility of the worst situated users must be taken into account. To cope with both objectives, we suggest a composed method. In the first phase, the disutility of the worst situated user is minimized. The second phase is based on the min-sum approach to optimize the average user's disutility. To formulate the mathematical models, the radial approach is used mainly for its excellent performance characteristics. Within this paper, we concentrate on effective usage of the radial approach to develop an algorithm for robust emergency service system design. A robust service system design is usually performed so that the design complies with specified scenarios so that the maximal objective function value of the individual instances corresponding with particular scenarios is minimized. To find the value paid for making the system resistant to catastrophic events, a new conception called the price of robustness is introduced.

Keywords: *Emergency service system, min-max approach, radial formulation, robustness.*

1. Introduction

Emergency service system design for a given road network in serviced area locates limited number of service centers at positions from a given set of possible locations to satisfy future system users' demands for service in case of emergency [1]. Different objectives can be applied on the design. A traditional one is minimal disutility perceived by an average user. In such a case, the perceived disutility is assumed to be proportional to the distance of a user location from the nearest located service center and then, sum of distances from particular system users to the nearest located service center is minimized. Such objective is referred to as min-sum criterion and is broadly used in private service system designs, where service delivering is provided and paid by system owner. In contrast to the private systems, users of an emergency public service system share cost of the system by paying tax, which approves them to claim the equal or fair access to the provided service. In general, the fairness emerges whenever limited resources are to be fairly distributed among participants [2, 3 and 4]. Plethora of fairness schemes were studied, but the strongest one applicable in the public service system design is so called lexicographic min-max criterion [5, 6, and 7]. As the lexicographical min-max approach produces such a system design, where the price of fairness, i.e. relative deterioration of the average user's disutility, is too high, composed approaches were designed to mitigate this drawback [8]. A composed approach usually performs two phases, where the first one strives to minimize disutility perceived by the worst situated users using min-max objective and the second phase applies the min-sum approach under condition that the disutility of the most exposed user must not be worsened.

As traversing time between service center and an affected user might be impacted by various random events following weather or traffic, the system designer must face the demand for system resistance to such critical events [9 and 10]. Most of the approaches to increasing the system resistance are based on making its design resistant to possible failure scenarios, which can appear in the road network as a consequence of random failures due to congestion, disruptions or blockages. An individual scenario is characterized by particular time distances between the users' and possible service center locations. A robust service system design has to comply with all the specified scenarios. The usual way of taking into account all scenarios is based on minimizing the maximal objective function of the individual instances corresponding with the particular scenarios. The min-max link-up constraints represent an undesirable burden in any integer programming problem due to bad convergence of the branch-and-bound method, which dominates solving tools of available IP-solvers. Thus these approaches to the robustness constitute a big challenge to family of operational researchers and professionals in informatics.

* **Jaroslav Janacek, Marek Kvet**
Faculty of Management Science and Informatics,University of Zilina, Slovakia
University Science Park, University of Zilina, Slovakia
E-mail: jaroslav.janacek@fri.uniza.sk

Within this paper, we focus on the emergency service system design, which is robust considering given finite set of scenarios. The basic design problem is formulated as min-max first and min-sum second composed approach, which means that the accented objective is to minimize disutility perceived by the worst situated user and the point of the average user is of secondary importance. The basic design is computed for the original scenario, which corresponds to the usual situation in the network, i.e. situation, which does not correspond to any of the above-mentioned possible failure scenarios.

Complexity of location problems with limited number of facilities to be deployed and the necessity to solve large instances of the problem led to searching for a suitable algorithm. It was found that in contrast to original location-allocation formulation, the radial formulation of the problem can considerably accelerate the associated solving process [11 and 12]. Simultaneously, an attention was paid to the radial formulation with homogenous system of radii [13]. As this approximate approach used for the "system optimal" public service system design proved to be a suitable and enough precise tool, we decided to apply the radial formulation with homogenous system of radii also on the robust emergency system design.

The remainder of the paper is organized as follows: Section 2 is devoted to the description of original composition of robust min-max and min-sum method including the radial formulation. The customization of the suggested method is described in Section 3 and the associated numerical experiments are performed in Section 4. The results and findings are summarized in Section 5.

2. Min-max optimal robust design of emergency service system

The min-max public service system design problem can be described by the following denotation. Let symbol J denote the set of users' locations and symbol I denote set of possible service center locations. We denote by b_j the number of users, who share the location j. To solve the problem, at most p locations must be chosen from I so that the maximal disutility perceived by the worst situated user be minimal. The value of user's disutility is given by the mutual positions of the users´ location and the location of the service center providing them with service. Let symbol U denote the set of possible failure scenarios. We assume that user's disutility grows with increasing distance between the user and the service center. Disutility following from the distance between locations i and j under a specific scenario $u \in U$ is denoted as d_{iju}. The decisions, which determine the designed public service system, can be modeled by further introduced decision variables. The variable $y_i \in \{0,1\}$ models the decision on service center location at place $i \in I$. The variable takes the value of 1 if a service center is located at i and it takes the value of 0 otherwise.

We use also the variable h_u as the upper bound of the all perceived disutility values under scenario $u \in U$. To obtain an upper or a lower bound of the original objective function, the range $[d_0, d_m]$ of all m+1 possible disutility values $d_0 < d_1 <...< d_m$ from the matrix $\{d_{iju}\}$ is partitioned into v+1 zones according to [13 and 14]. The zones are separated by values from a finite ascending sequence of so called *dividing points* $D_1, D_2 ... D_v$ chosen from the sequence $d_0 < d_1 <...< d_m$, where $0 = d_0 = D_0 < D_1$ and also $D_v < D_{v+1} = d_m$. The zone s corresponds to the right-closed interval $(D_s, D_{s+1}]$. The length of the s-th interval is denoted by e_s for $s = 0 ... v$. Further, auxiliary zero-one variables x_{jus} for $j \in J$, $u \in U$ and $s = 0, ..., v$ are introduced. The variable x_{jus} takes the value of 1, if the disutility of the user located at $j \in J$ under scenario $u \in U$ from the nearest located center is greater than D_s and it takes the value of 0 otherwise. Then the expression $e_0 x_{ju0} + e_1 x_{ju1} + e_2 x_{ju2} + ... + e_v x_{juv}$ constitutes an upper approximation of the disutility d^*_{ju} perceived at user location j from the nearest located service center under scenario $u \in U$. If the disutility d^*_{ju} belongs to the interval $(D_s, D_{s+1}]$, then the value of D_{s+1} is the upper estimation of d^*_{ju} with the maximal possible deviation e_s. Let us introduce a zero-one constant $a_{iju}{}^s$ under scenario $u \in U$ for each triple $[i, j, s]$, where $i \in I, j \in J$, $s = 0, ..., v$. The constant $a_{iju}{}^s$ is equal to 1, if the disutility d_{iju} perceived at the user location j from the possible center location i is less or equal to D_s, otherwise $a_{iju}{}^s$ is equal to 0. Then the radial-type min-max public service system design problem under given scenario $u \in U$ can be formulated as follows:

$$Minimize\ h_u \tag{1}$$

$$Subjet\ to: x_{jsu} + \sum_{i \in I} a^s_{iju} y_i \geq 1 \quad for\ \ j \in J, s = 0, 1, \ldots, v \tag{2}$$

$$\sum_{i \in I} y_i \leq p \tag{3}$$

$$\sum_{s=0}^{v} e_s x_{jus} \leq h_u \quad for\ \ j \in J \tag{4}$$

$$y_i \in \{0,1\} \quad for\ \ i \in I \tag{5}$$

$$x_{jus} \geq 0 \quad for\ \ j \in J, s = 0, 1, \ldots, v \tag{6}$$

$$h_u \geq 0 \tag{7}$$

In this model, the objective function (1) represented by single variable h_u gives the upper bound of the all perceived disutility values. The constraints (2) ensure that the variables x_{jus} are allowed to take the value of 0, if there is at least one center located in radius D_s from the user location j and constraint (3) limits the number of located centers by p. The link-up constraints (4) ensure that each perceived disutility is less than or equal to the upper bound h_u.

As concerns the obligatory constraints (6), only values zero and one are expected in a feasible solution, but it can be seen that the model has integrality property regarding the variables x_{jus}. It can be noticed that in the optimization process all relevant values of x_{jus} are "pushed down" and the constraints (2) and (6) bound the variable x_{jus} from below by value of one or zero. It follows that the relevant values of x_{jus} stay at one of these values.

The above-described way of modeling the min-max problem will be called "standard" min-max approach in the remainder of the paper.

Having solved the above problem, we denote h_u^* the optimal value of the objective function (1), then we can formulate the second phase of the composed approach to the design problem under the given scenario u by the following model:

$$Minimize \sum_{j\in J} b_j \sum_{s=0}^{v} e_s x_{jus} \tag{8}$$

$$Subject\ to: \sum_{s=0}^{v} e_s x_{jus} \le h_u^* \ \ for\ j \in J \tag{9}$$

and (2), (3), (5), (6).

Let us denote the set of all feasible solutions $(\boldsymbol{x}_u, \boldsymbol{y})$ of the constraint system (2), (3), (5), (6) by Q_u, then the first phase of the robust approach for the set U of scenarios can be modeled according to [9 and 10] as:

$$Minimize\ h \tag{10}$$

$$Subject\ to: \sum_{s=0}^{v} e_s x_{jus} \le h \ \ for\ j \in J, u \in U \tag{11}$$

$$(x_u, y) \in Q_u \ \ for\ u \in U \tag{12}$$

$$h \ge 0 \tag{13}$$

The second phase of the approach to the robust emergency system design can be performed in several ways due to the point of the average user is of secondary importance. We will study here the simplified approach, which consists in minimization of the average user disutility, when each combination user location and scenario is taken into account. This approach makes use of the optimal value h^* of (10) subject to (11)-(13) and the associated model can be stated in the following form.

$$Minimize \sum_{u\in U} \sum_{j\in J} b_j \sum_{s=0}^{v} e_s x_{jus} \tag{14}$$

$$Subject\ to: \sum_{s=0}^{v} e_s x_{jus} \le h^* \ \ for\ j \in J, u \in U \tag{15}$$

$$(x_u, y) \in Q_u \ \ for\ u \in U \tag{16}$$

3. Bisection radial approach to design of robust emergency service system

The bisection radial approach makes use of the radial model, but it uses only its reduced form to find whether there is any solution with the objective function value less than or equal to a prescribed disutility value D_s. The bisection is performed according to subscript s from the subscript range from 0 to v with the goal to determine the lowest subscript, for which the objective function (17) equals to 0.

In the following model, the zero-one variables $y_i \in [0, 1]$ for $i \in I$ are also used to the decisions on locating or not a service center at the location i. Next, the variables x_{ju} are introduced to indicate, whether user's disutility at location $j \in J$ following from the nearest located center under scenario $u \in U$ is greater than D_s. If user's disutility at location $j \in J$ under scenario $u \in U$ is greater than D_s, then the variable x_{ju} takes the value of 1, and it takes the value of 0 otherwise. The corresponding model can be formulated as follows.

$$Minimize \sum_{u\in U} \sum_{j\in J} x_{ju} \tag{17}$$

$$Subject\ to: x_{ju} + \sum_{i\in I} a_{iju}^s y_i \ge 1 \ \ for\ \ j \in J, u \in U \tag{18}$$

$$\sum_{i\in I} y_i \le p \tag{19}$$

$$y_i \in \{0, 1\} \ \ for\ i \in I \tag{20}$$

$$x_{ju} \ge 0 \ \ for\ \ j \in J, u \in U \tag{21}$$

In this model, the objective function (17) represents the number of user locations, where the perceived disutility is greater than D_s under scenario $u \in U$. The constraints (18) ensure that the variables x_{ju} are allowed to take the value of 0, if there is at least one center located in radius D_s from the user location j under scenario $u \in U$ and constraint (19) limits the number of located service centers by p.

Having performed the above bisection process, we denote s^* the minimal value of the subscript s, for which the objective function (17) equals to zero, we can formulate the second phase of the composed approach by the following model:

$$Minimize \sum_{u\in U} \sum_{j\in J} b_j \sum_{s=0}^{s^*-1} e_s x_{jus} \tag{22}$$

$$Subject\ to: x_{jus} + \sum_{i\in I} a_{iju}^s y_i \ge 1 \ \ for\ j \in J, s = 0, \ldots, s^* - 1, u \in U \tag{23}$$

$$\sum_{i\in I} a_{iju}^{s^*} y_i \ge 1 \ \ for\ j \in J, u \in U \tag{24}$$

$$\sum_{i\in I} y_i \le p \tag{25}$$

$$y_i \in \{0,1\} \quad for \quad i \in I \tag{26}$$

$$x_{jus} \geq 0 \quad for \quad j \in J, s = 0,\ldots,s^* - 1, u \in U \tag{27}$$

4. Numerical experiments

To compare studied approaches based on the radial formulation and their usage for basic and robust design of the emergency service system, we performed the series of numerical experiments. To solve the problems described in the previous sections, the optimization software FICO Xpress 7.9 (64-bit, release 2015) was used and the experiments were run on a PC equipped with the Intel® Core™ i7 5500U processor with the parameters: 2.4 GHz and 16 GB RAM.

The used benchmarks were derived from the real emergency health care system, which was originally implemented in eight regions of the Slovak Republic. For each self-governing region, i.e. Bratislava (BA), Banska Bystrica (BB), Košice (KE), Nitra (NR), Presov (PO), Trencin (TN), Trnava (TT) and Zilina (ZA), all cities and villages with corresponding number of inhabitants b_j were taken. The coefficients b_j were rounded to hundreds. These sub-systems cover demands of all communities - towns and villages spread over the particular regions by a given number of ambulance vehicles. In the benchmarks, the set of communities represents both the set J of users' locations and also the set I of possible center locations. The cardinalities of these sets vary from 87 to 664 according to the considered region. The number p of located centers was derived from the original design and it varies from 9 to 67. The network distance from a user to the nearest located center was taken as an individual user´s disutility.

The first set of numerical experiments was performed to compare the standard and bisection radial approaches to the emergency service system design. This comparison was performed for the basic situations (basic scenarios) in all eight self-governing regions. The obtained results are summarized in Table 1. Studied parameters of the experiments were:

CT – computational time in seconds, which was recorded for the individual phases
h –maximal disutility perceived by the worst situated users
minSum – min-sum objective function value according to (8) for the basic scenarios of the individual self-governing regions.

Since convergence of the standard optimization process is very slow and the demanded computational time exceeds acceptable limit, the process was prematurely terminated after 1 hour. In this case, the standard approach did not reach the optimal solution, but only a near-to-optimal one. To estimate the possible difference between the obtained value of the maximal disutility perceived by the worst situated users and the optimal one, the associated lower bound LB on the optimal solution is presented.

The achieved results reported in Table 1 show that the bisection radial approach enables to overcome the weakness of the standard approach as regards the computational time and the min-max objective as well. In addition, the bisection approach yields exact min-max solution contrary to prematurely terminated standard approach. It must be noted that the min-sum objective is a secondary indicator, which does not enter the above comparison of min-max approaches. Therefore, the bisection approach was used in the following experiments aimed at the robust emergency service system design.

As far as the robust design is concerned, various scenarios for each solved instance must be defined. Due to the lack of common benchmarks for study of robustness, the scenarios used in our computational study were created in the following way. We chose 25 percent of matrix rows so that these rows correspond to the biggest cities concerning the number of users. Then we chose randomly from 5 to 15 rows and the associated disutility values in the individual rows were multiplied by the randomly chosen constant from the range 2, 3 and 4. This way, 20 different

Comparison of the standard and bisection approach to emergency service system design applied on basic scenarios Table 1

region	$\|I\| = \|J\|$	p	STANDARD MIN-MAX APPROACH					BISECTION APPROACH			
			minMax			composed		minMax		composed	
			CT	h	LB	CT	minSum	CT	h	CT	minSum
BA	87	9	3599.65	14	13.00	4.16	26229	0.05	14	0.02	26229
BB	515	52	3599.70	15	8.90	24.76	18285	0.37	13	0.25	21780
KE	460	46	3600.39	13	8.00	6.01	21982	0.53	12	0.25	24117
NR	350	35	3599.25	14	9.00	2.13	24432	0.25	13	0.56	26894
PO	664	67	3600.17	14	8.00	7.72	21241	1.08	12	0.27	24467
TN	276	28	3600.15	13	9.00	2.45	19993	0.15	12	0.05	23476
TT	249	25	3599.92	14	9.00	0.74	20227	0.20	13	0.20	21067
ZA	315	32	3599.33	14	9.76	1.81	24424	0.23	14	0.19	24424

scenarios were generated for each self-governing region. It must be noted that the robust design was computed also for the whole road network of Slovakia, but only 10 scenarios were generated in this case due to the size of disutility matrix.

The comparison of the basic and robust designs of emergency service system for all self-governing regions of Slovakia is reported in Table 2. Table 3 contains the results obtained for the self-governing region of Zilina, where different values of parameter p were considered. The parameter p limits the number of service centers to be located. Both tables follow the same notation as used in Table 1. As above, CT denotes the computational time in seconds. It must be noted that the reported value contains the computational time of both phases, i.e. min-max and min-sum optimization processes. The maximal disutility perceived by the worst situated users is denoted by h_b for the basic design and the symbol h_r is used for the robust design. The associated value of the min-sum objective function (8) is given in the columns denoted by mS_b and mS_r respectively.

An individual experiment was organized so that the basic design was computed first. This way, the values h_b and mS_b were obtained. The resulting design, especially the values of location variables y_i were applied on all generated scenarios and for each scenario the maximal disutility perceived by the worst situated users was computed. The worst (highest) maximal disutility value out of all scenarios is denoted by *wh*. The associated value of the min-sum objective function (8) is given in the column denoted by *wmS*.

After the basic approach was computed and applied on all scenarios, the robust problems were solved. Here, the computational time is several times bigger, because the models of the solved problems contain much more variables and structural constraints as well (all scenarios are taken into account simultaneously). The resulting vector of location variables y_i was substituted into the objective functions (primary and secondary) associated with the basic scenario and the maximal disutility perceived by the worst situated users as well as the associated min-sum objective function were computed. These values are denoted by h_r and mS_r respectively. Similarly to the basic approach, the worst (highest) maximal disutility value out of all scenarios is reported in the column denoted by *wh*. The associated value of the min-sum objective function (8) is given in the column *wmS*.

The basic and robust approaches to the emergency service system design were compared from two points of view. First, the vectors of location variables y_i were compared by the Hamming distance HD, which is defined as follows. Let $\mathbf{y}^b$ denote the vector

Results of numerical experiments comparing the basic and robust design of emergency service system for the self-governing regions of Slovakia

Table 2

region	$\|I\| = \|J\|$	p	BASIC DESIGN					ROBUST DESIGN							
			CT	h_b	mS_b	wh	*wmS*	CT	h_r	mS_r	wh	*wmS*	HD	PoR_1	PoR_2
BA	87	9	0.07	14	26229	41	41207	1.24	15	53567	15	53567	14	7.14	104.23
BB	515	52	0.62	13	21780	28	22640	17.69	14	26890	14	27867	68	7.69	23.46
KE	460	46	0.78	12	24117	26	27864	51.67	13	28391	13	30076	52	8.33	17.72
NR	350	35	0.81	13	26894	28	29497	14.29	14	31584	14	31994	50	7.69	17.44
PO	664	67	1.34	12	24467	29	24901	35.19	13	29137	13	30365	78	8.33	19.09
TN	276	28	0.20	12	23476	24	27496	6.85	14	23489	14	24000	40	16.67	0.06
TT	249	25	0.40	13	21067	23	23931	4.10	13	30332	13	30783	44	0.00	43.98
ZA	315	32	0.42	14	24424	29	30859	6.02	15	26517	15	27758	34	7.14	8.57
SR	2916	273	47.07	13	193715	22	194599	573.22	13	203988	13	204450	86	0.00	5.30

Results of numerical experiments comparing the basic and robust design of emergency service system for the self-governing region of Zilina and different numbers of located service centers

Table 3

$\|I\| = \|J\|$	p	BASIC DESIGN					ROBUST DESIGN							
		CT	h_b	mS_b	wh	*wmS*	CT	h_r	mS_r	wh	*wmS*	HD	PoR_1	PoR_2
315	158	0.10	4	3213	12	3446	4.12	4	4329	4	4389	50	0.00	34.73
315	105	0.21	6	6512	14	6882	4.28	6	10539	6	10842	58	0.00	61.84
315	79	0.21	7	12428	20	13637	4.41	8	15090	8	15339	92	14.29	21.42
315	63	0.35	8	20960	23	22996	5.05	10	15297	10	15873	74	25.00	27.02
315	32	0.44	14	24424	36	30113	5.56	15	29522	15	30666	42	7.14	20.87
315	21	0.44	16	37887	36	43540	7.13	18	44550	18	46190	34	12.50	17.59
315	16	0.50	20	46647	42	49636	8.67	21	53884	21	53884	26	5.00	15.51

of location variables for the basic design and let $\boldsymbol{y}^r$ denote the vector for the robust one. Then the Hamming distance HD takes the form of (28).

$$HD = \sum_{i \in I} \left(y_i^r - y_i^b \right)^2 \tag{28}$$

Hamming distance evaluates the structural difference between two designs in the sense that it informs of number of locations, in which the designs differ, but it does not refer to the quality of the designs. Therefore, we have compared the basic and robust design also from the viewpoint of other characteristics. Similarly to the price of fairness introduced and studied in [4 and 14] to evaluate the loss of min-sum objective function value caused by application of measures for fairness improvement, we introduce here so-called *price of robustness*. The *primary price of robustness* PoR_1 expresses the difference between the maximal disutility values perceived by the worst situated users in the solutions obtained by the basic and robust approaches applied on the basic scenario. The value of PoR_1 is given in percentage and it can be computed according to (29).

$$PoR_1 = 100 * \frac{h_r - h_b}{h_b} \tag{29}$$

The *secondary price of robustness* PoR_2 (30) is similar to the primary one, but it takes into account the min-sum objective function values mS_b and mS_r.

$$PoR_2 = 100 * \frac{mS_r - mS_b}{mS_b} \tag{30}$$

As concerns the comparison performed in Tables 2 and 3, it must be noted that the primal criterion in the designed emergency systems is the maximal disutility perceived by the worst situated user. It means that the primary objective of the robust system design should be resistant to changing scenarios as much as possible. Therefore, the values of *wh* in the basic and robust designs give the substantial information for the comparison. The value *wh* gives the highest maximal disutility out of all scenarios.

5. Conclusions

We have suggested and verified a useful tool for robust design of emergency service system. The suggested tool is able to comply with the problem which size is several times bigger than the size of the standard emergency system design problem. The bigger size of the robust problem is caused by cardinality of the set of studied scenarios. Good time performance of suggested approach follows from smart bisection process applied in the first phase of the suggested algorithm. Usage of the radial approach proved its usefulness especially in the second phase, where the min-sum problem is solved. As can be noticed, the computational time stays acceptable even if the size of the problem is several times swollen. To be able to answer the question what we have paid for the robustness of the designed system, we have introduced the price of robustness. The primary price of robustness expresses the relative difference between the maximal disutility values perceived by the worst situated users in the standard solution and the robust solution. The robustness measure has similar meaning as the price of fairness commonly used to evaluate the loss of min-sum objective function value caused by application of measures for fairness improvement. The price of robustness might help the designer to find the value paid for making the system resistant to catastrophic events.

The future research in this field may be aimed at finding relevant scenarios, which can significantly impact the performance of emergency service system. In connection with possible high price of robustness, it can be valuable to focus future research on development of such compromising approach, which can limit the price of robustness not to spoil the standard solution too much. From the practical point of view, we will try to find such method, which allows to change only limited number of service center locations in comparison with the standard solution. This algorithm may be useful when current service system is subjected to a reengineering process and stability of current service center deployment is required.

Acknowledgement

This work was supported by the research grants VEGA 1/0518/15 “Resilient rescue systems with uncertain accessibility of service”, VEGA 1/0463/16 “Economically efficient charging infrastructure deployment for electric vehicles in smart cities and communities”, APVV-15-0179 “Reliability of emergency systems on infrastructure with uncertain functionality of critical elements” and by the project University Science Park of the University of Žilina (ITMS: 26220220184) supported by the Research & Development Operational Program funded by the European Regional Development Fund. We would also like to thank to “Centre of excellence for systems and services of intelligent transport” (ITMS 26220120050) for built up the infrastructure, which was used.

Európska únia
Európsky fond regionálneho rozvoja

Operačný program VÝSKUM a VÝVOJ

Agentúra
Ministerstva školstva, vedy, výskumu a športu SR
pre štrukturálne fondy EÚ

References

[1] JANOSIKOVA, L.: Emergency Medical Service Planning. *Communications - Scientific Letters of the University of Zilina*, ISSN 1335-4205, vol. 9, No. 2, pp. 64-68, 2007.

[2] NASH, J.: The Bargaining Problem. *Econometrica*, vol. 18, No. 2, pp. 155-162, 1950.

[3] MARSH, M., SCHILLING, D.: Equity Measurement in Facility Location Analysis. *European J. of Operational Research*, 74, pp. 1-17, 1994.

[4] BERTSIMAS, D., FARIAS, V. F., TRICHAKIS, N.: *The Price of Fairness*. Oper. Res., 59, pp. 17-31, 2011.

[5] OGRYCZAK, W., SLIWINSKI, T.: *On Direct Methods for Lexicographic Min-Max Optimization*. In: Gavrilova M. et al. (Eds.): ICCSA 2006, LNCS 3982, (pp. 802-811). Berlin: Heidelberg: Springer, 2006.

[6] BUZNA, L., KOHANI, M., JANACEK, J.: Proportionally Fairer Public Service Systems Design. *Communications - Scientific Letters of the University of Zilina*, ISSN 1335-4205, vol. 15, No. 1, pp. 14-18, 2013.

[7] CEBECAUER, M., BUZNA, L.: Re-Aggregation Heuristics for the Large Location Problems with Lexicographic Minimax Objective. *Communications - Scientific Letters of the University of Zilina*, ISSN 1335-4205, vol. 17, No. 2, pp. 4-10, 2015.

[8] JANACEK, J., KVET, M.: *Emergency System Design with Temporarily Failing Centers.* SOR 15: Proc. of the 13th Intern. Symposium on Operational Research, Ljubljana: Slovenian Society Informatika: Section for Operational Research, ISBN 978-961-6165-45-7, pp. 490-495, 2015.

[9] PAN, Y., DU, Y., WEI, Z.: Reliable Facility System Design Subject to Edge Failures. *American J. of Operations Research*, 4, pp. 164-172, 2014.

[10] CORREIA, I., SALDANHA da GAMA, F.: Facility Locations under Uncertainty. *Location Science*, eds. Laporte: Nikel, Saldanha da Gama, pp. 177-203, 2015.

[11] GARCIA, S., LABBE, M., MARIN, A.: Solving Large p-median Problems with a Radius Formulation. *INFORMS - J. on Computing*, vol. 23, No. 4, pp. 546-556, 2011.

[12] JANACEK, J.: *Approximate Covering Models of Location Problems*. Lecture Notes in Management Science: Proc. of 1st Intern. Conference ICAOR ´08, vol. 1, Sept. 2008, Yerevan, pp. 53-61, 2008.

[13] JANACEK, J., KVET, M.: *Relevant Network Distances for Approximate Approach to the p-median Problem*. Operations Research Proceedings 2012: Selected Papers of the Intern. Conference of the German operations research society (GOR), September 2012, Leibniz: Univesitat Hannover : Springer, pp. 123-128, 2014.

[14] KVET, M., JANACEK, J.: *Price of Fairness in Public Service System Design*. Mathematical Methods in Economics 2014, Olomouc, ISBN 978-80-244-4209-9, pp. 554-559, September 2014.

Zuzana Borcinova - Stefan Pesko*

NEW EXACT ITERATIVE METHOD FOR THE CAPACITATED VEHICLE ROUTING PROBLEM

The aim of the Capacitated Vehicle Routing Problem is to find a set of minimum total cost routes for a fleet of capacitated vehicles, to serve a set of customers. This problem belongs into category of NP hard problems that can be exactly solved only for small instances. In this paper we propose an exact iterative method for the CVRP based on the solution of Mixed Linear Programming model, which is able to solve problem for relatively larger instances in acceptable computing time.

Keywords: *Capacitated vehicle routing problem, mixed linear programming model.*

1. Introduction and problem description

The Capacitated Vehicle Routing Problem (CVRP) is one of fundamental problems in combinatorial optimization with a number of practical applications in transportation, distribution and logistics. The aim of CVRP is to find a set of minimum total cost routes for a fleet of capacitated vehicles, based at one depot, to serve a set of customers under the following constraints:

(1) each route begins and ends at the depot,
(2) each customer is visited exactly once,
(3) the total demand of each route does not exceed the capacity of the vehicle [1].

Figure 1 shows an example of a feasible CVRP solution with seven customers and equal capacity $Q = 50$ of vehicles.

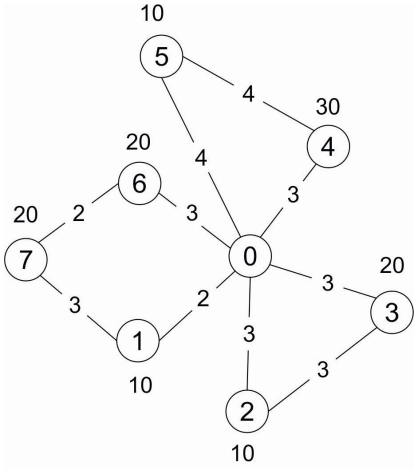

Fig. 1 A feasible CVRP solution, n = 7, Q = 50

2. Mathematical formulation

Let $G = (V, H)$ be a complete directed graph with $V = \{0, 1, 2,..., n\}$ as the set of nodes and $H = \{(i, j): i, j \in V, i \neq j\}$ as the set of arcs, where node 0 represents the depot for a fleet of identical vehicles of capacity Q and remaining n nodes represent geographically dispersed customers. Each customer $i \in V - \{0\}$ has a certain positive demand $d_i \leq Q$ and $d_0 = 0$. Non negative travel cost c_{ij} is associated with each arc $(i, j) \in H$. The cost matrix is symmetric, i.e. $c_{ij} = c_{ji}$ for all $i, j \in V, i \neq j$ and satisfies the triangular inequality, $c_{ij} + c_{jk} \leq c_{ik}$ for all $i, j, k \in V$ [2]. The minimum number of vehicles needed to serve all customers is

$$p = \left\lceil \sum_{i=1}^{n} d_i / Q \right\rceil.$$

Two-index decision variables x_{ij} are used as binary variables equal to 1 if arc (i, j) belongs to optimal solution and 0 otherwise. For all pairs of nodes $i, j, i \neq j$ we calculate the savings s_{ij} for joining the cycles $0 \rightarrow i \rightarrow 0$ and $0 \rightarrow j \rightarrow 0$ using arc (i, j):

$$s_{ij} = c_{i0} + c_{0j} - c_{ij}$$

as in Clarke and Wright's saving method [3]. The saving s_{ij} is illustrated in Fig. 2. In the left part, customers i and j are served by their own vehicle, in the right part, customers i and j are served by one vehicle.

Now, instead of minimizing the total cost, we can maximize the total saving. To ensure continuity of route and to elimininate subtours we define an auxiliary continuous variable y_i, $d_i \leq y_i \leq Q$ for $i \in [1, 2, ..., n]$, $y_0 = 0$, which shows (in the case of collection of the goods) the vehicle load after visiting customer i [4]. To make modeling easier, each feasible route $0 \rightarrow v_1 \rightarrow v_2 \rightarrow ... \rightarrow v_k \rightarrow 0$ we

* **Zuzana Borcinova, Stefan Pesko**
Department of Mathematical Methods and Operations Research, Faculty of Management Science and Informatics, University of Zilina, Slovakia
E-mail: zuzana.borcinova@fri.uniza.sk

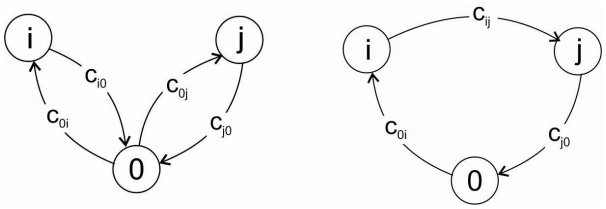

Fig. 2 Saving s_{ij} for customers i and j in cycle $0 \rightarrow i \rightarrow j \rightarrow 0$

replace by path from node 0 to node v_k, i.e. $0 \rightarrow v_1 \rightarrow v_2 \rightarrow ... \rightarrow v_k$.

For example, a feasible solution is illustrated in Fig. 3. Two values are assigned to each node i – the demand d_i of customer i and the load y_i of vehicle after visiting customer i (inside the brackets). The path $0 \rightarrow 2 \rightarrow 3$ represents the route $0 \rightarrow 2 \rightarrow 3 \rightarrow 0$, where customer 2 has demand $d_2 = 10$, customer 3 has demand $d_3 = 20$, the value of the vehicle load in node 2 is $y_2 = 10$ and in node 3 is $y_3 = 30$.

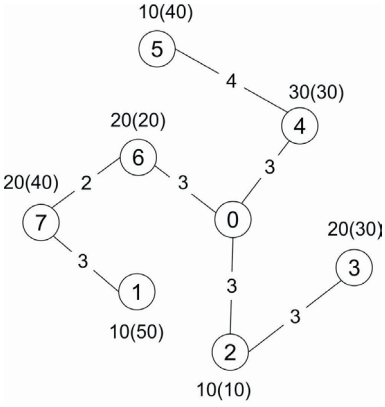

Fig. 3 n = 7, Q = 50

The CVRP can be formulated as following mixed linear programming model (*CVRP 1*):

$$\text{Maximize} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} s_{ij} x_{ij} \tag{1}$$

$$\text{Subject to} \sum_{j=1}^{n} x_{0j} = p \qquad \forall j \in \{1,2,\ldots,n\} \tag{2}$$

$$\sum_{i=0, i \neq j}^{n} x_{ij} = 1 \qquad \forall j \in \{1,2,\ldots,n\} \tag{3}$$

$$\sum_{i=0, i \neq j}^{n} x_{ij} \leq 1 \qquad \forall j \in \{1,2,\ldots,n\} \tag{4}$$

$$y_i + d_j x_{ij} - Q(1 - x_{ij}) \leq y_j \quad \forall j \in \{1,2,\ldots,n\}, i \neq j \tag{5}$$

$$d_i \leq y_i \leq Q \qquad \forall j \in \{1,2,\ldots,n\} \tag{6}$$

$$x_{ij} \in \{0,1\} \qquad \forall (i,j) \in H \tag{7}$$

In this formulation, the objective function (1) maximizes the total travel saving. The constraints (2) impose that exactly p vehicles leave the depot, (3) and (4) are indegree and outdegree constraints for customers. Constraints (5) are route continuity and subtour elimination constraints, ensuring that the solution contains no subtour disconnected from the depot and that the vehicle load is non decreasing step function in accordance with the demand of the customers which are on the route of the vehicle. Constraints given in (6) are capacity bounding constraints and restrict the upper and lower bounds of y_i.

The Capacitated Vehicle Routing Problem belongs into category of NP hard problems that can be exactly solved only for small instances of the problem. We propose an exact iterative method for the CVRP based on the solution of Mixed Linear Programming model, which is able to solve problem for relatively larger instances in acceptable computing time.

3. Exact iterative method

The essence of our method is the iterative improvement of initial feasible solution S by replacement of some of the arcs with the other ones, obtained by exact solution of the simpler problem.

Algorithm:

Step 1: Apply a fast heuristic method to find an initial (feasible) CVRP solution S_0. Set $S = S_0$.

Step 2: Create a list E of arcs which belong to solution S and are not incident with the depot. Let $m = |E|$

Step 3: Set values $k = 1$, $m_1 = m - \delta$ and $m_2 = m - 1$, where δ is a predetermined integer value $1 < \delta < m$.

Step 4: Find an optimal solution S_k, so that the set E_k of arcs which belong to solution S_k and are not incident with the depot contains minimally m_1 and maximally m_2 of arcs from E, i.e. we add to (*CVRP 1*) the constraints:

$$\sum_{(i,j) \in E}^{n} x_{ij} = z \tag{8}$$

$$m_1 \leq z \leq m_2 \tag{9}$$

where integer variable z determines how many arcs from E are retained in E_k.

Step 5: If the solution S_k is better than the solution S, then set $S = S_k$ and continue to *Step 2*.

Else set values $k = k + 1$, $m_1 = m_1 - \delta$ and $m_2 = m_2 - \delta$ and go to *Step 4*.

The algorithm can be stopped after a predetermined computational time or number of iterations without improvement.

4. Computational experiments

Our improved model was coded in Python 3.4 [5] and solved using Gurobi 6.5 [6] on PC with Intel Xeon 32 cores, 2.4 GHz, 256 GB RAM. We have done our experiments on 7 instances

taken from classical sets of CVRP benchmark from Augerat et al. (1995). The input data is available online at [7].

For finding the starting solution we applied the well known heuristic proposed by Clarke and Wright (1964). Each instance data was solved to optimality for various values of parameter δ. In Table 1 the corresponding computational times are shown.

Results of computational experiments with parameter δ Table 1

						time (sec.)		
Instance	n	Q	p	Start	Opt.	δ = 2	δ = 3	δ = 4
A-n32-k5	31	100	5	901	784	9	**6**	7
A-n34-k5	32	100	5	886	778	335	136	**50**
A-n36-k5	35	100	5	870	799	12 048	3 458	**1 847**
A-n38-k5	37	100	5	828	730	81	**76**	134
A-n44-k6	43	100	6	1100	937	718	748	**302**
A-n53-k7	52	100	7	1167	1010	**1 165**	5 608	18 134
A-n55-k9	54	100	9	1291	1073	37 884	**4 964**	165 540

n - number of the customers, *Q* - capacity of the vehicles, *p* - number of the vehicles, *Start* - cost of the initial solution, *Opt* - cost of the optimal solution from literature

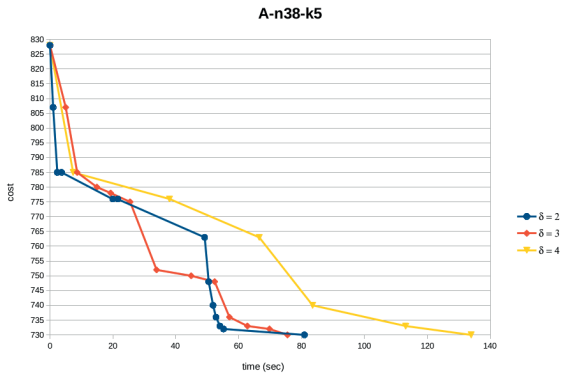

Fig. 4 Time evolution of solution improvement for various δ

Figure 4 illustrates time evolution of solution improvement for A-n38-k5 instance with 37 customers and 5 vehicles, from initial cost 828 to optimal cost 730.

5. Conclusions

We have proposed an exact iterative method for the CVRP based on the solution of Mixed Linear Programming model. From our computational experiments is evident that it is not possible to determine which value of parameter delta is the most appropriate for given instance. This led us to the idea of parallel computation with respect to the parameter delta. Our method seems to be promissing, because it is able to exactly solve problem for relatively large instance and we plan to develop it in the further research.

Acknowledgment

The research of Z. Borcinova was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences under project VEGA 1/0518/15 *"Resilient rescue systems with uncertain accessibility of service"*, and the research of S. Peško was supported by the Slovak Research and Development Agency under projects APVV-0760-11 *"Designing of fair service systems on Transportation networks"* and APVV-14-0658 *"Optimization of urban and regional public personal transport"*.

References

[1] LAPORTE, G.: What You Should Know about the Vehicle Routing Problem, *Naval Research Logistics*, 54: 811-819, 2007.

[2] TOTH, P., VIGO, D.: Models, Relaxations and Exact Approaches tor Capacitated Vehicle Routing Problem, *Discrete Applied Mathematics,* 123: 487-512, 2002.

[3] LAPORTE, G.: The Vehicle Routing Problem: An Overview of Exact dnd Approximate Algorithms, *European J. of Operational Research* 59: 345-358, 1992.

[4] KARA, I.: *Tightening Bounding Constraints of the Miller-Tucker-Zemlin Based Formulation of the Capacitated Vehicle Routing Problems and Some Extensions,* Proc. of the 2nd intern. conference on Manufacturing Engineering, Quality and Production Systems, edited by C. Panait et al., WSEAS Press, Constantza, 2010, pp.137-142.

[5] The Python Language Reference, 1990 - 2016, Python Software Foundation, http://docs.python.org/3/.

[6] Gurobi Optimizer Reference Manual, version 6.5, 2016, Gurobi Optimization, Inc., http://www.gurobi.com.

[7] Branch Cut and Price Resource Web, Vehicle Routing Data Sets USA, http://www.coin-or.org/SYMPHONY/branchandcut/VRP/data/.

Vojtech Graf - Dusan Teichmann - Michal Dorda*

MATHEMATICAL MODEL FOR CHARTER FLIGHTS PLANNING AT AIRPORTS WITH HIGH AIR TRAFFIC VOLUME

The article is devoted to a problem of flight timetable planning for charter airlines if air transport is operated at airports with high air traffic volume. It is typical for such airports that air transport is subject to time slots. That means aircraft can take-off or land only in given time intervals. A suitable tool for solving such type of the task is linear programming which has been successfully employed for planning in other transport systems. In the article a mathematical model which is able to plan a flight timetable for a given set of flights is presented.

Keywords: *Aircraft scheduling, linear programming, timetable planning.*

1. Introduction

To provide reliable and regular operation of air transport all round the world, it is necessary to plan flight timetables. The flight timetables are planned not only by airlines that provide regular air transport but also airlines providing charter air transport.

Planning flight timetables is a complex and time-consuming process. The process of flight timetable preparation is usually started a year before its year of validity. The whole process is influenced by many factors. The most important factors are demands of travel agencies, operating times of airports and time slots for taking-off and landing. By means of the time slots (time intervals) airports dictate to airlines when it is possible to plan taking-off or landing at the airport. The problem of flight timetable planning can be considered to be a task of planning service of a given set of flights under different time constraints. Therefore, mathematical modelling can be successfully employed for solving the task.

The article continues in article [1]; in the article a mathematical model that enables to plan flight timetables if two time slots are available for each flight was presented. The original mathematical model was generalised as regards the number of available time slots for the flights; the improved mathematical model is described in the article. The improvement lies in the fact that for each flight any finite number of time slots can be given. Please note that the number of the available time slots can differ for the individual flights.

2. State of the art

Transport processes in air transport that have to be planned by airlines can be divided into several groups of elementary problems:

- Schedule design problems.
- Fleet assignment problems.
- Aircraft maintenance routing problems.
- Crew scheduling problems.
- Rostering problems.

Approaches that are used for solving the basic above mentioned problems in air transport can be divided into two basic groups – approaches that are used only for solving one of the above mentioned problems (single purpose approaches) and approaches that solve several basic problems at the same time (integrated approaches).

From the broad spectrum of publications devoted to the single purpose approaches we can mention, for example, publications [2, 3, 4 or 5].

Source [2] presents a model for daily operational flight planning. The total costs of an airline including the costs of passengers caused by flight delays are used as an optimisation criterion. To solve the model the authors proposed a special four-step heuristic.

Publication [3] is devoted to aircraft maintenance planning. The total unused legal flying time of the critical aircraft is an optimisation criterion. The publication presents a mathematical model that minimises the value of the optimisation criterion. To solve the model a metaheuristic named compressed annealing

* **Vojtech Graf, Dusan Teichmann, Michal Dorda**
Institute of Transport, Faculty of Mechanical Engineering, VSB - Technical University of Ostrava, Ostrava-Poruba, Czech Republic
E-mail: vojtech.graf@vsb.cz

is employed, the metaheuristic is based on simulated annealing metaheuristic.

The crew scheduling problem is discussed in publication [4]. The publication is based, in contrast to other publications employing mathematical programming methods, on an unconventional approach that uses artificial neural networks.

The last publication devoted to the single purpose approaches we would like to mention is work [5]. The publication presents a model of the crew scheduling problem. The publication does not apply the model in air transport but the results can be also used for air transport. To solve the model an approach based on column generation technique is applied.

The integrated approaches include models and algorithms for solving different combinations of the above mentioned elementary problems. We can mention, for example, publications [6, 7 and 8].

An approach published in [6] is the simplest approach; the approach is sequential. It is based on sequential optimisation of the individual problems in a pre-defined order. The individual problems are solved in the order as given above in the section.

Publication [7] includes an integrated approach to the flight scheduling problem together with maintenance planning and crew scheduling within a period of a week. To solve the task a heuristic named cycle checking strategy is employed. The applied strategy splits flight plans in subsets and the subsets are matched one another. Maintenance limitations are checked when splitting the flight plans.

Publication [8] discusses an integrated approach to aircraft routing and crew scheduling. To solve the problem a three-step algorithm based on Benders' decomposition, column generation technique and dynamic constraint generation procedure is employed.

3. Time slots in air transport

The article deals with a modification of the mathematical model which enables to plan flight timetables in situations for which it is typical that several time slots for each flight are given. Firstly, let us discuss the time slots in air transport.

In general, time slots are defined as time intervals in which certain tasks connected with flights (taking-off or landing) or some aircraft services have to be carried out. The fact whether the airline has to make a request for the time slot or not depends on the airport where the task should be done. From the point of view of assigning the time slots, all the airports can be divided into uncoordinated, partially coordinated and fully coordinated airports. For the uncoordinated airports it is typical that the time slots are not applied for the tasks. The time slots are used for the partially or fully coordinated airports. For such airports high air traffic volumes are typical. The partially coordinated airports can apply the time slots only for some days or seasonally. The fully coordinated airports assign the time slots to the airlines during the whole year; that means each airline needs a free time slot for taking-off or landing at the airport.

Each airport can employ several types of the time slots. We can have the time slots for arrivals, departures, using passenger boarding steps or ramps, refuelling and so on. In the mathematical model presented in the article it is not necessary to consider such dividing the time slots. We can assume that each time slot used in the model corresponds to the intersection of all the above mentioned categories of the time slots. That means during the time slot the plane can land or take-off, the passengers can disembark from the plane or board the plane and the plane can be refuelled.

From the point of view of charter airlines that plan flight timetables for the whole season in advance, it is decisive that some airports (especially during the peak season or in seaside resorts) may be loaded by higher air traffic volumes. Therefore, the airline has to take into account the fact that the airport permits arrivals and departures only in the assigned time slots. Because the numbers of the time slots and their lengths may be different for the individual flights, it is necessary to incorporate the fact into the mathematical model.

Demands for the time slots are submitted to the airports in advance. The time slots at the airports are assigned to the airlines by an airport slot coordinator. After evaluations of all the demands submitted before the season the airport slot coordinator creates a preliminary flight plan to coordinate the individual time slots one another and assigns the time slots to each applicant (the charter airlines). That means the charter airline already knows its assigned time slots when planning the flight timetable and, therefore, has to use the time slots in order not to be sanctioned [9].

4. Problem formulation and mathematical model

Let a set I of flights be given, the set contains all the flights that have to be dispatched. For each flight $i \in I$ a set J_i is given; the set contains all the possible time slots. The time slots define time periods in which the flight $i \in I$ can depart (that means the flight has to be dispatched in one of the pre-defined time slots). Each time slot is defined by two values – a value d_{ij} is the lower bound of the time slot $j \in J_i$ assigned to the flight $i \in I$ and a value h_{ij} is the upper bound of the time slot $j \in J_i$ of the flight $i \in I$.

For each flight $i \in I$ its flight time T_i is given. It corresponds to the time between taking-off and landing. A so-called preparation time is included in the model. The preparation time is a time that is needed for carrying out all tasks after landing the flight and for preparing a consecutive flight $j \in I$. In order to simplify it, the preparation time consists of three sub-times. The first sub-time τ_1 includes stopping the aircraft, placing wheel chocks, putting passenger boarding steps or ramps, connecting the

aircraft to a ground power unit (GPU), getting off the passengers and unloading the baggage. The second sub-time depends on services which are demanded by the aircraft's operator and on the flight time of the consecutive flight. The dependence is expressed by a product $T_j \cdot k$, where T_j is the flight time of the flight $j \in I$ and k is a relative constant. The relative constant expresses the ratio of the flight time T_j of the consecutive flight and the time that is needed for preparation the flight $j \in I$. The second sub-time contains cargo unloading, cleaning the aircraft, catering replenishment, cabin service (water replenishment, toilette emptying and so on), refuelling, a technical inspection and cargo loading. The third sub-time τ_2 includes loading passengers' baggage, boarding the passengers, removing passenger boarding steps or ramps, disconnecting the ground power unit and taxiing the aircraft from a terminal to a runway.

In order to model required decisions and to create necessary logical links 3 groups of variables t_i, x_{ij} and z_{ij} are defined in the model. The variables t_i model the departure times of the individual flights $i \in I$; the variables t_i can take non-negative real values. The variables x_{ij} model decisions about links between the flights $i \in I$ and $j \in I$. If $x_{ij} = 0$, then the aircraft is not assigned to the flight $j \in I$ after serving the flight $i \in I$. If $x_{ij} = 1$, then the aircraft is assigned to the flight $j \in I$ after serving the flight $i \in I$. That means both flights are served by the same aircraft in the order *i→j*. Please note that if $x_{0j} = 1$ then a new aircraft is assigned to the flight $j \in I$. The last group of the variables z_{ij} gives information which time slot $j \in J_i$ is chosen for dispatching the flight $i \in I$ (that means the departure time of the flight $i \in I$ lies in the interval $\langle d_{ij}, h_{ij} \rangle$). If $z_{ij} = 0$, then the time slot $j \in J_i$ is not used for dispatching the flight $i \in I$. On the other hand, if $z_{ij} = 1$, then the departure time of the flight $i \in I$ lies in the time slot $j \in J_i$. P is a very large positive number.

Our goal is to assign the aircraft to the individual flights so that the number of the assigned aircraft is as minimal as possible. The mathematical model of the problem can be defined in the following form:

$$\min \sum_{j \in I} x_{0j} \tag{1}$$

$$\sum_{i \in I \cup \{0\}} x_{ij} = 1 \qquad j \in I \tag{2}$$

$$\sum_{j \in J}^{n} x_{ij} \leq 1 \qquad i \in I \tag{3}$$

$$t_j - \left(t_i + T_i + \tau_1 + T_j \cdot k + \tau_2\right) \geq P \cdot \left(xi_j - 1\right) \qquad i \in I, j \in I \tag{4}$$

$$\sum_{j \in J_i} d_{ij} \cdot z_{ij} \leq t_i \qquad i \in I \tag{5}$$

$$\sum_{j \in J_i} h_{ij} \cdot z_{ij} \geq t_i \qquad i \in I \tag{6}$$

$$\sum_{j \in J_i} z_{ij} = 1 \qquad i \in I \tag{7}$$

$$x_{ij} \in \{0,1\} \qquad i \in I \cup \{0\}, j \in I \tag{8}$$

$$z_{ij} \in \{0,1\} \qquad i \in I, j \in J_i \tag{9}$$

$$t_i \geq 0 \qquad i \in I \tag{10}$$

Formula (1) corresponds to the optimisation criterion of the mathematical model. As written earlier in the text, we try to minimise the number of the aircraft we need to serve all the planned flights. Constraints (2) ensure that each flight $j \in I$ has to be dispatched. Constraints (3) model that only one of two possible tasks is assigned to the aircraft after serving the flight $i \in I$ - the aircraft can be assigned to the consecutive flight $j \in I$ or is idle. Constraints (4) ensure that if a consecutive flight is assigned to the aircraft, the assignment is admissible in terms of time. The term on the left side of constraint (4) equals to the departure time of the consecutive flight $j \in I$ minus the sum of the departure time of the preceding flight $i \in I$ (t_i), the flight time of the flight $i \in I$ (T_i) and the preparation time before serving the flight $j \in I$ $(\tau_1 + T_j \cdot k + \tau_2)$. Constraints (5) and (6) ensure the admissible departure times of the flights $i \in I$. That means the departure times t_i have to lie in the pre-defined time slots $\langle d_{ij}, h_{ij} \rangle$. Constraints (7) assure that only single time slot $j \in J_i$ is chosen for dispatching the flight $i \in I$. Constraints (8), (9) and (10) define the domains of definition of the variables used in the model. In general, the total number of the variables of the model is equal to $m^2 n + m^2 + 2m$, where m is the number of the planned flights and n is the number of the time slots for the flight with the maximum number of the pre-defined time slots. The number of the constraints in the model is $m^2 n + 2m^2 + 7m$.

5. Experiments

Calculation experiments were carried out on model examples. The experiments were aimed at finding a dependence of the calculation time on the number of the planned flights and the length of the pre-defined time slots. In addition, we also tried to find out solvability limits of the mathematical model – to estimate for how many flights the model is able to find an optimal solution within a given time limit. In total 39 optimisation calculations were carried out with the model. All the experiments were run on a student (demo) version of optimisation software Express-IVE [10]. To run the experiments we employed a personal computer with processor AMD-8300 Eight-Core 3.3 GHz and 8GB RAM.

For each flight $i \in I$ its flight time T_i, the length of the time slots and their bounds d_{ij}, h_{ij} are known. A summary of all the values is provided in Table 1. The length of the time slots was not constant during the experiments. For the first group of the experiments the length was 5 minutes (see Table 1), for the

Input data

Table 1

Flight	Flight time	d_{ij}					h_{ij}				
		1	2	3	4	5	1	2	3	4	5
1	100	0	600	1000	-	-	5	605	1005	-	-
2	150	400	1150	-	-	-	405	1155	-	-	-
3	180	0	300	700	950	-	5	305	705	955	-
4	120	0	350	800	900	1300	5	355	805	905	1305
5	110	450	1200	-	-	-	455	1205	-	-	-
6	200	200	550	650	-	-	205	555	655	-	-
7	180	600	-	-	-	-	605	-	-	-	-
8	100	150	340	900	-	-	155	345	905	0	0
9	130	500	1100	-	-	-	505	1105	-	-	-
10	160	1300	-	-	-	-	1305	-	-	-	-
11	90	0	250	350	450	650	5	255	355	455	655
12	195	180	620	850	-	-	185	625	855	-	-
13	120	860	1050	1200	1350	-	865	1055	1205	1355	-
14	140	1260	1320	1400	-	-	1265	1325	1405	-	-
15	115	50	230	300	450	520	55	235	305	455	525
16	175	90	470	830	1370	-	95	475	835	1375	-
17	95	490	-	-	-	-	495	-	-	-	-
18	135	360	480	1230	-	-	365	485	1235	-	-

second group the length was 10 minutes and for the third group 15 minutes. Please note that all the values are expressed in minutes with beginning at midnight. That means, for example, 6:30 is expressed as 390 minutes.

For each group of the experiments 13 optimisation calculations were carried out; the calculations differed in the number of the planned flights – each group of the experiments was started with 6 flights and ended with 18 flights. For all the optimisation calculations we observed the calculation time.

5.1 Results of the experiments

Within the experimental part 39 optimisation calculations were carried out. Table 2 presents results of the experiments for which the length of the time slots was 5 minutes. Table 3 summarises results for 10-minute time slots and Table 4 for 15-minute time slots. Each table consists of three columns. The first column gives information about the number of the flights that have to be served. The second column shows orders in which the planned flights should be served. And finally, the third column presents how many aircraft we need to serve all the planned flights. One can notice that in the tables results for 18 flights are missing. That is because for 18 flights the limitations of the student (demo) version of Xpress-IVE were exceeded; the limitations are given by the maximum number of constants and constraints.

Results of experiments for 5-minute time slots

Table 2

Number of flights	Order of flights	Number of aircraft
6	0-1-6 0-3-5-4-2	2
7	0-1-5-6-3-4 0-7-2	2
8	0-1-5-6-8-2 0-3-7-4	2
9	0-1-6-9-4-5 0-3-7-8-2	2
10	0-3-2-1-4-9-10 0-6-7-8-5	2
11	0-3-11-7-4-9-10 0-8-2-6-1-5	2
12	0-1-12-7-4-2 0-3-11-5-6-8-9-10	2
13	0-1-2-3-5 0-8-4-7-13-9 0-12-11-6-10	3
14	0-3-8-9-12-10 0-4-7-1-5 0-11-6-13-2-14	3

Results of experiments for 5-minute time slots Table 2

15	0-1-12-5-6-3-10 0-8-9-4-2 0-15-11-7-13-14	3
16	0-3-11-7-13-2-14 0-4-6-15-8-9-10 0-16-1-12-5	3
17	0-1-8-15-6-3-5 0-12-17-11-13-9-10 0-16-4-7-2-14	3

Results of experiments for 10-minute time slots Table 3

Number of flights	Order of flights	Number of aircraft
6	0-1-4-2 0-6-3-5	2
7	0-4-2-6-1-5 0-7-3	2
8	0-1-5-3-2 0-6-7-8-4	2
9	0-1-3-7-8-2 0-6-9-4-5	2
10	0-1-8-7-4-9-10 0-6-5-3-2	2
11	0-1-6-5-3-2 0-8-11-7-4-9-10	2
12	0-1-12-7-4-5 0-3-11-2-6-8-9-10	2
13	0-1-7-8-2 0-4-3-6-9 0-11-12-5-13-10	3
14	0-1-12-5-11-8-10 0-2-6-3-4 0-7-13-9-14	3
15	0-4-6-7-3-10 0-11-12-9-8-5 0-15-2-1-13-14	3
16	0-3-2-1-16-5-14 0-6-7-13-10 0-11-8-15-12-9-4	3
17	0-3-5-6-1-10 0-11-8-15-17-12-2-14 0-16-7-13-9-4	3

Results of experiments for 15-minute time slots Table 4

Number of flights	Order of flights	Number of aircraft
6	0-1-4-6 0-2-3-5	2
7	0-1-4-7-5 0-3-2-6	2
8	0-1-8-7-4-5 0-6-3-2	2
9	0-1-7-8-2 0-3-5-6-9-4	2
10	0-1-6-7-8-5 0-4-2-3-9-10	2
11	0-3-2-6-1-5 0-8-11-7-4-9-10	2
12	0-1-8-4-7-12-2 0-11-6-5-3-9-10	2
13	0-1-8-5-6-4-9-10 0-3-11-7-12-2-13	2
14	0-1-6-4-5-14 0-11-8-2-3-9-10 0-12-7-13	3
15	0-1-12-5-13-9-14 0-11-6-8-10 0-15-3-7-4-2	3
16	0-4-15-7-12-16 0-8-2-1-5-14 0-11-6-9-3-13-10	3
17	0-1-12-7-4-9-10 0-3-11-15-16-2-13 0-6-17-8-5-14	3

Fig. 1 The dependence of the calculation times on the number of the flights and the length of the time slots

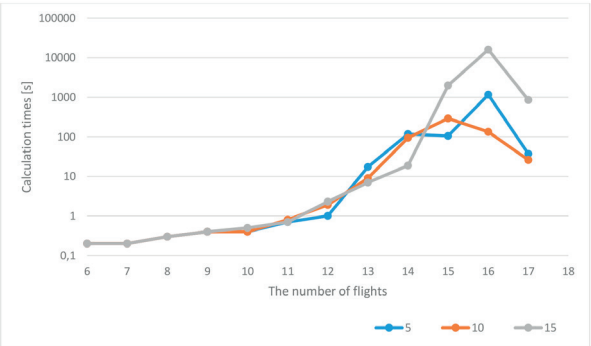

Table 5 and Fig. 1 summarise results of the experiments as regards the dependence of the calculation time (in seconds) on the number of the flights and the length of the time slots. Please note that the length of the time slots is expressed in minutes in Table 5 and the calculation times in seconds. Due to large range

Results of experiments – calculation times (in seconds) Table 5

Time slot length	Number of flights											
	6	7	8	9	10	11	12	13	14	15	16	17
5	0.2	0.2	0.3	0.4	0.4	0.7	1.0	17.3	118.6	105.1	1162.2	37.2
10	0.2	0.2	0.3	0.4	0.4	0.8	1.9	9.0	93.9	290.7	133.0	26.0
15	0.2	0.2	0.3	0.4	0.5	0.7	2.3	7.0	18.8	1972.8	15943.5	851.7

of the calculation times the logarithmical scale is used for the vertical axis of Fig. 1.

6. Conclusions

The presented article is devoted to planning flight timetables for charter airlines using mathematical modelling. In the paper the mathematical model that enables to plan flights in situations where due to capacity limitations at the airport the time slots for serving the flights are applied. The model is universal in terms of the number of the time slots. That means the individual flights can have different numbers of the possible time slots.

The calculation experiments carried out with model were focused on the calculation times needed to get an optimal the experiments the number of the planned flights and the length of the time slots were changed. The length of the time slots was assumed equal to 5, 10 and 15 minutes and the number of the flights was equal from 6 up to 17. For 18 flights the limitations of the academic (demo) version of Xpress-IVE were exceeded.

In our future research we would like to implement some other constraints following from real operation. We can mention, for example, a more detailed analysis of times needed for doing the tasks connected with aircraft services or planning maintenance of the aircraft. Our next target is to test the model in full version of optimisation software Xpress-IVE in order to find out calculation limits for the number of the planned flights.

References

[1] GRAF, V., TEICHMANN, D., DORDA, M.: *Mathematical Model of Charter Flight Scheduling with Predefined Time Slots.* Proc. of international conference Aplimat, February 2016, Bratislava, pp. 417-423, ISBN 978-80-4531-4.

[2] DOZIC, S., KALIC, M., BABIC, O.: Heuristic Approach to the Airline Schedule Disturbances Problem: Single Fleet Case. *Social and Behavioral Science,* vol. 54, 2012, pp. 1232-1241.

[3] BASDERE, M., BILGE, U.: Operational Aircraft Maintenance Routing Problem with Remaining Time Consideration. *European J. of Operational Research,* vol. 235, 2014, pp. 315-328.

[4] LAGERHOLM, M., PETERSON, C., SODERBERG, B.: Airline Crew Scheduling Using Potts Mean Field Techniques. *European J. of Operational Research,* vol. 120, 2000, pp. 81-96.

[5] JANACEK, J., MARTON, P., KONIORCZYK, M.: The Column Generation and Train Crew Scheduling. *Communications - Science Letters of the University of Zilina,* vol. 17, No. 1a, 2016, pp. 21-27.

[6] BARNHART, C., BELOBABA, P., ODONI, A. R.: Applications of Operations Research in the Air Transport Industry. *Transportation Science,* vol. 37, No. 4, 2003, pp. 368-391.

[7] DIAZ-RAMIREZ, J., HUERTAS, J. I., TRIGOS, F.: Aircraft Maintenance, Routing and Crew Scheduling Planning for Airlines with a Single Fleet and a Single Maintenance and Crew Base. *Computers & Industrial Engineering,* vol. 75, 2014, pp. 68-78.

[8] MERCIER, A., SOUMIS, F.: An Integrated Aircraft Routing, Crew Scheduling and Flight Retiming Model. *Computers & Operations Research,* vol. 34, 2007, p. 2251-2265.

[9] KERNER, L., KULCAK, L., SYKORA, V.: *Operational Aspects of Airports* (in Czech), Praha : CVUT, 2003, 270 p., ISBN 80-01-02841-0.

[10] FAIR ISAAC CORPORATIN, [online], ©2014 [11-2-2016]. http://www.fico.com/en/products/fico-xpress-optimization-suite/.

Dalibor Gottwald - Libor Svadlenka - Hana Pavlisova*

CUSTOMER CAPITAL AS A KEY FACTOR OF E-COMMERCE MARKET DEVELOPMENT

The importance of customer satisfaction has always been crucial to development of every single market in general. E-commerce market has gained importance during the last few years – especially in terms of annual turnover of e-commerce market. This trend is the result of changes in customer behavior. Increasing customer loyalty is one of the keys to business success. Active customer policy is necessary for the development of the market in individual states. The central point of the issue is to find concrete social constructs which can explain customer behavior in recent years. The main objective of the research is to define the area of selected variables which can be regarded as crucial in the process of predicting development of e-commerce market. By using Theory of Planned Behavior (TPB), the dependence of development of e-commerce market on customer capital has been confirmed.

Keywords: *Communication, customer capital, e-commerce, B2C market.*

1. Introduction

The effort of each company and of any market is to guarantee the highest possible level of customer satisfaction [1]. The issue of consumer behavior in relation to possible prediction of the behavior of market development is dealt with by the means of intellectual capital and customer capital in particular, as it is a part of the intellectual capital [2 and 3].

E-commerce market has gained importance during the last few years, especially in terms of annual turnover of e-commerce market of individual states. This trend is important especially when taking into account the fact that the transition to this new way of doing business is associated with a wide range of activities [4, 5 and 6]. Within these activities, there exists a considerable scope for reduction of individual cost elements [7, 8, 9 and 10]. E-commerce market can be divided in terms of direction of the transaction – individual subjects stand on the side of supply or demand [11, 12 and 13]. Timeliness of this research is evident from the number of studies dealing with the analysis of e-commerce market trends on global basis. Dieke et al. from the company WIK Consult GmbH have dealt with the issues of designing and development of initiatives to promote the growth of e-commerce market via better functioning parcel delivery systems [14]. Burkl from the Gfk company addresses in his study the current state of the consumer surroundings in selected countries of European Union (EU) [15].

The main objective of this article is associated with the above defined topics. It is to define the area of selected variables which can be regarded as crucial in the process of predicting development of e-commerce market. It means to carry out analysis of indicators of development of turnover B2C (business to customer) e-commerce market from selected member states of the EU in relation to dependent variables which define the social constructs that may affect customer capital of e-commerce market. The general model called Theory of Planned Behavior (TPB) will be used for verification of the defined area. At present, this model is not used for research of components of intellectual capital.

2. Material and methods

TPB is often used as a theoretical basis for prediction of human (social) behavior. TPB is based on the Theory of reasoned action [16 and 17]. Both theories are based on the assumption that individual's intention to perform a specific act of behavior is the result of the determinants of that behavior. The intent is understood as the "expression of the difficulty of performing a specific act of behavior" or the "expression of the effort which

* **[1]Dalibor Gottwald, [1]Libor Svadlenka, [2]Hana Pavlisova**
[1]Department of Transport Management, Marketing and Logistics, Jan Perner Transport Faculty, University of Pardubice, Czech Republic,
[2]Department of English and American Studies, Faculty of Arts, Palacky University Olomouc, Czech Republic
E-mail: libor.svadlenka@upce.cz

individuals plan to make in order to perform this act of behavior." For the general expression of TPB, it is necessary to define the following areas:

- attitudes towards behavior: the degree of individual's favorable or unfavorable evaluation of performance of a specific act of behavior;
- subjective norms determining the intent: the social pressure perceived in regards to the performance of a specific act of behavior [18].

Over the past few years, the TPB and the Theory of reasoned action were applied in a number of research articles [19, 20, 21 and 22]. Their aim was to predict the future intentions and specific behaviors. Later on, the TPB theory was expanded by the construct of Perceived Behavioral Control, which can be understood as a general attitude of the individual or a society towards the investigated act of behavior [19 and 23]. Now the TPB theory is used as a general theoretical framework for prediction of behavior (Fig. 1).

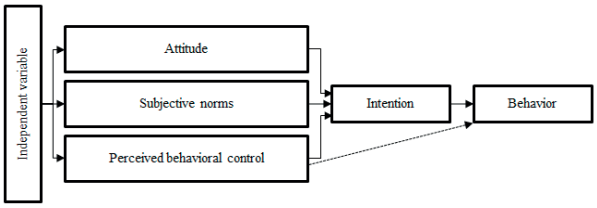

Fig. 1 General model Theory of Planned Behavior
Source: [18]

It is clear from the above mentioned model that behavior is determined by the intent, which is constituted by three components. To ensure the maximum relevance of predicted expected behavior, it is necessary to compare the current state of area under consideration with findings from the previous periods of time [24].

Nowadays, TPB is frequently used especially in social sciences, especially in research of predicting the behavior of individuals in terms of health straight behavior [25 and 26].Given the fact that TPB is used to predict the behavior of individuals, it will be necessary to take over a general methodology with an overlap to allow the prediction of the behavior of e-commerce market in selected member states of the EU.

Attitude: Variables which can be considered adequate in regards to compliance with the general TPB model are "Households with internet access at home – percentage of households with at least one member aged 16 to 74" and "Last internet use: within last 12 months". Using these indicators, it is possible to gain information about society-wide attitudes towards online shopping in individual states. The Internet is an infrastructure of online shopping and the access to Internet can be seen as society's attitude towards a specific act of behavior (online shopping).

Subjective norms: "Annual Net Earnings: single person without children, 50 % of average wage" variable can be understood as a subjective norm for the possibility of online shopping (the higher the society's earnings, the bigger the assumed total turnover of e-commerce market). Also the "Enterprises selling via Internet: all enterprises without financial sector (10 persons employed or more)" variable is relevant, this time for its relation

B2C turnover (bn EUR) Table 1

	2011	2012	2013	2014
United Kingdom	84.107	96.193	110.890	127.190
France	37.700	45.000	51.100	56.800
The Netherlands	10.927	11.869	12.877	13.961
Ireland	3.040	3.800	4.600	5.300
Belgium	2.200	3.040	3.820	4.368
Germany	41.085	50.000	63.400	71.200
Austria	8.405	9.800	10.970	11.685
Poland	3.349	4.183	5.225	6.541
Czech Republic	1.505	1.800	2.160	2.874
Spain	10.916	12.969	14.414	16.900
Italy	8.078	9.565	11.268	13.278
Greece	1.800	2.560	3.200	3.850
Portugal	2.000	2.300	2.600	2.945
Sweden	6.418	7.221	8.622	9.938
Denmark	6.172	7.339	8.367	9.886
Finland	6.005	6.137	6.500	7.290

Source: [27].

to the number of such companies (the bigger the number of such companies, the bigger the assumed total turnover of e-commerce market). These defined variables comply with the assumptions of the general model, i.e. the possibility, and not the necessity, of performance of a specific act of behavior (online shopping).

Perceived behavioral control: The "Last online purchase in the last three months (percentage of individuals)" variable is a factor informing about the attitude of an individual towards the area under consideration and can also be understood as a factor which directly influences the resulting behavior. It is, however, essential to examine its influence in combination with factors from previous areas.

Due to the nature of the research, one independent variable was determined as a meaningful variable for determining the behavior of e-commerce market, namely B2C e-commerce turnover in 2011, 2012, 2013, and 2014 (Table 1 and Fig. 2).

Data presented in Table 1 reflect the results of research into total turnover of B2C market in years 2011, 2012, 2013, 2014, which were taken from various sources:

- National associations and International associations: Netcomm (Italy), FDIH (Denmark), KAUPPA (Finland), SDH (Sweden), Adigital (Spain), BeCommerce (Belgium), Händlerbund (Germany), FEVAD (France), Apec (Czech Republic), etc.
- Corporate sources: Deloitte, Forrester, Google, Innopay, Social Bakers, I-Research, GfK, Planet Retail, etc.
- Other sources: Digital Hub Development Agency (DHDA), Eurostat, European Central Bank (ECB), National Statistic offices, World Economic Forum, etc.

EU member states which had turnover less than 1 billion EUR were excluded from the research, as well as the European states which are not members of the EU. Defined dependent variables were used to demonstrate their dependence on the independent variable in years 2011-2014. This was done to ensure the maximum relevance of the intention, which is the basis for the expected behavior prediction.

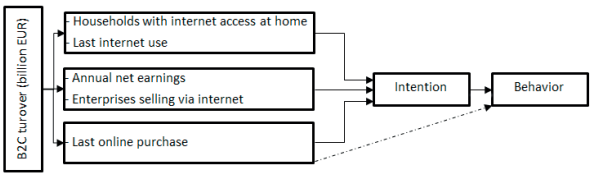

Fig. 2 Modified model Theory of Planned Behavior
Source: authors

Pearson's correlation coefficient was used to find the correlation between the independent variable determining the behavior of e-commerce of the market and the defined dependent variables, which are guaranteed by the respective indicators in the three basic categories of TPB model.

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}} \quad (1)$$

The final determination of significance of correlation was established on the basis of calculation of P value with significance level at 5 %.

3. Results and discussion

The indirect measurement method was used in the research of three components of TPB, i.e. the stated results reflect values taken from studies of international non-governmental organization OECD and the Directorate-General of the European

Measurement of Attitudes — Table 2

ATTITUDE	Households with internet access	Last internet use
Country	P value	P value
	Pearson's (r)	Pearson's (r)
United Kingdom	0.0514	0.0205
	0.9486	0.9795
France	0.0261	0.0120
	0.9739	0.9880
The Netherlands	0.0473	0.0662
	0.9527	0.9338
Ireland	0.1069	0.0103
	0.8931	0.9897
Belgium	0.0477	0.3173
	0.9523	0.6827
Germany	0.0043	0.0126
	0.9957	0.9874
Austria	0.041	0.0212
	0.9599	0.9788
Poland	0.0076	0.1695
	0.9924	0.8305
Czech Republic	0.0730	0.0096
	0.9270	0.9904
Spain	0.0017	0.0013
	0.9983	0.9987
Italy	0.0410	0.0106
	0.9590	0.9894
Greece	0.5810	0.0083
	0.9419	0.9917
Portugal	0.0142	0.0739
	0.9858	0.9261
Sweden	0.7387	0.6542
	-0.2613	-0.3458
Denmark	0.1166	0.0250
	0.8834	0.9750
Finland	0.1512	0.1253
	0.8488	0.8747

Source: authors

Commission - Eurostat between years 2011 and 2014 (Tables 2, 3 and 4).

Measurement of Subjective norms Table 3

SUBJECTIVE NORMS	Annual net earnings	Enterprises selling via internet
Country	P value	P value
	Pearson's (r)	Pearson's (r)
United Kingdom	0.1963	0.1158
	0.8037	0.8842
France	0.0212	0.0489
	0.9788	0.9511
The Netherlands	0.0038	0.8372
	0.9962	0.1628
Ireland	0.0445	0.2406
	0.9555	0.7594
Belgium	0.0069	0.2997
	0.9931	-0.7003
Germany	0.0229	0.1386
	0.9771	0.8614
Austria	0.0183	0.6700
	0.9817	0.9330
Poland	0.0615	0.0542
	0.9385	0.9458
Czech Republic	0.1050	0.2574
	-0.8950	0.7426
Spain	0.0237	0.0436
	0.9763	0.9564
Italy	0.0630	0.0698
	0.9370	0.9302
Greece	0.0189	0.2406
	-0.9811	0.7594
Portugal	0.1439	0.2407
	0.8561	-0.7593
Sweden	0.2165	0.6444
	0.7835	-0.3556
Denmark	0.0388	0.9448
	0.9612	0.05521
Finland	0.2131	0.1759
	0.7869	-0.8241

Source: authors

It was found in the research that in the area of "Attitude" (Table 2), the dependence of defined variables on the independent variable (B2C turnover) has been confirmed as statistically significant in both variables ("Households with Internet access" and "Last Internet use") in France, Germany, Austria, Spain, and Italy. Defined dependence has then shown to be significant in at least one of the variables in the United Kingdom, The Netherlands, Ireland, Belgium, Poland, Czech Republic, Greece, Portugal, and Denmark. Results of the statistical testing of the area of "Subjective norms" in Table 3 confirm the dependence of both variables ("Annual net earnings" and "Enterprises selling via Internet") on the independent variable in France and Spain. Defined dependence has then shown as significant in at least one of the variables in The Netherlands, Ireland, Belgium, Germany, Austria, and Denmark.

From the above tables it is clear that Sweden and Finland (both states are insular states in the north of Europe) do not confirm this dependence in either variable in both areas, which could be used as input data for further research.

Measurement of Perceived behavioral control Table 4

PERCEIVED BEHAVIORAL CONTROL	Last online purchase	
Country	P value	Significance of correlation? (alfa = 0.05)
	Pearson's (r)	
United Kingdom	0.0748	NO
	0.9252	
France	0.0436	YES
	0.9564	
The Netherlands	0.1102	NO
	0.8898	
Ireland	0.0797	NO
	0.9203	
Belgium	0.0431	YES
	0.9569	
Germany	0.0205	YES
	0.9795	
Austria	0.1070	NO
	0.8930	
Poland	0.0160	YES
	0.9840	
Czech Republic	0.0037	YES
	0.9963	
Spain	0.0134	YES
	0.9866	
Italy	0.0272	YES
	0.9728	
Greece	0.0135	YES
	0.9865	
Portugal	0.0087	YES
	0.9913	
Sweden	0.1214	NO
	0.8786	
Denmark	0.0478	YES
	0.9522	
Finland	0.0141	YES
	0.9859	

The research of statistical dependence of the behavior of e-commerce market (B2C turnover) on the variable defining Perceived behavioral control has shown dependence in France, Belgium, Germany, Poland, Czech Republic, Spain, Italy, Greece, Portugal, Denmark, and Finland.

Only one variable has been established as a dependent variable in Perceived behavioral control. It was due to insufficient data basis of other potential variables which could define this area. On the other hand, there is no requirement for minimal amount of dependent variables defining a given area in the application of TPB [18 and 19].

General result of application of TPB model can be expressed via the correlation of independent variable on the behavior of e-commerce market expressed by the B2C turnover indicator. In Tables 1 and 2, states which show positive correlation in both variables defining the given area of TPB are marked in yellow, while states that show positive correlation at least in one of these variables are marked in green.

Table 3 shows in yellow-green color the states that show positive correlation with the indicator defining behavior of e-commerce market.

Based on the results it can be stated that defined model TPB can be applied to the markets of France and Spain (where correlation was shown in all areas of research), and for Belgium, Germany and Denmark (where positive correlation was shown at least in one variable of the three researched areas of TPB).

The values found in other states can be considered partial results and cannot be applied to prediction of e-commerce market behavior through TPB model. With regards to the fact that the given theory has not been used for prediction of e-commerce market of the EU states so far, these results can be seen as innovative and can be used as a basis for further research of application of this theory on the given area of research.

4. Conclusions

The main objective of this article was to highlight the key variables which can be determined as crucial for the prediction of development of e-commerce market from the point of view of the customer capital. It also establishes selected variables which can be understood as relevant in the prediction. TPB is commonly used in social sciences to predict development of a specific phenomenon and in this research was used to define the area of research. TPB was used in this context to innovate the scientific approach towards prediction of e-commerce market behavior.

It can be concluded that TPB has proved to show results in some states. The results can be used in further research, especially in establishment of other variables applicable to the general model of TPB. Another topic for further research would be a closer look at results in a more general context of the currently turbulent e-commerce market.

The results have shown that the influence of customer capital in the selected group of states is statistically significant, and thus the prediction of development of e-commerce market must take into account selected social constructs (determined by specified variables) which influence the individual customer.

Acknowledgements

The article was supported by student grant - SGSDFJP_2015001.

References

[1] LUO, X., GRIFFITH, D. A., LIU, S. S., SHI, Y. Z.: *The Effects of Customer Relationships and Social Capital on Firm Performance: A Chinese Business Illustration, J. of International Marketing,* 12(4), 25-45, 2004, http://dx.doi.org/10.1509/jimk.12.4.25.53216.

[2] HORIBA, F.: *Managing Knowledge Workers - New Skills and Attitudes to Unlock the Intellectual Capital in your Organisation*, 1999, Toronto: John Wiley & Sons, ISBN 0471643181.

[3] SVEIBY, K. E.: *The New Organizational Wealth: Managing and Measuring Knowledge-based Assets,* 1997, Barrett-Kohler: San Francisco, ISBN 1-57675-014-0.

[4] ESHGHI, A., HAUGHTON D., TOPI, H.: Determinants of Customer Loyalty in the Wireless Telecommunications Industry. *Telecommunications Policy*, 31(2), 93-106, 2007, DOI10.1016/j.telpol.2006.12.005.

[5] MORGANTI, E., DABLANC, L., FORTIN, F.: Final Deliveries for Online Shopping: The Deployment of Pickup Point Networks in Urban and Suburban Areas, *Research in Transportation Business & Management,* July, 23-31, 2014, ISSN 2210-5395, DOI doi:10.1016/j.rtbm.2014.03.002.

[6] PRADHAN, R. P., ARVIN. M. B., NEVILLE, N., BELE, S. K.: Economic Growth and the Development of Telecommunications Infrastructure in the G-20 Countries: A Panel-Var Approach, *Telecommunications Policy,* 38(7), 634-649, 2014, doi:10.1016/j.telpol.2014. 03.001.

[7] FRANCISCO, M., WILLIAM, F., JAY, B.: *Information Technology and Sustained Competitive Advantage: A Resource-Based Analysis*. MIS Quarterly, 19(December), 487-505, 1995, ISSN 2162-9730, DOI 10.2307/249630.

[8] HAUCAP, J.: Editorial: *Consumer behavior and telecommunications policy*. *Telecommunications Policy*, 39(8), 625-626, 2015, DOI:10.1016/j.telpol.2015.07.013.

[9] LA PAZ, A., RAMAPRASAD, A., SYN, T., VASQUEZ, J.: *Editorial: An Ontology of E-Commerce - Mapping a Relevant Corpus of Knowledge, J. of Theoretical and Applied Electronic Commerce Research*, 10(2), 1-9, 2015, DOI 10.4067/S0718-18762015000200001.

[10] ZHAO, H., MORAD, B.: From e-commerce to Social Commerce: A Close Look at Design Features, *Electronic Commerce Research and Applications*, 12(4), 246-259, 2013, DOI 10.1016/j.elerap.2012.12.003.

[11] HAQUE, A., KHATIBI, A.: E-Shopping: Current Practices and Future Opportunities Towards Malaysian Customer Perspective. *J. of Social Sciences*, 1(1), 41-46, 2005, ISSN 1549-3652, DOI 10.3844/jssp.2005.41.46.

[12] HAQUE, A., SADEGHZADEH, J., KHATIBI, A.: Identifying Potentially Online Sales in Malaysia: A Study on Customer Relationships Online Shopping, *J. of Applied Business Research*, 22(4), 119-131, 2006, ISSN 0892-7626.

[13] VANHOOSE, D. D.: *E-Commerce Economics*, 2011, New York: USA, ISBN 978-0-415-77897-8.

[14] DIEKE, A., HILLEBRAND, A., JOYCE, C., NIEDERPRUM, A.: *Design and Development of Initiatives to Support the Growth of E-commerce via better Functioning Parcel Delivery Systems in Europe,* 2014, WIK-Consult: Bad Honnef, ISBN 978-92-79-35322-2, DOI 10.2780/29929.

[15] BURKL, R.: *Consumer Climate Significantly Improves Across Europe.* GfK SE, Nuremberg, Retrieved October 15, 2015, from https://www.gfk.com/news-and-events/press-room/press-releases/pages/consumer-climate-significantly-improves.aspx.

[16] FISHBEIN, M., AJZEN, I.: *Belief, Attitude, Intention and behavior: An Introduction to Theory and Research,* Reading, Massachusetts, Addison-Wesley, 1975, ISBN 0-201-02089-0.

[17] FISHBEIN, M., AJZEN, I.: Attitudes and Voting behaviour: An Application of the Theory of Reasoned Action, *Progress in Applied Social Psychology*, 1, 95-125, 1981, ISBN 0471279544.

[18] AJZEN, I.: The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes,* 50(2), 179-211, 1991, ISSN 0749-5978, DOI 10.1016/0749-5978(91)90020-T.

[19] MADDEN, T. J., ELLEN, P. S., AJZEN, I. A.: Comparison of the Theory of Planned behavior and the Malaysian Customer Perspective, *J. of Social Sciences*, 1(1), 41-46, 1992, ISSN 1549-3652, DOI 10.3844/jssp.2005.41.46.

[20] ORBELL, S., HODGKINS, S., SHEERAN, P.: Implementation Intentions and the Theory of Planned behavior. *Personality and Social Psychology Bulletin*, 23(9), 945-954, 1997, ISSN 01461672, DOI 10.1177/0146167297239004.

[21] RANDALL, D. M.: Why Students Take Elective Business Ethics Courses: Applying the Theory of Planned behavior, *J. of Business Ethics*, 13(5), 369-378, 1994, ISSN 0167-4544, DOI 10.1007/BF00871764.

[22] SHEPPARD, B. H., HARTWICK, J, WARSHAW, P. R.: The Theory of Reasoned Action: A Meta-Analysis of Past Research with Recommendations for Modifications and Future Research, *J. of Consumer Research,* 15(3), 325-43, 1988, ISSN 0093-5301, DOI 10.1086/209170.

[23] VALLERAND, R. J., DESHACES, P., CUERRIER, J., PELLETIER, L. G., MONGEAU, C.: Ajzen and Fishbein's Theory of Reasoned Action as Applied to Moral behavior: A Confirmatory Factor Analysis, *J. of Personality and Social Psychology*, 62(1), 98-109, 1992, DOI 10.1037/0022-3514.62.1.98.

[24] MONTANO, D. E., KASPRZYK, D.: The Theory of Reasoned Action and the Theory of Planned behavior, In K. Glanz, F. M. Lewis, & B. K. Rimer (Eds.), *Health behavior and health education: Theory, research and practice* (pp. 67-98), ISBN-13: 978-0787996147.

[25] FALAGAS, M. E., ZARKADOULIA, E.: Factors Associated with Suboptimal Compliance to Vaccinations in Children in Developed Countries: A Systematic Review. *Current Medical Research and Opinion*, 24(6), 1719-1741, 2008, ISSN 0300-7995, DOI 10.1185/03007990802085692.

[26] MILLS, E., JADAD, A. R., ROSS, C., WILSON, K.: Systematic Review of Qualitative Studies Exploring Parental Beliefs and Attitudes Toward Childhood Vaccination Identifies Common Barriers to Vaccination, *J. of Clinical Epidemiology*, 58(11), 1081-1088, 2005, ISSN 0895-4356, DOI: 10.1016/j.jclinepi.2005.09.002.

[27] ECOMMERCE EUROPE: *European B2C E-commerce Report 2015*, Retrieved October 18, 2015, from https://www.ecommerce-europe.eu/facts-figures/free-light-reports.

Zoran Cekerevac – Zdenek Dvorak – Ludmila Prigoda – Petar Cekerevac*

BIG VS SMALL DATA IN MICRO AND SMALL COMPANIES

This article deals with the challenges that are faced by micro and small enterprises using data. After naming main hypotheses of the research, the attention is paid to definitions of big and small data and to their comparison. In the paper, analyses of the connections between big data and Internet are given. The main attention is paid to the use of big data in micro and small enterprises through the analyses of needed skills, organizational changes and big data management including possible pros and cons. As a special item, the analysis of risks and implementation is given. The analyses are supported with a lot of research results and numerous literature sources. At the end of the article there are conclusions about directions how micro and small enterprises should use big data and how to invest rationally in big data.

Keywords: *Big data, small data, micro enterprises, small enterprises, cloud computing.*

1. Introduction

Today's globalized society is developing together with the development of information and communication technologies. Development of information technologies, especially in the fields of big data, cloud computing and data mining is very dynamic. This is well doccumented by publications of Asay, M. [1], Thomas, J. W. [2], Carter, K. B. [3] which defines technical and technological aspects of big data, cloud computing and datamining. These technologies are gradually used in real life and all institutions and companies try to implement these technologies.

Every company, independently of its size, tries to collect as much information as possible, not only in the field of its activities, counting that the possession of data is a capital that will pay off sooner or later. Possession of data represents significant potential. So one can conclude that, based on reliable data, manager is able to timely make right decisions. However, is it always possible? Many companies, although able to collect a number of relevant data, are not able to successfully exploit them. In their analyses, there remain a lot of unused data, not only from the group big data, but, even, data from the group small data.

In the very beginning of the collection of big data, the main beneficiaries were big companies, which used data primarily for the prevention of fraud, or for tracking and retail management. Big companies had to develop complex sophisticated methods and software for processing of big data. Nowadays, big data are available also to micro and small enterprises (next MSEs) and they also have to develop systems and procedures for their proper and effective exploitation. Due to their sizes and the limited area of interest, it is quite possible that MSEs find a new specific, and more convenient and efficient methods to benefit from the data collected. Computer resources are no longer a challenge, because MSEs can use cloud computing at an affordable price. The main challenge is now of the strategic and organizational nature. MSEs do not need to deal with creating of their own systems to handle big data. It is enough to properly choose open source software that will suit their current needs and be scalable enough to accommodate their future growing needs. Although it looks so simple, the challenge lies in the word „properly". The question is, whether the MSEs are able to „properly" choose and apply the software.

In this paper, after a brief comparison of small and big data, here will be discussed micro and small enterprises' abilities to use big data and the Cloud.

Microenterprises are categorized here in accordance with the Commission Recommendation 2003/361/EC [4] as: "A microenterprise is defined as an enterprise which employs fewer than 10 persons and whose annual turnover and/or annual balance sheet total does not exceed EUR 2 million". According to the same recommendation, "A small enterprise is defined as an enterprise which employs fewer than 50 persons and whose annual turnover and/or annual balance sheet total does not exceed EUR 10 million" [4].

* **[1]Zoran Cekerevac, [2]Zdenek Dvorak, [3]Ludmila Prigoda, [4]Petar Cekerevac**
[1]Faculty of Business and Industrial Management, "Union – Nikola Tesla" University in Belgrade, Serbia
[2]Faculty of Security Engineering, University in Zilina, Slovakia
[3]Faculty of Economics and Service, Maykop State Technological University, Maykop, Russia
[4]Hilltop Strategic Services, Belgrade, Serbia
E-mail: zoran@cekerevac.eu

In Serbia 95% of all enterprises are micro enterprises, with more than 270,000 private entrepreneurs and micro companies [5]. According to Forbes: "Small businesses are driving the U.S. job recovery"... "Small businesses have led our comeback from the downturn," Bagley writes. "For 15 straight quarters, small firms have contributed to employment growth – accounting for as much as 80 percent of job gains in any given quarter" [6]. For this reason, this issue is becoming more important.

2. Research question – hypotheses

The creation and organization of infrastructure largely depends on the type of activities of the company and may be substantially different from industry to industry. This is why it can be difficult to formulate general conclusions. In order to obtain more general and realistic picture, this paper will deal with the general principles of using small and big data.

The working hypothesis was that MSEs due to their limited resources are not able to effectively use big data for the benefit of their business.

$\mathbf{H_0}$ The null hypothesis of this analysis is: MSEs are not able to form their own infrastructure in order to be able to take advantage of big data in order to form their strategies of development and doing business.

$\mathbf{H_1}$ The alternative hypothesis is: There are MSEs that can form their infrastructure so that they are able to take advantage of big data in order to form their strategies of development and doing business.

In the research, there were applied methods of induction and generalization, as explained in [7].

3. Small and big data definition and comparison

From the definition of small and big data arise their essential differences, but also achievements that by their use can be achieved. According to Techopedia [8]: "Small data describes data use that relies on targeted data acquisition and data mining", and "Big data refers to a process that is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data."

Big data acquisition is a result of the desire of the company to gather as much information as possible about anything and everything, and then to process the data collected to come to the conclusion what its customers want, may want and when. A company wishes to get an opportunity to project needs and desires of its customers in the future and define its own strategy and action. A very demanding task in every respect: in collecting data, in resources for data storage and processing, and in the models and software for processing and presentation of results. This task can be compared with the weather forecast. As with weather forecasting, on some level, collecting of data must be stopped, because the further collection of data, below this level, brings enormous costs and practically no impact on the result obtained. In weather forecasting, there will not be taken into account, for example, the impact of heat given by individual buildings, or damping the wind by a single tree, or causing of air movement from a single man breathing. Consequently, when processing mass data, data with minor impacts will not be considered. Or, maybe, they will? The answer depends of the decision of one who defines the strategy. In contrast to this approach, companies may limit the collection to the specific data which will allow them, according to their beliefs, to make the right decisions with considerably less effort and cost. Big data provide a greater potential to drive business intelligence in key ways, but no one can guarantee that the results won't be similar to the results of weather forecasts. Because of the sudden changes in business environment, it is possible that the predictions won't be met.

Reviewing small data IBM concluded that small data are connected with [9]:

- low volumes,
- batch velocities,
- structured varieties

On the other hand, big data are characterized by [9]:

- into Petabyte volumes,
- real-time velocities,
- multistructured varieties.

Both, small and big data, must be first brought under control. There is no doubt, that using of small data has a longer history and is better developed. The use of big data is recent date task and tools for their processing are still in the phase of developing with implementing the best practices. According to [9] in using small data dominant data platforms are OLTP &EDW built on relational DBMS. For big data it is recommended to use big transactional, data warehouse systems. And it is a current state of development. In the near future it is to expect the massive use of emerging platforms: Apache Hadoop[1], NoSQL[2], stream computing[3], in-memory[4].

4. Big data and cloud computing

Each company has its own specifics in resources and in the organization of work. Therefore, it is impossible to find an ideal technological solution that would satisfy everyone, but from past experiences it is possible to clarify some guidance. For small data it is easy to obtain appropriate tool, but the technological ways of big data utilization can vary from company to company. Often,

[1]About Apache Hadoop see in [10]
[2]About NoSQL (originally referring to "non SQL" or "non-relational" database) see more in [11]
[3]About stream computing see in [12]
[4]About In-memory computing see in [13], [14], and [15]

standard software packages are not good enough and customized solutions are better to use to wholly utilize the data potentials.

When we talk about the cloud, we consider it as a whole, as the Internet. But, there are several deployment models of clouds. According to NIST [16], there are: private, community, public, and hybrid clouds. Each of them can be a good solution for MSEs. According to Microsoft, as shown in Fig. 1, in the year 2014 most of the companies ran dedicated servers (48%), but this percentage is expected to decrease to the end of 2016 with an increase of hosted private cloud users [17].

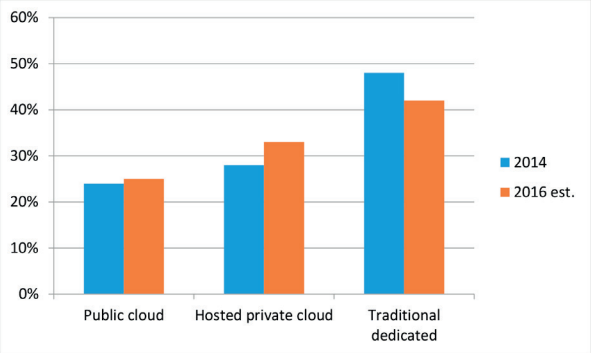

Fig. 1 Hosted infrastructure services of deployment models Source: [17]

A company needs to choose which of the service models it will use, and who will be a provider. As well as the other companies, MSEs can use three service models, three classes of services: Infrastructure (IaaS), Platforms (PaaS), and/or Software (SaaS). In the case of IaaS, beside the pure infrastructure, MSEs can expect plenty of common software that can be used for preparing of company's own databases, and software. Consumer "has control over operating systems, storage, and deployed applications; and possibly limited control of selected networking components (e.g., host firewalls)" [16, 18]. Using PaaS, MSEs have an environment where they can create and deploy desired applications without any concern about the hardware, but such environment can be usually too wide for their needs. "The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment" [19]. By using SaaS MSEs can have at their disposal different software, which can be accessed through Web portals. Instead of traditional desktop applications end user can use them as a service in the Cloud. Full control of the hardware and software is on the provider's side. User can in some cases change specific application configuration settings.

Beside the three service models mentioned above, MSEs can use outsourcing and collocation services, as well. Figure 2 shows the use of current IT services for companies with less than 100 Emps, based on the sample of n=480 micro and small companies. It can be seen that the use of SaaS and IaaS are dominant with more than 60% each. The use of other models is far less pronounced. According to the same research, website hosting is leader between services. 70% of participants confirmed that their companies pay for this service. The next in line of the most common services are storage and backup/restore services with 64% each.

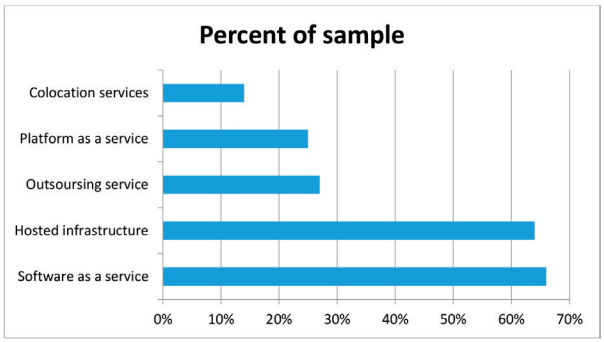

Fig. 2 Current IT Services for companies with less than 100 Emps (n=480) Source: [17]

Figure 3 shows the use of IT services by region. It is easy to see that SaaS and IaaS are the most used services in all continents, and that North America and Asia are leaders. Among all continents, outsourcing services are the most popular in South America.

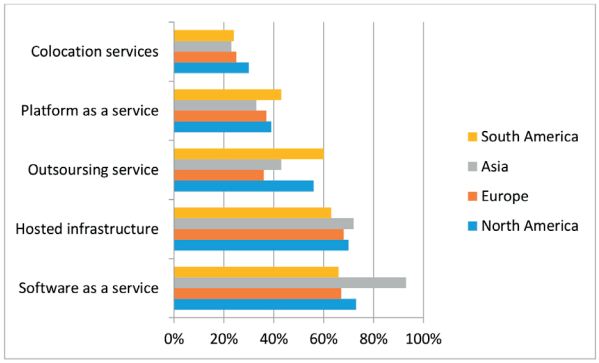

Fig. 3 Current IT services by region (n - the number of respondents in the poll) Source: [1]

According to Asay [10]: "Big data is all about the cloud". Cloud can offer:

- multiple ways to store big data from S3 storage to data warehousing solutions;
- real time analytics; and
- dedicated task clusters.

With all these components, Cloud gives to users an opportunity to choose right tool for their job. Using Cloud there

is no single choice "or". Clients have to their disposal the option "and". They can always choose appropriate tool for their current job [1, 20].

5. BIG DATA and MSEs

Good comparison between big and small data was given by Thomas [2]: „You are not feeling well, so you visit your friendly family doctor. He puts you in a new electronic scanner and generates 28 trillion measurements of your temperature all over the surface of your body. He then saves all of these measurements and, using advanced statistical algorithms and supercomputers, announces that your temperature is 98.6 degrees Fahrenheit. What a relief! Big Data to the rescue." From our experience, it is to expect that the very similar data can be obtained by classic measuring of temperature in one single point, e.g. under the armpit. It is obvious that the application of big data in this case is not practical. Is it the same and in doing business? Why do big data attract the companies, and where are the micro and small enterprises in such environment? The answers are not so simple. MSEs can also use basic premises of big data:

- more is always better;
- data by their quantity create new sources of knowledge and possibilities for predictions; and.
- by using big data, one can answer all questions.

According to Bagley [6] 75% of small businesses, in 2015-6, plan to invest in analytics of big data. Because of big interest of small businesses, even powerful software packages intend to become available at very moderate prices. At the same time they will be easy to use by any employee. It can give an opportunity to micro and small enterprises that by using advanced analytics techniques they can analyze previously untapped data sources. Using the analysis of big data independently or together with their existing data, MSEs will gain new views on their doing business which can consequently result in more appropriate decisions.

According to Microsoft research [17] more than one third of MSEs don't plan to add any service from service providers until the end of 2016. From the rest, 26% plan to invest in PaaS, 12% plan to invest in IaaS, 12% in colocation services, 16% in outsourcing services and 14% in SaaS. Percentage is given to the number of 480 MSEs. Total sum is bigger than 100% because some of MSEs respondents planned to invest in more than one service.

People transform resources into products or services. Data-savvy employees are critical for creating value from big data. Data experts are in huge demand and in low supply [21]. Enterprises must adapt to this quickly changing landscape to establish an analytical competitive advantage [22]. This raises several questions:

No. 1. How can MSEs act in order to obtain the necessary skills?
No. 2. How can MSEs organize themselves to extract value?
No. 3. How is information obtained, and governed?

5.1. MSE obtains necessary skills

The answer to the question No. 1 is very unfavorable for MSEs. According to Davenport, Barth and Bean [23] a big data company does not need a "data analyst", but a "data scientist". That it is not just the mere change of the name of the workplace, Davenport and Patil told us in their article [24]. They describe a data scientist as a hybrid of data hacker, analyst, communicator, and trusted advisor. David Sims suggested three main skills that a data scientist needs [25]:

- The first skill is a base in statistics, algorithms, machine learning, and mathematics;
- Second, a good data scientist is handy with a collection of open-source tools – Hadoop, Java, Python, among others. Knowing when to use those tools, and how to code, are prerequisites.
- The third set of skills focuses on making products real and making data available to users.

In his discussion about these skills, Iqbal R. J. [26], from different points of view, considered a possible scenario: "Shop-Mart and Bulk-Mart are two competitors in selling retail. Some higher up in the management chain asks this question: 'How many Shop-Mart customers also go to Bulk-Mart?'" He discussed how different data-related roles could approach the problem. The discussion included: traditional BI/reporting professional; data analyst; business analyst; data mining or big data engineer; statistician (a traditional one); and program/project manager. "A data scientist should have the skills of all the mentioned individuals. In addition to the skills mentioned above, a data scientist should have rapid prototyping and programming, machine learning, visualization and hacking skills" [26]. Because of many reasons, for MSE it is difficult to hire some data scientist. If a company wants to use big data it seems that the best solution is to find someone from the staff who's willing to deal with it. Even when MSE wants to find such employee, there is a big probability that it can't find appropriate person. Then there are two possibilities: outsourcing or reduction of demands to the acceptable level.

5.2 MSE organizes itself for big data

The answer to the question No. 2 demands a deeper analysis. Can MSE be organized in a manner different from large enterprises, and be effective? Who in a company deals with big data? According to Bean and Kiron [27] "ownership for big data initiative sometimes resides within the business side

and sometimes within the technology side". In the former case business management manages big data, and IT only receives and executes orders. In the latter case the entire responsibility about big data lies on the IT department. It is easy to conclude that there exist also two extreme possibilities: That no one cares about big data, and, the most favorable one, that a company is organized on matrix principle where business and IT work together in the way that follows the company's strategy. It is logical that each Head of MSE wants to apply the best variant, but practice shows that this is often not feasible. Problems may lie in the lack of strategy, in inability of competent managers, in lack of personnel, in lack of knowledge...

5.3. MSE manages big data

"Many organizations are setting their big data analytic efforts up for failure by plunging ahead without proper planning upfront" [3]. Many MSEs as well as big companies move in collecting of data that are directly related to doing business, but soon appetites are growing and small data in short time exceeds the class of big data. It is possible that the speed of data collection becomes greater than the capacities for their systematization and analysis. This way, the company continues to collect data making costs, but (practically) doesn't use them. The accumulation of data, as well as their placing into the databases and creating of spreadsheets provides no better picture of the analyzed occurrence than in the case of small data.

What to do with that data? The company can filter them out, arrange them, and classify. Although the picture will be clearer, it is unlikely that the information will be able to be used to solve some specific problem. The reason is simple. It's hard to find a real situation which will coincide with the data entered in the database.

6. Riks and implementation of BIG DATA

In general, big data and cloud services provide new functionality to users, can improve utilization of resources, and can help to generate revenue faster. But a lot depends of the kind of business that MSE conducts. According to International Data Corporation (IDC), the top 5 industries to benefit from the Cloud are [28]: IT operations (52%), Operations (35%), Financial and Accounting (33%), Customer support (31%), and Program Management (29%). The benefits are specific for each of the segments.

But, the use of big data brings some risks. There is only one convenient scenario, when business department and IT are headed by a single leader, competent in both areas, and numerous negative scenarios. Let us consider some of them:

- IT department possesses appropriate technical skills, but IT experts may suffer from the lack of the business knowledge. So, they can focus on the improving the technological solution. Such activity doesn't provide a business value.
- For alignment of business strategy with big data can be responsible a business group. In that case it is very likely that the business group may not fully be able to leverage the technology, and there is the risk of silo mentality or bad architectural solutions.
- Even in the currently positive scenario, which possesses the highest potential, there are risks from an organizational view. It is difficult to keep interdisciplinary collaboration on the demanded high level. Such organization demands a complex and expensive structural setup. Also, if cooperation between business and IT departments is not standardized and well-established, problems can arise at any change or absence of a responsible manager.

„It isn't surprising to find that 55% of big data projects aren't completed and many more fail to achieve their objectives – often those charged with implementation are the last consulted," said Jim Kaskade, CEO [29]. According to the survey of the same source, the reasons lie in: inaccurate scope (58%), technical roadblocks (41%), and 39% in siloed data and non-cooperation.

Why the majority of big data implementation tends to fail? Sicular [30] has quantified eight causes of failures, and she framed the problem in: "Learning Hadoop is easier than learning the business". People want to be driven by data, but first of all they must understand doing of business. Which data will be collected depends from the decision of men. If the decision is not based on understanding of business needs, there is big probability that data will be useless.

Many companies collect data indiscriminately, but raw data couldn't say anything to most business users, and provides no value to them. Even when company categorizes data it is hard to make a business case if the data doesn't provide answers to real business questions [3].

Also, all data are historical data, independently of their kind: financial -, sales -, customer behavioral -, or inventory data. All data are connected with the past, and they naturally tend to be backward-looking. Their use is analogous to watching in the rearview mirror when driving a car forward. This is an omnipresent limitation. Data analysis can give only trends, and some prediction of the near-term future, „but most historical data are of limited value in predicting the future" [2].

While collecting data, one tends to see them as equal, of the same value and trust, but rarely is that true. The world and Internet are awash in data. The amount of information rises constantly, and this flood means more confusion. How to extract trusted, correct and useful information? This is a problem that needs permanent monitoring and troubleshooting.

7. Conclusions

The research didn't disprove the alternative hypothesis H_1: "There are MSEs that can form their infrastructure so that they are able to take advantage of big data in order to form their strategies of development and doing business". The research gave a lot of possibilities for MSEs to use big data.

This research disproved the hypothesis H_0 and has shown that there are some micro and small enterprises that can successfully use big data when certain conditions are met. Many MSEs just starts to use big data, data mining and cloud computing technology. Therefore it is very important task of ensure a high level of security for sensitive data.

Bearing in mind the whole analysis, one can certainly conclude that many managers in decision making prefer to work with small data and already proven tools, rather than to rely on the illusion obtained from complicated, often „black box", software tools and uncontrolled vast mass of data. And these managers can be right because, as Thomas claimed, it is proven that an analysis of a poll with the sample of 1,500 is enough precise to predict who will be a next president. A poll with the sample of 300 is sufficient to predict how much the whole population will like a new product or service. A sample of 200 users can test a new product in-home for a week, and from that it can be precisely determined whether it is optimal and what its market share will be.

But why do companies of all sizes intend to use big data? Today's data processing platforms allow to big data users tools to work with very different types of data easily, so MSEs don't need highly specialized staff. Working in the Cloud MSEs can obtain speed, capacities and scalability without big investments. Big data tools allow creating of dashboards that can visualize data in a friendly and appropriate way. Big data analytic tools are still not fully developed and that is a great new business opportunity for MSEs to participate in new tools creation. There are big expectations from new tools close to artificial intelligence, use of natural language, text mining etc. There is only one big data challenge: New people, data scientists.

When a company decides to go into the big data environment, it has to ask itself: Which is the right way to use big data? The answer can be in forming of business strategy, and then in starting to acquire data needed to support this strategy. And finally, for MSEs it is of crucial importance to start with small investments and to form scalable system that will satisfy MSEs' needs in every moment.

References

[1] ASAY, M.: Big data is all about the cloud. *InfoWorld.* [Online] Apr 6, 2015. http://www.infoworld.com/article/2905917/big-data/big-data-is-all-about-the-cloud.html.

[2] THOMAS, J. W.: Little Data vs. Big Data: Nine Types of Data and How They Should Be Used. *MarketingProfs.* [Online] March 17, 2014. http://www.marketingprofs.com/articles/2014/24670/little-data-vs-big-data-nine-types-of-data-and-how-they-should-be-used.

[3] CARTER, KEITH B.: Managing Big Data: The Two Biggest Mistakes Companies Make. *TechTarget.* [Online] Sept 2014. http://searchbusinessanalytics.techtarget.com/ feature/Managing-big-data-the-two-biggest-mistakes-companies-make.

[4] *Commission Recommendation 2003/361/EC: Definition of micro, small and medium-sized enterprises.* EC. May 20, 2003, Official Journal L 124.

[5] CA: Mikro preduzeca. *Credit Agricole.* [Online] October 26, 2015. https://www.creditagricole.rs/mali-biznis/krediti/mikro-preduzeca.198.html.

[6] BAGLEY, R. O.: 6 Small Business Trends To Watch Out For In 2015. *Forbes.* [Online] Dec 19, 2014. http://www.forbes.com/sites/rebeccabagley/2014/12/19/6-small-business-trends-to-watch-out-for-in-2015/.

[7] BEIC, M. S.: *Methodology of political science with statistics.* Belgrade: Fakultet politickih nauka, 2008.

[8] Small data, *Techopedia.* [Online] Oct 27, 2015, https://www.techopedia.com/definition/ 29539/small-data.

[9] IBM: Taming Big Data: Small Data vs. Big Data. *IBM Big Data & Analytics Hub.* [Online] Oct 28, 2015. http://www.ibmbigdatahub.com/infographic/taming-big-data-small-data-vs-big-data.

[10] Hadoop - What is it and Why Does it Matter? *SAS.* [Online] Oct 28, 2015. http://www.sas.com/en_us/insights/big-data/hadoop.html?gclid=Cj0KEQjw5MGxBR DiuZm2icXX2-BEiQA619bq3LvL2T9ARLXa766CXlZhEOq6Qe09tyqOrcLOd1OuPYaAtHG8P8HAQ.

[11] EDLICH, S. N.: NoSQL - Your ultimate guideto the non-relational universe! *NoSQL.* [Online] Oct 28, 2015. http://nosql-database.org/.

[12] ANON: What is Stream Computing? *IBM Analytics.* [Online] Oct 28, 2015. http://www.ibm.com/analytics/us/en/technology/stream-computing/.

[13] NIKITA, I.: In-Memory Computing: In Plain English. *GridGain.* [Online] March 17, 2014. http://www.gridgain.com/in-memory-computing-in-plain-english/.

[14] VASUDEVA, A.: System Architecture for In-Memory Database. *IMEX Research.* [Online] 2012, http://www.imexresearch.com/IMEXPresentation/InMemory Computing .pdf.
[15] GRANDPIERRE, M., BUSS, G., ESSER, R.: In-Memory Computing technology: The holy grail of analytics? *Deloitte.* [Online] 2013. http://www2.deloitte.com/content/ dam/Deloitte/de/Documents/technology-media-telecommunications/TMT_Studie_In_Memory_Computing.pdf.
[16] *The NIST Definition of Cloud Computing - Recommendations of the National Institute of Standards and Technology.* Mell, Peter and Grance, Timothy. NIST Special Publication 800-145, s.l. : NIST, Sept 2011, Computing security, pp. 1-7.
[17] Microsoft: Hosting and Cloud Study 2014. *Microsoft.* [Online] 2014. http://news.microsoft.com/download/presskits/cloud/docs/hostingstudy2014.pdf.
[18] HUDAKOVA, M., BUGANOVA, K., DVORSKY, J., BELAS, J., DANA, L.: Analysis of the risks of small and medium – sized enterprises in the Zilina region, *Communications - Scientific Letters of the University of Zilina*, vol. 17 (2015), pp. 34-39.
[19] JUNOVA, E.: Public interaction of web self-presentation, *Communications – Scientific Letters of the University of Zilina*, vol.16 (2014), pp. 109-113.
[20] KRAJCOVIC, M., STEFANIK, A., DULINA, L.: Logistics processes and systems design using computer simulation, *Communications – Scientific Letters of the University of Zilina*, vol.18 (2016), pp. 87-94.
[21] RISING, C. J., KRISTENSEN, M., TJERRILD-HANSEN, S.: Is Big Data too Big for SMEs? *Stanford University.* [Online] summer 2014. http://web.stanford.edu/class/msande238/projects/2014/GainIT.pdf.
[22] *Predicts 2014: Big Data.* Heudecker, Nick, et al. s. l. : Gartner Research, Stanford, Connecticut, Nov 20, 2013, Gartner Insight.
[23] *How "big data" is different.* Davenport, T. H, Barth, P. and Bean, R.: July 30, 2012, MIT Sloan Management Review, vol. 54.
[24] DAVENPORT, T. H, PATIL, D. J.: *Data Scientist*, 90, 2012, Harvard Business Review, pp. 70-76.
[25] SIMS, D.: 3 Skills a Data Scientist Needs. *O'REILLY Radar.* [Online] Jan 25, 2011. http://radar.oreilly.com/2011/01/3-skills-of-data-scientists.html.
[26] IQBALL, R. T.: *What are The Key Skills of a Data Scientist? Quora.* [Online] Jan 10, 2014. https://www.quora.com/What-are-the-key-skills-of-a-data-scientist.
[27] *Organizational Alignment Is Key to Big Data Success.* Bean, R, Kiron, D. 3, Jan 28, 2013, MIT Sloan Managemen, vol. 54, pp. 1-6.
[28] IDC: *Midsize Enterprises Leading the Way with Cloud Adoption*. *SlideShare.* [Online] 2014, http://www.slideshare.net/CiscoBusinessInsights/idc-cloud-adoption-trends-infographic.
[29] INFOCHIMPS: Survey: What IT Teams Want Their CIOs to Know About Enterprise Big Data. *PR Newswire.* [Online] Feb 26, 2013. http://www.prnewswire.com/news-releases/survey-what-it-teams-want-their-cios-to-know-about-enterprise-big-data-188190311.html.
[30] SICULAR, S.: Glossary. *Gartner.* [Online] Aug 14, 2014. http://www.gartner.com/it-glossary/big-data.

Slavomir Kavecky - Penka Martincova*

AD HOC GRID RESOURCE MANAGEMENT: GRID SECURITY

The purpose of security in ad hoc grid environments is to support secure execution of tasks on shared resources and to protect the resources from malicious user actions. The concept of trust management is capable to solve the security issues by incorporating trust into the process of decision making. The quality of the decisions is dependent on a correct assessment and representation of trustworthiness assigned to the potentially collaborating parties. In most cases the value of trustworthiness is derived at least from direct trust and recommendations, but other factors as risk, uncertainty, context dependent information and attributes characterizing the task and the shared resource should be included in the derived value as well. This paper presents a specification of parameters relevant for an accurate trust evaluation.

Keywords: *Ad hoc grid, trust management, trust aware grid security, trust aware job scheduling.*

1. Introduction

Generally, the grid security protects shared resources against malicious actions of users and other entities that could damage the resources or corrupt the integrity of data stored and processed on the resources. However, in many situations the users of the ad hoc grid have to be protected from those who offer the resources, so the issue is also vice-versa [1].

Authentication and authorization, which are referred to as hard security mechanisms, do not allow any occurrence of risk or uncertainty (the user is either authenticated and authorized to access a shared resource or is not), but collaborations in an open environment are necessarily coupled with potential dangers that necessitate reasoning about risk and uncertainty. Trust was recognized as an important aspect of decision making in many distributed systems and it is used as a mechanism for managing dangers and learning from past interactions in order to reduce the risk exposure. For example, trust and reputation systems support decision making on the Internet provided services, which are based on a trust derived from ratings assigned to a certain provider by customers after completion of a transaction. Other parties can use the trust and reputation derived from the aggregated ratings to decide whether or not to run a transaction with the rated party in the future. Trust management, which is referred to as a soft security, represents the shift from attempting to provide absolute protection against potential dangers to accepting dangers as an intrinsic part of any global computing [1 and 2].

The aim of the paper is to present a detailed classification and specification of parameters that can be used for an accurate evaluation of trustworthiness assigned to a grid entity. The reminder of the paper is organized as follows: Section 2 presents a short overview of trust models integrating trust management into ad hoc grid computing; Section 3 describes the process of job scheduling performed in ad hoc grid environment; Section 4 provides classification of parameters needed for trust evaluation, describes relations between the parameters and proposes a procedure inferring the parameters into a final trust value; The verification of the proposed trust management integration into the ad hoc grid infrastructure is described in section 5; and finally, the section 6 concludes the paper.

2. Related work

The incorporation of trust management into the grid infrastructure has been a subject of research for several years. Model proposed by Azzedin and Maheswaran [3] is one of the first models introducing trust management as a part of the grid computing. In the model authors classify trust into two categories: identity trust and behavior trust. However, the evaluation and update of trust is derived only from the behavior trust, which is concerned with observations of past collaborations. The identity trust dealing with verification of entity's identity and determination of assigned authorization permissions is left out. The model also omits the integration of risk and uncertainty into the trust value, but as already stated there is no need for reasoning about trust if risk and uncertainty are not involved. The main asset of the model is the introduction of trust between

* **Slavomir Kavecky, Penka Martincova**
Faculty of Management Science and Informatics, University of Zilina, Slovakia
E-mail: slavomir.kavecky@fri.uniza.sk

two collaborating entities as a bidirectional relationship, where the trust of one entity in other entity differs from the trust of the other entity in the first entity. Therefore, the collaboration among entities is not executed unless both parties see each other as trustworthy.

The trustworthiness of a resource entity in the model proposed by Song et al. [4 and 5] is dependent on its self defense capability and reputation determined from prior collaborations and is referred to as a trust index. On the other hand, the user demands minimal required security assurance, which may appear as a request for authentication, data encryption, access control, etc., and is referred to as a security demand. During the process of assigning jobs onto resources the condition *Security Demad* ≤ *Trust Index* must be satisfied in order to start the jobs on the selected resources. The trust index assigned to the resource corresponds to a wider notion of trustworthiness and is derived from behavior trust as well as from the attributes of the resource. Risk and uncertainty are still omitted and are not part of the final trust index value. The main drawback of this model is evaluation of trustworthiness from only the user's point of view. The resource has no means to determine the trustworthiness of the user and to make trust based decision on whether or not to collaborate with the user.

Explicit usage of uncertainty as a part of trust value is presented in the model proposed by Lin et al. [6]. Trust is evaluated as a combination of belief and disbelief in the entity's trustworthiness and uncertainty as a filling of the absence of both belief and disbelief. The value of trust is deduced from the user's and also from the resource provider's point of view. The model also states that the user and the resource provider are interested in different types of trust. The user is interested in execution trust, which represents the ability of the provider to faithfully allocate appropriate resources to enable successful completion of the job. From the provider's point of view trust in the user is defined as a belief in the ability of the user to produce competent user code and it is referred to as code trust.

The model proposed by Shi et al. [7] introduces several novelties previously not considered in the trust models. Trust of one entity in another entity (in the model determined as a combination of direct experiences and experiences of other subjects) is influenced by a particular situation, in which the collaboration is about to take place, and is referred to as a situational trust. Different situations require different considerations with regard to trust and result in different values of trust. In case two entities have just encountered, the model introduces the initial trust as means to represent the basic trusting disposition of unknown entity and is derived from all previous experiences in all situations through the entire life time of the trusting entity. Despite the novelties the model is introducing, it has a few drawbacks. It lacks integration of risk into the trust value and uncertainty is expressed only indirectly through initial trust evaluated in case full information about trusted entity is missing.

Over the last decade more models emerged that express trust value as a combination of behavioral trust corresponding to the observations of past collaborations and trust derived from attributes describing the current state of evaluated entity. A more detailed overview of these models is presented in our previous work [8].

3. Trust Aware Ad Hoc Grid Scheduling

The purpose of the trust management in the ad hoc grid environment is to guarantee the quality of services provided by the grid nodes and the quality of user's behavior. The integration of trust into the ad hoc grid infrastructure is coupled inseparably with the scheduling of jobs on the provided resources. However, there are some differences between the ad hoc and the traditional grid scheduling when the trust management is involved.

In order to integrate trust management into the ad hoc scheduling process, the steps executed during the resource discovery, system selection and job execution must perform the following additional tasks: *(i)* definition of minimal trustworthiness needed to begin the collaboration between the user and the resource provider, *(ii)* determining the current trustworthiness of the involved nodes, *(iii)* and update of trustworthiness after the job completion.

It is evident that these tasks impose new requirements on the ad hoc grid architecture. The architecture depicted in Fig. 1 introduces the trust manager module as a solution to meet the imposed requirements.

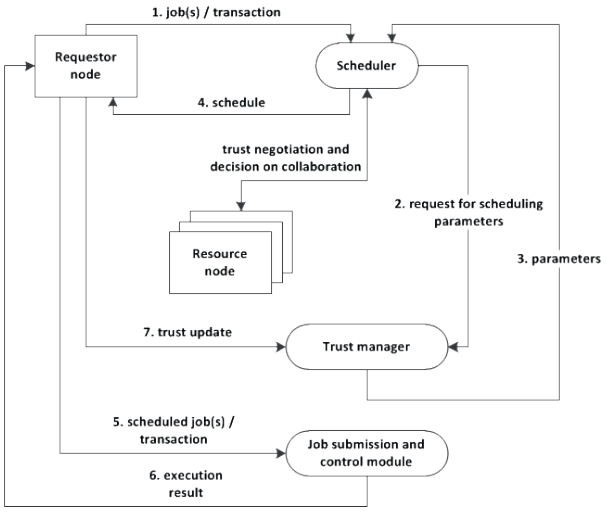

Fig. 1 Trust manager integration into the ad hoc grid infrastructure from the requester's point of view

During the phase of resource selection the user defines the job and the requirements needed for the job to run. To select a trustworthy resource, the user defines the security demand [4 and 5], which is taken as a constraint during the system selection step. The security demand is determined either directly by the user as one of the job requirements, or by the trust manager according to the parameters provided by the user.

The resources that passed the authorization and minimal filtering requirements are assigned a trust index [4 and 5] determined by the trust manager on the basis of static and dynamic information about the resources, job definition parameters and other factors managed by the trust manager. The trust index is a combination of more parameters, but which parameters and how exactly they are used for the trust index evaluation depends on the used trust model. The minimal components of the calculated value are the direct trust and the recommendations. However, other factors as risk, uncertainty and context dependent information should be included in the trust index as well. It is important to note that the scheduler uses the security demand and the trust indexes obtained from the trust manager only to exclude the untrustworthy resources. The schedule optimization itself is not affected and still corresponds only to the quality of services demanded by the user.

The resource provider demands a certain level of trustworthiness as well as the user. Therefore, after the exclusion of untrustworthy resources the scheduler requests the most optimal and trusted resource to consent to the future collaboration. The decision whether or not to accept the collaboration is based on the resource's security demand and the trust index assigned to the requesting node. Both values are obtained from the resource's trust manager on basis of job parameters included in the request, recommendations, previous experiences, uncertainty, risk and other factors. The decision on the collaboration is responded back to the scheduler. In case of negative response the scheduler sends the request for consent to the next most optimal resource until an affirmative answer is received.

The job scheduled with the help of the trust manager is forwarded to a module responsible for job submission and execution. After the job completion the result of the execution is transferred to the requester node. The trust update is the final step involving the trust manager module on the requester node as well as the resource node. The update is performed according to a positive or a negative experience resulting from the job execution and is necessary for correct representation of trust in the collaborating parties.

4. Trust Aware Ad Hoc Grid Security

The collaborations in the grid environment are executed by two types of entities: user and resource provider. User and resource provider require protection against malicious behavior that can take form of user's program containing malicious code capable to compromise the provider's resource node or it can take a form of a malicious resource node capable to harm the user's job running on the provided resource.

The security infrastructure incorporating trust should be based on a trust model that is capable to support or enhance the functional aspects of the grid infrastructure. The model should be also capable to process evidence of the previous collaborations and to transform it together with other relevant parameters into a trust value that is part of the security decisions for both the user and the resource provider protection.

4.1 Parameters Classification

Each participant of collaboration in the grid environment has his own set of expectations for the quality and performance of the collaboration and is satisfied with the executed collaboration only if the required expectations are met. Trust in this context can be used to express the confidence of the relying entity that a collaborating party will meet the desired expectations. The expectations for the quality of collaboration placed by the users and resource providers are mapped to system parameters and capabilities that can be abstracted into three groups of trust component: *(i)* behavioral parameters, *(ii)* system attributes *(iii)* and descriptive attributes.

Behavioral parameters (e.g. accessibility, availability, competence and reliability) describe the behavior of collaborating entities and are used to create history of data obtained from past interactions. By analyzing the history of the collected data using statistical methods together with the entity's personalized notion of normal or anomalous behavior it is possible to predict the outcome of future collaborations.

System parameters (e.g. authentication and authorization mechanism, utilized security mechanisms, maintenance of data integrity, etc.) describe the technical parameters and capabilities of the provider's shared resources and the user's node serving as access point into the grid community. The system attributes are characterized by a slow change over time. Over a period of time the attribute values do not change gradually, but the change is made suddenly and is noticeably large.

In contrast with behavioral and system attributes the **descriptive parameters** (e.g. benefit and loss associated with a particular collaboration, amount of observed behavioral parameters, time passed since last collaboration, etc.) do not describe the trusting disposition of the relying entity in the collaborating party, but they indicate the level of security assurance required by the relying entity. In a particular collaboration context the security assurance corresponds to the minimal trustworthiness of the collaborating party required by the relying entity.

4.2 Determining Trust from Parameters

The incorporation of trust management into the grid infrastructure should support the fundamental functional aspects of the grid as resource allocation and execution of tasks. The scheduling of tasks is responsible for finding an appropriate resource node meeting the required security assurance expressed as a security demand. Similarly, the resource node declares its own security demand that must be fulfilled by the user in order to process his request by the resource node.

The **security demand** is dependent on the risk and uncertainty (as depicted in Fig. 2) perceived by the relying party in the context of a particular collaboration. In a risky situation the relying party requires a high level of security assurance provided by the collaborating party in order to start a collaboration. Of course, the required security assurance is lower in case of a less risky situation. The uncertainty influences the security assurance in a similar manner. The higher the level of uncertainty the less certain about a collaboration execution the relying party becomes. Therefore, the required level of security assurance increases as well.

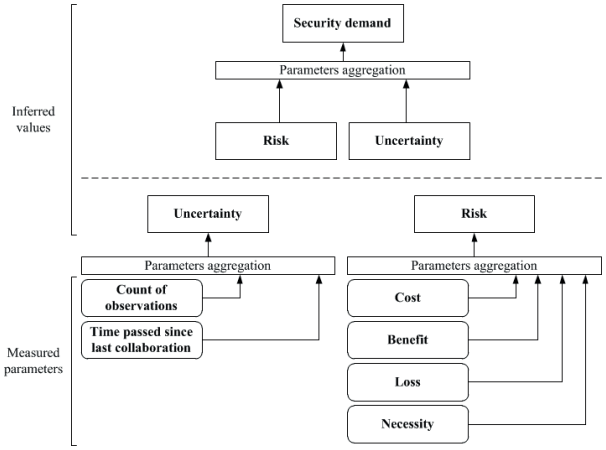

Fig. 2 Security demand inferred from trust components

The more important a flawless collaboration the more severe the damage will become in case of failure. The likelihood of failure occurrence and the cost incurred to the relying party is referred to as the **risk**. Risk and trust are related in the sense that there is no need for a trusting decision unless there is a risk involved. The measurable parameters used for inferring the value of risk as depicted in Fig. 2 are [1, 9 and 10]: *(i)* cost of collaboration, *(ii)* benefit *(iii)* loss, *(iv)* and necessity of a collaboration execution.

Uncertainty refers to a situation where the relying party cannot be fully sure about the accuracy of the decision. For example, a situation can occur where two completely unknown entities have to collaborate, but they have neither the experiences with each other, nor recommendations from other entities are available. A similar situation can also occur if only a part of the information is available and other decision factors are missing. The measurable parameters used for inferring the value of uncertainty as depicted in Fig. 2 are: *(i)* count of observations *(ii)* and time passed since the last collaboration.

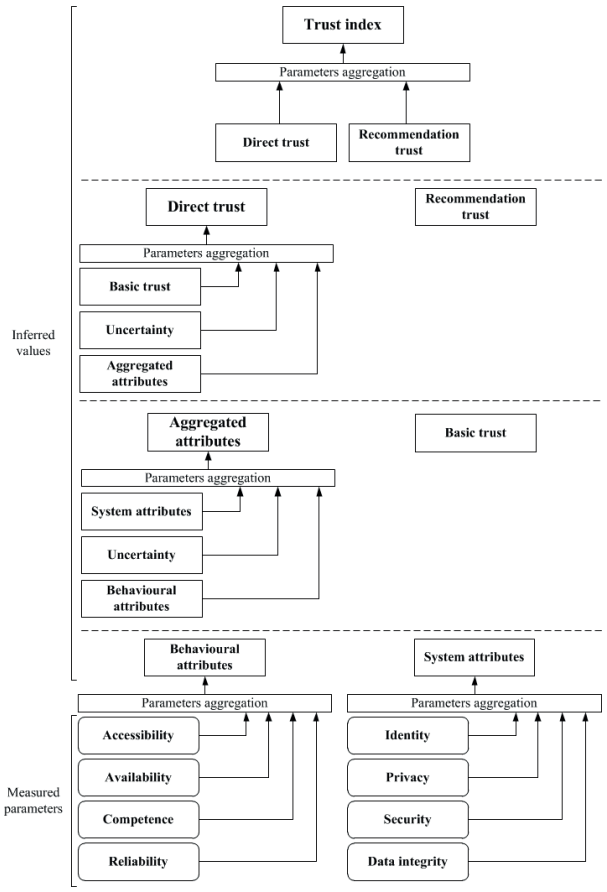

Fig. 3 Trust index inferred from trust components

Trust index, which specifies the trustworthiness of a collaborating party, is dependent on direct trust and recommendations (as depicted in Fig. 3). Recommendations are obtained from other entities of the grid environment and correspond to the reputation of the recommended entity. Reputation can be described as everything that is generally said or believed about the entity's character or standing. If the relying party is aware of the collaborating entity's reputation it can base its trust on that reputation, i.e. the collaborating entity is trusted because of its good reputation. Similarly, the entity becomes distrusted in case of its bad reputation.

Direct trust represents the private knowledge the relying party has about the collaborating entity and is formed from previous interactions, current context of the collaboration and attributes characterizing the collaborating entity. The direct trust and recommendations have different effect on the inferred trust index. The private knowledge of the relying entity in form of direct trust

is capable to overrule the reputation of the collaborating entity, i.e. in case of high direct trust the collaborating entity is trusted despite its bad reputation and similarly, in case of low direct trust the entity is distrusted despite its good reputation. The capability of the direct trust to overrule the recommendations is dependent on the weights assigned to these two parameters. As depicted in Fig. 3, the parameters used for inferring direct trust are: *(i)* basic trust, *(ii)* aggregated attributes, *(iii)* and uncertainty.

The **aggregated attributes** (as depicted in Fig. 3) are inferred from system and behavioral attributes. Uncertainty represents a weighting factor determining the relative importance of the considered attributes. The impact of system attributes is higher than the impact of behavioral attributes in case of high uncertainty and the behavioral attributes affect the inferred value more significantly in case of low uncertainty. Exact value of aggregated attributes can be inferred from the following system attributes (as depicted in Fig. 3): *(i)* identity, *(ii)* privacy, *(iii)* security, *(iv)* and data integrity. According to the models dealing with behavior trust [11, 12 and 13] the behavior of entity can be described with the following attributes (as depicted in Fig. 3): *(i)* accessibility, *(ii)* availability, *(iii)* competence, *(iv)* and reliability.

5. Experimental Results

The verification of the proposed model for trust value calculation and the trust management integration was carried out by a computer simulation. The simulation was performed using the GridSim simulation toolkit [14]. The verification of the modeled ad hoc grid infrastructure is evaluated according to the metrics specified in the section 5.3. The verification consists of two experiments. The first experiment was performed without the trust management integration and provides reference values describing the capabilities of the modeled ad hoc grid infrastructure. The latter experiment was performed with the trust management integration and shows the effect of the proposed trust value calculation and trust management integration on the secure execution of collaborations.

The model of the ad hoc grid infrastructure executed by the simulation toolkit consists of ten user entities and ten resource provider entities. The modeled entities were assigned several system capabilities and forms of behavior (Table 1, 2, 3 and 4). The capabilities and forms of behavior assigned to the modeled entities were used for trust calculation according to the model described in the section 4. Security demand and trust index, which are used for decisions making whether or not to start a collaboration, are values inferred from various system attributes, parameters describing the behavior of collaboration participant and context dependant parameters. As depicted in the Figs. 2 and 3, only the parameters used for inferring behavioral attributes, system attributes, risk and uncertainty are obtained as the result of measurement. All other parameter values are inferred from parameters placed one level below (except recommendations and basic trust, which values are obtained in a separate manner).

Each of the measurable parameters is measured directly or can be broken up in measurable elements. The parameters used for inferring the value describing the behavior of grid entity are measured directly according the entity's behavior elements observed over multiple collaborations. Considering the collaboration participant X as trustor and Y as trustee, the calculated value of inferred behavioral attributes V_x is based on the entirety of N behavior elements under observation and is expressed according to the following formula:

$$V_x = \frac{\Sigma_{i=1}^{N} E(Y)_i}{N} \qquad \text{Formula 1}$$

where $E(Y)_i$ represents value assigned to the i-th behaviour element under observation. Each observed element is calculated as the number of "positive" observations (the good behavior was observed) divided by the total number of observations, as generalized in the formula 2:

$$E(Y)_i = \frac{\text{"positive" observations of the } i-th \text{ element}}{\text{all observations of the } i-th \text{ element}} \qquad \text{Formula 2}$$

The parameters used for inferring the value describing the system properties of grid entity either are evaluated directly, or they are broken up in measurable and evaluable elements. Considering the collaboration participant X as trustor and Y as trustee, the value of inferred system attributes V_x is based on the entirety of N measured system parameters and is expressed according to formula 1 where $E(Y)_i$ represents value assigned to the i-th evaluated system parameter. The value of identity and privacy parameters is calculated as the value assigned to measurable element associated with the system parameter divided by the maximal attainable value as generalized in formula 3:

$$E(Y)_i = \frac{\text{value assigned to the } i-th \text{ measurable element}}{\text{maximal attainable value of the } i-th \text{ element}} \qquad \text{Formula 3}$$

In case of security and data integrity parameters, the value $E(Y)_i$ is calculated as the sum of n values assigned to the measurable element associated with the system parameter divided by the maximal value attainable by the measurable element as generalized in formula 4:

$$E(Y)_i = \frac{\Sigma_{j=1}^{n} j-th \text{ value assigned to the } i-th \text{ measurable element}}{\text{maximal attainable value of the } i-th \text{ element}} \qquad \text{Formula 4}$$

A more detailed measurement and evaluation procedure of modeled behavioral attributes and system parameters is presented in our previous work [15].

Characteristics of the users (1 - 5) modeled in the computer simulation Table 1

Characteristic	**User**				
	1	2	3	4	5
Antivirus protection	Yes	Yes	Yes	Yes	Yes
Firewall	Yes	Yes	Yes	Yes	Yes
Intrusion detection system	No	No	No	No	No
Transport layer security	Yes	Yes	Yes	Yes	Yes
IPsec	No	No	No	No	No
Rate of faulty tasks	No faulty tasks	Very low occurrence rate	Low occurrence rate	Low occurrence rate	Common occurrence rate

Characteristics of the users (6 - 10) modeled in the computer simulation Table 2

Characteristic	User				
	6	7	8	9	10
Antivirus protection	Yes	Yes	Yes	Yes	Yes
Firewall	Yes	Yes	Yes	Yes	Yes
Intrusion detection system	No	No	No	No	No
Transport layer security	Yes	Yes	Yes	Yes	Yes
IPsec	No	No	No	No	No
Rate of faulty tasks	Common occurrence rate	High occurrence rate	High occurrence rate	Very high occurrence rate	Very high occurrence rate

Characteristics of the resource provider nodes (1 - 5) modeled in the computer simulation Table 3

Characteristic	**Resource provider**				
	1	2	3	4	5
Antivirus protection	Yes	Yes	Yes	Yes	Yes
Firewall	Yes	Yes	Yes	Yes	Yes
Intrusion detection system	No	No	No	No	No
Transport layer security	Yes	Yes	Yes	Yes	Yes
IPsec	No	No	No	No	No
Sandbox	No	No	No	No	No
Authentication type	X509 certificate	X509 certificate	X509 certificate	X509 certificate	X509 certificate
Authorization type	Role based authorization	Role based authorization	Role based authorization	Role based authorization	Role based authorization
MIPS	2700	2350	2000	2350	2350
Rate of resource non-availability	Always available	Always available	Always available	Very low occurrence rate	Low occurrence rate
Rate of task execution failure	No failures	No failures	No failures	Common occurrence rate	Common occurrence rate

Characteristics of the resource provider nodes (6 - 10) modeled in the computer simulation Table 4

Characteristic	Resource provider				
	6	7	8	9	10
Antivirus protection	Yes	Yes	Yes	Yes	Yes
Firewall	Yes	Yes	Yes	Yes	Yes
Intrusion detection system	No	No	No	No	No
Transport layer security	Yes	Yes	Yes	Yes	Yes
IPsec	No	No	No	No	No
Sandbox	No	No	No	No	No
Authentication type	X509 certificate	X509 certificate	X509 certificate	X509 certificate	X509 certificate
Authorization type	Role based authorization	Role based authorization	Role based authorization	Role based authorization	Role based authorization
MIPS	2350	2700	2000	2700	2000
Rate of resource non-availability	Common occurrence rate	High occurrence rate	High occurrence rate	Very high occurrence rate	Very high occurrence rate
Rate of task execution failure	Common occurrence rate	High occurrence rate	High occurrence rate	Very high occurrence rate	Very high occurrence rate

5.1 Experiment 1 – Ad Hoc Grid without Trust Management

The first experiment was carried out by the simulation toolkit according to the modeled ad hoc grid infrastructure without incorporation of the trust management. The values measured during the simulation represent reference values describing the capabilities and qualities of the modeled infrastructure. The values were used for comparison to values measured during other experiments.

Count of tasks executed without trust management integration into the modeled ad hoc grid infrastructure Table 5

Task type	Measured values	
		Percentage of the executed tasks [in %]
All executed tasks	5833.10	100.00
Successful tasks	4880.36	83.67
Failed tasks	952.74	16.33

During the experiment 1000 simulation runs were executed. The count of all tasks, successful tasks and failed tasks are given as an arithmetic mean calculated from the values measured in each simulation run. Table 5 shows the count of all tasks executed during the first experiment, as well as the count of successful and failed tasks. The table also shows the percentage of the executed tasks. The count of all tasks is 5833.10, count of successful tasks is 4880.36 (83.67% of all executed tasks) and count of failed tasks is 952.74 (16.33% of all executed tasks).

5.2 Experiment 2 – Ad Hoc Grid with Trust Management

During the experiment 1000 simulation runs were executed. The count of all tasks, successful tasks and failed tasks are given as an arithmetic mean calculated from the values measured in each simulation run. The second experiment was carried out by the simulation toolkit according to the modeled ad hoc grid infrastructure with the incorporation of trust management. Table 6 shows the count of all tasks executed during the second experiment, as well as the count of successful and failed tasks. The table also shows the percentage of the executed tasks. The count of all tasks is 5829.20, count of successful tasks is 5292.12 (90.79% of all executed tasks) and count of failed tasks is 537.09 (9.21% of all executed tasks).

Count of tasks executed with trust management integration into the modeled ad hoc grid infrastructure Table 6

Task type	Measured values	
		Percentage of the executed tasks [in %]
All executed tasks	5829.20	100.00
Successful tasks	5292.12	90.79
Failed tasks	537.09	9.21

5.3 Evaluation Results

The verification of the modeled ad hoc grid infrastructure is evaluated according to the following quantitative metrics: *(i)*

competence *(ii)* and reliability. The competence corresponds to the capability of the modeled ad hoc grid infrastructure to support execution of user tasks. The competence is measured as count of all tasks executed during the simulation. The reliability corresponds to the capability of the modeled ad hoc grid infrastructure to support secure execution of user tasks. The reliability is measured as count of failed tasks observed during the simulation.

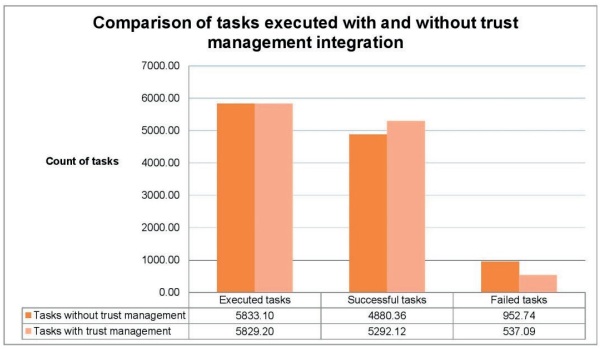

Fig. 4 Comparison of tasks executed with and without trust management integration into the modeled ad hoc grid infrastructure.

Figure 4 shows that there is almost no difference in the count of all tasks executed during the first and second experiment. The integration of trust management into the ad hoc grid infrastructure has no negative effects on the infrastructure competence to execute user tasks. On the other hand, the integration of trust into the ad hoc grid infrastructure affects the capability of the infrastructure to support secure execution of user tasks in a considerable manner. Figure 4 shows that the proposed incorporation of trust management results in reduced count of failed tasks and improves the reliability of the ad hoc grid infrastructure. The percentage of improvement in the modeled ad hoc grid reliability is equal to 43.62%. This proves the capability of the proposed trust model to ensure the secure execution of collaborations mediated through the ad hoc grid infrastructures.

6. Conclusion

The paper presents an ad hoc grid architecture integrating trust manager module taking over tasks as trustworthiness assessment of collaborating nodes and update of trustworthiness after a job completion. The procedure of trust evaluation is a complex process including procession of various trust components and relations among these components. The paper describes in detail these relations and their impact on the inferred values produced by the trust management inference system. The values are deduced from parameters describing the most significant system attributes and behavioral traits of evaluated grid entities. The verification of the proposed trust management integration into the ad hoc grid infrastructure was carried out by a computer simulation proving the capability of the proposed trust management integration to execute collaborations in a more secure manner.

References

[1] JOSANG, A., KESER, C., DIMITRAKOS, T.: Can We Manage Trust? *Trust Management*, vol. 3477 of Lecture Notes in Computer Science, pp. 93-107, Springer: Berlin : Heidelberg, 2005.

[2] JOSANG, A., ISMAIL, R., BOYD, C.: A Survey of Trust and Reputation Systems for Online Service Provision, *Decision Support Systems*, 43(2), 618-644, 2007.

[3] AZZEDIN, F., MAHESWARAN, M.: Evolving and Managing Trust in Grid Computing Systems, Canadian Conference on Electrical and Computer Engineering, 2002, pp. 1424-1429, 2002.

[4] SONG, S., HWANG, K., KWOK, Y. K.: Trusted Grid Computing with Security Binding and Trust Integration. *J. of Grid Computing*, 3(1-2), 53-73, 2005.

[5] SONG, S., HWANG, K., MACWAN, M.: Fuzzy Trust Integration for Security Enforcement in Grid Computing. *Network and Parallel Computing*, vol. 3222 of Lecture Notes in Computer Science, pp. 9-21. Springer: Berlin : Heidelberg, 2004.

[6] LIN, CH., VARADHARAJAN, V., WANG, Y., PRUTHI, V.: Enhancing Grid Security with Trust Management, Proceedings of the 2004th IEEE International Conference on Services Computing, 2004, pp. 303-310, 2004.

[7] SHI, J., BOCHMANN, G., ADAMS, C. A.: Trust Model with Statistical Foundation, IFIP International Federation for Information Processing: Formal Aspects in Security and Trust, pp. 145-158, 2005.

[8] KAVECKY, S., MARTINCOVÁ, P.: Overview of Trust Models Integrating Trust Management into Grid Computing, *International Journal of Computer Applications*, 129(7), pp. 1-6, 2015.

[9] JOSANG A., PRESTI, S. L.: Analysing the Relationship between Risk and Trust. *Trust Management*, vol. 2995 of Lecture Notes in Computer Science, pp. 135-145, Springer: Berlin: Heidelberg, 2004.

[10] ENGLISH, C., TERZIS, S., WAGEALLA, W.: Engineering Trust Based Collaborations in a Global Computing Environment. *Trust Management*, vol. 2995 of Lecture Notes in Computer Science, pp. 120-134, 2004.

[11] DIONYSIOU I., GJERMUNDROD, H., GUTS, S: *Simplified Grid user Trust Service for Site Selection.* The 7th Intern. Conference on Internet Monitoring and Protection, 2012, pp. 40-46, 2012.
[12] MANUEL, P., THAMARAI SELVI, S., BARR, M. E.: Trust Management System for Grid and Cloud Resources. 1st Intern. Conference on Advanced Computing, 2009, pp. 176-181, 2009.
[13] PAPALILO E., FREISLEBEN, B.: Managing behaviour Trust in Grid Computing Environments. *J. of Information Assurance and Security*, 3, 27-37, 2008.
[14] BUYYA R., SULISTIO, A.: *Service and Utility Oriented Distributed Computing Systems: Challenges and Opportunities for Modeling and Simulation Communities.* Simulation Symposium, ANSS 2008, 41st Annual, pp. 68-81, April 2008.
[15] KAVECKY, S., MARTINCOVA, P.: Specification of Parameters Relevant for Trust Evaluation in an Adhoc Grid Environment, *Intern. J. of Computer Applications*, 132(11), pp. 1-8, 2015.

Peter Czimmermann*

LOCATION PROBLEMS IN TRANSPORTATION NETWORKS

It is known that many optimisation problems on networks are NP-hard. However, it seems that the real transport networks have some interesting properties which allow us to find a „good" solution in reasonable time. In this paper, we suggest and study some new parameters of the transportation networks which could be useful in optimisation problems. We define the evenness and the robustness of the solution. We also concern ourselves with the statistical distribution of distances and edge values in transportation networks.

Keywords: *Transportation networks, Euclidean networks, robustness, distribution of distances.*

1. Introduction and preliminaries

It is well known fact that many optimisation algorithms work more successfully in real networks than in random graphs. For example, the greedy method for the traveling salesman problem gives, in real transportation networks, better results than in general graphs [1]. The difference is caused by triangle inequality which usually holds in transportation networks. We can also mention networks with Euclidean metric. Their properties allow us to suggest more efficient algorithms [2].

We will need the following definitions from [3 and 4]. Network is an ordered quadruple $G=(V,E,c,w)$, where $V\neq\emptyset$ is the set of vertices, $E\neq\emptyset$ is the set of edges (oriented or un-oriented), $c:E\rightarrow R_0^+$ is a function which represents the length of edges and $w:V\rightarrow N_0$ represents the weights of vertices. Let $d_G(u,v)$ be the distance between vertices u and v in network G and let $d_G(v,D)=\min\{d_G(v,u)\,|\,u\in D\}$ be the distance between a set $D\subset V$ and a vertex v in network G.

The eccentricity of a set D is
$ec_G(D)=\max\{d_G(v,D)\,|\,v\in V\}$.

The weighted eccentricity of D is
$ecc_G(D)=max\{w(v)\cdot d_G(v,D)\,|\,v\in V\}$.

The total distance of vertices of G from D is
$f_G(D)\sum_{v\in V} d_G(v,D)$

The total weighted distance of vertices of G from D is
$f_G^w(D)=\sum w(v)\cdot d_G(v,D)$.

Now we are able to define various location problems on network G.

Set $D\subset V$ is the p-center of G, if $|D|=p$ and $ec_G(D)\leq ec_G(D')$ for any p-element subset $D'\subset V$.

Set $D\subset V$ is the weighted p-center of G, if $|D|=p$ and $ecc_G(D)\leq ecc_G(D')$ for any p-element subset $D'\subset V$.

Set $D\subset V$ is the p-median of G, if $|D|=p$ and $f_G(D)\leq f_G(D')$ for any p-element subset $D'\subset V$.

Set $D\subset V$ is the weighted p-median of G, if $|D|=p$ and $f_G(D)\leq f_G(D')$ for any p-element subset $D'\subset V$.

Set $D\subset V$ is the anti-p-center of G, if $|D|=p$ and $ec_G(D)\leq ec_G(D')$ or any p-element subset $D'\subset V$.

Set $D\subset V$ is the p-maxian of G, if $|D|=p$ and $f_G(D)\leq f_G(D')$ for any p-element subset $D'\subset V$.

Definitions of weighted anti-p-center and weighted p-maxian are similar. It is known that the mentioned problems are *NP*-hard (if p is a part of the input) [5 and 6]. This is the reason why heuristic algorithms are usually used for finding a suboptimal solution [7, 8 and [9]. The only exception is the anti-p-center problem, which is polynomial – as it was shown in [10].

2. Robustness of a solution

In this section we introduce the robustness of a solution of various location problems. Let $D\subset V$ (where $|D|=p$) be a solution (not necessarily optimal) of a given location problem with objective function $\min h_G$ (where $h_G\in\{ec_G,ecc_G,f_G,f_G^w\}$). We can suppose that k edges of G are not rideable. We denote the set of these edges by F. It is reasonable to investigate the changes of objective function in network G - K , where G - F

* **Peter Czimmermann**
Department of Mathematical Methods and Operations Research, Faculty of Management Science and Informatics, University of Zilina, Slovakia
E-mail: peter.czimmermann@fri.uniza.sk

denotes the network obtained from G by deleting the edges from F. The k-robustness of set D for a given objective function h_G is

$$\min = \left\{ h_G(D)/h_{G-F}(D) \middle| F \subset E, |E| = k \right\}.$$

The problem of finding a k-robustness of D for function $ec_G(D)$ seems to be similar to the problem of edge k-connectivity in graphs which is solvable in polynomial time [11]. An exact polynomial algorithm or heuristic based on edge k-connectivity with tests on real data will be published in the next paper. The application of k-robustness is in facility location problems. For example, we need to locate p emergency medical stations. We have solutions $D_1, D_2, \ldots, D_a$ of similar quality. Then the robustness can be the next criteria for choosing one solution, because it says more about stability of the solution when there are parts of the network that are not rideable.

Similar concepts are studied for example in [12], where the authors deal with a facility (vertex) disruption, and [13], where the author studies the network reliability in general.

3. Evenness of a solution

In [2], the Euclidean metric, the triangle inequality and its impact on the solution of the p-median problem have been studied. It follows from the results that practical algorithms are more efficient in networks with these properties. We would like to continue in the research started in the above mentioned work. Hence, we suggest the study of three new parameters of solutions of various location problem (two of them are taken over from mathematical statistics). We would like to compare these parameters in networks and in graphs without the above mentioned properties. These parameters are the mean of vertex distances, variance and evenness. Let the set $D \subset V$ ($|D| = p$) be given. The arithmetic mean of distances (of vertices) in D is

$$\bar{d}_G(D) = \frac{2}{p(p-1)} \cdot \sum_{\substack{u \in D \\ v \in D}} d_G(u,v).$$

The variance of distances (of vertices) in D is

$$Var(D) = \frac{2}{p(p-1)} \cdot \sum_{\substack{u \in D \\ v \in D}} (d_G(u,v) - \bar{d}_G(D))^2.$$

The evenness of D in G is

$$\varepsilon_G(D) = Var(D)/Var(V).$$

We suppose that it is possible to approximate the mean, variance and evenness of the optimal solution from the properties of networks with Euclidean metrics. If we have appropriate approximations of these values, then we can restrict the set of admissible solutions. Computation of the optimal solution could be faster than computation in general graphs. It is convenient to test the properties of optimal solutions in large networks for the anti p-center problem, since this problem is solvable in polynomial time [10].

4. Distribution of distances

In this section we develop some ideas from the work [2], where the authors solve the p-median model using standard linear programming (plus branch and bound) method and the vertex substitution heuristic. The types of networks studied in [2] include Euclidean networks (distances between points are measured by Euclidean metric), path networks (with triangle inequality) and random distance networks. It follows from the results that the above mentioned methods are more efficient in Euclidean networks and path networks (which are similar to the actual transportation networks) even if the methods were not suggested for these special types of networks. We suppose that the algorithms using specific properties of these networks could provide significantly better results.

Important properties could be the distribution of distances and the distribution of edge lengths. Our tests on the road network of Slovakia show the distribution of edge lengths which is in Fig. 1.

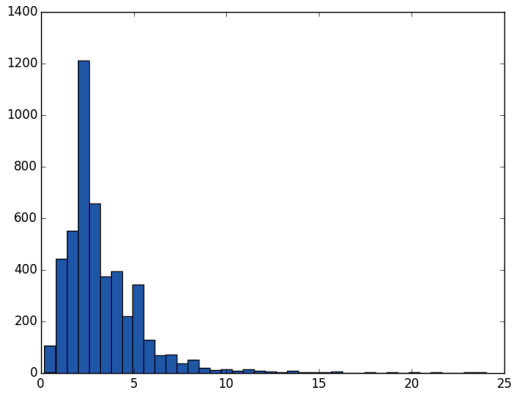

Fig. 1 The distribution of edge lengths in the road network of Slovakia

This distribution is similar to the Maxwell-Boltzmann distribution of molecule speeds in idealised gas [14] with the distribution function $f(x) = \sqrt{2/(\pi a^6)} \cdot x^2 \cdot e^{-x^2/(2a^2)}$, where a is an appropriate constant. The proof, that the distribution of edge lengths could be approximated by the Maxwell-Boltzmann distribution function, involves finding a special, continuous embedding of the road network into the three dimensional Euclidean space which preserves the incidence relation of a network and maps every edge to line segment with the same length. The existence of such embedding remains an open question. The next step of the proof involves finding the representation of the road network by the above mentioned

physical model, where edge lengths mean speeds of the particles. We are preparing a paper which is devoted to this topic.

If we consider the distribution of distances in the road network, then we achieve a similar situation. The results for distance distribution in Euclidean networks can be found in [2] and, for distance distribution of bus stops and stations in China, can be found in [15], where the authors work with a gamma distribution. We think that the best approximation of the distance distribution can be found in [16], where the probability distribution for the distance between two random points in a rectangle with given sides is determined. Its probability density function for a rectangle with sides *a, b* is

$$f(x) = \frac{2x}{ab}\left(-\frac{2x}{a} - \frac{2x}{b} + \pi + \frac{x^2}{ab}\right), \text{ for } 0 < x \leq a,$$

$$f(x) = \frac{2x}{b}\left(-\frac{2x}{a^2} - \frac{1}{b} + \frac{2}{a}\arcsin\left(\frac{a}{x}\right) + \frac{2}{a^2}\sqrt{x^2 - a^2}\right),$$

for $a < x \leq b$,

and

$$f(x) = \frac{2x}{ab}\left(2\arcsin\left(\frac{a}{x}\right) + \frac{2}{a}\sqrt{x^2 - a^2} + 2arcin\left(\frac{b}{x}\right) + \right.$$
$$\left. + \frac{2}{b}\sqrt{x^2 - b^2}\right) - \frac{2x}{b^2} - \frac{2x}{a^2} - \frac{2\pi x}{ab} - \frac{2x^3}{a^2 b^2},$$

for $b < x \leq \sqrt{a^2 + b^2}$.

We suppose that appropriate solutions of the *p*-center and *p*-median problem have the same distance distribution as the whole network. Our goal is the suggestion of a method for solving location problems which would be more effective on Euclidean and path networks than existing methods. Information about behaviour of solutions (distance distribution and evenness) can help us to find some new algorithms.

5. Conclusions

This paper is an introduction to our study of parameters and properties of transportation networks. We suppose that new knowledge in this area of research could bring more efficient methods for solving the problems that are too hard for exact computation in general.

Acknowledgment

The research was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic under project VEGA 1/0518/15 *"Resilient rescue systems with uncertain accessibility of service"*.

References

[1] COOK, W. J.: *In Pursuit of the Traveling Salesman: Mathematics at the Limits of Computation*, Princeton University Press, 2012, ISBN 978-0691152707.

[2] SCHILLING, D. A., ROSING, K. E., REVELLE, C. S.: Network Distance Characteristics that Affect Computational Effort in P-Median Location Problems, *European J. of Operational Research* 127, p. 525 - 536, 2000.

[3] PALUCH, S.: *Graph Theory* (in Slovak), EDIS: University of Zilina, p. 173 - 177, 2001.

[4] PALUCH, S., PESKO, S.: *Quantitative Methods in Logistics* (in Slovak), EDIS: University of Zilina, 2006.

[5] KARIV, O., HAKIMI, S. L.: An Algorithmic Approach to Network Location Problems, Part I: The p-centers, *SIAM J. of Applied Mathematics*, 37, p. 513 - 538, 1979.

[6] KARIV, O., HAKIMI, S. L.: An Algorithmic Approach to Network Location Problems, Part II: The p-medians, *SIAM J. of Applied Mathematics*, 37, p. 539 - 560, 1979.

[7] JANACEK, J., JANOSIKOVA, L., BUZNA, L.: *Optimized Design of Large-Scale Social Welfare Supporting Systems on Complex Networks,* Chapter 12 in Handbook of Optimization in Complex Networks: Theory and Applications, Springer Optimization and Its Applications 57, editors: Thai, M. T., Pardalos, P. M., 2012.

[8] JANACEK, J., Radial approach to the emergency public service system design with generalized system utility, *International J. of Applied Mathematics and Informatics*, 8: 7-14, 2014.

[9] JANOSIKOVA, L.: Emergency Medical Service Planning, *Communications - Scientific Letters of the University of Zilina*, 9(2): 64-68, 2007.

[10] CZIMMERMANN, P., PESKO, S.: *A Polynomial Algorithm for a Particular Obnoxious Facility Location Problem,* Proc. of the 13th Intern. Symposium on Operational Research in Slovenia, SOR'15, Bled, 2015.

[11] PLESNIK, J.: *Graph Algorithms* (in Slovak), VEDA: Bratislava, 1983.

[12] SCAPARRA, M., P., CHURCH, R., L.: *Location Problems under Disaster Events,* Chapter 24 in Location Science, Springer, editors: **Laporte**, G., **Nickel**, S., **Saldanha da Gama**, F., 2015.

[13] SHOOMAN, M., L.: *Reliability of Computer Systems and Networks: Fault Tolerance, Analysis, and Design,* John Wiley & Sons : New York, 2002.
[14] KREMPASKY, J., *Physics (in Slovak),* ALFA, Bratislava, 1988
[15] LU, H., SHI, Y.: Complexity of Public Transport Networks*, Tsinghua Science and Technology*, vol. 12, No. 2, 2007.
[16] PHILIP, J.: The Probability Distribution of the Distance between two Random Points in a Box, *Trita Mat, 7(10)*, 2007.

Jiri Cee - Oleksii Dieiev - David Holman - Radim Lenort - David Stas - Pavel Wicher*

SYSTEM ORIENTED SUSTAINABLE SUPPLY CHAIN MANAGEMENT INNOVATIONS IN AUTOMOTIVE INDUSTRY – SKODA AUTO CASE STUDY

System thinking and system theory are currently widely used in a variety of subjects and fields. Sustainability Supply Chain Management is defined as the strategic, transparent integration and achievement of an organization's social, environmental, and economic goals. The automotive industry is known as one of the most competitive branches in the industrial sphere. Efficient and effective deliveries of complete cars are fundamentally dependent on the performance of the delivery concept in the whole supply chain. The aim of the article is to research and define the fundamental principles and assumptions of system oriented innovations utilized in the Sustainable Supply Chain Management of automotive industry. These principles are derived on the basis of the case study related to the Efficient, Electronic, and Ecology Delivery Concept developed by SKODA AUTO Logistics.

Keywords: *System thinking, Sustainable Supply Chain Management, Innovations, Automotive industry.*

1. Introduction

The traditional approach to develop and implement innovations in automotive supply chain is based on productivity improvements of each individual part without considering the mutual interactions and the properties of the whole supply chain. Two critical assumptions of the proper system thinking application: improving mutual interactions between particular parts of the system and understanding of the essential properties of the wholes – systems, which derive properties of the parts and their interactions, haven not been applied either in academics nor practice so far. System thinking could rapidly increase the productivity of supply chain processes not only in automotive industry. The most important recent concept of Supply Chain Management (SCM) built by leading automotive companies is the sustainability. Practical example of Sustainable Supply Chain Management (SSCM) solution in automotive industry is Ecology, Efficient, Electronic and Ecology Delivery Concept (EDC) developed by SKODA AUTO Logistics (SAL). The aim of the article is to research and define the fundamental principles and assumptions of system oriented innovations derived on the basis of the EDC innovation in SKODA AUTO case study.

2. Literature review

2.1. System thinking

System thinking is one of the core methodological approaches to SCM research. Activities in organizations are best understood and developed when seen as an interaction of various subsystems and processes constituting a whole [1]. Based on the current practical experiences with social-economic systems (international teaching and research activities, solving practical projects for leading companies in the manufacturing industry) in SCM development including systems thinking utilization and literature research in understanding of systems thinking in relations to the wholes, in social system management, two attitudes could be identified:

Environmental free systems thinking - properties of the parts and their interactions creating the whole and deriving its properties without consideration for the environment.

Environmental full systems thinking - understanding of the essential properties of the whole by identifying its functions and properties in the containing (superior) system of which the examined system is a part. These essential properties derive the properties of the parts and their interactions. The environment is included.

The first attitude has dominated the theory and practice so far (from the introduction of General System Theory in 50's).

* **Jiri Cee, Oleksii Dieiev, David Holman, Radim Lenort, David Stas, Pavel Wicher**
Skoda Auto, Mlada Boleslav, Czech Republic
E-mail: David.Holman@savs.cz

It could be described either as self-centred orientation of social system management or fragmentation of the whole into individual parts and its separated management. In the supply chains, businesses, universities, governments, hospitals of any social system see the environmental free system thinking orientation as to divide the social system into parts, attempt to maximize the performance of the parts or their relations and hope that the sum and interactions will create an efficient whole, explicitly expressed by Christopher, one of the leading scientists in the branch of SCM in general. After almost 30 years of SCM concept development, he suggests a new direction of the concept, replacing the term SCM by demand chain management. Self-centred orientation of businesses and academic research should be replaced by customer-centred orientation, properties derived outside the system. Outside orientation has been seen only in the marketing departments so far and should be rolled out systematically to the whole supply chain [2]. The second attitude and its impact and implications have stood aside mainstream academic and practical attention so far. Particularly mentioned by Ackoff's: Systems thinking definition – the essential properties of the system are properties which none of its parts have, the essential properties are derived outside the system and could be seen only in the containing system [3].

2.2. Supply Chain Management (SCM) in Automotive industry

2.2.1. Supply Chain Management

Supply Chain Management encompasses the planning and management of all activities involved in sourcing and procurement, conversion, and all logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third-party service providers, and customers. In essence, supply chain management integrates supply and demand management within and across companies. Supply Chain Management is an integrating function with primary responsibility for linking major business functions and business processes within and across companies into a cohesive and high-performing business model. It includes all of the logistics management activities noted above, as well as manufacturing operations, and it drives coordination of processes and activities with and across marketing, sales, product design, finance and information technology [4]. Typical structure of Supply chain starts with final customer, who is delivered by final product or service. In general, the chain contains retailer, distributor, producer, logistics service provider, forwarder, supplier. Each of the chain is the semi-final customer of the previous chain. Each chain of the supply chain contains the inbound logistics flow (flow of resources into the production e.g. assembly line) and outbound logistics flow (flow of resources outside the chain e.g. delivery of final product to distributor, retailer or final customer). The concrete structure depends on the product, market or the particular strategy of the producer to complete customer demand.

2.2.2. Specifics of SCM in automotive industry

Automotive industry belongs to the leaders of innovation in the industrial sphere because of its complexity, production volumes, demanding customers and the worldwide competitive environment. The typical automotive supply chain contains thousands of 1st, 2nd and 3rd tier suppliers, OEM´s, distributors and dealers connected by forwarders, logistics services providers and cross-dock centres. The management of these chains in supply chain is characterised by integration, process management, logistics and production value added analysis and pull principle. The leading management system of innovations utilized in automotive supply chains in general is LEAN including tools and principles such as JIT, Kanban, SMED, Kaizen, 5S, Jidoka, Andon etc. Actual situation in automotive SCM, described later by Original delivery concept, haven´t used the potential of LEAN improvement yet and have been still focused on particular optimization striving for short term, cost reduction results. Latest trends in automotive industry consider the impact of the production to the natural environment and the human labour. Together with serious accent to costs and productivity, there was defined the new SCM solution including efficiency, social aspects and natural environmentally friendly attitudes called Sustainable Supply Chain Management.

2.2.3. Sustainable Supply Chain Management

Although there is a divergence of definitions of sustainability, these differences are not too great. Most definitions of sustainability incorporate a consideration of environmental, economic, and social dimensions:

- Sustainability is a wise balance between economic development, environmental stewardship, and social equity [5].
- Sustainability includes equal weightings for economic stability, ecological compatibility and social equilibrium [6].

This definition is based on the Elkington's [7] triple bottom line (the intersection of environmental, social and economic performance) and the four supporting facets (facilitators) of SSCM [8]:

1. Strategy – holistically and purposefully identifying individual SSCM initiatives which align with and support the organization's overall sustainability strategy.
2. Risk management, including contingency planning for both the upstream and the downstream supply chain.

3. An organizational culture, which is deeply ingrained and encompasses organizational citizenship and which includes high ethical standards and expectations along with a respect for society and the natural environment.
4. Transparency in terms of pro-actively engaging and communicating with key stakeholders and having traceability and visibility into upstream and downstream supply chain operations.

SSCM is the SCM concept oriented on balancing the economic, environmental and social supply chain performance.

3. Case study - SKODA AUTO

SKODA AUTO is a significant worldwide producer of cars. The daily production of 3,000 cars in the Czech Republic, need to be supported either by inbound flow of components or outbound flow of complete cars. The general numbers describing daily performance of the total inbound flow of resources are 2,500 trucks, 100 containers, 200 wagons. All the material flow must be supported by information flow and human resources considering the living and social environment. The case study is focused on inbound logistics flow of SCM.

3.1. Original delivery concept

The original delivery concept could be described by particular optimization of chains in supply chain focused on the lowest possible costs of each individual chain with limited coordination, cooperation and natural environment consideration. Based on the current state analysis the following problems have been defined:

1. Necessity to deal with an increasing number of truck deliveries, which are the result of both growing production and a rise in part numbers. The truck has to go through a tangled arrival process, making the route to the unloading area in the plant, or into the logistics centre more complicated. Furthermore, just-in-time and just-in-sequence deliveries are difficult to optimize due to the considerable paperwork and manual check-in processes.
2. Difficulties in optimization of the truck loading to achieve the optimum amount of materials in warehouses. Deliveries in this regard take place 2-3 times a day, thus the amount of material loaded on trucks should not exceed the specified maximum.
3. There is a need to unify two separate transport labels of C and B types in order to reduce the cost of printing, improve the time management among workers and spend less paper and ink to care about the environment.
4. Currently used trucks with diesel engines produce a lot of carbon dioxide, which is harmful to the environment. It is necessary to find a solution that corresponds to the diesel engine as far as technical characteristics, but is more sparing in relation to the environmental situation.

3.2. Efficient, Electronic, and Ecology Delivery Concept

The reaction of the SAL to above mentioned challenges in supply chain was proposal of system oriented solution including the SSCM elements. The new innovative concept (EDC) developed by SAL is based on the system oriented integration of the four pillars (see Fig. 1):

1. QCI (Quick Check-In) is the geo-fencing project, which involves a simple smartphone app that uses GPS technology to verify the position of incoming trucks, and connects to the control system for checking in material at the plant. The system allows trucks that adhere to delivery and production schedule to bypass the entry gate and move directly to the plant dock or the logistics centre adjacent to the factory without any extra paperwork or stopping [9].
2. FOLAB is the material requirement calculation program for delivery instructions which always runs once a week, now including the vehicle build-program for the 3rd production week in line. Single calls are shown in the delivery instructions from Monday of the following week. If no delivery has been recorded for the material requirement that was scheduled for a delivery date before the current requirement period, this will be entered as backlog in the delivery instruction. The total material amount entered as backlog and urgent requirement (plus any material requirement for the following delivery day) should be shipped immediately. If necessary, the mode of shipment should be agreed separately with the responsible material follow-up analyst [10].
3. GTL (Global Transport Label) was created in order to substitute the outdated system of separated B and C labels, constituents of VDA 4902 (Odette transport label). GTL has been created in accordance to ISO 15394. GTL label is the most universal solution to ensure a stable delivery in the transport chain from the supplier via the intermediate storage to the point of assembly or directly to the customer [11].
4. CNG (Compressed Natural Gas) is natural gas under pressure which remains clear, odourless, and non-corrosive. Although vehicles can use natural gas as either liquid or gas, most vehicles use the gaseous form. Natural gas is produced worldwide at relatively low cost and is cleaner than gasoline or diesel fuel. Natural gas vehicles show an average reduction in ozone-forming emissions of 80 percent compared to gasoline vehicles [12].

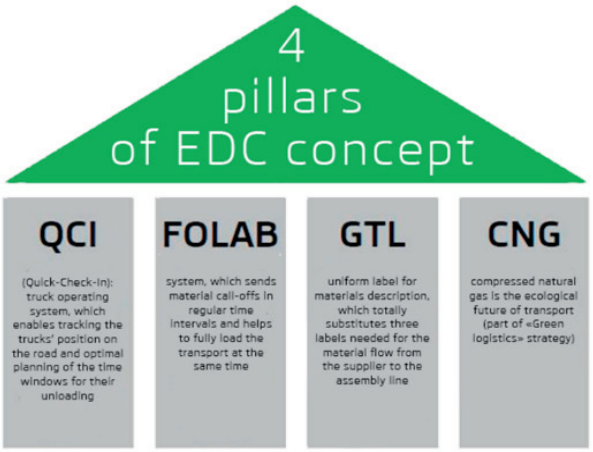

Fig. 1 Pillars of EDC concept [13]

The benefits of EDC against original delivery concept are [13]:

1. Logistics expenses, including fuel and pallets costs, are reduced by 25% in comparison with the existing system.
2. Expenses on material inventories are 55% lower.
3. Expenses associated with logistics space are also decreased by 55%.
4. CO_2 emissions are 50% lower through the use of CNG in trucks instead of diesel engine. Also reduction of CO_2 emissions is caused indirectly by the use of the FOLAB system due to optimal loading of truck resulting to less number of round trips.

4. Findings and discussion

Based on the research and analysis of the case study, especially the EDC concept, two basic principles of environmental full system oriented innovations utilized in the SSCM in automotive industry could be defined:

1. **Improving mutual interactions between particular parts of the system (P1):**
- Efficient flow of active and passive components between particular chains in the supply chain.
- Integration of physical and non-physical resources in the SSCM system.
2. **Derivation properties of interactions of SSCM from the essential properties of the SSCM system (P2):**
- The essential properties of interactions of the whole SSCM were derived by the upper system in which the SSCM system is working – business markets and society conditions of 21st Century. The list of properties which should be fulfilled is: the reduction of the negative impact on the natural environment, better working conditions, delivery quality, reasonable costs, and availability of resources, agility, and resilience.
- The performance (intensity of innovations) of interactions is derived from the essential properties of the SSCM system as a whole.

The particular optimization of separated chains in supply chain (suppliers, forwarders and SKODA AUTO plant was replaced by unified goal of all participants supported by improving of interactions (information and material flow) of all chain in inbound logistics supply chain. Instead of bottom up, the top-down approach was used to manage the quality and productivity of interaction between chains in supply chain and the productivity of the whole, inbound logistics system, rather than separated productivity of the each individual part.

Comparison of traditional and environmental full system oriented approach to innovations in SSCM is shown in Fig. 2.

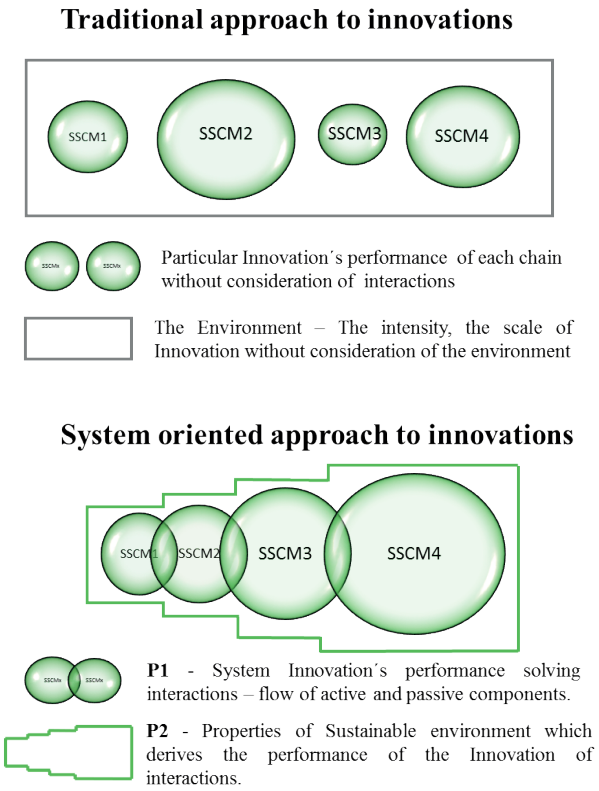

Fig. 2 Traditional and environmental full system oriented approach to innovations in SSCM

The traditional attitude on the implementation of innovations is focused on the maximum performance (the lowest possible cost) of each individual element in the sustainable supply chain. This approach does not take into account the environment and consequences. All these circumstances, influencing the output of the whole supply chain dramatically, need to be considered. Particular optimization creates serious barriers to efficiency and effectiveness of the whole.

System Oriented SSCM Innovations matrix Table 1

	Principles	Assumptions	EDC Pillars
System oriented SSCM innovations	P1	Utilization of modern ICTs	QCI, GTL
		Compatible AISs in the supply chain (supplier, forwarder, warehouses, assembly line)	QCI, GTL
		Vertical integration of the supply chain (supplier, forwarder, producer)	QCI, FOLAB, GTL
		Shared planning	FOLAB, GTL
		Shared information flow	FOLAB, GTL
	P2	Reduction of negative impact on the natural environment (CO_2)	QCI, FOLAB, GTL, CNG
		Better working conditions	FOLAB, GTL, CNG
		Good delivery quality	QCI, FOLAB, GTL, CNG
		Reasonable costs	QCI, FOLAB, GTL
		Availability of reseources	QCI, GTL
		Agility	QCI, FOLAB, GTL
		Resillience	QCI, FOLAB, GTL

The analogy could be seen in any team sport, or more precisely world championship of the concrete sport. There are generally two teams at the end of the championship. The one, who defeats all the opponents during the tournament and the one, created from the best possible players on each individual position from all the participating teams during the tournament (as goalkeeper, defender, striker in football), all stars team. Which one would be working as a team better and whose performance creates better output? The one that could improve its interactions during the tournament, although the opponent could have a better individual player at each position. This example advocates the importance of interaction – the first principle of environmental full system oriented innovations in SSCM (P1).

The same example could be used for explaining the importance of the second principle (P2). As the coach of the team, should the same tactics be used - the performance of cooperation of all members of the team, against all opponents? Certainly not. The optimal performance of the team members' interaction needs to be derived from the qualitative properties of each individual opponent (better offensive skills, defensive skills, etc.), which means considering the environment.

Literature research and the consultation of its findings with automotive professionals enable to identify the best possible innovative solutions for the current supply chain in SAL. More efficient particular solutions than the ones used in EDC, could be found in the current technological world for each part of the supply chain interaction. However, the goal of system oriented innovations is not the maximal efficiency of particular parts (sustainability does not mean minimum costs + minimum CO_2 production + minimum social negative impact of each chain in the SSCM) but the maximal performance of the whole system including the sustainable elements. The optimal output of EDC innovation could be reached and measured only as a whole, not only by evaluating the performance of particular parts and its interactions.

The detailed analysis of the four main pillars of the EDC enables to create the System Oriented SSCM Innovations (SOI) matrix, which combines the two principles and twelve assumptions necessary for successful system oriented innovation in SSCM (see Table 1). Significant improvements in interaction of the elements in SSCM are reached by EDC pillars utilizing ICTs (Information and Communications Technologies improving integration, interaction and cooperation between chains in supply chain), AIS (Automatic Identification Systems connecting the production resources without human support) and integration focusing on shared planning and information flow. The environment creating the upper system of SSCM derives the properties - performance of interactions. The triple bottom line: environmental, social and economic aspects are specified in the reduction of CO_2, improving working conditions for employees and reasonable cost and delivery quality, availability of resources, agility and resilience corresponding with the requirements of final and semi-final customers of the automotive industry.

5. Conclusion

SKODA AUTO is currently one of the most successful companies of the Volkswagen group and SKODA AUTO Logistics belongs to the leaders in logistics and supply chain management innovations. One of the innovative concepts developed by SAL is the EDC. The concept is oriented to make interactions in the SSCM efficiently, quickly, smoothly and environmentally friendly. Based on the case study research it is evident, the connection of more subjects (supplier, forwarder and automotive producer with all of its departments) to an efficient whole - sustainable supply chain, requires implementation of two key principles. Focus on mutual interactions between particular elements of SSCM and the derivation of the performance of these interactions from the essential properties of the whole SSCM. The upper system, from which the SSCM system is a part of, creates business markets and society conditions for the 21st Century. The combination of the principles and derivation of necessary assumptions for system oriented SSCM innovation can be presented as SOI matrix (see Table 1).

Innovations are traditionally focused on improving particular interactions in supply chain processes. More efficient particular solutions, than used in the EDC concept, could be found for each part of the supply chain interaction in the case study in the current technological world. However, the effectiveness of the whole system – SSCM in the automotive industry is not composed of the best particular solutions. Innovations carried out by system thinking and sustainable attitude reached far better presented results, than innovation improving only particular parts, without considering of interactions and the environment. The proposed SOI matrix could be used in the inbound or outbound processes of the SSCM not only in the automotive industry.

Acknowledgement

This article and associated research was carried out as a part of the project financed by Internal Grant Agency of SKODA AUTO University No. SGS/2015/02.

References

[1] KOTZAB, H., SEURING, S., MULLER, M.; REINER, G.: *Research Methodologies in Supply Chain Management.* Heidelberg : Springer Science and Business Media, 2005.
[2] CHRISTOPHER, M., RYALS, L. J.: Supply Chain Becomes the Demand Chain. *J. of Business Logistics,* vol. 35, No.1, 2014, pp. 29-35.
[3] ACKOFF, R. L.: *Ackoff's Best: His Classic Writings on Management*. New York : John Wiley & Sons. 1999.
[4] COUNCIL OF SUPPLY CHAIN MANAGEMENT PROFESSIONALS: CSCMP Supply Chain Management Definitions and Glossary, [cit. 14/06/2016], available from: https://cscmp.org/supply-chain-management-definitions[5] SIKDAR, S. K.: Sustainable Development and Sustainability Metrics. *AIChE J.*, vol. 49, No. 8, 2003, pp. 1928-1932.
[6] GONCZ, E., SKIRKE, U., KLEIZEN, H., BARBER, M.: Increasing the Rate of Sustainable Change: A Call for a Redefinition of the Concept and the Model for its Implementation. *J. of Cleaner Production*, vol. 15, No. 6, 2007, pp. 525-37.
[7] ELKINGTON, J.: *Enter the Triple Bottom Line*. In Henriques, A.; Richardson, J. (Eds), The Triple Bottom Line: Does It All Add up? London: Earthscan, 2004, pp. 1-16.
[8] CARTER, C. R., EASTON, P. L.: Sustainable Supply Chain Management: Evolution and Future Directions. *Intern. J. of Physical Distribution & Logistics Management*, vol. 41, No. 1, 2011, pp. 46-62.
[9] LUDWIG, C.: Audi Geo-Fencing Project Takes Top VW Group Logistics Prize, [cit. 23/08/2015]. Available from: http://www.automotivelogisticsmagazine.com/news/audi-geo-fencing-project-takes-top-vw-group-logistics-prize.
[10] SKODA AUTO: Annual Report 2014, [cit. 04/08/2015]. Available from: http://www.skoda-auto.com/en/company/investors/annual-reports.
[11] VOLKSWAGEN: Implementation Guideline: Global Transport Label (GTL), [cit. 23/08/2015]. Available from: http://goo.gl/OkRSvQ.
[12] Consumer Energy Center. Compressed Natural Gas (CNG) as a Transportation Fuel, [cit. 23/08/2015], available from: http://www.consumerenergycenter.org/transportation/ afvs/cng.html.
[13] SKODA AUTO: TOP 50 Innovative Logistics Solutions in SKODA AUTO a. s., 2015.

Peter Futas - Alena Pribulova - Jozef Petrik - Marcela Malindzakova - Andrea Rosova - Beata Grabowska*

ELIMINATION OF NEGATIVE IMPACTS OF STEEL SCRAP AS A CHARGE COMPONENT DURING THE PRODUCTION OF SYNTHETIC CAST IRON

The article presents the research results of omitting the steel scrap component from the production process of synthetic gray cast iron. The quality of synthetic iron produced in electric induction furnaces (EIF) is influenced by numerous factors. Its characteristic attributes are high values of mechanical properties (Rm, HB), but also its tendency to form chill out and shrinkages, foundry stress and sensitivity to hardness change at different casting wall thicknesses. Possibilities of offsetting and eliminating such negative effects were tested by introducing metallurgical countermeasures based on the properties of flake graphite alloys.

Keywords: *Synthetic cast iron, charge materials, steel scrap, mechanical properties, microstructure.*

1. Introduction

Production of synthetic cast iron was enabled by the development of sleeve-type EIF. In addition to using up to 100 % of the relatively cheaper steel scrap in the steel preparation, the main benefit is the possibility to produce liquid metal of specific and precise chemical composition. Right from the very beginning of its production, the better mechanical properties (*R*m - tensile strength, and HB - hardness) of synthetic iron were discovered, in comparison to iron cast with high volume of expensive pig iron, compared at the same carburization degree (Sc).

It is explained as the effect of high nitrogen content in the steel, although this has not yet been verified in practice. The increase of cast iron mechanical properties (on average by about 20 %) in practice represents a considerable advantage.

During the formation of solid solutions and phases, the nitrogen inhibits graphitization and increases pearlite stability. At the same time, due to the modifying effect, nitrides of the AlN, BN, TiN type can soften the structure and graphite during the initial stage of crystallization [1-5].

Carbon and silicon reduce the solubility of nitrogen in pig iron [6-7]. The alloying with metallic Ti, as well as of FeTi oxides and TiO_2 ($FeTiO_3$) neutralizes the effect of nitrogen. Increasing the temperature of molten metal improves the solubility of nitrogen in the liquid iron [8-10].

2. Experimental tests

The main goal of the experimental tests was to produce cast iron alloy with properties matching the EN-GJL-250 cast iron alloy standard (Sc carburizing degree of 0.87 - 0.93; C content of 3.1 – 3.3; Si content of 1.7 – 1.9) under operating conditions in foundries in high-volume electric induction furnaces. Overall six experimental melts have been designed and realized. The tests examined the impacts of the increased steel scrap levels in the charge. They also evaluated the possibilities to eliminate the negative impacts of high steel scrap content on the final iron quality by:

a) increasing the temperature of heat treatment (1 500 °C, typically 1 420 °C) and inoculation,
b) alloying with titanium (by heat treatment to 1 500° C and inoculation),
c) increasing the carbon content (+ 0.5 %) and by reduction of the Si content.

The cast iron alloy properties from these experimental melts were compared to conventionally prepared semisynthetic cast iron (with the ratio of steel scrap in the charge of about 33 %). The castings from each melt were compared and analyzed. In order to properly compare the properties of both alloys, the following tests were performed:

- Chemical analysis of both samples.

* [1]Peter Futas, [1]Alena Pribulova, [1]Jozef Petrik, [2]Marcela Malindzakova, [2]Andrea Rosova, [3]Beata Grabowska
[1]Faculty of Metallurgy, Technical University of Kosice, Slovakia
[2]BERG Faculty, Technical University of Kosice, Slovakia
[3]Faculty of Foundry Engineering, AGH University of Science and Technology, Poland
E-mail: marcela.malindzakova@tuke.sk

- Stairs test (R-Block with three thicknesses of 100, 50 and 20 mm – standard thickness of the walls castings, Fig. 1), to determine the sensitivity difference between the walls' thickness and the castings' hardness (cooling rate).
- Rm - tensile strength comparison – performed on test bars with diameter of 30 mm.
- Comparison of wedge-shaped test bar properties [11].
- Comparison of the samples to determine the difference in the propensity towards shrinkage occurrences (tested on cylinders with diameter of ø 95 mm and height of 150 mm). The cylinder dimensions and volumes were measured before and after the shrinkage occurrence.
- Samples comparison in regard to determining the nitrogen content.

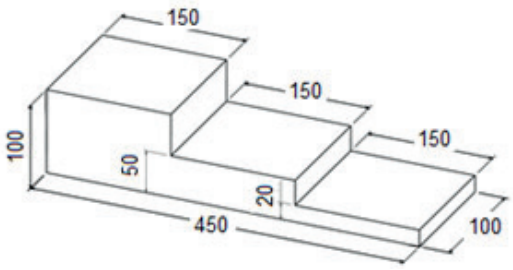

Fig. 1 Stairs-test (R-block)

The samples for metallographic analysis were prepared in a form of test bars; theywere prepared in a traditional manner.

R-Block was cut in each of the cross sectional areas. For every section the Brinell hardness was measured. The hardness was measured using the HPO 3000 durometer (setting: 10/3000/10). Tensile strength was measured on test-bars (ø 30 mm) on a standard testing machine of ZWICK brand. The percentages of the charge materials content for each melt are shown in Table 1.

3. Results and discussion

The results of the chemical analysis and the results of mechanical tests comparing the properties for each melt are shown in Table 2.

The chemical composition of cast alloy matched the requirements of identical Sc 0.848 – 0.869 (slightly different). Melt No. 4 with high carbon content of C = 3.79 %, and a low content of Si = 1.026 %; Sc = 0.951 - which corresponds to the GJL – 250 standard. The nitrogen content in the experimental melts showed the lowest amount in melt No. 4 (N_2 = 0.0073 %) where there was a high carbon content (C = 3.79 %), and in melt No. 3 (N_2 = 0.0091 %) where there was titanium added. Low nitrogen content of these melts was caused by the reduced solubility of N_2 by increasing the content of C and Si, where carbon has a significant effect [12]. There was an unexpected effect, namely high levels of gaseous nitrogen in the electric induction furnace, caused by the large surface area of the many steel scrap fragments (thin metal sheet cuttings). There were nitrogen levels in the charge of 60 to 80 ppm, whereas for the synthetic iron alloy (for melts No. 5 and 6) the levels were 175 and 205 ppm respectively. The change in the HB hardness depending on the wall thickness of the casting was verified by the Stairs test on an R-block (Fig. 1). The results are presented in Fig. 2.

Charge materials content for each melt — Table 1

Charge materials	Melt No.					
	1. (standard melt)	2.	3.	4.	5.	6.
	wt / %					
Steel scrap (sheets)	32.8	35.5	35.5	82.7	97.8	97.5
Return material	53.3	52.9	52.9	-	-	-
PIG iron	10.0	9.8	9.8	13.0	-	-
FeSi 75 %	-	0.12	0.12	0.22	0.59	0.6
FeMn 80 %	0.4	0.5	0.5	0.47	0.4	0.47
Recarburizer	1.30	0.98	0.98	3.45	0.67	0.67
Inoculant (added during reladling)	0.2	0.2	0.2	0.2	0.34+0.2**	0.5
FeTi 60 %	-	-	0.28*	-	-	0.26*
Heat treatment temperature / °C	1 420	1 500	1 500	1 450	1 450	1 520
*FeTi60 – into the furnace **SiC – primary inoculant (into the furnace)						

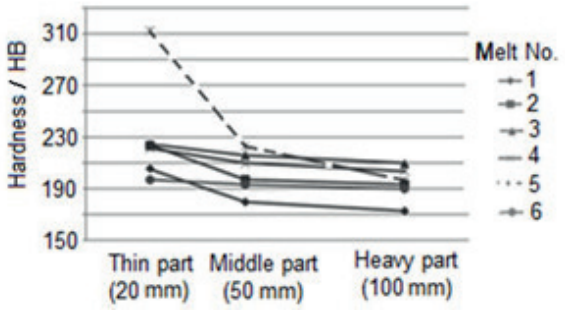

Fig. 2 Change of HB hardness on R – block

The most significant change between the HB hardness of the thinnest (20 mm) and the thickest (100 mm) casting wall was found on a sample from melt No. 5. In this melt, there was also the highest hardness recorded (on average by 243 HB) and the steepest increase in HB hardness in accordance to the decreasing wall thickness of the cast, with a difference of 59 %. It is clear that a high content of steel scrap in the charge, that is an increased nitrogen content, increases the hardness and the tensile strength of alloy. A more substantial hardness increase occurred on the thin wall of the castings where the hardness reached 312 HB for melt No. 5. In contrast, in melt No. 2, where the proportion of steel scrap reached only 35.5 %, there was the smallest variance of hardness between the thinnest and thickest wall of the casting - only 15%.

Thus, the impact of nitrogen in the charge was not as pronounced. Also, a sufficient level of heat treatment temperature (1 500 °C) affected the refinery processes. The results of the measured chill out are documented in Table 3.

The greatest shrinkages occurred in melt No. 5 where shrinkage penetrated to a depth of 15 mm and had the greatest cross section – 30 mm (volume of 1.5 ml). This case also demonstrates the negative impact of nitrogen in the alloys, as a carbonizing element with a tendency to cause shrinkages. In other melts, flat shrinkages occurred.

The microstructure of all melts was pearlitic with a 92 – 96 % portion of pearlit, Fig. 3 - melt No 1. In melt No. 5 (synthetic gray iron) there was a fully pearlitic microstructure and there were carbides detected, Fig. 4. The occurrence of carbides was the reason for the increased hardness in this gray iron alloy.

Depth of the chill out Table 3

Melt No.	Chill out depth / mm
1.	5
2.	1
3.	1
4.	6
5.	10
6.	2

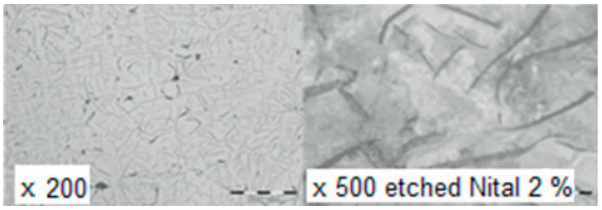

Fig. 3 Microstructure of gray iron alloy in melt No. 1

Results of chemical analysis and of mechanical properties tests Table 2

No.	Steel scrap ratio / %	wt. / %							Sc	CE / %	*R*m / MPa	*H*B
		C	Si	Mn	P	S	Ti	N_2				
1.	33.0	3.23	1.612	0.657	0.024	0.025	-	0.0113	0.848	3.721	297	186 *(205)
2.	35.5	3.33	1.52	0.664	0.024	0.021	-	0.0113	0.868	3.793	352	204 *(223)
3.	35.5	3.32	1.486	0.658	0.023	0.021	0.183	0.0091	0.863	3.773	268	217 *(224)
4.	81.7	3.79	1.026	0.773	0.018	0.012	-	0.0073	0.951	4.103	256	211 *(221)
5.	97.8	3.28	1.69	0.84	0.06	0.061	-	0.0175	0.869	3.805	258	243 *(312)
6.	97.5	3.13	1.61	0.79	0.016	0.011	0.192	0.0205	0.821	3.618	220	197 *(209)

* hardness HB measured on the thin wall of the casting (20 mm)

Besides the standard evaluation of cast iron quality by measuring the tensile strength, and hardness, the quality criteria were also evaluated and compared to chemical composition [13 and 14]:

$$RG = \frac{R\mathrm{m}_{measured}}{R\mathrm{m}_{calculated}} \cdot 100\% \quad (1)$$

where $R\mathrm{m}_{calculated}$ = 1 000 – 800. Sc

$$R\mathrm{H} = \frac{H\mathrm{B}_{measured}}{H\mathrm{B}_{calculated}} \cdot 100\% \quad (2)$$

where $H\mathrm{B}_{calculated}$ = 100 + 0.44 x $R\mathrm{m}_{measured}$

A value higher than 100 % means a high quality of gray iron.

The quality number (*GZ*) or the quality factor (m) is calculated by dividing the *R*G/*R*H or by *R*m/*H*B – measured values. Quality gray iron shows high strength at low hardness values.

The calculated values of the quality criteria are shown in Table 4. The quality criteria values quantitatively describe the effects of particular production conditions on the produced cast iron properties in comparison to the optimal statistically valid conditions. A cast iron alloy of truly high quality should show high maturity values RG and low relative hardness values RH.

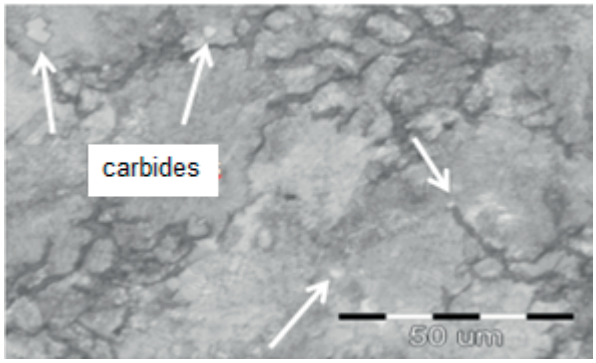

Fig. 4 Microstructure of gray iron in melt No. 5

The quality criteria Table 4

Melt No.	Quality criteria			
	RG / %	RH	m	GZ
1.	94.25	0.805	1.599	117.036
2.	117.4	0.803	1.719	146.166
3.	88.3	0.996	1.235	88.643
4.	109.2	0.996	1.210	109.643
5.	86.2	1.143	1.057	75.427
6.	65.38	1.001	1.117	65.317

Based on the calculated quality criteria (Table 4) it can be stated that melt numbers 1, 3, 5 and 6 are of 100 % maturity level and that their tensile strength is smaller than that of a corresponding sample of such chemical composition. For melt number 2, the heat treatment temperature of 1 500 °C showed its positive influence on the final quality, also reflected by the highest quality number "*GZ*" (146.166), by the maturity level "*R*G" (117.4 %) and also by the highest quality factor "m" (1.719). Melt No. 5 (synthetic iron) demonstrated a high relative hardness "*R*H" (1.143), which reduces the quality number "*GZ*" (75.427). A similar effect was also observed for melt number 6 where a high relative hardness was suppressed by the Ti micro-alloying.

4. Conclusion

The achieved results demonstrate, in particular:

- A significant increase in mechanical properties and quality criteria values for heat treated and inoculated synthetic iron, its low dispersion HB for varying thickness and very good foundry properties.
- A slight decrease in the mechanical properties of cast iron with higher C content and lower Si, but improvement of its mechanical properties, in particular the reduction of HB variance.
- More significant decrease of mechanical properties of cast iron alloyed with titanium, but also significant improvement of monitored mechanical properties, particularly change of hardness for different wall thickness.

The best results were achieved by the heat treatment of liquid metal at 1 500 °C and a high-quality vaccination. The alloy showed the highest tensile strength (*R*m 352 MPa, for melt No. 2) and excellent quality criteria. Melts alloyed with titanium (Ti = 0.2 to 0.3 %) experienced a reduction in mechanical properties, but, on the other hand, showed improved foundry properties, especially a reduced tendency to chill out. Such melts are suitable especially for thin walled castings. There was also an improvement of the casting properties even after the increase of the C content (3.79 % - melt No.4) compared to the current practice (3.1 to 3.3 % C) in GJL-250. The cast alloy exhibited lower hardness differences depending on the wall thickness (a change of 8.73 % from 20 to 100 mm).

A new finding indicates a major presence of nitrogen gasses, in particular if small steel scrap with large surface area is used (small metal sheet cuttings used in the melts No. 6). A double or even a threefold increase in nitrogen amount was observed.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under the grants VEGA No. 1/0216/13.

References

[1] VOZDVIZENSKIJ, V. M.: *Casting Alloys and Technology of their Melting Treatment* (in Russian), Moskva: Masinostrojenije, 1984.

[2] CALLISTER, D. W.: *Materials Science and Engineering,* Department of Metallurgical Engineering : University of Utah, 2009.

[3] COLLINI, L. G., KONECNA, R.: Microstructure and Mechanical Properties of Pearlitic Gray Cast Iron, *Material Science and Engineering*, series A, 2008, 529-539.

[4] KAGAWA, A., OKAMOTO, T.: *Partition of Alloying Elements in Freezing Cast Irons and its Effect on Graphitization and Nitrogen Blowhole Formation*, Osaka, 2011.

[5] WILBERFORS, F., INGVAR, SVENSSON, I.: *The Effect of Nitrogen and Inoculation on the Tensile Properties and Microstructure of Cast Iron Lamellar*, Sodertalje : Jonkopings university, 2010.

[6] PLACHY, J., NEMEC, M.: *Metallurgy and Technology of Casting Alloys* (in Czech), CVUT : Praha, 1986.

[7] HANQI, H., QUJIE, ZHAI: Effect of Nitrogen on Matrix Structure of Gray Cast Iron, *Acta Metallurgica Sinica* (English Ed.), series A, 6, 1993, 370-372.

[8] KUMANIN, I. B.: *Problems in the theory of Foundry Processes* (in Russian), Moskva : Masinostrojenije, 1976.

[9] KERN, R.: *Controlling Cast Iron Gas Chill out, Charge Materials Chemistry, Melting, Practices and Molten Metal Handling Procedures are Root Causes, but Chill out can be Avoided*, Modern Casting, Birmingham : Alabama, 1993.

[10] KERN, R.: *La Maitrise Des Defauts Dus Au Gaz Dans Les Fontes*, La Technique, Hommes Et Fonderie, 1994.

[11] ROUCKA, J.: *Metallurgy of Cast Iron* (in Czech), PC - DIR : Brno, 1998.

[12] JOHNSON, B. D., HEINE, R. W.: *Nitrogen in Molten Iron Processing and Its Effects*, AFS Transactions, 81, 2000, 163-169.

[13] RAFAEL, N.: *Metals HandBook*, Gray Iron, 15, 1988.

[14] FUTAS, P., PRIBULOVA, A.: *Quality Criteria as Implement for Advisement of Metallurgical Quality of Grey Iron*, Evaluation of People and Products Features, Celje : University of Maribor, 2014, 99-118.

Jaroslav Smutny - Lubos Pazdera – Vladimir Tomandl - Herbert Seelmann*

ASSESSMENT OF DYNAMIC PARAMETERS OF RAIL FASTENING

The paper is devoted to the analysis and comparison of the dynamic parameters of the rail fastening used in a standard way in the core railway network of the Czech Republic, i.e. the elastic rail fastening Vossloh W 14 and Pandrol FC. The paper also includes the description of the measurement and evaluation method. The conclusion of the paper is devoted to recommendations for further measurement and practical applications.

Keywords: *Rail fastening, signal analysis, impulse excitation, frequency response function.*

1. Introduction

The constant increase in demands on the load capacity of railway structures leads to greater stiffness of construction layers and subgrade. Concrete sleepers are also almost exclusively used, because have much higher bending stiffness and less flexibility than wooden sleepers. All these factors lead to greater loading on track bed, which changes its shape under increasing load and thereby affects the track geometry parameters. In the ballast track, for the reasons described above, gravel bench behind the heads of sleepers tends to crash. Open spaces between the sleeper and gravel are created, due to the high bending stiffness of concrete sleepers. When a train passes the sleepers rest on the collapsed gravel bench, which creates uneven sleeper support and enhance the dynamic effects, which accelerates the degradation of track bed [1].

To avoid such undesirable phenomena as much as possible, considerable financial resources must be expended to new designs research, as well as to diagnostic quality of track geometry and their potential maintenance [2]. Still, there are some relatively simple measurement procedures, which can well predict the characteristics of the rail fastening and thus also its suitability. The aim of this paper is to point out the use of selected methods of experimental modal analysis to determine the characteristics of rail fastening.

2. Methods of measurement

The basic objective of methodology was comparison of rail fastening test samples in a laboratory with respect to the vibration response of the results obtained in situ. The method of excitation with impulse hammer was selected. The measured data were analysed using the Frequency Response Function (FRF) [3]. The dynamic responses were measured on the foot rail, on the head of the sleeper and in gravel ballast (except in-situ measurement).

A part of a sleeper with installed fastening system was used as a test sample in the laboratory. The sleeper was placed in gravel bed in shape of a truncated pyramid, which was built in a test box with dimensions of 2.0 x 2.0 m. The test box was insulated from the base plate by an interlayer of cork.

Furthermore, additional measurements were carried out in situ. For these purposes was selected a test section on line no 260 in the km 166.140 in the area of Bilovice nad Svitavou [4]. For obvious reasons, the exciting with impulse hammer was used in situ too.

3. Test samples

In the laboratory, the following types of baseplatesless flexible rail fastening for system UIC 60 and gauge extension + RK 0 mm were tested [5 and 6]:

1. The part of the sleeper B91P with Fastclip fastening and side insulator 7049
   - The fastening Pandrol (flexible clips, FC1501 with insulators 8494, pad under bottom rail 6530)
2. The part of the sleeper B91S/1 with polyamide opened plug
   - The fastening system Vossloh W14 (flexible clamps Skl14, angled guides Wfp 14K-12, sleeper screw R1 with pad Uls 7, pad under bottom of the rail WU 7)

* [1]Jaroslav Smutny, [1]Lubos Pazdera, [1]Vladimir Tomandl, [2]Herbert Seelmann
[1]Institute of Railway Structures and Constructions, Faculty of Civil Engineering, Brno University of Technology, Brno, Czech Republic
[2]Rail Data Services Austria GmbH & Co KG, Wien, Austria
E-mail: smutny.j@fce.vutbr.cz

- The fastening system Vossloh W14NT (flexible clamps Skl14, angled guides Wfp 14K NT, sleeper screw R1 with pad Uls 7, pad under bottom of the rail Zw 900 NT)
- The fastening system Vossloh W21NT (flexible clamps Skl21, angled guides Wfp 21K NT-12, sleeper screw R1 with pad Uls 7, pad under bottom of the rail Zw 1000/150 HS-35)
- The fastening system Vossloh E14 (flexible clamps Skl 21, angled guides Wfp 16G/Wfp 16F, sleeper screw R3 with pad Uls 7, pad under bottom of the rail Zw 693/150, steel spread plate Grp 22/150, highly flexible pad Zwp E14 NT/150).

Verifying measurements in situ were performed on fastening Vossloh W14 in track No. 1, or on fastening Pandrol FC I in track No. 2 of the given section of the track. The rail fastening Pandrol FC I and especially Vossloh W14 are the most commonly used without-clip-plate flexible fastenings in the Czech Republic. These can be found in all the upgraded sections of transit railway corridors and tracks of national importance.

Fastening W14NT is equipped with angled guide inserts Wfp 14K NT provided with a special guide rail on the underside. This guide rail in the mounting position enters the area under the track base, avoiding excessive tipping or decline of the track. For this reason, the fastening W14NT is preferably used in track with directional arcs of small radii. Fastening E14 belongs among the fastenings with increased vertical flexibility.

Fastenings E14 is suitable where there is a need to increase the damping of dynamic effects or reduce the emitted noise. It was originally developed for slab track where highly elastic pad Zwp E14 NT/150 replaced the dampening effects of the entire track bed. Currently the trend for fastenings E14 can be observed also in places where it is not possible to establish a sufficient track bed thickness. It is a fastening with a highly elastic underlay under the track base Zw 1000/150 HS-35. This increased flexibility is also consistent with vertical higher fatigue limit of clamps Skl21. Field of application of fastening W21NT is similar to fastening W14NT or E14.

4. The test arrangement

Laboratory samples were fitted with rail grid vibration acceleration sensors as shown in Fig. 1. From the perspective of the cross section all sensors were placed in the longitudinal axis of the sleepers. Location of sensors *Ar* and *As* was due to the rail web chosen conceptually in accordance with previous measurements [7 and 8]. A special measuring stone was developed at the Institute of Railway Structures and Constructions, due to the location of the sensor *Agb* to the gravel beneath the mattress platform of sleeper. The *Agb* sensor was not used in measuring in-situ. The in-situ test arrangement is shown in Fig. 2.

Suitable Bruel&Kjær accelerometers (type BK 4507B004) were used as sensors. Impact was driven with the impact hammer Bruel&Kjær, type 8210, with excitation tip hard tip black. Data from measurement were analysed using PULSE 3560D modular analyser from Bruel & Kjær.

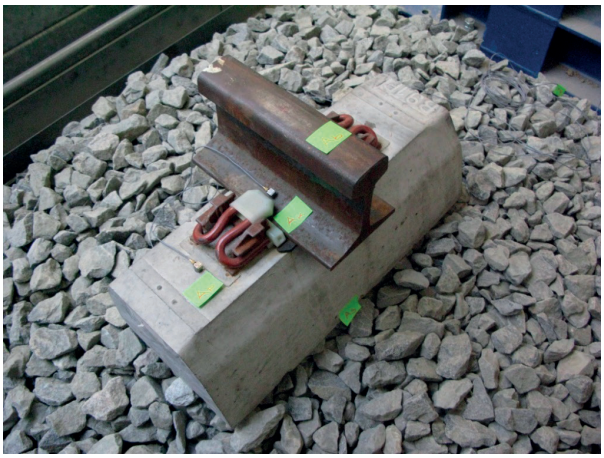

Fig. 1 The test arrangement in-labo, impact excitation

Fig. 2 The test in-situ, impact excitation

5. Evaluation of the data

The results are partially shown in Tables 1 - 2 and Fig. 3. Assessed value of impulse excitation was the accelerance. Frequency range 1 ÷ 1000 Hz was set after the initial testing. In the case of impulse evaluation three repeated laboratory measurements of accelerance were performed for each type of rail fastening, each derived from ten hammer hits. Impulse hammer was used also in the case of the verification in situ. Frequency and survey methodology of accelerance measurement was similar to that in the laboratory. However, in order to guarantee safe

operation of the track, the influence of the tightening of the propeller was not measured at Vossloh fastening.

The results of individual measurements were averaged and displayed in graphs of frequency response functions H3. Relative damping (damping ratio) of tested structures was determined by half-power bandwidth method. Evaluation was done in software ME'Scope. It is worth noting that the Tables 1 and 2 contain full results for rail fastenings Pandrol FC I and Vossloh W14 only.

The evaluation focused on the low-frequency range of 1 ÷ 80 Hz, medium-frequency range 80 ÷ 400 Hz and high-frequency range of 400 ÷ 1000 Hz. This division is based on previous work and is based on the principle that in the low frequency range mainly the effects of the operation of bogies and wheelsets of moving railway vehicles can be observed, middle frequency range is a manifestation of imperfections on the wheel rail contact area, as well as a high frequency range up to 1000 Hz. Here, however, the influence of acoustic phenomena caused by moving vehicles is also manifested.

In the low-frequency range all fastening systems show a significant response between 60 and 80 Hz. The most significant relative damping on the rail 5.3% is achieved with Vossloh fastening W14NT, the smallest 0.8% with fastening Vossloh E14. High damping ratio of fastening W14 NT is most likely caused by the guide rail on the bottom surface of the angular guide insert that prevents larger rail movements. On the contrary, the fastening E14 is characterised by an increased vertical flexibility, which may result in increased vibration of the rail. On the sleeper there was the largest structural damping 9.3% recorded at fastening Vossloh W21NT, the smallest at Pandrol FC I and at Vossloh W14NT (both 3.1%). The difference between the damping on the rail and on the sleeper can indicate vertical fastening flexibility. Vossloh fastening E14 W and Vossloh W21NT can therefore be defined as the fastening with largest vertical flexibility. The resulting damping 4.3 ± 2.2% in the track bed suggests that the densified layer of gravel created stable surface for all tested sets.

There are no significant frequencies for fixing Pandrol FC I on the rail at the medium frequency range, except for frequency 277 Hz in situ, where the tested structure shows a damping of 3.1%. This frequency is equivalent to a frequency of 430 Hz, which was recorded in laboratory tests. On the sleeper there are significant frequency ranges of fastening Pandrol frequencies of 145 ÷ 165 Hz, where it reaches an average damping of 5.3%. Greater damping at high frequencies in the track bed was not reached. Neither at Vossloh rail fastening W14 significant frequency was found. This fastening achieves an average damping of 3.7% on sleeper In this respect, the range of 110 ÷ 215 Hz is particularly important. Damping of 6.4% at frequencies of 115 ÷ 155 Hz in gravel was calculated. Fastening W14NT shows damping 4.0% on the rail, 3.4% on the sleeper and 4.2% in the track bed at selected frequencies in the range 110 ÷ 340 Hz. For fastening Vossloh W21NT the essential resonance frequency is 290 Hz. Fastening here shows damping of 7.3% on the rail, 6.3% on the sleeper and 7.4% in the gravel. Another frequency spectrum of 130 ÷ 145 Hz is not so important; the construction here achieves damping 4.5%. The last tested construction Vossloh E14 shows a high resonant range 120 ÷ 130 Hz with damping 3.6% on the rail, 3.1% on the sleeper and 3.9% in the track bed. Even more significant damping of about 5.2% was obtained at a frequency of 220 ÷ 265 Hz. Overall, the comparison of the behaviour of structures reveals that with increasing vertical fastening flexibility the size of the resonance range of medium frequencies of 80 to 400 Hz increases as well. The least favourable in this respect seems to be W21NT Vossloh fastening, fixing Pandrol FC I, or Vossloh fastening E14.

For fastening Pandrol FC I the frequency 430 Hz is important in the higher frequency range of 400 ÷ 1000 Hz, where the average damping is 2.6% and in track bed also frequencies 630 ÷ 730 Hz with the calculated damping of 1.9%. For fastening Vossloh W14 damping was found at the large number of high-frequency range. Only the values of range 450 ÷ 500 Hz and 600 ÷ 700 Hz are, however, higher than 1.0%.

For fastening Vossloh W14NT damping was found in the frequency range 510 Hz to 600 Hz in the average of 1.5%.

The relative damping in dependence on the frequency, Pandrol FC I fastening — Table 1

Pandrol FC I fastening				
Sensor	Laboratory, Impact excitation		In situ, Impact excitation	
	f [Hz]	Damping [%]	f [Hz]	Damping [%]
Rail Ar	69	3.0	83	3.3
	167	1.3	277	3.1
	430	3.1	359	0.4
	573	0.1	518	0.6
	650	0.1	774	0.5
	701	0.1	931	0.1
Sleeper As	73	4.0	64	2.1
	114	1.7	147	7.7
	165	2.6	278	4.6
	422	5.6	591	0.1
	651	0.1	773	0.1
	874	0.1	798	0.1
Gravel ballast Agb	70	2.8	-	-
	89	2.8	-	-
	198	2.4	-	-
	222	1.3	-	-
	437	5.3	-	-
	733	4.5	-	-
	850	0.2	-	-

Internal damping of 3.3% of the corresponding frequencies 410 ÷ 445 Hz was also found in track bed. The average damping value of 1.9% was calculated for Vossloh fastening system W21NT. This damping was achieved at frequencies of 415 ÷ 445 Hz. The second major frequency is 562 Hz, where the damping of 3.3% was found. Fastening E14 showed in the selected frequency range 400 ÷ 1000 Hz no significant damping values exceeding 0.5%. In the range of higher frequencies all tested fastenings show the diminished ability of vibration damping than in the range below 400 Hz. The least favourable of all tested structures seems to be Vossloh fastening W 14NT that in 510 Hz to 600 Hz shows the highest amplitude of the vibration acceleration on the sleeper, hence in the track bed.

From the waveform of frequency response functions it is possible to determine the characteristics indicative for the test of rail fastening in terms of their mounting rigidity and in terms of relative acceleration of the rail towards the sleeper in a vertical direction by quantifying surface between curves A_k and A_p. The presence of the guides on the underside of the guide angle insert fastenings W14NT and W 21NT using low frequency range up to 80 Hz which causes a very small area between the given curves in this frequency range - relative vertical vibration acceleration transmission from the track to the sleeper is minimal.

In the overall frequency scale, the largest differences between the curves A_r and A_s are at fastenings E14, the smallest at fastenings W14. Measurement tentatively confirmed the static cutting stiffness of tested sets of rail fastenings listed in company regulations.

Fastening Pandrol FC I shows interesting properties in the resonant frequency range of around 430 Hz. This frequency is a manifestation of corrugation with a period of waves about 7 ÷ 9 cm, depending on the driving speed, which is considered in the range typical for corridor route, i.e. 120 to 160 km·h $^{-1}$. Using the fastenings Pandrol FC I on routes where there is emergence of corrugation with that wavelength, can accelerate the development of rail defects. Similar characteristics are also shown by fastenings Vossloh W21NT at high frequencies around 290 Hz or by Vossloh fastenings E14 at frequencies in the range of 220 ÷ 265 Hz.

At the tested fastenings with maximum static secant stiffness Vossloh W14 and Vossloh W14NT there is the most noticeable shift of resonant range towards higher frequencies. The presence of waves of shorter wavelengths up to 5 cm can be seen on the corridor lines in the high frequency range. Therefore it follows that in areas where we can expect the development of periodic defects of the rail surface of small and medium-sized lengths it is, due to general properties, suitable to use vertical flexible fastening systems, such as Vossloh E14 or Vossloh W21NT. On the contrary, in places where we can expect rail defects like gliding or long waves, it is preferable to use fastenings with higher static secant stiffness of the rail pad. These include fastening system

The relative damping in dependence on the frequency, Vossloh W14 fastening Table 2

Vossloh W 14 fastening				
Sensor	Laboratory, Impact excitation		In situ, Impact excitation	
	f [Hz]	Damping [%]	f [Hz]	Damping [%]
Rail Ar	72	2.3	-	-
	128	2.3	119	1.0
	497	0.4	595	0.6
	646	0.2	752	0.1
	772	0.3	781	1.0
	830	0.2	802	1.8
	905	0.1	906	0.1
Sleeper As	74	6.1	-	-
	96	3.2	149	6.2
	112	3.7	305	0.3
	215	3.1	473	0.1
	499	6.9	583	0.1
	633	0.5	792	0.1
	772	0.3	803	0.1
	831	0.2	-	-
	834	0.1	-	-
Gravel ballast Agb	40	1.6	-	-
	72	4.5	-	-
	85	2.2	-	-
	146	4.2	-	-
	222	0.3	-	-
	500	0.4	-	-
	699	0.2	-	-
	751	0.1	-	-
	821	0.1		

Vossloh W14, whether made with a guide rail on the underside of the angular guide insert (type NT) or without.

6. Conclusion

Carried out laboratory and operational measurements showed stable performance of all tested types of fastenings for the range of testing. Dependence on the size of the static cutting stiffness and resonant field of fastening systems was revealed. The firmer the fastening the more is the range of the most important frequencies shifted towards higher frequencies.

From the above it can be concluded that, for example, in the Czech Republic operationally validated fastening type Vossloh

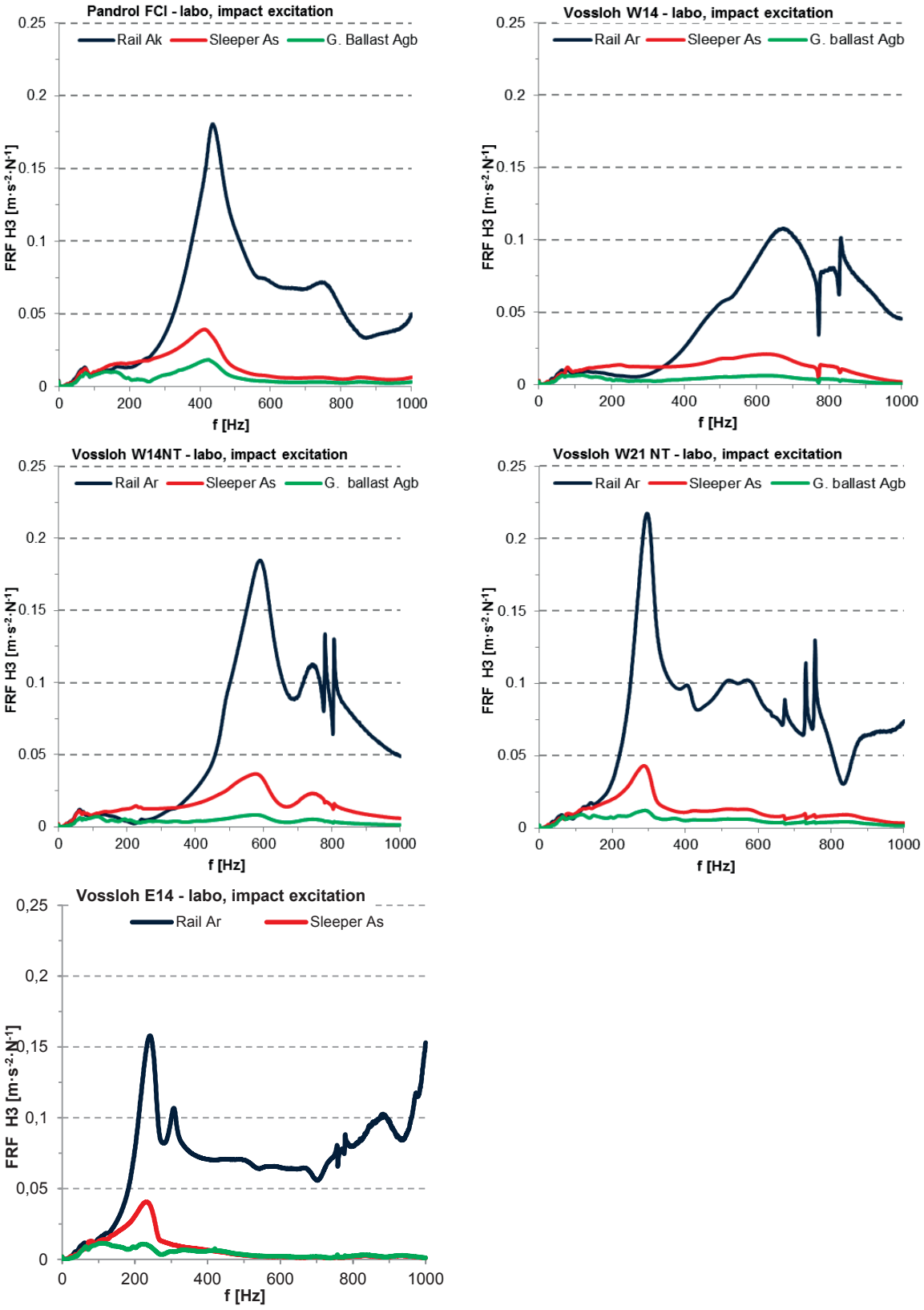

Fig. 3 Frequency characteristics of the testing samples rail fastening, standard tightening torque

W21NT cannot be recommended to bends with small radii, where there is a risk of emergence of slip waves. These waves are at speeds of 120 to 160 km·h^{-1} significant contributors to the level of acceleration in the medium frequency range of 80 ÷ 400 Hz. For its properties, fastening W21NT could therefore contribute to faster development of the rail defect.

However, based on the data obtained, we can confirm the suitability of each type of tested fastening for the area of application described in the section "test samples". Comparing the results of laboratory measurement with measurements in the field, it could be generally stated that in the in situ measurement a downward trend of significant medium frequency ranges on the rail and on the sleeper of approximately 100 to 150 Hz toward the beginning was apparent. This was true for both compared the fastenings Pandrol FC I and Vossloh W14.

The cause of this phenomenon has not been shown by measurements. Possible connection is most likely to be seen in the continuously supported rail and the effect of axial forces and stresses in continuous welded rail, which was not taken into account in measurements under laboratory conditions. Slightly different characteristics of both analysed fastenings in higher frequency ranges are then probably also related to the generally complex ties rail - sleeper - rail bed in situ, which cannot be fully simulated under laboratory conditions.

Based upon the analyses made it is possible to say that the used methods offer good results and conclusions. The measured and calculated values prove to be sufficiently accurate and have excellent testing ability.

Acknowledgments

This paper has been supported by the research project FAST-S-15-2806, "The analysis of dynamic response of the railway line structure".

References

[1] ESVELD, C.: *Modern Railway Track*, MRT Production: Delft, ISBN 90-800324-3-3, 2001.
[2] SMUTNY, J., PAZDERA, L.: New Techniques in Analysis of Dynamic Parameters Rail Fastening, *INSIGHT*, ISSN 1354-2575, 2004.
[3] SMUTNY, J., TOMANDL, V., PAZDERA, L.: Dynamic and Acoustic Analysis of Elastic Rail Fastening without Sole Plates, *Civil Engineering J.*, ISSN 1210-4027, Praha, 2009.
[4] SMUTNY, J., PAZDERA, L.: *The Experimental Analysis of Dynamic Processes Related to Railway Transport*, Monography, ISBN 978-80-7204-827-4, Academic publisher CERM: Brno, 2012.
[5] Czech Technical Standards series CSN EN 13481-1 till 5 and CSN EN 13481-7. Railway applications - Track - Performance requirements for fastening systems, 2013.
[6] Czech Technical Standard CSN EN 13146-8. Railway applications - Track - Test methods for fastening systems 2013.
[7] MORAVCIK, M.: Analysis of Vehicle Bogie Effects on Track Structure-Nonstationary Analysis of Dynamic Response, *Communications - Scientific Letters of the University of Zilina*, ISSN 1335-4205, 2011.
[8] SMUTNY, J., SADLEKOVA, D.: The Vibration Analysis by Margenau-Hill Transformation Method, *Communications - Scientific Letters of the University of Zilina,* ISSN 1335-4205, 2014.

Jan Moravec*

CUTTING METAL SHEET WITH A PUNCH WITH INTERNAL HEATING

This paper describes issues related to the process of cutting with a punch with internal heating. The theoretical part explains the effects related to thermal expansion, conduction and transfer of heat in the body of punches. The experimental part refers to the design of a tool with an internally heated punch. Experimental results are summarised and the process is analysed in the discussion.

Keywords: *Metal sheet, internal heating, cutting, cutting tool, forming.*

1. Introduction

Forming technologies have long used various ways and methods based primarily on physical knowledge. For a longer time, the method has been used of deep drawing using heating the active parts of drawing tools. The first person to describe this method was Romanovskij in 1968 (p. 191, and he published the process picture No. 114 B) [1]. In 1965, Srp, et al. published this solution (p. 132, Figure 199) [2]. This method is also presented by Cabelka, et al. in 1967 (pp. 261-262, Figure 3.509) [3], Blascik, et al. in 1988 (pp. 297-298, Figure 1298) [4] and Baca, in 2000 (p. 131, Figure 108) [5]. This paper originated on the basis of the above information. We will discuss the possibility of using internal heating in metal sheet cutting using a cutting machine. In the past we have already published a paper on the propagation of heat in the cut zone, in which we dealt with this issue - Moravec in 2000 [6].

2. Theoretical part

First approximation of the problem

If a cylindrical cutting punch is considered a rod clamped at one end, then a first approximation can read as follows: rods, when heated, expand especially in length. When heating, the amplitudes of oscillating molecules increase and the molecules fill a larger space. Therefore, solids stretch in all directions. A cylinder-shaped cutting punch can be considered the surface area S_1, which is the area prior to a change in temperature. When performing calculations, we can understand surface expansion as linear expansion in two directions/dimensions (Fig. 1) [7].

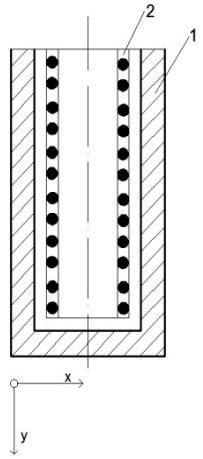

Fig. 1 Scheme of a cutting punch with internal heating
1 – punch body, 2 – heating body,

The variables are as follows: S_1 – the surface area prior to temperature change, S_2 – the surface area following temperature change, ΔS – the surface area difference = $S_2 - S_1$, Δt – the temperature difference = $t_2 - t_1$.

The following also applies:

$\Delta S = S_2 - S_1 = l_2^2 - l_1^2 = l_1^2(1 - \alpha \Delta t^2) - l_1^2 = l_1^2[1 + 2\alpha\Delta t + \alpha^2(\Delta t)^2] - l_1^2$

Since α is small, the member with the square number in the equation can be neglected. Then, the following will apply:

$\Delta S = l_2^2(1 + 2\alpha \Delta t) - l_1^2 = l_1^2 2\alpha \Delta t$, and also: $\Delta S = S_1 2\alpha \Delta t$,

* **Jan Moravec**
Department of Technological Engineering, Faculty of Mechanical Engineering, University of Zilina, Slovakia
E-mail: jan.moravec@fstroj.uniza.sk

as well as: $S_2 = S_1 + \Delta S = S_1 + S_1 2\,\alpha\,\Delta\, t^2$, and finally:
$S_2 = S_1 (1 + 2\ \alpha\ \Delta\ t^2)$.

Note: *The coefficient* α *depends on temperature to a very limited extend only. Within the range 0 to 100 °*C*, the values apply as given in physical tables, and with sufficient accuracy.*

Heat conduction and transfer in a rod

The rod is made of a mono-material, and is of a constant cross-section. The base of the rod has heat conducted in a way that at all points of the initial cross-section the temperature T_a (Fig. 2) is the same. The rod is situated in an environment with the same unchanging temperature T_0. Assuming that the material thermal conductivity coefficient as well as the heat transfer coefficient are constant all across the rod surface, we are to determine temperature distribution over the rod length, and determine heat transfer [8 and 9].

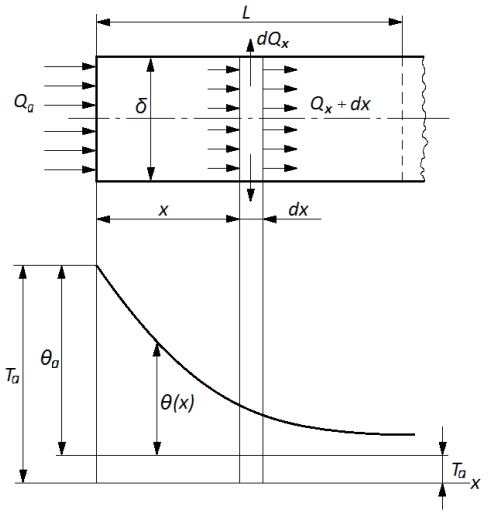

Fig. 2 Diagram for determining heat conduction and transfer through the rod

Solving can be done assuming a one-dimensional heat flow. This assumption will be all the closer to reality, the lower will be the following values:

Biot number for a rod ($\left(Bi \equiv \frac{\alpha\delta}{\lambda} \ll 1\right)$, *α* is the heat transfer coefficient from the rod surface into the environment; *ô* – the rod thickness; λ – the thermal conductivity coefficient of the rod material).

Biot number represents the ratio of values $\frac{\alpha}{\lambda/\delta}$, which are characteristic of the intensity of heat removal from the rod surface into the environment, compared to heat input intensity to the rod surface from the rod material. The lower the *Biot number*, the lower heat removal from the cooled body surface, compared to a possible heat supply from the inside. Due to the above facts, for *Bi* « **1** we can approximately consider, when solving heat conduction in a rod, that the temperature gradient in the rod cross section be sufficiently small compared to the temperature gradient along the rod axis. Using this assumption/condition, we can execute the equation of rod heat conduction equation solving as a task with one-dimensional heat flux.

We execute an elementary layer of the thickness *dx*, and for this rod thickness we write an equation of heat balance. Heat flux in the elementary layer changes under stationary conditions due to the heat transfer on the outer surface:

$$Q_x = Q_x + dx + dQ_x \quad \text{or} \quad S\lambda\frac{d^2\Theta}{dx^2}dx = u \cdot dx \cdot \alpha \cdot \Theta,$$

where $\theta = T(x) - T_0$ is the temperature difference between the rod and the environment.

In this way we arrive to a differential equation of heat conduction along a rod:

$$\frac{d^2\Theta}{dx^2} - m^2\Theta = 0$$

where $\frac{d^2\Theta}{dx^2} - m^2\Theta = 0$ is the equation's dimensional parameter.

The integral of this differential equation is in the form: $\Theta(x) = C_1 e^{+mx} + C_2 e^{-mx}$.

3. Experimental part

A cavity is formed inside the cutting punch body. A heating body is placed into the cavity. The heating body heats the face of the cutting punch. It was necessary to find out what parameters correspond to the required temperature. A 40W tiny heating body (device) from a micro-solder proved to be suitable for experimental work. As shown in Fig. 3, we connected the assembly to a 230V source, which ensured heating of the active part of the cutting punch. Figure 4 shows a view of the experimental workplace of temperature sensing. The ambient temperature was 26 °C. We found out the following: at *U = 116 V*, and after 25 minutes, we achieved the temperature of 97 °C, and after 40 minutes the temperature increased to 112 °C – the value that had to be obtained (to be more precise – we needed to achieve temperature between 100 and 110 °C). The real temperature in experimental conditions varied between 108 and 113 °C. We chose this temperature to avoid initiation of phase transformations. The voltage U_{term} *= 4.435 mV →t = 110 °C,* U_{term} *= 4.280 mV →t = 105.7 °C* U_{term} *= 4.404 mV →t = 109 °C*. The final values were as follows: *U = 116 V*, current *I = 78 mA,* U_{term} *= 4.404 mV.*

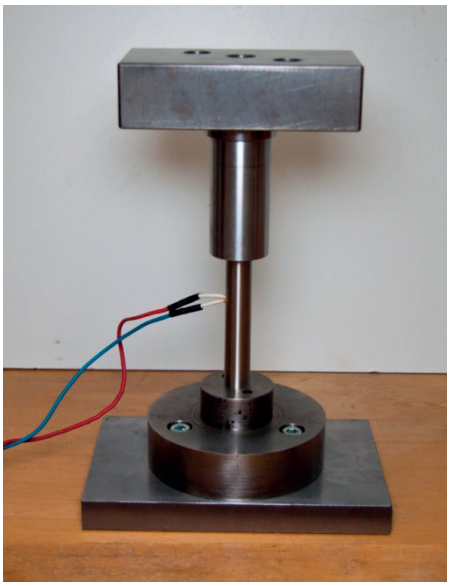

Fig. 3 Assembly of the experimental cutting tool

Prior to heating, the cutting punch diameter was 15.91 mm, and after heating it reached 15.922 mm, which was measured directly and does not differentiate significantly from the calculated value. This laboratory verification was followed by incorporation of the modified cutting punch with internal heating into the cutting tool assembly. Since it is an open tool without any cutting punch guidance, we had to precisely adjust the tool and consistently clamp it on the press. Figures 5 and 6 show the experimental cutting tool and its clamping on the press.

The presented tool was clamped on a LESP 63 press, and used to cut 1-mm metal sheet. The metal sheet was made of conventional STN 41 1321 material, with the composition of *0.10% C, 0.45% Mn, 0.035 %.P, 0.035% S.* The other values included R_m = 420 MPa, R_e = 235 MPa, A = 26%. The number of blanks was chosen in thousands of pieces. The cutting punch was heated to 108 °C (± 5°). From this perspective, we can assess the cutting process with heating as stable. Attention was focused on the effect of heating during the cutting process.

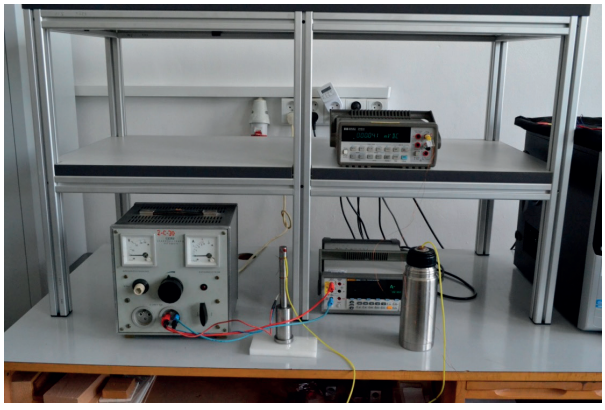

Fig. 4 Experimental workplace

Fig. 5 Workplace for cutting

Technological aspect

Using a heated cutting punch, we identified some factors in experimental work that are instrumental in deeper understanding of the above issues. The condition is that heating does not replace the forming tool sharpening, alignment or repair, but only helps prolong the phase of economic wear (Fig. 6). The size of the production batch is of significant influence in the manufacture of blanks (Fig. 7) [10, 11 and 12].

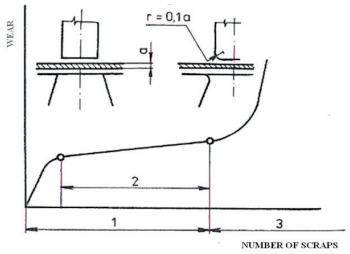

Fig. 6 Criterion of the cutting edge wear
1 - sharp tool, 2 - economic wear, 3 - supercritical wear

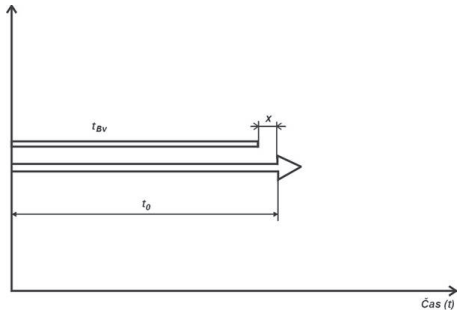

Fig. 7 Optimum distribution of continuous production time and tool operation time

Figure 7 shows the optimum distribution of the continuous blank production time and the quality tool operation. The difference x enables re-sharpening and re-aligning of the tool in this section of time value difference. Therefore, it is important to determine the optimum size of the production batch, whereas it should apply that during the time of manufacturing a certain batch, the tool should ensure the quality of parts manufactured in the given batch:

$$t_{Bv} \leq t_o ,$$

where:

t_{Bv} – the time for manufacturing a batch with n pieces,
$t_o = t_{pn}$ the time of quality operation of the tool between two acts of re-sharpening or adjustment.

Condition: the degree of precision of the manufacturing process is directly dependent on the quality and precision of the forming tool.

When applying a heated punch in the *proces*s of metal sheet cutting, we have to consider what can be achieved with this solution. The following applies generally to sheet cutting, which is a common operation in the preparation of blanks, such as bending or deep drawing: cutting tolerance plays a major role. This tolerance is directly dependent on the thickness and mechanical properties of the material being cut. Cutting gap (*CG*) is half the cutting tolerance. Optimal *CG* is an important measure of quality of the cutting tool, and it contributes significantly to the overall quality of the desired result of the cutting process.

The geometry of punches and punching dies includes the shape, dimensions and tolerances of the working parts of the cutting tools. The following applies: the size of the punched hole determines the punch dimensions, and the blank size determines the punching die and punch dimensions. In hole-punching operations (openings and blanks), the part of material that falls from the opening (waste) determines the size of the cut-out hole, whereas the size of the punch and punching die is enlarged by the cutting tolerance. The size of a blank resulting from the cutting operation depends on the punching die size, therefore the cutting gap size is on the expense of the cutting punch. The cutting punch size will be smaller by the cutting tolerance value. Hence, it follows that the cutting tolerance is the difference in dimensions of the punching die and the punch in the corresponding point of cross-section. It is the distance of the punch from the punching die after its insertion at any point. The ideal cutting gap is uniform all along the punching die (punch) circumference, and is equal to half the cutting tolerance (Fig. 8).

Optimum cutting gap is an important measure of quality of cutting tools. It has a decisive influence on the durability of the cutting edge and the service life of the tool. Its size depends mainly on the thickness and mechanical properties of the material to be cut. An optimum cutting gap is considered the one that is uniform along the entire cut curve and that achieves the required quality of the cutting surface with a minimum exercised force and labour. A reduction in the cutting gap leads to an increase in cutting force as well as labour. Increase in the force is not large, but increase in labour can reach up to 40 %.

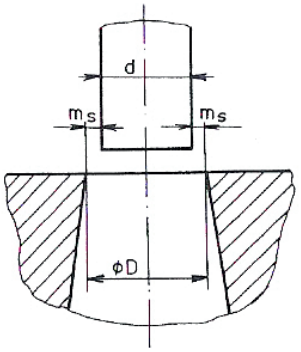

Fig. 8 Cutting gap and cutting tolerance

Since the size of the cutting gap changes during the operation due to the cutting edge wear, new tools are made with minimum acceptable cutting gap. Cutting edges gradually wear out, the cutting gap grows, and when it exceeds the region of economic wear, the forming force increases by 55%. It applies for the upper limit of the economic wear region that the cutting edge radius R = 0.1 a, where a is the thickness of the material being cut (Fig. 6). Such wear is followed by a necessary renovation of cutting edges [13 and 14].

However, the *cutting gap size* can be determined also theoretically from the following formula:

$$m_s = (a - h_v)\, tg\, \beta$$

where:

h_v – the indentation depth; $h_v \approx$ (0.2 to 0.5) a,
β – the cutting surface angle relative to the direction of the cutting force acting, which has the following values according to the properties of the material being cut: mild (soft) steel β = 5 to 6°, medium-hard steel β = 4 to 5°, and hard steel β = 4°.

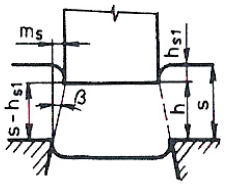

Fig. 9 Determining the cutting gap size

According to Fig. 9, the punching die profile is characterised by relief grinding (clearance) by the angle α. This profile is used in blanks with lower shape accuracy requirements and in larger metal sheet thickness.

It follows from the above considerations that using a heated punch possibly results in quite properly corrected size of the cutting tolerance. As mentioned above, the increase in applied force at reduced cutting gap is small, but the labour increment is significant. This can be easily read in the graph in Fig. 6. Therefore, we are dealing with the maximum length of section *2* that represents the economic wear. The amount of labour in this section increases rather gradually, not in leaps. Certainly, there is also a significant impact of the amount of heating of the active part of the cutting tool.

When applying a heated cutting punch, one must consider such a structure as early as in the design stage of the forming tool, with the economic aspect of the process coming to the fore here. According to Fig. 10, the appropriate cutting tolerance is exceeded at a certain stage of the cutting process. And it is exactly at this moment that a heated cutting punch can be applied so that the acting part of the forming tool gets heated and quite simply corrects the cutting tolerance in order to ensure that the process runs as if with a re-adjusted tool and corrected cutting tolerance. Using this solution can be generally recommended in the cases when the cutting tolerance increases to the extent that the blanks feature visible burrs with a height exceeding the permissible level of cut tolerance. The size (dimensions) of cutting punches also plays its role here, since the larger the size, the larger the linear expansion coefficient. Of course, the important in the process is only the direction in the *x* axis, elongation in the *y* direction does not affect the cutting process (Fig. 1).

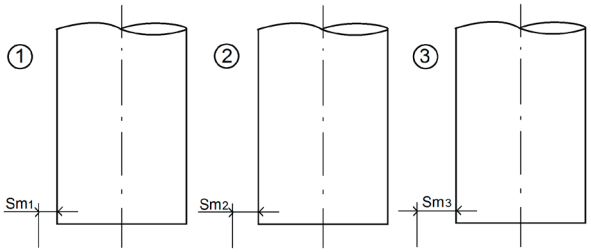

Fig. 10 Change in the cutting gap resulting from the cutting punch wear

$s_{m1} < s_{m2} > s_{m3}$

Quality of cutting surface

Field tests did not find any significant difference on the cutting surfaces between the quality of the surface being cut traditionally (without any heating), and with the applied heating of the active part. The above facts had been expected, and the assumption was fully confirmed. Surface roughness measurement was carried out using MITUTOYO Surftest. The surface roughness was R_a = 0.77 µm (Fig. 11).

4. Discussion

Heat conduction and transfer in a hollow cylinder with heat sources

In a hollow cylinder body there are evenly distributed heat sources with a specific output q_{zdr} [W/m^3]. The wall of the cylinder transfers heat from the inside and the outside to the environment that is of unchanging (in time) temperatures T_1 and T_2. To solve this problem we need to use the following boundary conditions:

For internal heat exchange at $r = d_1/2$:

$$T_{d_1/2} = -\frac{q_{zdr}}{\lambda}\frac{1}{4}\left(\frac{d_1}{2}\right)^2 + C_1 \ln\frac{d_1}{2} + C_2.$$

Due to the boundary condition:

$$-\left.\frac{dT}{dr}\right|_{r=d_1/2} = -\frac{\alpha_1}{\lambda}\left(T_{d_1/2} - T_1\right)$$

the equation is:

$$-\frac{q_{zdr}}{\lambda}\frac{1}{2}\frac{d_1}{2} + C_1\frac{1}{d_1/2} =$$
$$-\frac{\alpha_1}{\lambda}\left[C_2 + C_1\ln\frac{d_1}{2} - \frac{q_{zdr}}{\lambda}\frac{1}{4}\left(\frac{d_1}{2}\right)^2 - T_1\right]$$

For external heat exchange at $r = d_2/2$:

$$T_{d_2/2} = -\frac{q_{zdr}}{\lambda}\frac{1}{4}\left(\frac{d_2}{2}\right)^2 + C_1 \ln\frac{d_2}{2} + C_2.$$

Due to the boundary condition of the third kind:

$$-\left.\frac{dT}{dr}\right|_{r=d2/2} = -\frac{\alpha_2}{\lambda}\left(T_{d_2/2} - T_2\right)$$

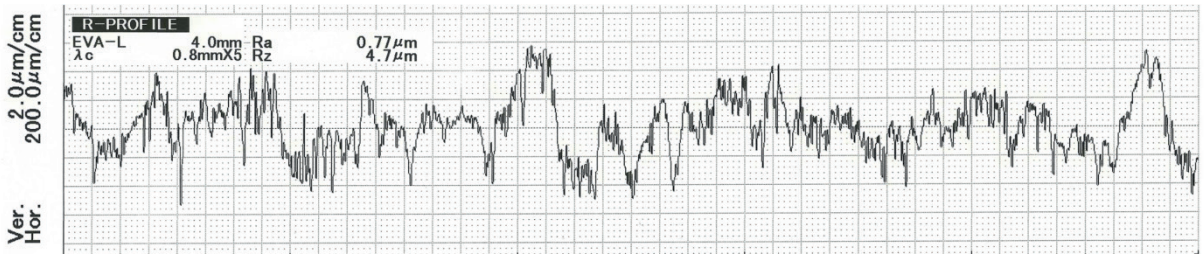

Fig. 11 Surface roughness when hot cutting after 850 strokes

the equation takes the following form:

$$-\frac{q_{zdr}}{\lambda}\frac{1}{2}\frac{d_2}{2}+C_1\frac{1}{d_2/2}=$$
$$-\frac{\alpha_2}{\lambda}\left[C_2+C_1\ln\frac{d_2}{2}-\frac{q_{zdr}}{\lambda}\frac{1}{4}\left(\frac{d_2}{2}\right)^2-T_2\right]$$

Using the two related equations of boundary conditions we determine the constants C_1 and C_2:

$$C_1=\frac{B_1\alpha_2+B_2\alpha_1}{A_1\alpha_2+A_2\alpha_1},$$

$$C_2=\frac{A_1B_2-A_2B_1}{A_1\frac{\alpha_2}{\lambda}+A_2\frac{\alpha_1}{\lambda}},$$

$$A_1=\frac{1}{d_1/2}+\frac{\alpha_1}{\lambda}\ln\frac{d_1}{2},$$

$$A_2=\frac{1}{d_2/2}+\frac{\alpha_2}{\lambda}\ln\frac{d_2}{2},$$

$$B_1=\frac{q_{zdr}}{\lambda}\frac{1}{2}\frac{d_1}{2}\left(1-\frac{\alpha_1}{\lambda}\frac{1}{2}\frac{d_1}{2}\right)-\frac{\alpha_1}{\lambda}T_1,$$

$$B_2=\frac{q_{zdr}}{\lambda}\frac{1}{2}\frac{d_2}{2}\left(1-\frac{\alpha_2}{\lambda}\frac{1}{2}\frac{d_2}{2}\right)-\frac{\alpha_2}{\lambda}T_2.$$

The equation of temperature distribution in the cylindrical wall is in the following form:

$$T(r)=-\frac{q_{zdr}}{\lambda}\frac{1}{4}r^2+C_1\ln r+C_2.$$

The total amount of heat Q transferred from the inside and the outside cylinder surfaces to the environment, related to the cylinder length unit, is determined as the sum of:

$$Q=-\pi d_2\lambda\left.\frac{dT}{dr}\right|_{r=\frac{d_2}{2}}+\pi d_1\lambda\left.\frac{dT}{dr}\right|_{r=\frac{d_1}{2}}=$$
$$\pi d_2\lambda\left(\frac{q_{zdr}}{\lambda}\frac{1}{2}\frac{d_2}{2}-C_1\frac{2}{d_2}\right)-\pi d_1\lambda\left(\frac{q_{zdr}}{\lambda}\frac{1}{2}\frac{d_1}{2}-C_1\frac{2}{d_1}\right)$$

and after the adjustment it will be:

$$Q=\frac{\pi}{4}(d_2^2-d_1^2)q_{zdr}\qquad[\text{W/m}],$$

i. e. the total heat transfer is equal to the aggregate source output in the cylinder volume.

5. Conclusion

This paper describes the issue of applying heating to the cutting process. We presented the theoretical background, construction solution (design) of an experimental tool, and the results of experimental tests. The discussion delivered an insight into the issue of using metal sheet cutting with heating the active component. Phase transformation is each qualitative change in the state of the same substance (in particular, steel), which occurs at least in a certain part of the observed system due to changes in conditions (temperature, pressure, electric field, magnetic field). The above processes may after all result in complete disappearance of the initial matrix phase, to be replaced by a new phase. In the case of an active component such cutting punch, this actually becomes disabled since the required hardness is reduced. In thermo-mechanical terms, phase transformations constitute a transition from the state when it is characterised by a higher value of free enthalpy to a state with its lower value. The loss of energy is represented by labour involved in carrying out the contemplated conversion:

$-\Delta G = (d A_{\max})_{\text{neobj}}$

where: $\Delta G = \Delta H - T\Delta S$, $(d A_{\max})_{neobj}$ is the maximum non-volume labour, ΔH is the system enthalpy change, T is the temperature (K), ΔS is the system enthalpy change [15].

This paper adds new knowledge and conclusions to the large sphere of the issue of metal sheet cutting.

The article was created in frame of VEGA 1/0551/14.

Acknowledgements

Hereby we would like to thank Mr. Prof. RNDr. P. Bury, Head of Department of Technical Physics, and Mr. F. Cernobil for facilitating the execution of laboratory tests. Our gratitude also goes to Mr. Ing. B. Melo, PhD of the Vienna International, spol. s r. o., Martin, for facilitating the execution of operational tests in their press shop.

References

[1] ROMANOVSKIJ, V. P.: *Guide for Cold Pressing* (in Czech), 1968, SNTL-SVTL : Praha.
[2] SRP, K. et al.: *Basics of Pressing* (in Czech), 1965, SNTL : Praha.
[3] CABELKA, J. et al.: *Mechanical Technology* (in Slovak), SAV : Bratislava, 1967.

[4] BLASCIK, F. et al.: *Forming Technology, Casting and Welding* (in Slovak), ALFA : Bratislava, 1988.
[5] BACA, J.: *Forming (in Slovak)*, SVST : Bratislava, 2000.
[6] MORAVEC, J. : Heat Transfer Zone Shear (in Slovak), *Materialove inzinierstvo - Material Engineering,* vol. 7, No. 2,2000, pp. 29-37, ISSN 1335-0803.
[7] MORAVEC, J. : *Unconventional Methods of Forming Metal* (in Slovak),EDIS : University of Zilina, 2011, 135 p., ISBN 978-80-8070-554-0233-8
[8] HOSFORD, W. F., CADDEL, R. M.: *Metal Forming, Mechanics and Metallurgy,* Cambridge University Press, 2011, ISBN 978-1-107-00452-8.
[9] JOHNSON, W., MAMALIS, A. G.: *Plasticity and Metal Forming*, 1998, Cambridge,
[10] MARCINIAK, Z., DUNCAN, J. L., HU, S. J.: *Mechanics of Sheet Metal Forming*, *Butterworth-Heinemann*, Jordan Hill : Oxford, 2002, p. 221, ISBN 0-7506-5300-9.
[11] BOLJANOVIC, V.: S*heet Metal Forming Processes and die Design*, 2005, ISBN 08311-3182-9.
[12] BILIK, J.: *Forming Technology* (in Slovak), STU: Bratislava, 2000.
[13] PEARCE, R.: *Sheet Metal Forming*, 2010, ISBN 0-7503-0101-5.
[14] SPISAK, E. et al.: *Engineering Technology* (in Slovak), 2011, ISBN 978-80-553-08-203.
[15] BELAN, J.: *Theory of Phase Transformations Selected Chapters* (in Slovak), 2015, EDIS : University of Zilina, ISBN 978-80-224-1034-0.

Petr Jonsta - Irena Vlckova - Zdenek Jonsta - Roman Heide*

MATERIAL ANALYSIS OF DEGRADED STEAM TURBINE ROTOR

This article deals with structural, fractographic and non-destructive analysis of a defective rotor made of 26NiCrMoV14-5 steel and assessment of possible technological failures resulting in the occurrence of cracking and subsequent defects. Metallographic analysis of the microstructure and evaluation of the steel metallographic purity were executed, including analysis of inclusions present and ultrasonic testing. Attention was paid to fractographic analysis of fracture areas. Degradation possibilities were monitored in relation to technological failures. The aim was to elucidate the cause of the rotor defect hindering its operational use.

Keywords: *Material analysis, production technology, metallurgical purity, hydrogen embrittlement, degradation.*

1. Introduction

An important factor, indisputably affecting steel quality, is its metallurgical purity. Therefore, great attention should be paid to the presence of inclusions in steel. As known generally, formation and composition of inclusions mostly derives from chemical composition of steel and the way of its deoxidation and alloying [1 and 2]. Moreover, the type and character of inclusions significantly depends on oxygen content and treatment of inclusions during steel processing outside the furnace by means of calcium [3].

Contamination of steel by excessive rate of inclusions constitutes a significant cause leading to initiation of degradation processes, and thus to the product deterioration [4 and 5]. To achieve the required purity, it is also necessary to observe the correct production technology so that during both the production process and the steel treatment there are no defects occurring on the pertinent equipment, which would also result in impaired steel purity as well as quality. In extreme cases, the quality impairment may result even in the product degradation.

Examinations were performed on samples taken from the rotor forging of steel 26NiCrMoV14-5 for a steam power plant, made from an ingot of the weight of 70 tons. After primary processing in an electric arc furnace, the object molten steel for the turbine rotor was transported in a steelmaking ladle for processing outside the furnace in the ASEA SKF device. At the bottom of the ladle, there is an eccentrically installed porous block on which an argon inlet is installed. This provides for steel flushing, setting it in vertical motion. In this way, the melt in the ladle is properly homogenized, inclusions flown out and the melt degassed simultaneously [6]. To check the internal product quality, ultrasonic and magnetic tests are stipulated for the final product.

2. Material and experimental technique

The rotor was subjected to ultrasonic test which indicated inadmissible defects. For the actual material analysis, a plate of approximate width of 20 mm was cut out across

Chemical composition of the steel melting and chemical composition of the experimental plate at the place of defects, in comparison with the production specification of 26NiCrMoV14-5 steel in wt. %, * ppm

Table 1

	C	Mn	Si	P	S	Cr	Ni	Mo	V	H*
Melting	0.26	0.37	0.20	0.003	0.007	1.61	3.44	0.40	0.07	1.20
Experimental plate	0.32	0.39	0.23	0.004	0.001	1.66	3.43	0.42	0.008	-
Specification	0.24 - 0.28	0.35 - 0.40	0.15 - 0.25	Max 0.007	Max 0.02	1.55 - 1.70	3.40 - 3.60	0.40 - 0.45	0.07 - 0.10	0.80

* [1]Petr Jonsta, [2]Irena Vlckova, [1]Zdenek Jonsta, [3]Roman Heide
[1]VSB-Technical university of Ostrava, Faculty of Metallurgy and Materials Engineering, Czech Republic
[2]RMTSC, Material & Metallurgical Research Ltd., Ostrava, remote site VUHZ, a. s., Dobra, Czech Republic
[3]Favea Europe, s. r. o., Prague, Czech Republic
E-mail: petr.jonsta@vsb.cz

the rotor body over the area with the defects indicated (Fig. 1). The actual experimental samples were prepared from it.

Table 1 shows comparison of the melting chemical composition of the object steel 26NiCrMoV14-5 with the production specification of the given steel and the chemical composition of the experimental plate.

The metallographic analysis was executed by means of the OLYMPUS GX51 light microscope; the electron microscopic analysis was executed on the Quanta FEG 450 scanning electron microscope with the TRIDENT-APEX 4 microanalytic system.

3. Results and discussion

During the ultrasonic test, inadmissibly large indications, bigger than 2 mm, were detected in the rotor examined around the forging axis on the head side of the ingot. The defective areas are encircled on the experimental plate (Fig. 1). In sum, 35 inadmissibly large indications of Dn 2.1 to 2.6 mm were found. Most of them were of Dn 2.1-2.2 mm, three of the indications had the dimension of Dn = 2.4 mm, one of them had the dimension of Dn = 2,6 mm. All indications were located at the depth of 427 – 558 mm below the surface of the rotor body. Almost the entire half of the rotor on the ground side of the ingot was without any inadmissible defects.

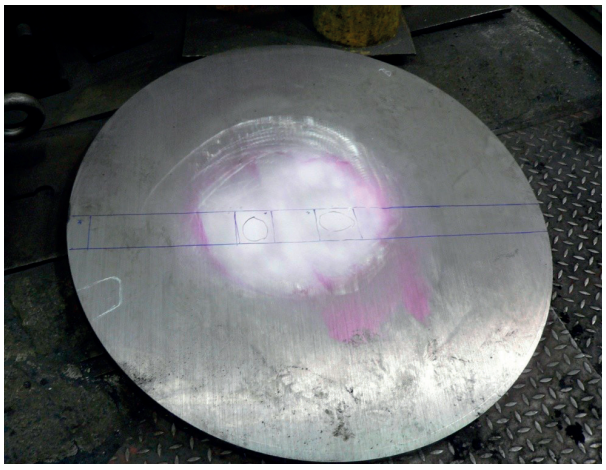

Fig. 1 Experimental plate

A Baumann print was executed on the rotor body cross section. Its appearance confirmed very low sulphur content in the steel (see Table 1). The occurrence of manganese sulphides was very low, particles were small, and sulphur did not concentrate at the macroscopic level very much either in the ingot head part or in the segregates where the segregation of elements is most evident.

Macrostructual analysis proved the presence of dendritic structure and dendritic segregation within the entire cross section of the forging machined, including its peripheral surface layers. The individual dendrites were relatively fine; their size ranged about 1 mm and their arrangement was chaotic. To the depth of approx. 200 mm below the surface, the dendritic segregation was manifested less intensively than in the other parts of the forging cross section (Fig. 2). It was the areas of more evident segregations in which the above mentioned defects were detected. Moreover, a distinctive phenomenon in these areas was the spots of existence of stem segregates. The chaotically arranged dendrites and cross sections of stem segregations, not exceeding the size of 2 mm, are then presented in Fig. 3.

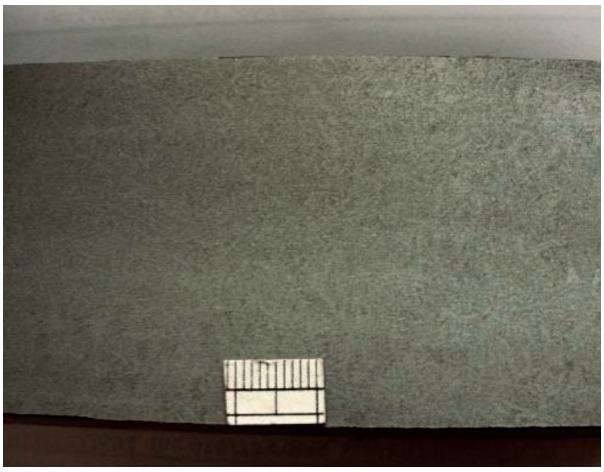

Fig. 2 Macrostructure below the surface

Fig. 3 Macrostructure with segregates

Throughout the entire volume, the analyzed steel had a fine-grained microstructure made of tempered heterogeneous bainite, which shows evidence of both high hardenability of the steel used for the rotor manufacture and the optimum hardening mode. The microstructure of the basic metal matrix was mostly of the acicular character (Fig. 4). More distinctive

dendritic segregations in higher depths below the rotor surface are presented in Fig. 5.

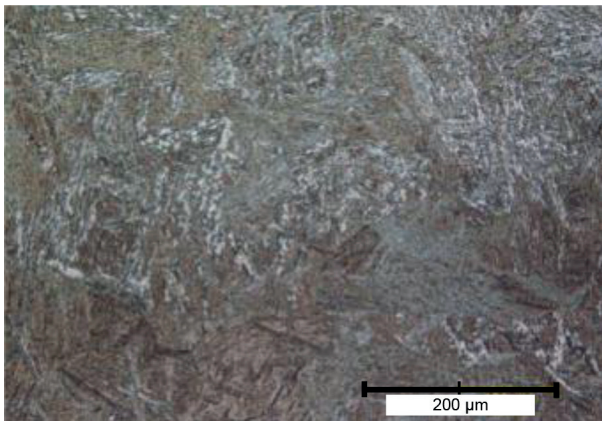

Fig. 4 Microstructure of the basic metal matrix

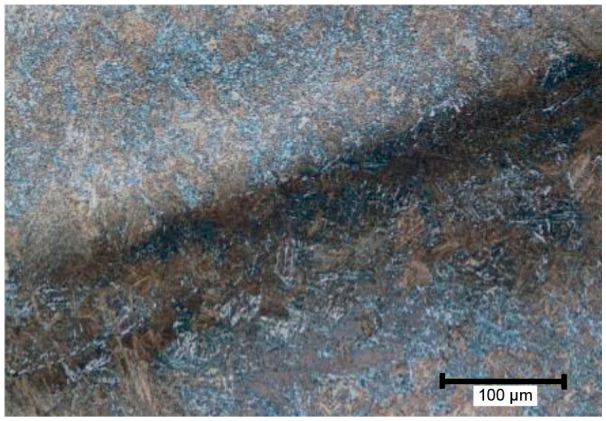

Fig. 5 Macrostructure with segregates

In addition to bainite, these areas, characterized by lower carbon content, probably contain probainitic acicular ferrite. The occurrence of carbides of globulitic morphology bears evidence of high tempering temperature after hardening [7].

In the zone located approximately at one half of the rotor body radius, that is, in the middle of the distance between the rotor surface and its axis, sporadic stem segregates were detected. They included clusters of manganese sulphides and sharp-edged pink-coloured titan carbonitride particles arranged in rows – see Fig. 6. In the microstructure, the stem segregate spots appear dark as a result of the locally increased carbon content and thus a higher amount of fine globulitic carbides in this area [7].

At the points of inadmissible ultrasonic findings around the forging axis, discontinuities were detected in the form of coarse inclusions, rounded elongated cavities and course inclusions situated inside the cavities. The length of the largest of these discontinuities at the metallographic samples exceeded 2 mm, as indicated above. The coarse inclusions formed conglomerates of non-metallic particles the dominant constituent of which was brittle sharp-edged complex oxides on the basis of $MgO.Al_2O_3.CaO$. An example of a complex inclusion at the place of ultrasonic indication is presented in Fig. 7. Besides these inadmissible discontinuities, tiny oxidic and oxysulphidic inclusions of micrometric dimensions as well as titan nitrides or carbonitrides were analyzed (Fig. 8).

Fig. 6 Microstructure with inclusions

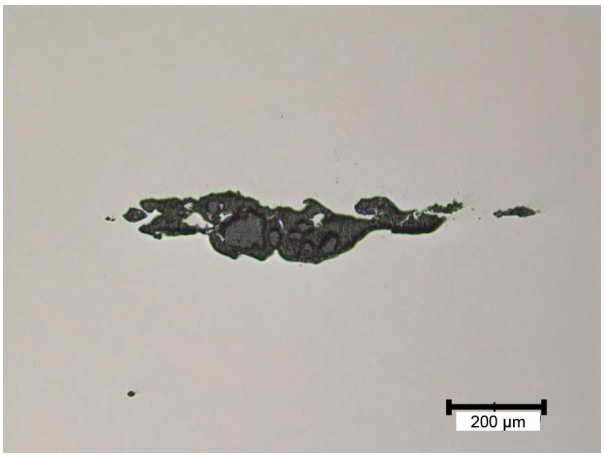

Fig. 7 Complex inclusion

The forging microstructure in the area where the defects occur consisted of tempered bainite; no visible structural changes were ascertained around the defects [8].

The following Fig. 9 presents the morphology of brittle transcrystalline cleavage. This was typical of substantial part of the fracture area. The fine cleavage facets are characteristic of the bainitic structure detected earlier by the metallographic analysis.

The fracture area of tensile tests featured defective zones which are typical as the so-called “fish eyes” (Fig. 10). The fish eye morphology is characteristic of defects resulting from hydrogen action [9]. This proposition is supported by the analysis of chemical composition. Apparently, the hydrogen content reached the value of 1.2 ppm, as opposed to the maximum admissible 0.8 ppm (see Table 1). Another

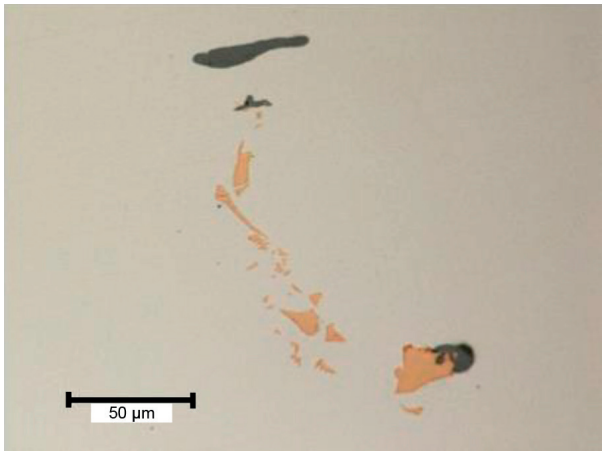

Fig. 8 Ti nitride and Mn sulphide

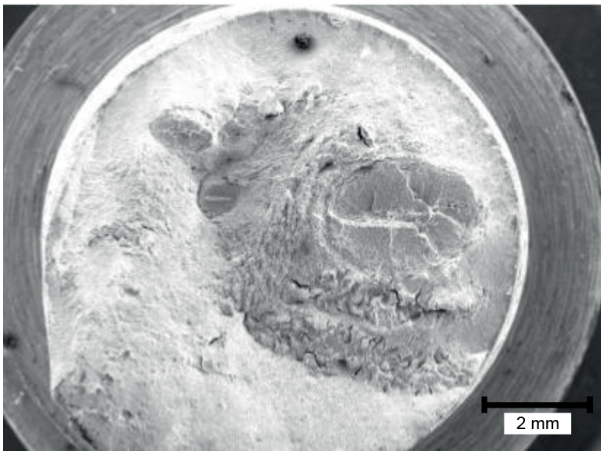

Fig. 11 Defective area with intercrystalline failure

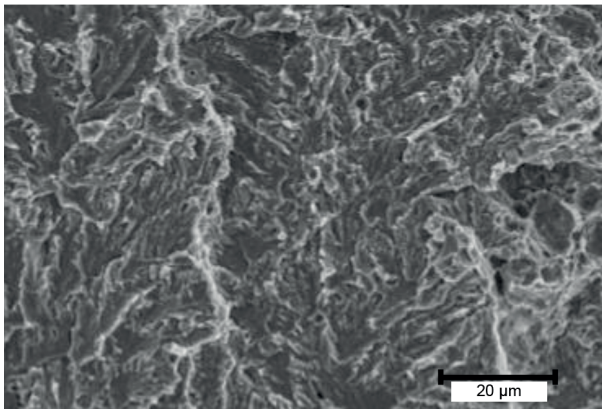

Fig. 9 Brittle transcrystalline cleavage

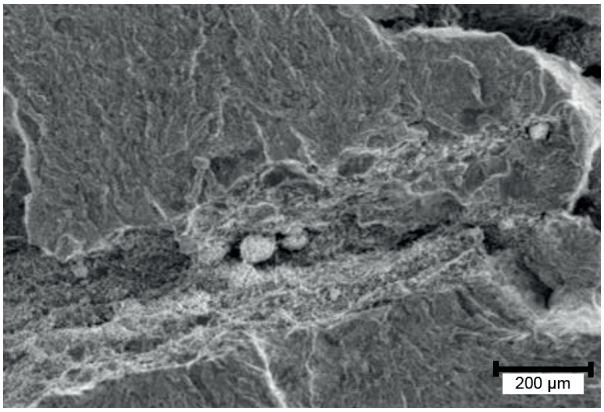

Fig. 12 Inclusions in rows

Fig. 10 Fish eyes in the fracture area

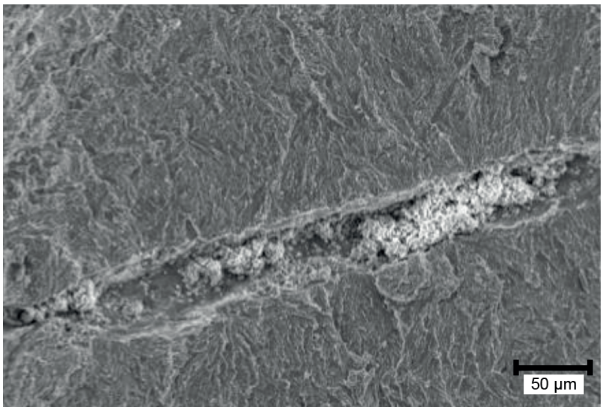

Fig. 13 Detail of inclusions in rows

significant fact is the presence of intercrystalline failure on the fracture area in the vicinity of the fish eyes, as apparent in Fig. 11.

Figures 12 and 13 present inclusions in these defective zones of fracture areas. Figure 12 documents oxidic and oxysulphidic inclusions arranged in lines in the centre of the defective zone of the fracture area, Fig. 13 then shows these inclusions in detail.

To the processing outside the furnace, as indicated above, the steel was transported in a ladle equipped in the bottom with an eccentrically installed porous block with argon inlet for steel flushing. Blasting of this inert gas sets the steel in vertical motion, which results in its proper homogenization in the ladle. Simultaneously, the inert gas bubbles cause flow-out of inclusions

and bind gases, such as hydrogen and oxygen, which leads to degassing of the melt.

After the ladle had been emptied, the porous block was found to be cracked over its entire width and completely clogged with steel. This caused the absence of inert gas flowing into the melt and thus also of the actual flushing process and melt movement in the vertical direction. Likewise, its proper degassing was hindered. The systems of steel movement during continuous flushing with argon, however, contribute to the high grade of its micropurity.

Therefore, one can state that the defect in the technological process, which was the broken porous block, hindered proper flushing and movement of the melt, which was degassed insufficiently at the same time. This resulted in both a substantial increase in the hydrogen content as well as increased content of oxygen. The increase in the hydrogen content by approx. 50 % above the maximum admissible level led to hydrogen embrittlement taking effect in typical fish eyes and intercrystalline steel failure. The increased oxygen content resulted in substantial increase of the steel inclusion rate, which led to significant deterioration of its micropurity.

Therefore, the broken porous block of the secondary metallurgy equipment was the technological cause of the radical decrease of the object steel quality and the subsequent degradation of the rotor made from it.

4. Conclusions

Within the paper submitted, detailed material analysis was executed of the steam power plant rotor made of 26NiCrMoV14-5 steel. The aim was to assess the cause of its degradation. Particular attention was paid to the issue of metallurgical purity of the steel, which presents the decisive factor in the area of its quality.

On the basis of detailed material analysis, one can state that the secondary metallurgy equipment of ASEA SKF, which contains the inductive stirrer, is not efficient enough in the case of a defect in the porous block for melt flushing with inert gas. In the event of the above discussed defect, the melt is homogenized sufficiently, as apparent from, for instance, the chemical analysis; with regard to the volume of exothermic inclusions, however, it is not sufficient for production of steel with a high demand on internal quality. Likewise, the degree of degassing, despite the deep vacuum, does not achieve the required level. Owing to the increase in the hydrogen content together with the exothermic inclusions, the steel produced was deteriorated; in our case, even to the degradation of the rotor.

Acknowledgement

This paper was created in the project No. LO1203 “Regional Materials Science and Technology Centre-Feasibility Program” funded by Ministry of Education Youth and Sports of the Czech Republic.

References

[1] PARK, J., H., TODOROKI, H.: Control of $MgO{\cdot}Al_2O_3$ Spinel Inclusions in Stainless Steels. *ISIJ Int.*, vol. 50, No. 10, 2010, pp. 1333-1346, ISSN 0915-1559.

[2] PARK, J., H., LEE, S., B., GAYE, H., R.: Termodynamics of the Formation of MgO-Al_2O_3-TiO. *Metall. Trans. B*, vol. 39B, 2008, pp. 853-861, ISSN 1073-5615.

[3] GELDENHUIS, J., M., PISTORIUS, P., C.: Minimisation of Calcium Additions to Low Carbon Steel Grades. *Ironmaking and Steelmaking,* vol. 27, 2000, p. 442, ISSN 0301-9233.

[4] BURSAK, M., BOKUVKA, O.: Fatigue Properties of Steel with Increased Atmospheric Corrosion Resistance. *Communications - Scientific Letters of the University of Zilina*, vol. 11, No. 1, 2009, pp. 27-30, ISSN 1335-4205.

[5] BURSAK, M., BOKUVKA, O.: Influence of Technological Factors on Fatigue Properties of Steel Sheets. *Communications - Scientific Letters of the University of Zilina*, vol. 8, No. 4, 2006, pp. 34-37, ISSN 1335-4205.

[6] SAKATA, K.: Technology for Production of Austenite Type Clean Stainless Steel. *ISIJ Int.,* vol. 46, No. 12, 2006, pp. 1795-1799, ISSN 0915-1559.

[7] MAZANCOVA, E, BUZEK, Z., MAZANEC, K.: Evaluation of Influence of Residual Non-ferous Metals on Structural Metallurgical Properties of Steels (in Czech), *Hutnicke listy - Metallurgical J.,* vol. LIII, No. 2, 1998, pp. 19-23, ISSN 0018-8069.

[8] ZHANG, L., THOMAS, B., G.: State of the Art in Evaluation and Control for Steel Cleanliness, *ISIJ Int.*, vol. 43, No. 3, 2003, pp. 271-291, ISSN 0915-1559.

[9] SOJKA, J: *The Resistance of Steels to Hydrogen Embrittlement* (in Czech), 1st ed., ES VSB-TU : Ostrava, 2007, 108 p., ISBN 978-80-248-2187-0.

Frantisek Novy - Robert Ulewicz - Otakar Bokuvka - Libor Trsko - Jan Lago*

RELIABILITY AND SAFETY OF STRUCTURAL ELEMENTS IN THE GIGACYCLE REGION OF LOADING

The authors of this paper, based on their own results, discuss fatigue properties of low-alloyed fine-grained steel STRENX 700MC obtained in the ultra-high-cycle region (gigacycle fatigue) with regard to the reliability and safety of structural elements. According to obtained data it is obvious that fatigue strength continuously decreases with increasing number of cycles to failure (in a range from $N = 10^7$ to $N = 10^{10}$ cycles). Fatigue limit σ_c determined usually at $N = 2 \times 10^6 \div 10^7$ cycles does not meet the requirements of reliability and safety of structural elements. Design of structural elements which will be working in the ultra-high-cycle fatigue region requires complete information about fatigue behavior of used structural materials.

Keywords: *Reliability, safety, steel STRENX 700MC, structural elements, gigacycle fatigue.*

1. Introduction

Fatigue lifetime of structural elements has been investigated for more than 170 years. The strain or stress vs. number of cycles plot (S - N, $\sigma_a = f(N)$ dependence) including conventional fatigue limit σ_c usually referred to $N = 2 \times 10^6 \div 10^7$ cycles (valid for steels and cast irons) are the main parameters used for evaluation of fatigue properties of structural materials [1-3].

The modern industry has set the direction for the expansion of new design solutions for higher performance of industrial devices. The new design solutions are connected with the proper selection of materials for individual structural elements. This fact is crucial from the point of view of their reliability and safety and has also a huge contribution to increasing their service life. With the aim to prolong the service life of structural elements, the ultra-long fatigue life ($10^7 < N < 10^{10}$ cycles) has, therefore, been studied very intensively during recent years [1, 4 and 5]. This reality is emphasized by the fact that fatigue fractures are observed after billion cycles of loading and more.

The advanced structural materials are developed with the aim to prolong the service life of structural elements; the high-strength fine-grained steels are advisable materials applicable for the ultra-long loading of these elements. Advanced steels often exhibit better mechanical properties compared to conventional steels [6] and, relevant to this fact, it is necessary to know also the fatigue characteristics of these advanced materials.

In this paper the authors provide their own results followed by a discussion about fatigue properties of low-alloyed fine-grained structural steel STRENX 700MC with reference to the reliability and safety of structural elements.

2. Experimental part

Experimental procedures, quantitative chemical analysis, metallographic analysis, tensile tests, fatigue tests and fractographic analysis were carried out for the low-alloyed fine-grained steel STRENX 700 MC. Chemical analysis was performed with the help of emission spectrometry on an ICP (JY 385) emission spectrometer using a fast recording system Image. A light metallographic microscope AXIO Imager A1m was used for metallography analysis. Tensile tests were carried out on a ZWICK Z050 testing machine at ambient temperature of $T = 20 \pm 5$ °C, with the loading range in interval $F = 0 \div 20$ kN and the strain velocity of $\dot{\varepsilon}_m = 10^{-3}$ s^{-1}. Round cross-section specimens were used; the shape and dimensions of the test specimens met the requirements of EN 10002-1 standard (5 specimens were used). Fatigue tests were carried out at high-frequency sinusoidal cyclic tension-compression loading ($f \approx 20$ kHz, $T = 20 \pm 5$ °C, $R = -1$, cooled by distilled water with anticorrosive inhibitor) and with the use of high-frequency loading equipment KAUP-ZU Zilina. Smooth round bar specimens (14 pieces) with 4 mm

* [1,2] **Frantisek Novy,** [3]**Robert Ulewicz,** [1]**Otakar Bokuvka,** [2]**Libor Trsko,** [1]**Jan Lago**
[1] Department of Materials Engineering, Faculty of Mechanical Engineering, University of Zilina, Slovakia
[2] Research Centre of the University of Zilina, Slovakia
[3] Czestochowa University of Technology, Inst. of Production Technology, Czestochowa, Poland
E-mail: frantisek.novy@fstroj.uniza.sk

diameter, ground and polished by metallographic procedures were used during the fatigue tests [1, 4, and 7]. The investigated region of number of cycles ranged from N ≈ 3×10^6 to N = 10^{10} cycles of loading. The fractographic analyses of fatigue fracture surfaces were carried out using a scanning electron microscope TESCAN LMU II.

3. Results and discussion

The results of quantitative chemical analysis (chemical composition), metallographic analysis (microstructure), tensile tests (yield point in tension R_e, ultimate tensile strength R_m) and high-frequency fatigue tests (stress amplitude vs. number of cycles, σ_a – N curve) are shown in Table 1 and Figs. 1 and 2 respectively. The experimental low-alloy steel STRENX 700MC had ultra-fine-grained ferrite microstructure, see Fig. 1.

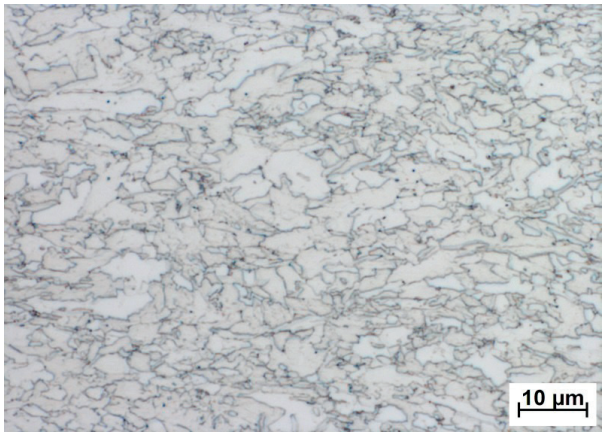

Fig. 1 Microstructure of tested steel STRENX 700MC

As is obvious from Fig. 2, the fatigue strength continuously decreases with increasing number of cycles to fracture in the whole tested region of cycles ($\Delta\sigma_a$ = 160 MPa; σ_a = 400 MPa at N = 3×10^6 cycles vs. σ_a = 240 MPa at N = 10^{10} cycles). This fact is in a good agreement with works [1, 4, 5, 8 and 9] where $\Delta\sigma_a$ is given from $\Delta\sigma_a$ = 20 MPa to $\Delta\sigma_a$ = 200 MPa whereas the higher values were observed at high-strength steels. The step-wise or duplex S – N curve was not observed and it cannot be said that there is a constant amplitude of loading under which the structural element can stand an infinite number of cycles. The values of fatigue limit σ_c obtained at N = $2\times10^6 \div 10^7$ cycles are overestimated and do not meet the demands of reliability and safety of structural elements [10]. These facts can be shown by comparison of N = 10^7 cycles vs. N = 10^{10} cycles in the simplified Smith's diagram. The area of Smith's diagram for fine-grained steel STRENX 700MC drawn according to σ_a = 384 MPa obtained at N = 10^7 cycles is much bigger than the area of the Smith's diagram drawn according to σ_a = 240 MPa obtained at N = 10^{10} cycles, Fig. 3.

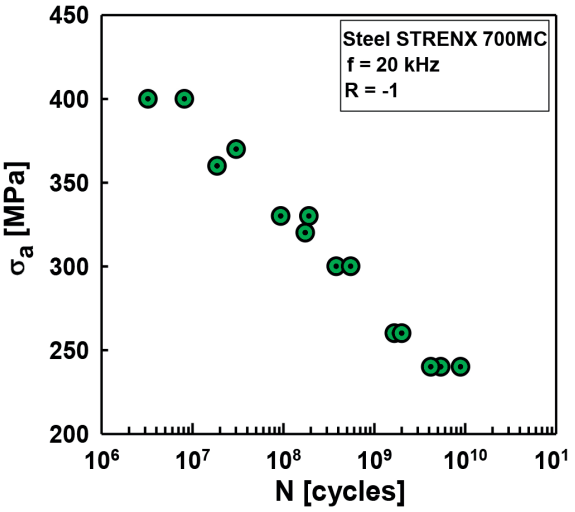

Fig. 2 σ_a - N curve of steel STRENX 700MC tested at high-frequency loading

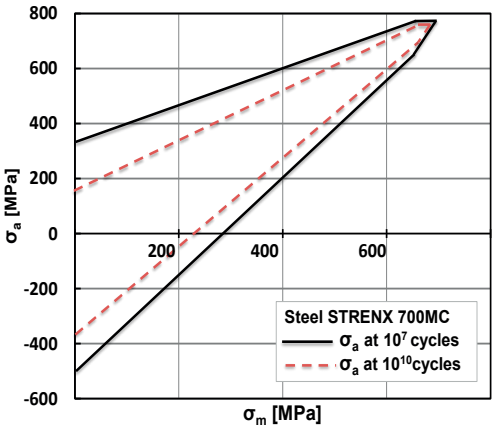

Fig. 3 Smith's diagram for steel STRENX 700MC drawn for σ_a obtained at N = 10^7 and N = 10^{10} cycles

The threshold value K_{ath} is very important information used by engineers for an optimal design of structural elements. K_{ath} represents the resistance of steel against the fatigue crack

Chemical composition (in weight %), Re and Rm of tested steel STRENX 700MC — Table 1

Element	C	Si	Mn	P	S	Cr	Ni	Mo	B	R_e (MPa)	R_m (MPa)
	0.20	0.39	0.80	0.005	0.005	0.45	0.05	0.01	0.001	796	849.4

growth. It is determined according to interatomic distance (crystallography lattice parameter). If the crack length growth for one cycle is smaller than one interatomic distance, then cyclic loading has no degradation effect from the physical point of view (it does not cause a macroscopic growth of fatigue crack). In practice it means that for values $K_a \leq K_{ath}$ cracks do not grow or the growth is extremely slow (slower than $da/dN = 10^{-10}$ m.cycle^{-1}) and it does not cause the breaking of steels by fatigue fracture for expected time of use [11 and 12]. The approximate threshold value K_{ath} of tested steel STRENX 700MC was calculated with following equation [13]

$$K_{ath} = -0.0052 \times R_m + 8.5906 \qquad (\text{MPa.m}^{1/2}) \qquad (1)$$

in which R_m is ultimate tensile strength (equation is valid in the region from $R_m \approx 360$ MPa to $R_m \approx 1040$ MPa and at high-frequency loading). The obtained value (K_{ath} = 4.18 MPa.m$^{1/2}$) is in an agreement with general conclusions published in works [1, 4, 7, 10 and 14] and with Lal's results which are valid for low-frequency loading [15]. Consequently, approximate intrinsic crack lengths a_{01} were calculated for N = 10^7 cycles (at σ_a = 384 MPa) and a_{02} for N = 10^{10} cycles (at σ_a = 240 MPa) of loading using the following equation [1, 4 and 11]

$$K_{ath} = \sigma_a \times \left(w \times tang \frac{\pi \times a_0}{w} \right)^{1/2} \qquad (\text{MPa.m}^{1/2}) \qquad (2)$$

in which σ_a is the stress amplitude (MPa), a_0 is the intrinsic crack length (m) and w = 0.016 m is the width of the used specimen [1 and 4]. The intrinsic crack lengths values were a_{01} = 0.0377 mm and a_{02} = 0.0965mm. The obtained results are in a good agreement with [14] where the author writes that $a_0 \leq 0.1 \div 1$ mm in the case of low-strength steels and $a_0 \leq 10^{-2}$ mm in the case of high-strength steels. From Kitagawa-Takahashi's diagram [16] drawn for N = 10^7 and 10^{10} cycles, Fig. 4, it is obvious that under cyclic loading with higher stress amplitude (K - T diagram for N = 10^7 cycles, intrinsic crack length a_{01} = 0.0377), shorter cracks start to propagate under cyclic loading with lower stress amplitude (K - T diagram for N = 10^{10} cycles, intrinsic crack length a_{02} = 0.0965 mm). Also the boundary between short and long cracks is moving to bigger crack lengths with decrease of the loading amplitude [17].

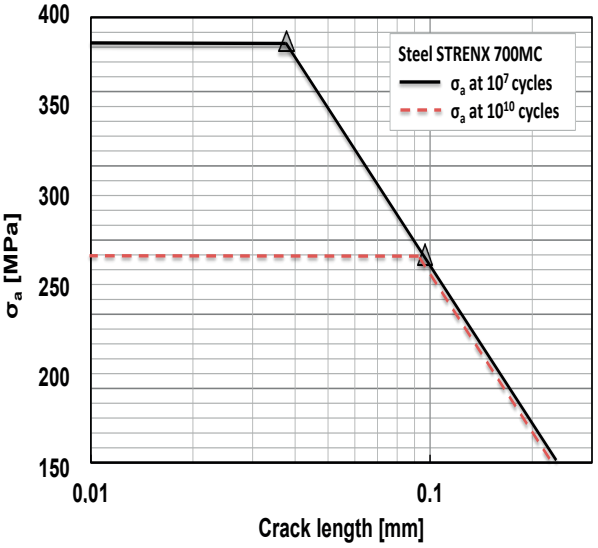

Fig. 4 Kitagawa-Takahashi's diagram for steel STRENX MC700 drawn for N = 10^7 and 10^{10} cycles

Many years ago it was reported that fatigue cracks initiation occurred from small defects on the surface at high stress amplitude levels and low number of cycles. Only in the last

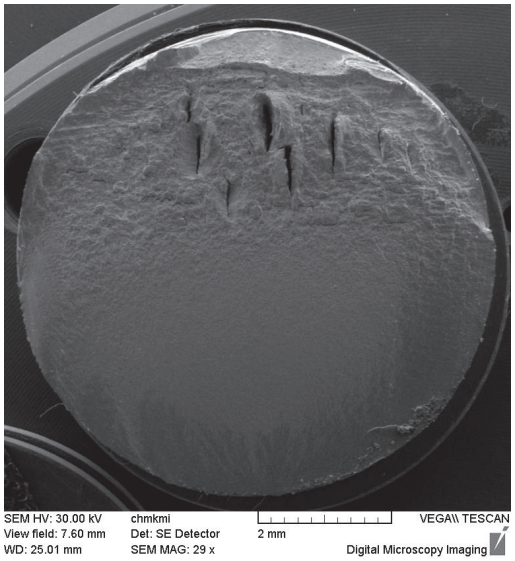

Fig. 5 Surface fatigue crack initiation, σ_a = 240 MPa and at N = 5.4 ×10^9 cycles

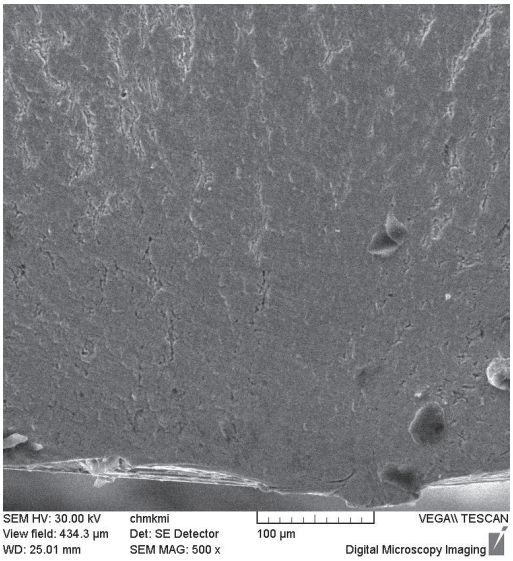

Fig. 6 Surface fatigue crack initiation, σ_a = 240 MPa and at N = 5.4 ×10^9 cycles

decades it was reported that fatigue cracks start in the subsurface zone at low stress amplitude level and ultra-high number of cycles. Subsurface fatigue crack initiation in the ultra-high number of cycles region is a microstructure related phenomenon which determines the fatigue lifetime, and is associated with structural heterogeneities such as inclusions, microphores, microshrinkages, large or very small grains region, etc. [1, 4, 10, 14 and 17 - 22]. The microstructure image of tested steel STRENX 700MC revealed sparse titanium nitride particles with a regular geometric shape as well as few inclusions in the form of black manganese sulphide particles. Despite of this fact only surface crack initiation was observed, Figs. 5 and 6. As the result of polished surface of the specimen, with the purpose to eliminate the notch effect after machining, it was possible that the surface initiated cracks could start only from the surface concentrators in the surface relief, which was created during the cyclic loading. During the cyclic loading there are microstructural changes observed in the material. The number and the configuration of structural defects are changing, whereby the most important change is the creation of the dislocation structure. The characteristics of the dislocation structure depend on the micro-plastic deformation which depends on the applied stress amplitude. At the same time, the cyclic micro-plastic deformation is localized into slip bands. In the places where slip bands come up to the surface, intrusions and extrusions are generated and create the surface relief; intrusions are strong stress concentrators and the surface fatigue cracks initiation then starts in the roots of intrusions [1, 3, 4, 10, 11 and 16 - 23].

4. Conclusions

With reference to the experimental results the following can be summed up:

- in the low-alloyed fine-grained steel STRENX 700MC a continuous decrease of fatigue strength in dependence on the number of loading cycles to fracture is observable in the gigacycle region of loading;
- the values of fatigue limit σ_c obtained at $N = 2 \times 10^6 \div 10^7$ cycles are overestimated and do not meet the demands of reliability and safety of structural elements;
- the safety area of Smith's and Kitagawa-Takahashi's diagrams is much smaller at $N = 10^{10}$ cycles than at $N = 10^7$ cycles;
- the intrinsic crack length decreases with the increase of loading amplitude;
- only surface fatigue crack initiation was observed;
- these facts must be taken into consideration with reference to reliability and safety when designing structural elements.

Acknowledgements

The research was supported by Scientific Grant Agency of Ministry of Education, Science and Sport of Slovak Republic and Slovak Academy of Sciences, grant VEGA No. 1/0123/15 and 1/0720/14, by the project APVV-14-0096 and project "Research Centre of the University of Zilina", ITMS 26220220183.

References

[1] BOKUVKA, O., NICOLETTO, G., GUAGLIANO, M., KUNZ, L., PALCEK, P., NOVY, F., CHALUPOVA, M.: *Fatigue of Materials at Low and High-frequency Loading.* EDIS ZU Zilina, 2015.
[2] HURTALOVA, L., TILLOVA, E., CHALUPOVA, M.: *Key Eng. Mat*., vol. 592-593, p. 433.
[3] KUNZ, L.: *Experimental Determination of Fatigue Characteristics of Materials* (in Czech). EDIS ZU Zilina, 2003.
[4] BOKUVKA, O., NICOLETTO, G., KUNZ, L., PALCEK, P., CHALUPOVA, M.: *Low and High Frequency Fatigue Testing*. EDIS ZU Zilina, 2002.
[5] BATHIAS, C., PARIS, P. C.: *Gigacycle Fatigue in Mechanical Practice*. M. Dekker : N. York, 2005.
[6] KUZLIAK, R., KAWALLA, R., WAENGLER, S.: *Arch. of Civil and Mech. Engng., v*ol. 8, No. 2, 2008, p. 103.
[7] PUSKAR, A.: *Highfrequency Fatigue of Materials* (in Slovak). EDIS ZU Zilina, 1997.
[8] ULEWICZ, R., MAZUR, M.: *Prod. Engng. Arch*., No. 1, 2013, p. 32.
[9] SZATANIAK, P.: *Fatigue Properties of Fine-grained Steels.* PhD. Thesis, ZU Zilina, 2016.
[10] NOVY, F., BOKUVKA, O., TRSKO, L., CHALUPOVA, M.: *Annals of Fac. Eng. Huneodora - Int. J. of Engng.,* vol. 10, No. 2, 2012, p. 231.
[11] KLESNIL M., LUKAS P.: *Fatigue of Metal Materials at Mechanical Loading* (in Czech). ACADEMIA Praha, 1975.
[12] TRSKO, L., NOVY, F., BOKUVKA, O.: *Dynamics Strength and Fatigue Lifetime.* EDIS ZU Zilina, 2013.
[13] RUZICKOVA, M., BOKUVKA, O., PALCEK, P., CHALUPOVA, M.: *Materials Engng.,* vol. 6, No. 15, p. 19.
[14] RITCHIE, R.O.: *Application of Fracture Mechanics to Fatigue Crack Propagation.* University of California, 1981.
[15] LAL, D. N.: *Fat. Frac. Engng. Mater. Struct.,* vol. 15, No. 12, 1992, p. 1199.
[16] KITAGAWA, H., TAKAHASHI, S.*: Proc. of the 2nd Int. Conf. on Mechanical Behavior of Materials,* ASM, 1976, p. 627.

[17] NOVY, F., BOKUVKA, O., MINTACH, R., TRSKO, L.: *Perner's Contacts*, vol. 6, No. 2, 2011, p. 131.
[18] MURAKAMI, Y., NOMOTO, T., UEDA, T.: *Fatigue Fract. Engng. Struct.,* vol. 22, 1999, p. 581.
[19] MURAKAMI, Y., NOMOTO, T., UEDA, T.: *Fatigue Fract. Engng. Struct.,* vol. 23, 2000, p. 903.
[20] MURAKAMI, Y.: *Metal Fatigue: Effects of Small Defects and Nonmetalic Inclusions,* 1st ed., Oxford : Elsevier Ltd., 2002.
[21] SZATANIAK, P., MAZUR, M., ULEWICZ, R., NOVY, F.: *Communications - Scientific Letters of the University of Zilina,* vol. 18, No. 1, 2016, p. 89.
[22] MURAKAMI, Y., NOMOTO, T., UEDA, T.: *Fatigue Fract. Engng. Struct.,* vol. 23, 2000, p. 893.
[23] PUSKAR, A., HAZLINGER, M.: *Failure and Fractures of Components* (in Slovak), EDIS : ZU Zilina, 2000.

Michaela Solanska - Miroslav Markovic - Milan Dado*

RESOURCE RESERVATION SCHEMES FOR OPTICAL BURST SWITCHING NETWORKS

Since the current optical communication networks are based on internet protocol and demands for transmission capacity and speed are growing, the optical burst switching presents the solution for future high-speed wavelength division multiplexing optical networks. In order to successfully transmit bursts over optical burst switching network, resource reservation schemes have to be implemented. According to the way of resource reservation, reservation schemes use one-way reservation or two-way reservation. The schemes with one-way reservation are used mainly for optical burst switching networks. The survey of one-way resource reservation schemes is presented. The basic operations and properties of one-way resource reservation schemes are described.

Keywords: *Optical burst switching, resource reservation scheme, one-way reservation.*

1. Introduction

With growing demands of Internet Protocol (IP) services for transmission capacity and speed, the Optical Burst Switching (OBS) presents the solution for future high-speed Wavelength Division Multiplexing (WDM) optical networks. OBS is a technology for transmitting large amounts of data bursts through a transparent optical switching network. OBS combines the best elements of optical packet switching and wavelength routing, but avoids their shortcomings [1].

An OBS network consists of the edge nodes and the core nodes interconnected with each other with WDM links (Fig. 1). Two types of edge nodes are differentiated in OBS network, the ingress (core) and the egress (destination) edge nodes that are based on interface of the classical IP network and the OBS network [1-3].

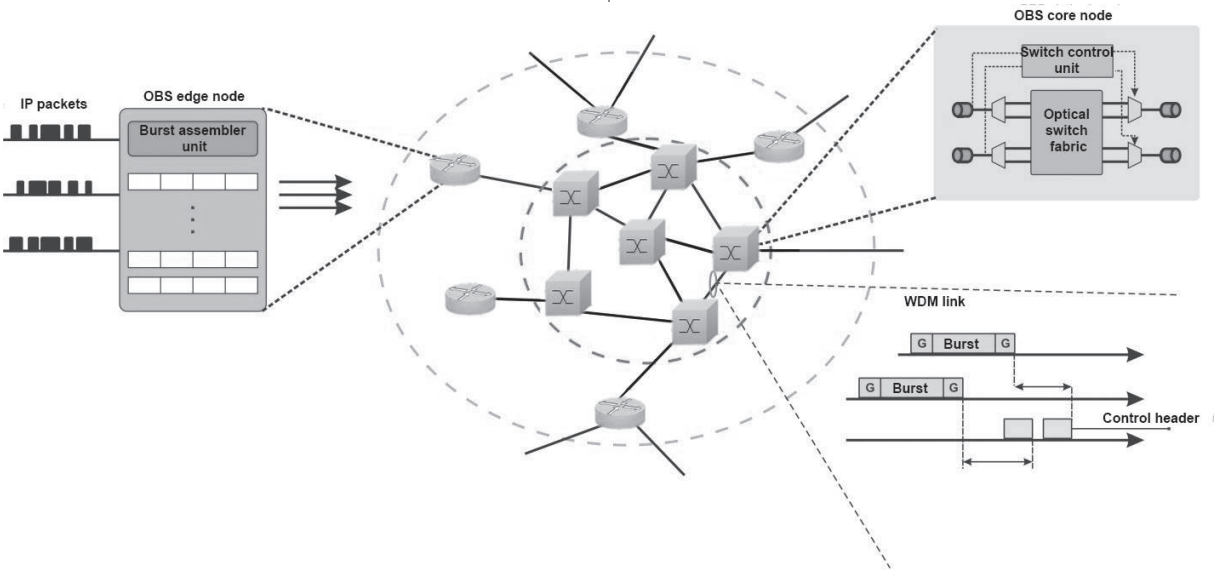

Fig. 1 The architecture of OBS network

* **Michaela Solanska, Miroslav Markovic, Milan Dado**
Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina, Slovakia
E-mail: michaela.solanska@fel.uniza.sk

Ingress edge nodes electronically assembly several incoming IP packets with the same destination into constant or variable length optical bursts, which stay in optical domain until they arrive to the egress edge node. To each optical burst is generated header, necessary to correct the switching structure setting in every core node, through which the burst will be passing. Just before the burst is transported, the control header is sent, to inform all the core nodes along the road about the burst arrival. The control header is sent through the independent channel, which is on different wavelength as transferred optical burst (out-of-band signaling). In each core node, the control header is converted by optical-electronic-optical (O/E/O) conversion, due to not available all optical control circuits. On the basis of an information contained in the control header is made the setting of switching structure in core node, which is processed in control circuit. The information contains the time delay between control header and optical burst (referred to as offset), size of the optical burst, required output port, incoming wavelength. Then the new header for transmitting burst is generated. This header is again sent from the core node in advance of the optical burst. This repeats until the optical burst does not arrive to the egress edge node. The purpose of the egress edge node is to disassembly the optical burst in a condition that has been set before entering into the edge node [1-9].

2. One-Way Resource Reservation Schemes

The main issue in OBS network is the random collision and burst loss. To successfully transmit bursts over OBS network and reach destination node, resource reservation schemes have to be implemented to allocate resources and configure optical switches for that burst at each node. According to the process of resource reservation, resource reservation schemes may be classified into two main classes: one-way reservation and two-way reservation. OBS networks use mainly reservation schemes with one-way reservation. By one-way reservation bursts are sent shortly after the control header and the source node does not wait for response about the success of the reservation of a full path sent by the destination node. A number of one-way resource reservation schemes have been proposed for OBS, including Tell-And-Go (TAG), Just-In-Time (JIT), Enhanced Just-In-Time (E-JIT), Enhanced Just-In-Time Plus (E-JIT$^+$), Just-In-Time Plus (JIT$^+$), Just-Enough-Time (JET), Horizon, JumpStart [4-9].

In the Tell-And-Go (TAG) reservation scheme a source node sends a control header that contains *Setup* message. Immediately after control header source node sends a data burst. This reservation scheme is rather idealistic and non-realistic, due to the lack of time of the receiving node to process the *Setup* message and to configure its switch structure, so it can switch the incoming burst to required output port. To apply these scheme to the real life, it is necessary to configure the node in advance to switch the burst, or equip it with optical buffers, which delay the bursts, while the node will process the *Setup* message and configures its switching structure. TAG is practical only if the switch processing time of the *Setup* message and the optical switch configuration time are very short. Therefore, due to the critical limitation of TAG the study is focused on aforementioned one-way resource reservation schemes [4-9].

Just-In-Time (JIT) reservation scheme was proposed by Wei and McFarland in [4]. JIT is resource reservation scheme with immediate reservation and it uses explicit releases to set free the switching structure resources. JIT operates as follows: an output wavelength is reserved for a burst immediately after the arrival of the corresponding *Setup* message. If a wavelength cannot be reserved immediately, then the *Setup* message is rejected and the corresponding burst is dropped (Fig. 2). JIT uses first-come, first- served (FCFS) service on each wavelength, because bursts are scheduled in the order in which their *Setup* messages arrive at the node [4-9].

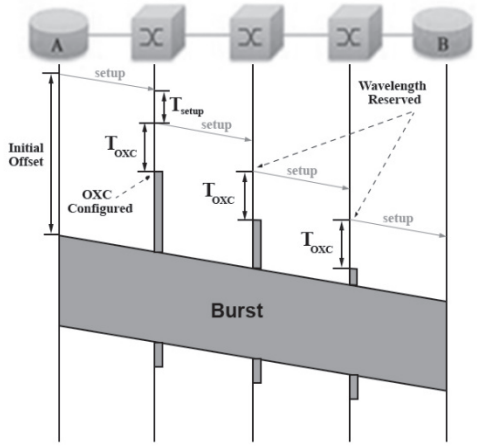

Fig. 2 Schematic representation of JIT reservation scheme [5]

Just-In-Time Plus (JIT$^+$) reservation scheme was proposed by Teng and Rouskas in [5]. JIT$^+$ is resource reservation scheme with immediate reservation and it uses estimated releases to set free the switching structure resources. JIT$^+$ is a modified and improved version of JIT reservation scheme. JIT$^+$ combines simplicity of JIT with the utilization of the time horizon that is used by delayed resource reservation schemes. JIT$^+$ adds limited burst scheduling (for a maximum of two bursts per channel). JIT$^+$ operates as follows: an output wavelength is reserved for a burst only if *(1)* the arrival time of the burst is later than the time horizon defined for the wavelength and *(2)* the wavelength has at most one other reservation. JIT$^+$ uses FCFS service on each wavelength [4-9].

Enhanced Just-In-Time (E-JIT) reservation scheme was proposed by Rodrigues, Freire and Monteiro in [7]. E-JIT is resource reservation scheme with immediate reservation and it uses estimated (or implicit) releases to set free the switching

structure resources, as the *Setup* message also carries the information of burst length and burst offset length, allowing each node to estimate the latest time the resources are no longer assigned to a burst transmission. E-JIT improves and optimizes JIT reservation scheme, keeping all the advantages of its simplicity in terms of implementation. Optimization is achieved by the improvement of data channel scheduling. The period of time in which wavelength remains in reserved status is reduced, so that it results in the optimization of the wavelength utilization and potentially reduce the burst loss probability. E-JIT operates as follows: an output wavelength is reserved for a burst immediately after the arrival of the corresponding *Setup* message, if *(1)* this wavelength is free or *(2)* if it is reserved, the end time of the last switched burst is smaller than the actual time to process the *Setup* message. If a wavelength cannot be reserved immediately, then the *Setup* message is rejected and the corresponding burst is dropped. E-JIT uses FCFS service on each wavelength [4-9].

Enhanced Just-In-Time Plus (E–JIT+) reservation scheme was proposed by Rodrigues, Gregorio and Vasilakos in [9]. E-JIT+ is resource reservation scheme with immediate reservation and it uses estimated (or implicit) releases to set free the switching structure resources. E-JIT+ is based on JIT+ reservation scheme, integrates functions of E-JIT reservation scheme in terms of scheme operations, and intends to increase the performance of OBS networks, maintaining the simplicity of JIT reservation scheme in terms of implementation. E-JIT+ uses the time horizon concept, which determines whether a channel is ready to accept new bursts (in order to make possible the calculation of the time horizon value, *Setup* messages carry information with the burst offset and the burst length values). E-JIT+ operates as follows: an output wavelength is reserved for a burst if the arrival time of the burst occurs later than the time horizon of that wavelength. Otherwise, the *Setup* message is rejected and the corresponding data burst is dropped. E-JIT+ uses FCFS service on each wavelength [4-9].

Just-Enough-Time (JET) reservation scheme was proposed in [10-12]. JET is the best known reservation scheme with delayed reservation that uses void filling. JET uses estimated releases to set free the switching structure resources. JET operates as follows: an output wavelength is reserved for a burst if the arrival time of the burst *(1)* is later than the time horizon defined for the wavelength, or *(2)* coincides with avoid on the wavelength, and the end of the burst (plus the configuration time of optical cross-connect) is sooner than the end of the void. If, after arrival of the *Setup* message, it is determined that none of these conditions are satisfied for any wavelength, then the *Setup* message is rejected and the corresponding burst is dropped (Fig. 3). JET uses non FCFS service on each wavelength [8, 10, 11 and 12].

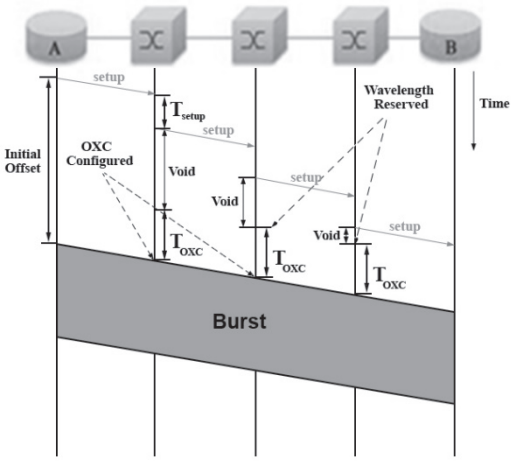

Fig. 3 Schematic representation of JET reservation scheme [5]

Horizon reservation scheme was proposed by Turner in [13]. Horizon is resource reservation scheme with delayed reservation that does not perform any void filling. Horizon scheme takes its name from the fact that each wavelength is associated with a time horizon for burst reservation purposes. This time horizon is defined as "the earliest time after which there is no planned use of the wavelength". Horizon operates as follows: an output wavelength is reserved for a burst only if the arrival time of the burst is later than the time horizon of the wavelength. If, after arrival of the *Setup* message, it is determined that the arrival time of the burst is earlier than smallest time horizon of any wavelength, then the *Setup* message is rejected and the burst corresponding burst is dropped. Horizon uses FCFS service on each wavelength [6, 8 and 13].

JumpStart is a joint project supported by Advanced Research and Development Agency (ARDA) developed by the North Carolina State University (NCSU) and MCNC Research and Development Institute [6]. JumpStart reservation scheme uses immediate and delayed reservation. JumpStart combines control and signalization simplicity together with multiple functions, including support for multicast and different types of connection. JumpStart operates as follows: a source node first sends the *Setup* message to core node with information related to the burst transmission, including the source and destination addresses. If the core node can switch the burst, it returns the *Setup ACK* message to the source node. Moreover, it forwards the *Setup* message to the next node. Otherwise, the core node refuses the *Setup* message and returns a reject message to the source node and the corresponding burst is dropped. In this case, the source node enters in an idle period waiting for another burst. When a new burst arrives, the edge node repeats the process [6, 14, 15 and 16].

Main properties of reservation schemes Table1

Reservation scheme	Reservation process	The release process of reservated resources	Burst service	T_{Setup}
JIT	Immediate	Explicit	FCFS	12,5μs
JET	Delayed	Estimate	Non FCFS	50 μs
E-JIT	Immediate	Estimate	FCFS	12,5 μs
E-JIT+	Immediate	Estimate	FCFS	12,5 μs
JIT+	Immediate	Estimate	FCFS	12,5 μs
Horizon	Delayed	-	FCFS	25 μs
JumpStart	Immediate/Delayed	-	-	12,5 μs

The basic features of reservation schemes are summarized in Table 1 (T_{Setup} is the amount of time it takes an OBS node to process the Setup message under reservation scheme).

3. Comparison of reservation schemes

The basic reservation schemes JET, JIT, Horizon are compared in terms of burst loss probability that is the main issue in OBS networks. The analytical burst loss probability of reservation schemes was calculated on the basis of study in [5]. Figure 4 shows the burst loss probability as a function of number of wavelengths (data channels) per optical fiber in a single node for reservation schemes. The burst loss probability decreases with then increasing number of wavelengths and the performance of each reservation schemes is very similar.

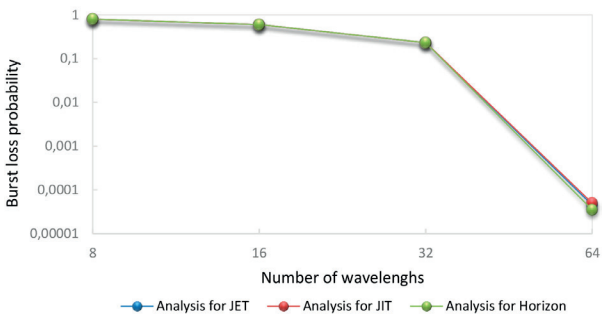

Fig. 4 Burst loss probability as a function of number of wavelengths per optical fiber in a single node for JET, JIT and Horizon reservation schemes

4. Conclusion

The reservation schemes play an important role for data transmission. The reservation schemes allocate resources along path from source to destination node. The one-way resources reservation schemes for OBS networks, the basic operations and properties of reservation schemes are described. The next goal is to compare analytical and simulation burst loss probability of JET, JIT and Horizon reservation schemes based on the OBS network model that was created using OMNeT++ simulation environment. To study the impact of the number of data channels, number of hops, number of nodes, traffic load, topology selection of OBS network on the performance of each reservation scheme.

Acknowledgements

This work is supported by the Slovak Research and Development Agency under the project APVV-0025-12 ("Mitigation of stochastic effects in high-bitrate all-optical networks") and the European Regional Development Fund and the Slovak state budget for the project "Research Centre of University of Zilina", ITMS 26220220183.

References

[1] PAPADIMITRIOU, G., PAPAZOGLOU, CH., POMPORTSIS, A.: *Optical Switching*, Hoboken : Wiley, 2007.

[2] UMALE, M. R., JADHAV, S. S., NARAVADKAR, S. D.: *Optical Burst Switching (OBS) for Next Generation Optical Transport Networks,* Emerging Trends in Computer Science and Information Technology, ETCSIT, 2011.

[3] SALEH, A. A. M., SIMMONS, J. M.: All-Optical Networking-Evolution, Benefits, Challenges, and Future Vision, *IEEE J. & Magazines*, vol. 100, No. 5, 2012.

[4] WEI, J. Y., MCFARLAND, R. I.: Just-In-Time Signaling for WDM Optical Burst Switching Networks, *J. of Lightwave Technology*, vol. 18, No. 12, 2019-2037, 2000.

[5] TENG, J., ROUSKAS, G. N.: A Detailed Analyses and Performance Comparison of Wavelength Reservation Schemes for Optical Burst Switched Networks, *Photonic Network Communications*, vol. 9, No. 3, 311-335, 2005.

[6] RODRIGUES, J. J. P. C., FREIRE, M. M., LORENZ, P.: *One-Way Resource Reservation Protocols for IP over Optical Burst Switches Mesh Networks,* Systems Communications, ICW'05, 229-234, 2005.

[7] RODRIGUES, J. J. P. C., FREIRE, M. M., GARCIA, N. M., MONTEIRO, P. M. N. P.: *Enhanced Just-In-Time: A New Resource Reservation Protocol for Optical Burst Switching Networks,* 12th IEEE Symposium on Computers and Communications, ISCC 2007, 121-126, 2007.

[8] RODRIGUES, J. J. P. C., VAIDYA, B.: *Evaluation of Resource Reservation Protocols for IP over OBS Networks,* 11th Intern. Conference on Transparent Optical Networks, ICTON'09, 1-4, 2009.

[9] RODRIGUES, J. J. P. C., GREGORIO, J. M. B., VASILAKOS A. V.: Enhanced Just-In-Time Plus Protocol for Optical Burst Switching Networks, *J. of Optical Engineering*, vol. 49, No. 7, 2010.

[10] QIAO, C.: *Optical Burst Switching - A Novel Paradigm,* Optical Internet Workshop, 1997.

[11] YOO, M., QIAO, C.: Just-Enough-Time (JET): *A High Speed Protocol for Bursty Traffic in Optical Networks,* IEEE/LEOS Conference on Technologies for a Global Information Infrastructure, 26-27, 1997.

[12] QIAO, C., YOO, M.: Optical Burst Switching (OBS) - A New Paradigm for an Optical Internet, *J. of High Speed Networks*, vol. 8, No. 1, 69-84, 1999.

[13] TURNER, J. S.: Terabit Burst Switching, *J. of High Speed Networks*, vol. 8, No. 1, 3-16, 1999.

[14] BALDINE, I., ROUSKAS, G. N., PERROS, H. G., STEVENSON, D.: JumpStart: A Just-In-Time Signaling Architecture for WDM Burst-Switched Networks, *IEEE Communications Magazine*, 82-89, 2002.

[15] ZAIM, A. H., BALDINE, I., CASSADA, M., ROUSKAS, G. N., PERROS, H. G., STEVENSON, D.: The JumpStart Just- In-Time Signaling Protocol: A Formal Description Using EFSM, *Optical Engineering*, vol. 42, No. 2, 568-585, 2003.

[16] BALDINE, I., ROUSKAS, G. N., PERROS, H. G., STEVENSON, D.: Signaling Support for Multicast and QoS within the JumpStart WDM Burst Switching Architecture, *Optical Networks Magazine*, vol. 4, No. 6, 68-80, 2003.

Martina Vackova - Lucia Kovacova - Maria Kancirova - Peter Losonczi*

THE NEED FOR INNOVATION OF SECURITY EDUCATION FOR STRENGTHENING THE RESULTS OF TRADITIONAL TEACHING AT UNIVERSITIES

This article deals with security education and the possibilities of innovative methods application in this kind of education at universities. The continuous development of science and technology, as well as constantly changing social needs which are closely tied to the emergence of new security risks and threats, dynamically affect the entire system of education. Therefore, an urgent need for the implementation of these requirements arises. There is a need to address the number of key roles in the framework of security education in the spirit of innovation methods in order to improve teaching and, subsequently increase the efficiency of the process.

Keywords: *Security education, universities, innovative methods, efficiency.*

1. Introduction

The tasks of educational work, which are supposed to shape the personality of a student more effectively have significantly expanded and became more complicated at universities. Requirements for the educational process in college education continue to grow. The frequently changing social needs, but also the continuing development of science and technology which are closely tied to the emergence of new security risks and threats, dynamically affect the entire education system. In relation to this, the reaction to achievement of better results in the process of security education should be the creation of realistic assumptions to prepare professionals that are broadly educated, creatively-thinking, and capable of rapid adaptability to new conditions.

The aim of education is to prepare a group of experts for the management of security institutions of different sizes and varying degrees of complexity. It is necessary to develop and clarify its own terminology of the relevant terms, used in the preparation of security experts - in education of security management specialists [1]. The objectives and tasks of the teaching process help create optimal conditions for the conscious and positive activity of well-educated and skilled security managers of different specializations [2].

Digital world is a natural part of the daily life of university students nowadays. New hardware and software school equipment, such as interactive whiteboards, voting machines, laptops and tablets, but also freely available educational software, raise questions about their effective use in the educational process. Classic and innovative educational methods and forms have a new dimension in the digital environment. However, new methods closely connected with the usage of digital technologies arise [3].

In addition to technical and scientific progress, or rather in close connection with it, there are also lifestyle changes, acceleration of development of young people, new knowledge about learning and a number of other factors that significantly affect pedagogical practices so far. The basic aim of college education is the formation of fully trained cognoscente who is capable of self-complementation of knowledge, and self-increasing of general-theoretical and professional level.

2. The essence of security education at universities

When facing the challenges of security, education and training of security services personnel and personnel at different levels of security and management as well as managing workers on which high demands for knowledge are placed, are important aspects. Within university education, improvement of the situation in given area can be achieved by increasing of security awareness of people coming out of school to practice, and by deepening of their knowledge and skills associated with security, by improving qualifications and skills of graduates and preparing them to

* [1]Martina Vackova, [1]Lucia Kovacova, [1]Maria Kancirova, [2]Peter Losonczi
[1]Institute of Humanitarian and Technological Science, University of Security Management in Kosice, Slovakia
[2]Institute of Civil Security, University of Security Management in Kosice, Slovakia
E-mail: martina.vackova@vsbm.sk

deal with risk, emergency and crisis phenomena in various aspects of security (civil, economic, environmental, technical and technological, logistical etc.). We can therefore speak about security education, which contains the theoretical and practical knowledge about the protection of property and individuals with the possibility of versatile use in the European security environment. The essence of security education lies in the analytical work which is aimed at averting or minimizing security risks of different forms and causes to individuals and to society as a whole. Security education does not only concern the increasing of education and skills of security services workers, experts in the field of security and management personnel. It is appropriate to understand security education as an effective education system building security awareness and shaping attitudes to safety behaviour and actions in lifelong learning from preschool education to education of economically active citizens. However, we focus on security education and its innovation only in terms of universities for the purpose of this article [4].

The system of security education must be channelled in such a way that security personnel is allowed to obtain and acquire knowledge and understanding of methods by which they will be able to analyse the security environment and its factors in relation to various objects, identify and assess security risks and threats and predict their evolution, determine the procedures and management measures, security risks and threats, to plan and organize the risk management measures, security and crisis management in line with available resources and capacities to design and manage complex security systems [5].

3. Streamlining the traditional teaching safety education in conditions of universities

Deficiencies arising in security education prevent the effective investigation of professional and social aspects of security problems, by which the basic cognitive element for effective management of security systems is limited. The deficiencies arising in this education need to be addressed examined and removed. Therefore, the need to evaluate teaching in terms of pedagogical efficiency arises and we need to focus on innovation of this process. Security education at universities is in the direct teaching system realized by system of all elements of the teaching process.

With some simplification, it can be concluded that the objectives of this process are reflected in the content of teaching, which is implemented in the organizational forms, through methods using teaching aids and teaching techniques and in particular conditions its security. Each element of the teaching process represents internal possibilities of increasing efficiency of education [6]. All are interrelated and contingent. There is an objective connection between them, which impacts the overall results of the teaching processes. Therefore, the rate of quality and efficiency of this process is always dependent on the use of the opportunities of individual elements, as well as on achieving the most favourable relations between them [7].

Traditional teaching is mainly characterized by expository-illustrative and reproductive methods primarily used on lectures, exercises, and seminars. Both methods ensure the acquisition of final knowledge and experience from implementation of known ways of actions.

These two cognitive teaching methods ensure the preparation of the young generation for reproduction but cannot teach creative activity. It is possible especially in troubleshooting teaching. Traditional cognitive methods cannot be criticized; they are still needed in modern teaching. Pedagogue should however innovate the forms of their implementation, for example through the utilization of computer technology in appropriate learning activities of students [8].

Possibilities of increasing of the efficiency of education are relatively broad, however it is important to choose the most suitable in the particular conditions. For selecting appropriate possibilities of increasing of the effectiveness of education, it is primarily necessary to assess the current status of education in those conditions, which provide the foundation for solving of research problems in this important area of pedagogy.

4. Appropriate innovative methods in the process of security education at universities

Even before the advent of digital technology, creative, innovative pedagogues promoted constructivist learning instead of instructionist learning of scientific subjects.

Instructionist learning is linear and systematic, the student is passive in such learning. Pedagogue demonstrates and explains, while student copies and integrates. The content of learning is seen as a closed system of knowledge and elements. Many pedagogues currently prefer this instruction. The reasons are different (lack of time, lack of quality textbooks and instructional materials, lack of creativity of the pedagogue etc.).

Constructivist theory of cognition and learning was developed by Swiss psychologist Jean Piaget. He starts from the premise that the student in active interaction with the environment gradually creates its internal system of knowledge. Learning process should take place in an inspiring educational environment that inspires students to research. Pedagogical constructivism requires that teaching should only deal with solutions of real life (specific and authentic) problems. It also requires creative thinking, work in groups, manipulation with objects, visual aids e.g. interactive computer programs. In constructivism, more than one current exists, but all of them present learning with comprehension. Understanding is created by student himself (by considering new information, comparing them with previous experience, knowledge and schemes). It is therefore a problem teaching.

Constructivist approaches include the **management method to investigate** or **method ERR** (evocation - realization of meaning - reflection). Extraordinary effective form of controlled examination is **Workshops method**, foundation of which is a group form of education with the principle of small steps, while respecting the principle of succession and performance, active student learning and immediate verification and own pace.

Method of creating bunches of problems also belongs to this group. It is based on identification of a suitable initial problem, which will be solved by students together with a pedagogue. Subsequently, after the resolution of the initial problem, students solve similar problems that do not differ significantly from the original. The method of solution of these problems is the same or very similar to the method of the solution of the original problem. After specific training, the aim is to draw students into a process of creation of the new problems that are becoming more and more distant from the original problem.

After certain training, the aim is to draw students into the process of creating new problems in such a way that they will gradually form problems that are more and more distant from the original one.

Project teaching is another form of mobilizing of teaching of science subjects that takes entirely new opportunities in the digital environment. Digital technologies such as computer, internet, digital camera, mobile phone, etc. are an excellent means when working on projects for students not only in collecting information, but also in data processing, real life situations modelling, in the calculations and also in the presentation of results.

The problems that students solve in project education are complex and their solution will require the knowledge of several sciences (traditional teaching subjects). Problems are solved by groups of students (cooperative teaching) mostly from self-interest and solutions lead to concrete results, product, written reference, etc. Such complex problems are called projects.

Nowadays, the usage of digital technology (computers, internet, software, mobile phones, digital cameras, etc.) in teaching is for students a certain connection between a subject (subjects) and real life, in which these technologies are commonly used. Even at home environment, students are often informed about solutions to various projects in real-life for example by their parents, older siblings etc. It is common for employees of different companies that within their employment they work on solution of various kinds of projects and to some extent use the ICT solutions. At schools, pedagogues are involved in projects in order to obtain grants and purpose finances. Therefore mimicking this activity by students within a well prepared project teaching with usage of digital technologies can be considered as one of the most current methods of teaching.

In order for project method with the use of ICT to work well, the presumption for creative, educated, innovative and digitally literate pedagogue must be fulfilled.

Pedagogue forms his students by his direct and indirect activity, qualification, expertise, pedagogical mastery and also personal example. If we want to nurture young people info information-skilled and creative people with the ability to solve problems in real life independently, pedagogues must have their lessons prepared competently, creatively, interestingly and modernly.

The onset of digital technology came in two forms: instructionist and constructivist approach. Pedagogues often use Microsoft PowerPoint presentations or prepared Flipcharts with interactive whiteboard as a form of "taking notes" digitally instead of the classical blackboard and chalk. Often it is the only electronic support of "classical", instructionist teaching. New constructivist approach to teaching with digital support for students is very interesting, motivating, and energizing, but from the perspective of pedagogue also very challenging for preparation (e.g. professionally). The biggest reserves can be seen in the lack of quality teachers and accessible electronic materials for constructionist teaching.

In the digital environment of the school, forms of teaching are interesting. The term **e-learning** can be understood in two ways. Most often, e-learning is understood as an electronic form of distance education. However, e-learning can be broadly understood as electronic learning, which means all that a student can learn through computer, or any other electronic way [9].

Blended learning is a combined teaching - a combination of a standard teaching (attendance, presentation, face-to-face form) with e-learning. This term arose only recently, but our pedagogues have long used a combination of classical forms of teaching with electronic support (for example e-materials are placed on their own or school websites, or are placed on paid portals).

Unlike traditional approaches to education, e-learning is not just one approach that works for everybody. In addition to a wide range of learning styles, which make up the target group, learners most often seek information informally, in a conventional manner. In order for schools to fulfil the needs of formal and informal education, they need different types of educational strategies. These strategies will bring rich possibilities of educational resources to students, in a manner which will be most convenient in a given time, while the content will reflect objectives of the individual, as well as the overall objectives of the university. The reasons for the introduction of e-learning are practical. With respect to the possibilities of Internet services and computer programs it is possible to include "unexpected" benefits that classical education cannot offer. Creation of educational texts for e-learning, which must meet the psychological-didactic aspects are crucial to the active involvement of students in the educational process.

E-learning brings information development into universities and supports their progress. E-learning is simply education of the future.

Expository - problem method does not only acquaint students with found solutions to certain scientific-cognitive or practical problems and with the method of their applications, but it also shows the logic behind the solution of these problems.

The structure of expository-problem method may be summarized in following stages:

1. Demarcation of problem.
2. Solution process and its logic.
3. Solution process, possible and real difficulties and contradictions.
4. Solution and rationale of its correctness.
5. Clarification of the importance of solutions for the further development of thought or spheres of activity.

In expository - problem method, students follow the logic of pedagogue's interpretation, and control the persuasiveness of his practice and his reasoning in their thoughts. In case of errors in interpretation, doubts and questions arise. The more experience you acquire from creative activity, the more questions and doubts arise.

Students discover not only the final results of the research and the stage of development, but also the relationship between the stages, the paths of progress from one stage to another, typical deviations and obstacles arising in form of new problems. This way, students acquire individual stages of solution of complex problems.

Specificity of expository-problem method consists in the fact that the student does not only perceives, realizes and stores the information but also follows the logic of argumentation, thinking pedagogy, or the means that represent it (information technology, books, etc.). Students learn to think, watch the cogency of the arguments, retrieve them, learn how to be learned, and what is the basis for the development of creativity.

The research method serves for complete acquisition of experience from creative activity. Pedagogue identifies a problem, but solution rests on the student. This method is very well applied in the conception of project teaching where student's activity by its nature approximates activity of the planner, designer, technologist, and economist. Its essence is the submission of a certain task with more comprehensive nature, which should be resolved by students. After introducing the project task (problematic task), and clarification of the problem, solving by students follows. Problems that are addressed in the project teaching are complex; they do not need to be based solely on the content of subject matter, but especially from life. Their solution requires knowledge of several sciences (teaching subjects) the problems are being solved by groups of students (in collaborative learning) out of self-interest. Projects may be proposed by the students themselves, by pedagogues, other entity, or by pedagogues and students together [8].

5. Conclusion

The current trend of the onset of digital technology in the lives of students as well as schools inevitably raises questions of their effective use. Experience shows that the educational research as well as in the area of theory of teaching must move in this direction. It is necessary to create quality e-materials and e-tests for utilization in the teaching process. Universities preparing future teachers of science subjects should deal with these questions more intensely in the near future. We believe that by qualified usage of innovative methods of teaching in the digital school, the decreasing trend of popularity of science subjects at schools will stop and perhaps turns around. Possibilities of increasing the effectiveness of security education in conditions of universities are therefore really wide, but it is important to choose the most appropriate in the circumstances given. The need for innovation, security education is also reflected on the theme concept of protection of the population by 2020, with a view to 2030 in the Czech Republic when one of the strategic points of is the targeted support of science and research, development, innovation, stressing the exploitation of the results achieved in the application sphere within the system education and training of professionals. [10].

References

[1] LOSONCZI, P., MESAROS, M.: Asymmetry of Understanding Access to the Information Protection (in Slovak), *Kosicka bezpecnostna revue*, vol. 2, No. 1, University of Security Management : Kosice, 2012, pp. 28-30, ISSN 1338-4880.

[2] KOVACOVA, L., KLIMO, V.: Fundamentals of Security Education in the Process of Globalization, *Odes´kyi Politechnichnyi Universytet Pratsi*, No. 2 (41), 2013, Odesa, pp. 217-222, ISSN 2076-2429.

[3] LOSONCZI, P.: Information Security in Primary Schools (in Slovak), *Kosicka bezpecnostna revue,* vol. 4, No. 2, University of Security Management : Kosice, 2014, p. 77-82, ISSN 1338-4880.

[4] KOVACOVA, L.: *Safety Education from the Perspective of University Pedagogy* (in Slovak)*,* 3. Intern. conference, University of Security Management : Kosice, 2010, ISBN 978-80-89282-44-9.

[5] KOVACOVA, L.: Effective and Innovative Trends of Education Security Personnel in Conditions of Universities (in Slovak), *Kosicka bezpecnostna revue*, vol. 3, No. 1, 2013, University of Security Management : Kosice, pp. 67-72. ISSN 1338-4880.

[6] PETLAK, E.: *Options to Streamline the Educational Process* (in Slovak), Technologia vzdelavania tretieho tisicrocia: Zborník z vedeckej konferencie, jun 1995, Nitra, 142 p., ISBN 80-967339-1-5.

[7] TUREK, I.: *Increasing the Effectiveness of Teaching* (in Slovak), Bratislava : Metodicke centrum, 1997, ISBN 80-88796-49-0.

[8] BLASKO, M.: *The Quality in the System of Modern Teaching - 1st ed.* (in Slovak), Department of Engineering Education : TUKE, 2013, 402 p., ISBN 978-80-553-1281-1.

[9] KORENOVA, I.: *Didactic Aspects of Digital Learning School Mathematics* (in Slovak), Proc. of VI conference Uziti pocitacu ve vyuce matematky, December 2013, Ceske Budejovice : Jihoceska unverzita, ISBN 978-80-7394-448-3.

[10] KAVAN, S.: *New Concept of Protection of the Population in the Czech Republic* (in Czech). Zbornik vedeckych prac Manazerstvo bezpecnosti, University of Security Management : Kosice, 2013, 198 p., ISBN 978-80-89282-92-0.

Katarina Valcova - Martina Pavlikova - Marie Roubalova*

RELIGIOUS EXISTENTIALISM AS A COUNTERMEASURE TO MORALISTIC THERAPEUTIC DEISM

Moralistic Therapeutic Deism (MTD) is a new spiritual trend that is increasingly changing the religious landscape of in the Euro-American cultural provenience. Though appealing to the generation of 'emerging adults' (age 18-25), MTD results in undesirable life-outcomes that prompt religious scholars, theologians, and sociologists to carefully study its roots, expressions, and possible alternatives. Using S. Kierkegaard's religious existentialism as a case study, the authors argue that religious existentialism has the potential to face the detrimental effects of MTD and to reinvigorate stagnating religious communities.

Keywords: *Religious Existentialism, Tradition, Moralistic Therapeutic Deism, consequential Faith, Kierkegaard.*

1. The Roots, Characteristics, and Life-Outcomes of Moralistic Therapeutic Deism

'Moralistic Therapeutic Deism' is a *terminus technicus* coined by the American sociologist, professor Christian Smith (University of Notre Dame), following his extensive nation-wide survey called: National Study of Youth & Religion [from now on cited as NSYR]. The project was designed to improve our understanding of the religious lives of American youth from adolescence into what Smith (among others) came to call "emerging adulthood" [1], using telephone survey and in-depth interview methods in several consecutive waves (2003-2015) [2]. Since the beginning of the NSYR, more than three thousand (precisely: 3,290) 13-17 year-old teenagers have been questioned through a telephone survey [3], and as these individuals aged, ca. 80% of them (2,458) were questioned again as 18-23 year-olds [1, p. 3]. In Smith' own words, the survey was "a national, random-digit-dial telephone survey of U.S. households containing at least one teenager age 13-17, surveying one household parent for about 30 minutes and one randomly selected household teen for about 50 minutes'; this was followed later in 2003 by 267 'in-depth, face-to-face interviews with a subsample of the telephone survey respondents in 45 states" [1, p. 6]. Other rounds of interviews were conducted in 2007 and 2008, then in 2010, and 2014-15. The uniqueness of the survey, in addition to its methodological cogency, rests primarily in its longevity that provides researchers the invaluable opportunity to make comparisons between different stages of young adulthood, "studying how socialization, culture, practices, and individual choices interact in shaping religious faith of this age group" [4, p. 117]. Summarizing the findings of his research team, Smith claims that "the de facto dominant religion among contemporary teenagers in the United States is what we might call 'Moralistic Therapeutic Deism'" [5, p. 46], the creed of which might be summarized in the following way: "(1) A God exists who created and orders the world and watches over human life on earth; (2) God wants people to be good, nice, and fair to each other, as taught in the Bible and by most world religions; (3) the central goal of life is to be happy and to feel good about oneself; (4) God does not need to be particularly involved in one's life except when God is needed to resolve a problem; (5) good people go to heaven when they die" [5, pp. 46-47].

Even though no teenagers will use the term "moralistic therapeutic deism" (from now cited as 'MTD'), their attitudes, life choices, as well as life outcomes seem to be profoundly influenced by it. [MTD] is about (1) adopting a moralistic approach to life, believing that central to living a happy life is being a good, moral person; (2) it is also about providing certain therapeutic benefits to its adherents; (3) and finally, it is about a belief in a God, who created the world, and the laws in it, but keeps away at a safe distance. God is "not one who is particularly personally involved in our affairs – especially affairs in which we would prefer not to have God involved. Most of the time, the God of this faith keeps a safe distance" [5, p. 49].

* **[1]Katarina Valcova, [2]Martina Pavlikova, [3]Marie Roubalova**
[1]Department of Religious Studies, Faculty of Humanities, University of Zilina, Slovakia
[2]Department of General and Applied Ethics, Faculty of Arts, Constantine the Philosopher University, Nitra, Slovakia
[3]Hussite Theological Faculty, Charles University in Prague, Czech Republic
E-mail: katarina.valcova@fhv.uniza.sk

The growth to maturity of those affected by the MTD is inhibited by their overall vision of life. The type of religious faith they embrace is closely connected to the kind of social, cultural and economic environments they grow up in. Extension of formal schooling, increase in average age of entering into marriage, high demand on flexibility and mobility in the globalized job market, and the extension of parental support well into the young adults' twenties – these factors contribute to a delay in 'settling down' and to the presence of significant uncertainty in the lives of emerging adults. Furthermore, emerging adults cyclically face the process of internal investigation, creating overwhelming angst in the eyes of Mike Hayes, a University of Buffalo Catholic campus minister [6, p. 141]. Such a self-orientation often withdraws concern for others and engages individuals in a race where personal fulfillment is the ultimate goal. As emerging adults attempt to maintain their choices, and believe that they do not stand in a position to judge others, religious devotion is often something that is of no, or very low priority [1, p. 48].

Embracing a "nice guy" image of a 'god' who is called upon only as needed, and who definitely makes no real demands on the daily lives of young adults, may, at first, seem like a desirable spiritual approach to religion. This spiritual outlook may be celebrated for his therapeutic benefits in terms of making people feel good about themselves and secure vis-à-vis numerous life challenges. Yet, despite its 'niceness' and therapeutic appeal, the MTD lacks genuine transformative power and, in crucial moments when difficult decisions are to be made, it fails to provide adequate guidance and motivation to its proponents. The consequences, or the so called 'life outcomes' of this attitude are less than desirable for many of the young people studied in the survey. As Smith insightfully summarizes at the end of his *Lost in Transition* (2011):

"In the popular imagination, these early adult years are filled with youthful fun and freedom enjoyed in the prime of life. For some, this image is true. The actual reality for many, however, is instead one of personal struggle, confusion, anxiety, hurt, frustration, and grief... the long transition from the teenage years to full adulthood in America has a dark side that is not often talked about or portrayed by mainstream media. Some cultural activists and even academic scholars seem eager to celebrate and promote what we consider to be troubling aspects of emerging adult life. We are far less sanguine" [7, pp. 226-227].

By asserting that everything is relative to the situation, emerging adults initiate no standard to fulfill. Such lenience in society devalues all entities and redistributes value to one's personal and subjective intuition. Unwittingly embracing a social constructionist outlook on reality, most emerging adults "presuppose that most things about the sociocultural world are not fixed or given facts of nature but rather human constructions invented through shared social definitions and practices that are historically contingent, changeable, and particular" [1, p. 50]. In addition to this, in their quest to achieve sociality (social inclusion), emerging adults are consuming large amounts of intoxicating substances, and are minimizing their involvement in the political as well as civic world [4, pp. 118-119]. They do not seem to have, or know how to operate, the intellectual instruments needed for an engaged discussion, since they live a world where "claims are not staked, rational arguments are not developed, differences are not engaged, ... and universals are not recognized. Rather, differences in viewpoints and ways of life are mostly acknowledged, respected, and then set aside as incommensurate and off limits for evaluation" [1, p. 51]. In his book *Lost in Transition* (third in the series of research monographs on the subject), Smith depicts concrete detrimental effects of MTD in the lives of emerging adults. In their attempt to improve their social status and prestige, the emerging adults are instead confronted by the negative implications that stem from regular intoxication. With social matters controlling the orientation of countless young adults, civic involvement has subsequently deteriorated to a ghostly level, which limits the possibilities for communal gains [7, p. 142]. Drunkenness and a lack of participation beyond social entities unquestionably harm the individual in question, but it also endangers innocent bystanders as well as the surrounding community [8].

2. The Responsibility and Potential of Institutional Religion

The MTD approach to religious belief is particularly evident among mainline Protestant and Catholic youth, though it can also be found among black and conservative Protestants, Jewish teens, and other religious teenagers in the United States and in Europe. Church representatives, theologians, and religious sociologists are noticing major shifts in institutionalized Christianity: "either Christianity is at least degenerating into a pathetic version of itself or, more significantly, Christianity is actively being colonized and displaced by a quite different religious faith" [5, p. 57]. Traditional, creedal faith and the corresponding piety of historic Christianity appears to be waning, while the old, ecclesiastical and social structures are outwardly still in place. Or, as Horton points out, "we are living out our creed, but that *creed* is closer to the American Dream than it is to the Christian faith" [9, p. 21]. MTD seems to by "colonizing many historical religious traditions and, almost without anyone noticing, converting believers in the old faiths to its alternative religious vision of divinely underwritten personal happiness and interpersonal niceness" [5, p. 56]. This is an interesting phenomenon because both in Europe and in the United States, most people who identify themselves as Christians are "actually only tenuously connected to the actual historical Christian tradition", yet in their religious outlook they have, according to Smith:

"Substantially morphed into Christianity's misbegotten step-cousin, Christian Moralistic Therapeutic Deism. This has

happened in the minds and hearts of many individual believers and, it also appears, within the structures of at least some Christian organizations and institutions. The language—and therefore experience—of Trinity, holiness, sin, grace, justification, sanctification, church, Eucharist, and heaven and hell appear, among most Christian teenagers in the United States at the very least, to be being supplanted by the language of happiness, niceness, and an earned heavenly reward" [5, pp. 56-57].

This emotive spirituality with a flavor of a watershed Christianity seems to be the result of a long term neglect on the side pastors and teachers in the church. Kenda C. Dean blames Christian leaders in the established churches, accusing them of being unfaithful by preaching and teaching a religion that is 'nice', but not transformative [10]. Youth is quick to follow their example. They think about what it means to have faith exactly as their faith models have taught them to think about it. Their belief content boils down to a 'benign positive regard' for one another, for those of other faiths who hold different convictions, and for the world in general. The MTD faith – a faith that serves to sooth us, to make us feel 'nice' and comfortable about ourselves, thus exhibits the traits of yet another post-modern version of religious idolatry.

On the other hand, it is precisely the organic, communal context of historic Christian communities, in which the mission is carried on and which has the necessary tools to help the emerging adults attain spiritual maturity that produces positive life outcomes. This can clearly be seen in Layton (et al.) who discusses seven anchors of religious commitment: (a) religious traditions, rituals, and laws; (b) God; (c) faith traditions or denominations; (d) faith community members; (e) parents; (f) scriptures or sacred texts; and (g) religious leaders [11]. In contrast to previous studies that depicted religious commitment as primarily a behavioral or attitudinal construct, the findings of Layton's study present a new conceptual understanding of commitment as primarily a relational construct. Teenagers who have highly religiously involved parents, and who grow up in healthy religious communities [12], who pray regularly, read Scripture frequently, and attend worship services routinely, are likely to be consistently involved in a religious community setting, hold faith in high regard, and practice prayer frequently. By developing homiletic, pastoral and catechetical ministries on the basis of relationships, learning, and personality development, emerging adults' level of consequential faith can escalate. According to Smith, "...every most-likely path to highly religious emerging adulthood must include combinations of distinctly different kinds of causal factors, almost always including groupings of relational, personal-subjective, and devotional-practice factors" [1, p. 227]. Whether we follow religious existentialists and call this type of faith "authentic faith", or whether we follow Smith's terminology and call it "consequential faith", the NSYR states that those with "authentic"/"consequential" faith are achieving at a notably higher level desirable life outcomes than those who identify themselves as "Christians" but internalize the spiritual outlook of MTD. Authentic, consequential faith produces more personal interactions with parents, a higher probability of abstaining from addictive substances and promiscuity, and a higher level of intentional involvement in the Christian community and in civic society [3, pp. 261, 274].

"Religion provides teenagers with moral directives, confirming spiritual experiences, role models, community and leadership skills, coping skills, cultural capital, social capital, greater network closure in relationships, and intercommunity links-all of which, solid social scientific reasoning indicates, can be expected to enhance their life experiences. Most of the same mechanisms apply equally well in the lives of emerging adults... Religion therefore matters and makes a significant difference in the life experiences, beliefs, and behaviors of American 18- to 23-year-olds" [1, pp. 277-278].

Smith' findings are further supported by Dollahite who conducted a valuable research about the willingness of religious adolescents to make personal sacrifices [13]. Based on the results of his research, adolescents indicated sacrifices in five major domains: (1) societal expectations; (2) popular culture; (3) comforts and pleasures; (4) time and activities; (5) and peer relations. More revealing, however, were the reasons the young people gave for their willingness to make sacrifices. Connecting to God or a higher meaning or purpose played a high importance, as well as connecting to the faith tradition or community they were coming from. Among other motivational factors were: feeling affective benefits, fulfilling expectations (of either their peers, parents, or faith community leaders), and avoiding problems.

Whether it is the question of inner motivation, or the question of identifying concrete, tangible life outcomes [14], people of all ages (but especially young adults) "make their decisions in social contexts that powerfully influence the time, direction, and impact of those choices" [7, p. 141] and these social contexts must be taken seriously and should be intentionally employed in shaping the emerging adults' lives.

3. The Promise of Existential Faith

This leads us back to the topic of "authentic", "consequential" faith. The constitutive characteristic that we wish to highlight in this respect is its existential nature. Neither a "cultural Christianity", nor an emotivistic MTD spirituality, nor an "exclusively humanist" (secularistic) agenda [15] will generate the kind of life outcomes that bring fulfilment and contentment to individual human subject and a general well-being to society.

A short note on why speak about "faith" at all is in order here. The authors of this study are convinced that the spiritual aspect of human life cannot be explained away by biological, psychological, or social influences, or any combination of these three. Christian theological anthropology understands human

being as a personal, holistic unity of the above mentioned aspects, integrating them with transcendent, spiritual realities. These provide an invaluable life orientation, inner motivational force, along with a structure of meaning and purpose. While the spiritual aspect of human existential experience can be expressed through biopsychosocial media [16], it should be distinguished as a unique, separate anthropological entity that overlaps into transcendent reality. Spiritual etiology, among other things, adds valuable insight into the multifaceted socio-ethical discourse in the contemporary debate between the secularists and those who point out a world-wide resurgence of religious traditions and new forms of spirituality. Though we no longer live in an "enchanted world", as Charles Taylor points out [17, p. 26], in which faith in God and the presence of a spiritual realm permeated daily life, humans (as individuals as well as societies) must find ways to constructively integrate experiences of the transcendent into the complex fabric of what constitutes our human predicament. The challenge is that, as postmodern thinkers, we are "so influenced by the age of reason that we cannot resist the urge to rationalize Christianity"; we are affected by the "general tendency of much of modem theology to rationalize, psychologize, or historicize some basic teaching of Christianity in order to defend it. The danger is that by so defending Christianity we compromise its very essence" [18, p. 197].

Restorationist sentiments in some contexts of institutionalized religion suggest that the best (or even only possible) way forward is to look backwards and simply perpetuate old forms of liturgy and catechetical instructions. Their expectation is that memorizing the correct answers from Catechism will restore people's faith commitment and enhance their spiritual vitality. Another proposed venue is to organize "exciting" and culturally "appealing" events during which young adults can experience intense feelings of belonging, happiness, fulfilment, etc. However, not extravagant events, nor mechanistic repetition of old doctrinal formulae or rituals, but a purposeful development of disciples through divine action is the proper and only adequate tool of existential transformation [10, p. 98]. Existential (consequential) faith emerges in historic, liturgical community of faith [19]. It is characterized by the Christian creed that summarized the vision and purpose of life as a gift of God in the narrative framework of the story of creation, redemption, and fulfillment. The current interest in grand narratives and fantasy stories across world-wide audiences connects imagination, human desire for meaning, with the concreteness of one's mundane life. These artistic renderings link past and present pieces of our shattered reality with a large supporting story – a "purpose framework". Many literary works attempt to return to the supporting story (an important aspect of modernism), but place them in a postmodern fantasy framework. Much of current literature reflects the human fear of nothingness, hopelessness, despair over the past, and anxiety about an uncertain future. Historic religious communities have the potential to "connect the dots", so to speak, to link the shattered pieces within the lives of individuals and put together (socially) the seemingly random mosaic of contradictory and competing desires and fears of human subjects into a community with a shared vision, ethos, and purpose. Consequential faith as an inner, existential disposition of the human subject, arising as a result of one's participation in the embodied narrative of the Gospel, has the potential to be this kind of bond of cohesion and life motivation.

Here, however, it is important to remember that "Faith does not mean mimicking Jesus, but participating in his self-giving love – not because we have somehow chosen to be like him, but because incredibly, God has chosen to become like us" [10, p. 104]. We can only imitate Him because we are embraced by Him; we can only follow because He is holding us, leading us, animating us (spiritually) from within. Transcendence lays its claim on humans, which enables humans to enter the process of "metanoia" (i.e., change of mind) and, subsequently, to recalibrate their moral compass and step forward on the path of selfless life (though never perfectly!). In any case, Evans' words on this topic serve as an important reminder that "[t]he quest for truth, at least the truth about the most important things, cannot be divorced from the quest to become the kind of person we need to become" [20, p. 26]. Personal transformation is an inevitable side effect of consequential, existential faith.

Soren A. Kierkegaard, above other religious existential thinkers, may serve as an incisive case study of religious existentialism. Kierkegaard, in fact, is considered the 'father' of religious existentialism [21]. This controversial Danish intellectual "feared that Christendom – the alliance of church, state, and middle-class culture – had falsely given the impression that one can have genuine faith simply by adhering to the norms of the allegedly Christian society, without undergoing a true transformation of one's passional life" [22, p. 105]. He was convinced that "Christendom bred the false and spiritually lethal attitude that faith is a matter of mere conformity. This attitude generated smugness, complacency, and militated against any passionate desire to be inwardly transformed" [22, p. 105-106]. The roots of Kierkegaard's disenchantment with the established religious culture of his day go to his fundamental understanding of anthropology. Kierkegaard defines the human self "relationally by means of an existential ontology of freedom and potentiality. The human self is thus being conceived of as an 'emerging' reality which is being constituted in the vary act of relating to itself. To become a fully constituted self, however, Kierkegaard maintains that the 'self-aware' self must relate its own relatedness to the divine origin of its being. For Kierkegaard, this divine source of each self's existence is a personal God, the God of Abraham, Isaac and Jacob, the Father of Jesus (Logos incarnated)" [23, p. 99]. Existential passion is both the product and vehicle (medium) of this self-aware relatedness to Transcendence [24, p. 699]. Existential faith, thus, must be passionate, otherwise it is not authentic faith. This "passion of faith", then, "requires the internalization of communal

convictions and values. Kierkegaard's basic complaint with the church was that it had ceased to make it clear that Christian concepts must be appropriated by individuals with personal passion and commitment. He saw his task as reintroducing the struggle and passion of personal transformation back into the lives of the individuals who constitute the community" [22, p. 106]. Kierkegaard's religious existentialism, including, above all, the "emphasis on the role that passion and inwardness of the human subject play in the acquisition of religious truth and in the living out of that truth may be considered one of Kierkegaard's most important and enduring contributions" [22, p. 106].

4. Conclusion: Lessons for the Ministry to the 'Emerging Adults' Generation

The Kierkegaardian type of religious existentialism points us to the essential questions of anthropology with the aim to call human individuals to authentic subjecthood. Experiencing a self-aware, authentic existence, however, is only possible relationally in a three dimensional interplay of personal being: the self's relation to itself; the self's relation to others; and the self's relation to personal Transcendence, which is constitutive to all other relations. "Man's relatedness to God is constitutive to human personhood" [25, p. 73]. It is not the MTD infected self that picks and choses relations based on whimsical preferences; it is, rather, a new existential outlook of the relational self, embraced by Transcendence (or Spirit) [26, p. 18] - defined Christologically and Trinitarilly - that should be the goal of Christian catechesis. To facilitate such catechesis, the passing down of authentic faith, emerging adults must be guided within the community, to personally acquire the Christian creed (i.e. a theocentric vision of life), purpose (i.e. inner motivation), and hope [27, p. 215] that correlates to consequential-existential faith. Engaging in religious conversations that aspire to spawn existential faith in the hearers is critical to the development of Christian identity [28], as is a robust proclamation and enactment (both liturgical and diaconal) of the constitutive religious narrative – the triune Gospel – in the context of a living, teaching, witnessing and serving faith community. Rather than predominantly speaking from behind the altar or sponsoring big events, emerging adult need to interact with individuals (who exemplify the authentic faith) in a small setting, building gradually their religious vocabulary and witnessing abilities along the way. Effective communication equates to intentional participation from both parties, and the pace is set by those being ministered to. Upon the creation of religious speaking abilities, consequential-existential faith "transcends into a communal zone, empowering participation that previously seemed futile. Partnerships with members of the Body of Christ grant emerging adults a personal ministerial resource that serves as a guide as well as an advocate" [4, p. 133].

References

[1] SMITH, C. & SNELL, P.: *Souls in Transition: The Religious and Spiritual Lives of Emerging Adults*. New York: Oxford UP, 2009. ISBN 978-0195371796. Smith described the following socio-cultural changes that formed the lives of contemporary young adults, creating a new social group – "the emerging adults": (a) the extension of formal schooling, with college as the norm; (b) increases in the average age of marriage for both men and women; (c) changes in the global economy, which demand flexibility and change, rather than the projected stability of a life-long career; (d) the extension of parental support well into one's twenties. Ibid., p. 5-6.

[2] SMITH, C. et al.: *National Study of Youth & Religion* [online]. Notre Dame: University of Notre Dame, 2003-2015. Available on the Internet: http://youthandreligion.nd.edu/. [15/01/2016].

[3] The first results of the survey were published in: SMITH, C. & DENTON, M. L.: *Soul Searching: The Religious and Spiritual Lives of American Teenagers. New York: Oxford University Press, 2005. ISBN 978-0195384772.*

[4] VALCOVA, K.: *Religiosity of the Emerging Adults and Cultural Challenges of the Fragmented Reality of Post-Modernism: Lessons to Learn from the us Context.* S. Sorys, T. Bratkowski (eds.), Wiedza, umiejetnosc, kompetencje wobec przemian XXI wieku: spoleczenstwo - tozsamosc - religia. Krakow: Drukarnia Eikon Plus, 2014, pp. 116-134. ISBN 978-83-60391-77-8.

[5] SMITH, C.: *On "Moralistic Therapeutic Deism" as US Teenagers' Actual, Tacit, de facto Religious Faith.* S. Collins-Mayo (ed.), Religion and youth. Aldershot: Ashgate, 2010, pp. 41-46. ISBN 9780754667681. This book further develops the interpretations published in: SMITH, C.: Is Moralistic Therapeutic Deism the New Religion of American Youth?: Implications for the Challenge of Religious Socialization and Reproduction. In J. L. Heft (ed.), Passing on the faith: Transforming traditions for the next generation of Jews, Christians, and Muslims. New York: S. M. Fordham University Press, 2006, pp. 55-74. ISBN 9780823226481.

[6] HAYES, M.: *Googling God: The Religious Landscape of People in Their 20s and 30s*. New York : Paulist Press, 2007. ISBN 978-0-8091-4487-7.

[7] SMITH, C., et al.: *Lost in Transition: The Dark Side of Emerging Adulthood*. New York: Oxford UP, 2011. ISBN 9780199828029.

[8] The NSYR showed that emerging adults are consuming alcohol and drugs extensively with intoxication being the direct result. By the age of 19, only 14% of young adults have not been drunk before, and of the 78% who do consume alcohol, 60% have indulged in binge drinking within the last two weeks [7, pp. 110-111]. High numbers of emerging adults are participating, or once did, in dangerous activities, as partiers, addicts, recovered addicts, and recovered partiers represent 55% of the entire NSYR, which clearly shows the dominant presence of intoxicating substances in emerging adult culture. Cf.: [4, p. 128].

[9] HORTON, M. S., WILKEN, T.: *Christless Christianity: The Alternative Gospel of the American Church*. Grand Rapids : Baker Books, 2008. ISBN 978-0-8010-1318-8.

[10] DEAN, K. C.: *Almost Christian: What the Faith of Our Teenagers is Telling the American Church*. New York: Oxford UP, 2010. ISBN 978-0-19-531484-7.

[11] LAYTON, E., DOLLAHITE, D. C., HARDY, S. A.: Anchors of Religious Commitment in Adolescents. *J. of Adolescent Research*, vol. 26, No. 3, 2011, 381-413. ISSN 0743-5584.

[12] Cabanova points out the importance of family context for a "transgenerational transfer of values" in: CABANOVA, V.: Social Attitudes and Values of Young People in the Context of Multicultural Education. *Communications - Scientific Letters of the University of Zilina,* No. 3, 2014, 4-8, ISSN 1335-4205. On the topic of parental involvement in the transfer of values, see also: Pala, G.: The Family in Media Ciphers. *European J. of Science and Theology*, vol. 11, No. 6, 2015, 45-56. ISSN 1841-0464; PETRO, M.: Clovek ako osoba - nadej pre Europu. In P. Dancak (ed.), Sapiencialny charakter antropologie Jana Pavla II. v kontexte europskej integracie, Presov: University of Presov, 2006, 96-111.

[13] DOLLAHITE, D. C., et al.: Giving Up Something Good for Something Better Sacred Sacrifices Made by Religious Youth. *J. of Adolescent Research*, vol. 24, No. 6, 2009, 691-725. ISSN 0743-5584.

[14] For a good study on the presuppositions and motives for toleration (as one of the most important life-outcomes), see: Jurova, J.: Ethical Aspects of Toleration from Liberal Perspective. *European J. of Science and Theology*, vol. 12, No. 1, 2016, 139-145. ISSN 1841-0464.

[15] Exclusive humanists, according to Taylor [13], promote human self-sufficiency and "expressive individualism" [13, p. 213] as the fulfillment of our human potential, without any reference to transcendent sources, values, or hopes.

[16] ENTWISTLE, D. N.: A Holistic Psychology of Persons: Implications for Theory and Practice. *J. of Psychology and Christianity*, vol. 28, No. 2, 2009, 141-149. ISSN 0733-4273.

[17] TAYLOR, C.: A Secular Age. Cambridge Massachusetts: Harvard University Press, 2007. ISBN 978-0-674-02676-6.

[18] VALCO, M.: Alternative Viewpoint: Edwards and the World Religions. McDermott, G. (ed.). Understanding Jonathan Edwards: An introduction to America´s theologian. Oxford: Oxford University Press, 2009, pp. 195-200. ISBN 978-0-19-537344-8.

[19] Lutheran reformers thus emphasized the constitutive character of liturgy, culminating in the Eucharistic feast (the Lord's Supper). In the words of Martin Chemntiz: "The Lord's Supper is closely connected to a person's faith and thus also to his/her salvation or condemnation." Valco, M.: Scripture as a source of Chemnitz' christology. *Communications: Scientific Letters of the University of Zilina*, vol. 9, No. 2, 2007, 5-8. ISSN 1335-4205.

[20] EVANS, S.: *Why Kierkegaard Still Matters - and Matters to Me. Why Kierkegaard Matters: A Festschrift in Honor of Robert L. Perkins.* Robert L. Perkins, Marc Alan Jolley & Edmon L. Rowell (eds.), Mercer University Press, 2010.

[21] The topic of religious existentialism, its origins as well as current intellectual potential, attracts high attention today. In relation to Kierkegaard, see: Kralik, R.: Marie Mikulova Thulstrupova (1923-2013) a jeji tvorba. *Filosoficky casopis*, vol. 61, No. 3, 2013, 439-442; the legacy of D. Bonhoeffer, the German Christian existentialist, strongly influenced by Kierkegaard, is discussed in: Ligus, J.: Obedience or resistance: The legacy of Bonhoeffer. *European J. of Theology*, vol. 24, No. 2, 2015, 173-182. ISSN 0960-2720, KONDRLA, P., KRALIK, R.: Authentic Being and Moral Conscience. *European J. of Science and Theology,* vol. 12, No. 4, 2016, 155-164. ISSN 1841-0464.

[22] VALCO, M., KRALIK, R., BARRETT, L.: Moral Implications of Augustine's Philosophical and Spiritual Journey in his Confessiones, *Communications - Scientific Letters of the University of Zilina*, vol. 17, No. 2, 2015, 103-108. ISSN 1335-4205. On this topic, see also: Kralik, R., Torok, L.: The Moment - Kierkegaard's Attack upon Christendom. *European J. of Science and Theology*, vol. 12, No. 3, 2016, 112-120. ISSN 1841-0464.

[23] VALCO, M.: Kierkegaard's 'Sickness unto Death' as a Resource in our Search for Personal Authenticity. *European J. of Science and Theology*, vol. 12, No. 1, 2016, 97-105. ISSN 1841-0464.

[24] REPAR, P.: Existential Dialectics of Choice and Scandalon (in Slovak). *Filozofia*, vol. 67, No. 8, 2012, 689-704. ISSN 0046-385X.

[25] KRALIK, R., TOROK, L.: Concept of Relationship God-Man in Kierkegaard's Writing 'What we learn from the lilies in the field and from birds in the Air'. *European J. of Science and Theology*, vol. 12, No. 2, 2016, 65-74. ISSN 1841-0464.

[26] HAJKO, D.: Sarvepalli Radhakrisnan and "Religion of Spirit" (in Slovak). *Filozofia*, vol. 59, No. 1, 2004, 18-19. ISSN 0046-385X. Summarizing Radhakrisnan's philosophy, Hajko points out the potential of a synthesis of mysticism and social action (social work).

[27] DANCAK, P.: Concreteness of Life as the Context of Thinking in the Philosophy of Jozef Tischner. *European J. of Science and Theology*, vol. 12, No. 2, 2016, 213-221. ISSN 1841-0464.

[28] KONDRLA, P.: Apophatic Theology, Philosophy of Dialogue and Communication Perspectives. *European J. of Science and Theology*, vol. 12, No. 1, 2016, 117-128. ISSN 1841-0464. Kondrla reminds us that "Immediacy of essential relationship with God is also at the center of interest for philosophers of dialogue. Philosophy of dialogue was based on Jewish and Christian roots, especially personalism and Cosmology." (Ibid., p. 124).

Peter Hanuliak - Michal Hanuliak*

HYBRID ANALYTICAL PERFORMANCE MODELS OF PARALLEL COMPUTERS

The paper is a follow up to the performance modelling of dominant parallel computers (NOW, GRID) with analytical hybrid models (mixed) based on combinations of a more precisely developed corrected standard analytical model (M/M/1and M/M/m) and on an improved analytical model (M/D/1and M/D/m). Firstly, the paper briefly describes the steps of development of parallel computer architectures and then it summarises the basic concepts of performance evaluation. In the case of using SMP parallel system as node computer, the suggested models use for all the activities of node combinations of M/M/m and M/D/1 or M/D/m and M/M/1different (hybrid) queuing theory systems. The achieved results of both developed analytical models have been compared with the results of other alternative evaluation method, based on simulation, to verify the accuracy of developed analytical models. All developed and presented analytical models could be used for various real ranges of input parameters which influence the final performance of analysed connected computing nodes for the practice.

Keywords: *Parallel computer, computing node, network of workstation (NOW), Grid, analytical modelling, queuing theory, performance evaluation, queuing theory system, simulation.*

List of used symbols and short cuts

- *M/M/1, M/M/m, M/D/1, M/D/m – queuing theory systems*
- *Grid – high integration of computer based networks (NOW, LAN, WAN etc.)*
- *FIFO – queue for service on a first-come first-served basis*
- *HPC – high performance computing*
- *LAN – local area network*
- *NOW – network of workstations*
- *p – number of computing nodes of parallel computer*
- *PC – personal computer*
- *SMP – symmetrical multiprocessor system*
- *WAN – wide area network*
- *ws – workstation as a computing node of NOW and Grid.*

Queuing theory symbols

- λ *– arrival rate at entrance to a queue*
- *m – number of servers in the queuing theory system*
- ρ *– traffic intensity (dimensionless coefficient of utilisation)*
- $\rho \equiv R = \lambda \,.\, E(t_s)$ *– traffic intensity ($0 < \rho < 1$) of one service*
- $\rho = \lambda \,.\, E(t_s) / m$ *– traffic intensity of m servicing servers*
- *q - random variable for the number of customers in a system at steady state*
- *w - random variable for the number of customers in a queue at steady state*
- γ *- sum of individual extern intensities γ_i to the i-th NOW module in the Grid*
- *N – number of used NOW modules*
- u_i *– number of used communication channel in i-th computing node*
- $\underline{U}$ *– number of network computing nodes*
- $\overline{x_i}$, $\overline{x_{ij}}$ *– constant service time in i-th – computing node resp. in i-th computing node and j-th communication channel*
- $E(t_s)$ *– the mean service time of a server*
- $E(q)$ *– the mean number of customers in a system at steady state*
- $E(w)$ *– the mean number of customers in a queue at steady state*
- $E(t_q)$ *– the mean time spent in system (queue + servicing) at steady state*
- $E(t_w)$ *– the mean time spent in the queue at steady state*
- c_{ij} *– correction factor*

* **Peter Hanuliak, Michal Hanuliak**
Dubnica Institute of Technology, Dubnica nad Vahom, Slovakia
E-mail: phanuliak@gmail.com

1. Periods of development of parallel computers

Scientific supercomputers which dominated in the first period of parallel computers between 1975 and 1995 were specially designed for high performance computing (HPC). These parallel computers mostly used computing models based on data parallelism. Those systems were ahead of standard common computers in terms of their performance and price. Increased processor performance was reached by the massive use of various parallel principles in all forms of produced processors. The parallel principles were used not only in single PCs and workstations (Scalar or super scalar pipeline or symmetrical multiprocessor systems SMP) [1 and 2] but also in extremely powerful PCs in form of variously connected network of workstations (NOW, cluster). Gained experience with the implementation of parallel principles and intensive extensions of computer networks led to the use of connected computers for parallel solution. We can define this period as the second period of development. Development trends were actually heading towards building widely spread connected NOW networks with high computation and memory capacity (Grid). Conceptually Grid comes to the definition of meta computer [3], where meta computer could be understood as a big computer network consisting of a massive number of computing nodes, memories and other needed resources that jointly create an illusion of one single powerful supercomputer. These highly integrated forms of NOWs create various Grid systems or meta computers that could be defined as the third period of parallel computers.

2. Typical architectures of modern parallel computers

2.1. Symmetrical multiprocessor system

Symmetrical multiprocessor system (SMP) is a multiple system using the same processors or cores which are implemented on motherboard in order to increase the whole performance of such system [4]. Typical common characteristics are as follows:

- each processor or core (Computing node) of the multiprocessor system can access the main memory (Shared memory),
- I/O channels or I/O devices are allocated to individual computing nodes according to their demands,
- integrated operation system coordinates cooperation of whole multiprocessor resources (Hardware, software etc.).

2.2. Network of workstations

Actually there has been an increasing interest in the use of networks of workstations (NOW) connected together by high speed networks for solving large computation intensive problems. This trend is mainly driven by the cost effectiveness of such systems as compared to massive multiprocessor systems with tightly coupled processors and memories (Supercomputers). Parallel computing on a network of workstations connected via high speed networks has given rise to a range of hardware and network related issues on any given platform [5 and 6]. With the availability of cheap personal computers, workstations and networking devices, the recent trend is to connect a number of such workstations to solve computation intensive tasks in a parallel way on such clusters. Network of workstations [7] has become a widely accepted form of high performance computing (HPC). Every workstation in a NOW is treated similarly as a processing element in a multiprocessor system. However, workstations are far more powerful and flexible than processing elements in conventional multiprocessors (Supercomputers).

Typical example of networks of workstations also for solving complex problems is shown in Fig. 1. The individual workstations are mainly extremely powerful personal workstations based on multiprocessor or multicore platform SMP (Symmetrical multiprocessor systems).

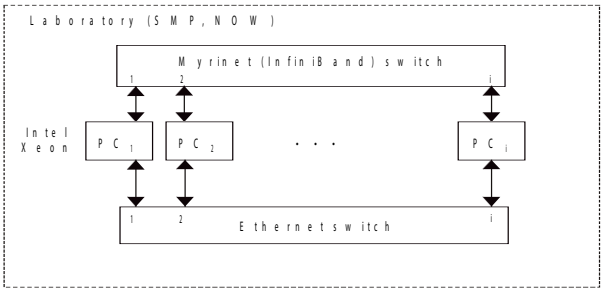

Fig. 1 Practical example of NOW module

We have studied such serious problems in parallel computing (Parallel and distributed computing) as load balancing, inter processor communication IPC, modelling and optimisation of parallel algorithms. etc. [8 and 9]. The coupled computing nodes PC_1, PC_2, ... PC_i (Workstations) could be single most powerful personal computers or SMP parallel computers. In this way, parallel computing on networks of conventional PC workstations (Single, multiprocessor, multicore) and Internet computing indicates advantages of unifying parallel and distributed computing [10].

2.3. Grid systems

In general, Grids represent a new way of managing and organising computer networks and mainly their deeper resource sharing [2]. Grid systems are expected to operate on a wider range of other resources as processors (CPU), as memory modules, data modules, network computing nodes, software files etc. All these resources typically exist within nodes that are geographically distributed and span multiple administrative domains. It is

obvious that existing HPC parallel computers (Supercomputers etc.) could be members of such Grid systems, too. Illustration of Grid modules based on integrated NOW networks are in Fig. 2.

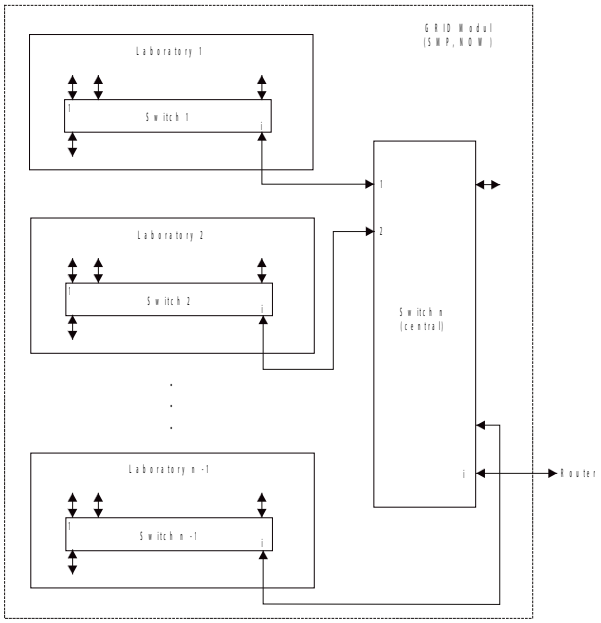

Fig. 2 Architecture of Grid module

Conceptually, Grids arise from a structure of virtual parallel computer based on computer networks (NOW). In general, Grids represent a new way of large integration and managing of resources as a large network of NOW modules. This expression is closely connected with the term massive computational Grid with the basic characteristics as follows:

- wide area network of integrated free computing resources - it is a massive number of inter connected networks, which are connected through high speed connected networks while the whole massive system is controlled by the network operation system, which makes an illusion of a powerful computer system (Virtual supercomputer) [11 and 12],
- grants a function of meta computing that means computing environment, which enables individual applications and functionality of all system resources,
- Grid system combines distributed parallel computation with remote computing from user workstations.

3. Abstract models of computing nodes

3.1. Abstract model of SMP computing node with shared memory

Basic abstract model of SMP parallel computer with shared memory is in Fig. 3.

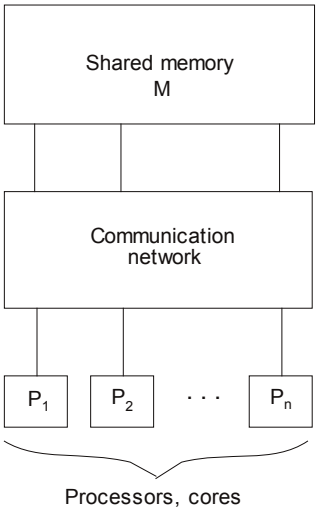

Fig. 3 Abstract model of SMP

3.2. Abstract model of SMP with distributed memory

Abstract model of parallel computer with distributed memory is in Fig. 4.

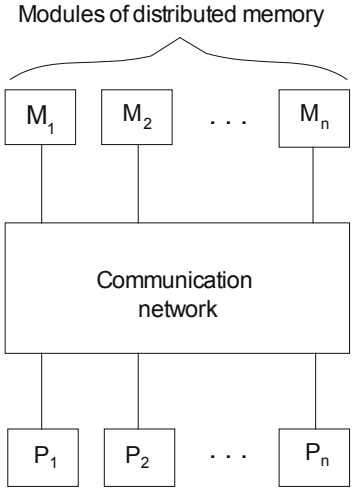

Fig. 4 Abstract model of NOW

4. The role of performance

Quantitative evaluation and modelling of hardware and software components of parallel systems are critical for the delivery of high performance [13]. Performance studies apply to initial design phases as well as to the procurement, tuning and capacity planning analysis. As performance cannot be expressed by quantities independent of the system workload, the quantitative characterisation of resource demands of application and of their behaviour is an important part of any performance evaluation

study [14 and 15]. Among the goals of parallel system performance analysis belong: to assess the performance of a system or a system component or an application, to investigate the match between requirements and system architecture characteristics, to identify the features that have a significant impact on the application execution time, to predict the performance of a particular application on a given parallel system, to evaluate different structures of parallel applications [16 and 17].

5. Application of queuing theory systems

The basic premise behind the use of queuing models for computer systems analysis is that the components of a computer system can be represented by a network of servers (Resources) and waiting lines (Queues) [18 and 19]. A server is defined as an entity that can affect, or even stop, the flow of jobs through the system. In a computer system, a server may be the CPU, I/O channel, memory, or a communication channel. Awaiting line is just that place where jobs (Customers) queue for service. To make a queuing model work, jobs or customers of data message blocks (Packets), which require any sort of processing, are inserted into the network. In such a system, jobs arrive at some rate, queue for service on a first-come first-served basis, receive service, and exit the system. This kind of model, with customers entering and leaving the system, is called an open queuing system model [14].

6. Results

6.1. Corrected standard analytical model

The described standard analytical model [1] supposes that the inter arrival time to the node's communication channels has the exponential distribution. This assumption is not true mainly in the important cases of high communication utilisation. The node servicing time of parallel processes (Computation complexity) could vary from nearly deterministic (in the case of balanced parallel processes) to exponential (in the case of unbalanced ones). From this, in the case of node's high processors utilisation, the outputs from individual processor of node's multiprocessor may vary from the deterministic interval time distribution to exponential one. These facts violate the assumption of the random exponential distribution and could lead to erroneous value of the whole node's delay computation. Worst of all, this error could be even greater if the node of utilisation were higher. From the above-mentioned we derived the correction factor which accounts for the measure of violation for the exponential distribution assumption. The inter arrival input time distribution to each node's communication channel depends on ρ_i, where ρ_i is the overall processor utilisation at the node i. But because only the part λ_{ij} from the total input rate λ_i for node i goes to the node's communication channel j, it is necessary to weigh the measure of influence of the whole node's processors utilisation through the value λ_{ij}/λ_i for channel j as $\rho_i \cdot (\lambda_{ij}/\lambda_i)$.

Mathematical abstract model of i-th node for NOW system is illustrated in Fig. 5.

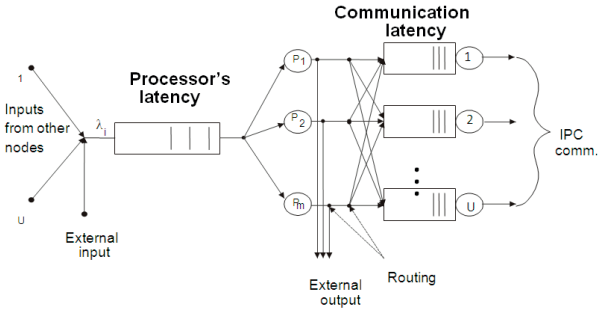

Fig. 5 Abstract model of i-th node of NOW

To clarify the influence of the node's processor utilisation on the average delay of communication channel we tested the 5-noded experimental parallel computer. The processing time was varied to develop various workloads of node's processors. Performed tests showed a remarkable fact - by decreasing the node's processor workload the assumption of the exponential inter arrival message time distribution to the communication

Achieved results for correction factor — Table 1

Processor utilisation at node 1	Average channel delay at node 1 - simulation [msec]	Standard analytical model		Corrected analytical model	
		Average channel delay [msec]	Relative error [%]	Average channel delay [msec]	Relative error [%]
0.6	21.97	22.27	1.4	22.03	0.3
0.7	21.72	22.27	2.5	21.92	0.9
0.8	21.43	22.27	3.9	21.70	1.3
0.9	21.05	22.27	5.8	21.45	1.9
0.95	20.91	22.20	6.5	21.31	1.9

channel is more effective. The achieved results are summarised in Table 1 for one of the communication channels at node 1.

Graphical illustration of achieved results is in Fig. 6.

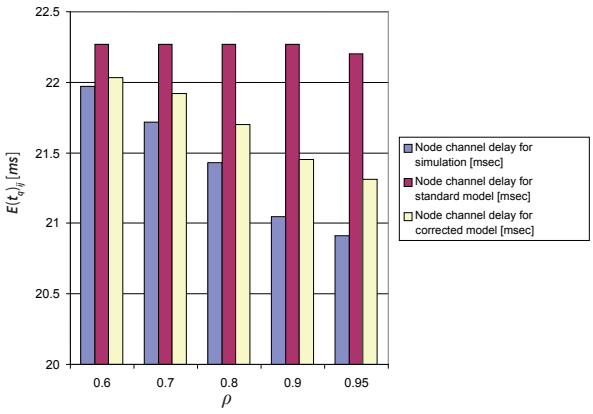

Fig. 6 The influence of the exponential time distribution and its correction

Extensive testing have proved that if we increase utilisation of communication channel and it develops saturation of the communication channel queue then the average queue waiting time is less sensitive to the nature of inter arrival time distributions. This is due to the fact that the messages (Communicating IPC data) wait longer in the queue, which significantly influences the increase of the average waiting time and the error influence of the non-exponential inter arrival time distribution is decreased. To incorporate this knowledge into the correlation factor we investigated the influence of weighting $p_i(\lambda_{ij}/\lambda_i)$ on the value $(1 - \rho_{ij})^x$ for various values x. The performed experiments showed the best results for the value x = 1. Derived approximation of the average queue waiting time of the communication channel j at the node i, which eliminates violence of the exponential inter arrival time distribution, is then given as

$$\frac{\rho_i \cdot (1-\rho_{ij}) \cdot \lambda_{ij}}{\lambda_i}$$

The final correction factor of the communication channel j at the node i, which we named as c_{ij} is as follows

$$c_{ij} = 1 - \frac{\rho_i \cdot (1-\rho_{ij}) \cdot \lambda_{ij}}{\lambda_i}$$

With the derived correction factor c_{ij} we can define the corrected average queue waiting time as

$$E(t_q)_{ij}'(LQ) = c_{ij} \cdot E(t_q)_{ij}(LQ)$$

where $E(t_q)_{ij}(LQ)$ is defined in [13] as

$$\frac{\lambda_{ij} \cdot E(t_q)_{ij}}{\gamma}$$

Based on the derived correction factor, the whole latency in modelled NOW is as

$$E(t_q)_{now} = \frac{1}{\gamma}\left[\sum_{i=1}^{U}\left(\lambda_i \cdot E(t_q)_i + \sum_{j=1}^{u_i} c_{ij}\left[\lambda_{ij} \cdot E(t_q)_{ij}\right]\right)\right] \quad (1)$$

where $\frac{\lambda_i \cdot E(t_q)_i}{\gamma}$ and $c_{ij}\frac{\lambda_{ij} \cdot E(t_q)_{ij}}{\gamma}$ define individual contribution of computation queue delay and communication channel delay of every node to the whole delay. We named this model based on M/M/m and corrected M/M/1 results as a corrected standard analytical model. The average delay values of the node's communication channel achieved through simulation are compared with the results of the standard analytical model (Exponential inter arrival time distribution) and with the results of the corrected standard model as illustrated in Fig. 7.

6.2. Other real analytical models

6.2.1. Analytical model with M/M/m and M/D/1systems

This model is a mixture of analysed models. We get the first part of the final total average time $E(t_q)_{now}$ from relation (1) of the previous chapter. The second part is the sum of all communication channels considered as M/D/1 system [17]. Then, for $E(t_q)_{now}$ we can finally get:

$$E(t_q)_{now} = \frac{1}{\gamma}\left[\sum_{i=1}^{U}\left(\lambda_i \cdot E(t_q)_i + \sum_{j=1}^{u_i}\left(E(t_w)_{ij}(LQ) + \overline{x_{ij}}\right)\right)\right] \quad (2)$$

6.2.2. Model with M/D/m and M/M/1 systems

In this alternative analytical model the first part of the final total average time $E(t_q)_{now}$ is the derived expression in [17] for M/D/m system. We get the second part from relation (1) of the previous chapter as the sum over all communication channels

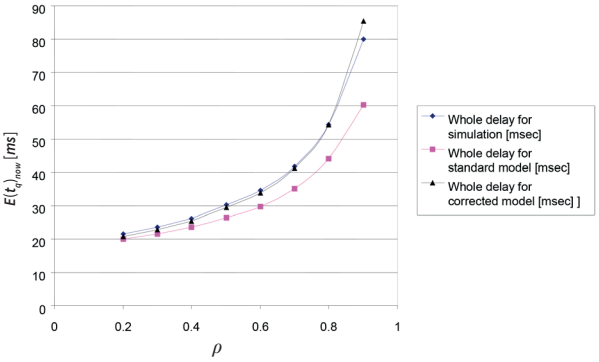

Fig. 7 Comparison of the used models

considering them as corrected M/M/1 system. Then, for $E(t_q)_{now}$ we finally get for this model:

(3)

$$E(t_q)_{now} = \frac{1}{\gamma}\left[\sum_{i=1}^{U}\left(\left(E(t_w)_{ij}(PQ)+\overline{x_i}\right)_i + \sum_{j=1}^{u_i} c_{ij}\left[\lambda_{ij}\cdot E(t_q)_{ij}\right]\right)\right]$$

Comparison of the relative errors (in relation to simulation results) is illustrated in Fig. 8.

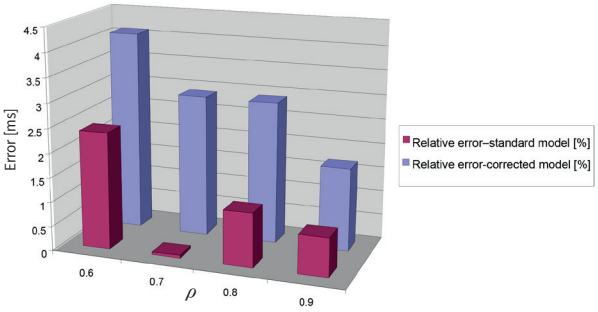

Fig. 8 Influence of channel utilisation on the accuracy of analytical models

6.3. Analytical model of real Grid systems

We defined the Grid system as a network of NOW network modules. Let N be the number of individual NOW networks or similar clusters. Then final total average time $E(t_q)_{grid}$

$$E(t_q)_{grid} = \frac{1}{\alpha}\left[\sum_{i=1}^{N} E(t_q)_{i\,now}\right] \quad (4)$$

where

- $\alpha = \sum_{i=1}^{N} \gamma_i$ represents the sum of individual total extern intensities to the i-th NOW module in the Grid
- $E(t_q)_{i\,now}$ corresponds to individual average times in i- th NOW module (i=1, 2, ... N).

7. Conclusion and perspectives

Performance evaluation of computers has generally been a very difficult problem since the birth of computers. It was very hard to apply any analytical methods to performance evaluation of sequential computers because of high number of non-predictable parameters. Secondly, endless user demands to increase computer performance were fulfilled more quickly via continuing technology improvements or computer architecture changes. For a long time incorporation of various forms of parallel principles has created more stable conditions to apply performance evaluation methods mainly for dominant parallel computers (SMP, NOW and Grid) to open more possibilities to apply mainly analytical methods (Queuing theory) in order to analyse performance of parallel computers. This implies to one known queuing theory result - many inputs to queuing theory system, which create shared stream and which are generated at various independent resources by chance, could be a very good approximation of Poisson distribution, as a basic assumption, to solve such coupled parallel systems in an analytical way. Therefore, we are able to model parallel computing nodes (Multiprocessors, multicores, workstations etc.) of any dominant or perspective parallel computer (SMP, NOW, Grid, meta computer) as M/D/m or M/M/m queuing theory systems and computing node's communication channels as M/D/1 or M/M/1 queuing theory systems respectively.

We have compared the relative errors of the developed models (standard, corrected) in relation to the developed simulation model. The relative errors of standard analytical model are from 7% to 25%. This is due to the influences of computing node queue delays, the nature of inter arrival inputs to the communication channels mainly in the case of high processor utilisation. The developed corrected analytical model provides more precise results in the whole range of input workload of processor utilisation, communication channels and network topologies with relative error, which does not exceed 14% and even in the most important cases of the high processor utilisation ($\rho > 0.6$) the relative errors according to Fig. 8 are in the range up to 9%.

Then, such a very flexible modelling tool based on preferred analytical solutions shows a real path to a very efficient and practical performance analysis tool including massive NOW networks or other types of massive computer networks (Grid, meta computer). To summarise it - all developed more precise analytical models could be applied to the performance modelling of dominant parallel computers in typical cases as follows:

- a single computing node based on SMP parallel computer,
- NOW based on workstations (single, SMP),
- Grid (Network of NOW network modules),
- mixed parallel computers (SMP, NOW, Grid),
- meta computers (Massive Grid).

For our further research work, in relation to dominant trends in parallel computers (SMP, NOW, Grid) based of powerful workstations, we will be looking for preferred analytical models [20] in which load balancing, inter-process communication (IPC) in both parallel and distributed computing [21], effective transport protocols, influence of various parallel computer architectures, performance prediction etc. could be studied. We would also like:

- to analyse the role of adaptive routing in considered analytical models,
- to prove or to indicate experimentally the role of the independence assumption if we are looking for higher moments of overhead latencies (IPC communication, synchronisation, parallelisation, parallel computer architecture etc.)

- to verify analytical models also for node's limited resources capacities – buffers, communication channels etc., and for other existing queue servicing algorithms as standard assumed FIFO (First in First out).

To model single computing nodes we can also use other more complicated single queuing theory systems as the analysed ones (M/M/1, M/M/m, M/D/1, and M/D/m). The reasons for choosing these single queuing theory systems are following:

- to finish performance analysis of network of coupled queuing theory systems we need results of chosen single queuing theory systems M/M/m and M/D/m [22],
- we need their results to compute approximation relation for M/D/m,
- M/M/1 and M/M/m models could be used to compare their results with other models M/D/1 and M/D/m respectively,
- the results of analysed models M/M/m and M/D/m are necessary to finish performance analysis of decomposed communication network of computing nodes.

Acknowledgement

This work was carried out within the project "Complete modelling, optimisation and prediction of parallel computers and algorithms" at the University of Zilina, Slovakia. The authors gratefully acknowledge the help of the project supervisor Prof. Ing. Ivan Hanuliak, PhD.

References

[1] HAGER, G., WELLEIN, G.: *Introduction to High Performance Computing for Scientists and Engineers*, 356 pp, 2010, CRC Press.
[2] PATTERSON, D. A., HENNESSY, J. L.: *Computer Organization and Design,* 4th ed., Morgan Kaufmann, 914 pp., 2011.
[3] WANG, L., JIE WEI., CHEN, J.: *Grid Computing: Infrastructure, Service, and Application,* CRC Press, 2009.
[4] DUBOIS, M., ANNAVARAM, M., STENSTROM, P.: *Parallel Computer Organization and Design,* 560 p., 2012, Cambridge university press.
[5] CENEK, J.: Evaluation of a Heuristic Algorithm for Transit Line Planning, *Communications - Scientific Letters of the University of Zilina*, No. 3, pp. 44-48, 2013.
[6] HANULIAK, J., HANULIAK, I.: To Performance Evaluation of Distributed Parallel Algorithms, *Kybernetes*, vol. 34, No. 9/10, 1633-1650, 2005.
[7] HANULIAK, P., HANULIAK, M.: Mod. of Single Comp. Nodes of Par. Comp., *AJNC (Am. J. of Networks and Comm.), Science PG,* vol. 3, Special No. 1, 57-69, 2014.
[8] RIANO, L., MCGINITY, T. M.: *Quantifying the Role of Complexity in A System's Performance,* Evolving Systems, Springer Verlag, 189-198, 2011.
[9] GABRISOVA, L., JANACEK, J.: Design of Capacitated Emergency Service System, *Communications - Scientific Letters of the University of Zilina*, No. 2, pp. 42-48, 2015.
[10] PETERSON L., DAVIE B. C.: *Computer Networks - A System Approach*, Morgan Kaufmann, 920 p., 2011.
[11] HARCHOL, B. M.: *Performance Modelling and Design of Computer Systems,* Cambridge University Press, 576 p., 2013.
[12] HANULIAK, M., HANULIAK, P.: Performance Modelling of Par. Comp. NOW and Grid, *AJNC - Am. J. of Networks and Communication*, Science PG, 112-124, 2013.
[13] ARORA S., BARAK B.: *Computational Complexity - A Modern Approach*, Cambridge University Press, 573 p., 2009.
[14] DATTATREYA, G. R.: *Performance Analysis of Queuing and Computer Network*, 472 pp., 2008, University of Texas: Dallas.
[15] LE BOUDEC J. Y.: *Performance Evaluation of Computer and Communication Systems,* CRC Press, 300 p., 2011.
[16] MEERSCHAERT, M.: *Mathematical Modelling*, 384 p., 2013, Elsevier.
[17] NATARAJAN, G.: *Analysis of Queues - Methods and Applications*, 802 p., CRC Press, 2012.
[18] HANULIAK P., HANULIAK M.: Modell. of Single Comp. Nodes of Par. Comp., *AJNC (Am. J. of Networks and Comm.), Science PG,* vol. 3, Spec. No. 1, 57-69, 2014, USA.
[19] KOSTIN, A., ILUSHECHKINA. L.: *Modelling and Simulation of Distributed Systems,* Imperial College Press, 440 p., 2010.
[20] BRONSON R., COSTA G. B., SACCOMAN J. T.: *Linear Algebra - Algorithms, Applications, and Techniques,* 536 p., Elsevier, 2014.
[21] HANULIAK, J.: Modelling of Comm. Complexity in Parallel Computing, *AJNC (Am. J. of Networks and Comm.), Science PG,* vol. 3, Spec. No. 1, 29-42, 2014.
[22] GAUTAM, N.: *Analysis of Queues: Methods and Applications*, CRC Press, 802 p., 2012.

COMMUNICATIONS – Scientific Letters of the University of Zilina
Writer's Guidelines

1. Submitted papers must be unpublished and must not be currently under review for any other publication.
2. Submitted manuscripts should not exceed 8 pages including figures and graphs (in Microsoft WORD – format A4, Times Roman size 12, page margins 2.5 cm).
3. Manuscripts written in good English must include abstract and keywords also written in English. The abstract should not exceed 10 lines.
4. Submission should be sent by e-mail – as an attachment – to one of the following addresses: komunikacie@uniza.sk or holesa@uniza.sk (or on CD to the following address: Zilinska univerzita, OVaV – Komunikacie, Univerzitna 1, SK – 010 26 Zilina, Slovakia).
5. Uncommon abbreviations must be defined the first time they are used in the text.
6. Figures, graphs and diagrams, if not processed in Microsoft WORD, must be sent in electronic form (as JPG, GIF, TIF, TTF or BMP files) or drawn in high contrast on white paper. Photographs for publication must be either contrastive or on a slide.
7. The numbered reference citation within text should be enclosed in square brackets - in numerical order. The reference list should appear at the end of the article (in compliance with ISO 690).
8. The numbered figures and tables must be also included in text.
9. The author's exact mailing address, full names, E-mail address, telephone or fax number, the name and address of the organization and workplace (also written in English) must be enclosed.
10. The editorial board will assess the submitted paper in its following session. If the manuscript is accepted for publication, it will be sent to peer review and language correction. After reviewing and incorporating the editor's comments, the final draft (before printing) will be sent to authors for final review and minor adjustments
11. Submission deadlines are: September 30, December 31, March 31 and June 30.

COMMUNICATIONS

SCIENTIFIC LETTERS OF THE UNIVERSITY OF ZILINA
VOLUME 18

Address of the editorial office:
Zilinská univerzita
Office for Science and Research
(OVaV)
Univerzitna 1
SK 010 26 Zilina
Slovakia

E-mail: komunikacie@uniza.sk

Each paper was reviewed by two reviewers.

Journal is excerpted in **COMPENDEX** and **SCOPUS**.

It is published by the University of Zilina in
EDIS – Publishing Institution of Zilina University
Registered No: EV 3672/09
ISSN 1335-4205

Published quarterly

Single issues of the journal can be found on:
http://www.uniza.sk/komunikacie

ICO 00397 563
September 2016