COMMUNICATIONS

Dear reader,

the 8th International scientific conference TRANSCOM 2009, was held in the University of Zilina, Slovakia, in June 2009.

The main purpose of the conferences TRANSCOM organised regularly every other year since 1994 is a presentation of scientific works (from the fields of transportation, telecommunications, mechanical, electrical, civil, security, forensic engineering and social sciences) of young research workers incl. PhD students up to the age of 35 from universities, scientific institutions and industry.

More than 365 contributions were published in 9 proceedings of the conference TRANSCOM 2009 (124 contributions were from abroad, Czech Republic, Germany, Great Britain, Hungary, Italy, Lithuania, Poland, Romania, Russia, Spain, Ukraine, 38 were from the universities of the Slovakia and 203 contributions were from the University of Zilina).

I should express my gratitude to all participants for contributing to the TRANSCOM 2009 and to the TRANSCOM 2009 scientific committee incl. organizing committee members for cooperative spirit, motivation and enthusiasm.

This volume of the Communications is devoted to the selected contributions (recommended by the scientific committee) of the 8th International scientific conference TRANSCOM 2009, Zilina, Slovakia.

Otakar Bokuvka

Maria Lucka – Stanislav Piecka *

MULTI-THREADED ANT COLONY OPTIMIZATION WITH ASYNCHRONOUS COMMUNICATIONS FOR THE VEHICLE ROUTING PROBLEM

In this paper we study behaviour of Ant Colony Optimization algorithm for solving the Vehicle Routing Problem implemented by POSIX Threads in parallel cluster environment. The algorithm is based on a fine-grained parallelism strategy which uses asynchronous communication for cooperation in finding solutions. Our aim is to analyze the effect of proposed method on speedup, execution and communication time with respect to the quality of solution.

Keywords: Ant Colony Optimization, Parallel Metaheuristic, Vehicle Routing Problem, POSIX threads.

1. Introduction

In the field of combinatorial problems, the Vehicle Routing Problem (VRP) introduced by [1] is one of the most challenging. This optimization problem and its variants have multiple applications in telecommunication, transportation and logistics. Unfortunately the majority of these applications belong to NP-hard problems so in the worst case the exponential time is required to find the optimal solution. The VRP problem determines a set of vehicle routes starting and ending at the depot where each customer is visited exactly once. The demand of each customer is satisfied and both maximum tour lengths and vehicle capacities cannot be violated. The objective of the VRP is to minimize the total travel costs. Exact algorithms can be used only for relatively small instances. In practice, all known solutions of larger instances come from heuristic or metaheuristic algorithms. It seems that especially metaheuristic methods produce good quality solutions in relatively short calculation time.

The Ant Colony Optimization method (ACO) developed by [3] has become very successful for solving the VRP problem. The idea of the method is inspired by behaviour of real ants where each ant deposits pheromone on the ground as information that other ants should follow it. Deposited pheromone evaporates in time. Successfulness of ant in finding the food causes that certain path is passed more often and more pheromone is deposited. Therefore it is more likely that other ants choose the same path as well. This behaviour of ants is in computer science modeled iteratively by repeatedly called procedures which create solutions by exploring fully connected graph of customers. The artificial ant makes decision about the way it will continue in every vertex. This decision making is specific for concrete problem and is influenced by two factors: joint memory and heuristic information. When created, best solutions are used for updating of common memory according to the achieved quality. This updating is done after all artificial ants have finished their search. The whole procedure is called repeatedly as many times as required.

In our work we have rewritten the parallel Savings based ACO algorithm for the VRP described in [4] using synchronous communication model with Message Passing Interface to a thread based asynchronous model. We study characteristics of the algorithm when decentralized asynchronous communication is used, on four larger instances: C5 published by [2] and G18, G19, G20 published by [6]. For these instances there are still not exactly calculated optimal solutions.

This paper is organized as follows. In the following two sections we shortly describe the Vehicle Routing Problem and the ACO parallelization strategies and propose the asynchronous algorithm. In Section 4 gained computational results are shown. We present dependence of the gained speedup and efficiency on the number of threads used, whereby the solution quality, execution and communication time are also presented. The last section concludes with several remarks and outlooks concerning the future work.

2. Formulation of the Vehicle Routing Problem

According to [2] we can describe the Vehicle Routing Problem (VRP) as follows: Let $G = (V, E, c)$ be a complete graph which has $n + 1$ nodes $(v_0, ..., v_N)$ corresponding to the customers $i = 1, ..., N$ and the depot $i = 0$, and the edge set $((v_i, v_j) \in E \; \forall \; v_i, v_j \in V)$. With each edge $(v_i, v_j) \in E$ is associated a non-negative weight c_{ij}, which refers to the travel costs between nodes v_i and v_j and a non-negative weight t_{ij}, which refers to the distance between the nodes.

* **Maria Lucka[1], Stanislav Piecka[2]**
[1]Faculty of Education, University of Trnava, Slovakia, E-mail: mlucka@truni.sk
[2]Faculty of Controlling and Informatics, University of Zilina, Slovakia

Furthermore, with each node v_i, $i = 1, ..., N$ is associated a non-negative demand d_i, which has to be satisfied, as well as a service time δ_i. The service time at the depot is set to $\delta_0 = 0$. At the depot a fleet of size K is available, where each vehicle has a capacity of Q^k and the maximum driving time for each vehicle is T^k.

Let x_{ij}^k denote the binary decision variables which equals 1 if vehicle k visits node v_j immediately after node v_i and 0 otherwise.

The objective can be written as

$$\min \sum_{i=0}^{N} \sum_{j=0}^{N} \sum_{k=1}^{K} c_{ij} x_{ij}^k \tag{1}$$

Using the following restrictions

$$\sum_{i=1}^{N} \sum_{\substack{j=0 \\ j \neq 1}}^{N} x_{ij}^k d_i \leq Q^k \quad 1 \leq k \leq K \tag{2}$$

$$\sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} x_{ij}^k (t_{ij} + \delta_i) \leq T^k \quad 1 \leq k \leq K \tag{3}$$

$$\sum_{\substack{i=0 \\ i \neq j}}^{N} x_{ij}^k - \sum_{\substack{l=0 \\ l \neq j}}^{N} x_{jl}^k = 0 \quad 1 \leq k \leq K, 0 \leq j \leq N \tag{4}$$

$$\sum_{\substack{i=0 \\ i \neq j}}^{N} \sum_{k=1}^{K} x_{ij}^k = \begin{cases} 1 & 1 \leq j \leq N \\ K & j = 0 \end{cases} \tag{5}$$

$$\sum_{i \in S} \sum_{\substack{j \in S \\ j \neq i}} x_{ij}^k \leq |S| - 1 \quad \forall S \subseteq \{1, \ldots, N\}, 1 \leq k \leq K \tag{6}$$

$$x_{ij}^k \in \{0, 1\} \quad 1 \leq k \leq K, 0 \leq i, j \leq N \tag{7}$$

The objective (1) is to minimize the total travel costs. Constraints (2) ensure that no vehicle can be overloaded. Constrains (3) require that for each vehicle the maximum driving time is respected. Constraints (4) ensure that vehicle which visits a customer also leaves the customer. Constraints (5) require that all customers are visited exactly once, and that the depot is left K times. Through constraints (6) sub-tour elimination is ensured. Constraints (7) are the usual binary constraints.

3. Parallelization of ACO

Interesting feature of the ACO is its feasibility of parallelization. Each ant makes relatively simple and independent task. The only dependence is caused by using the same pheromone matrix and the same best solution found so far. The first issue is required for decision making process and the second one for measurement of quality of generated solution. We can identify several goals that can be achieved by parallelization of the ACO such as reduction of calculation time, increase of solution quality or speed of convergence. To achieve the reduction of calculation time we can split the colony of ants between processors and let each processor calculate its part of colony. This approach is known as functional decomposition [10]. Instead of this we can apply domain decomposition by dividing customers into subsets and let processors calculate solutions of sub-problems. Generally there are possibilities of synchronous or asynchronous communication between processors. Classification of parallelization of the ACO for the VRP can be found in [4], and [5]. In short, we can use: fine-grained, coarse-grained and mixed parallelization. Fine-grained parallelization splits ants of a colony between processors which often communicate when update the pheromone matrix and exchange the best solutions. By course-grained parallelization schemes run more colonies of ants in parallel whereby each colony is calculated on one processor. The information exchange between processors take place in certain time intervals and concern specified parts of information. The mixed approach is a combination of the first two.

In our approach we used fine-grained parallelization strategy with decentralized approach whereby one ant colony is proportionally divided among several computing threads. An execution thread [9] is a fork of a computer program into more tasks which can run concurrently. Those threads share memory and other resources, but they run independently. Considering the shared memory computational model of threads does not need to change the address space, inter-process communication of threads is faster than that for processes.

We suppose that all threads in our implementation have the same behavior and calculate homogeneous parts of the colony of ants. Number of threads is specified by the number of the processor's cores used. Each core runs exactly one thread. Every thread computes its own pheromone update by using information received from other threads. When possible, an inter-thread communication is done by using shared memory, otherwise the network is used. Concurrent access to the shared memory is secured by critical sections which are implemented by PThread mutexes [9]. Instead of using shared files proposed in [7] we used the user datagram protocol as a communication layer. If a thread has found better solution and has to use the network layer, the user datagram packet is sent only once per cooperating node. All received packets are stored in system buffers and are processed by first thread which reads them. After the packet is processed the thread publishes the best gained solution to all threads in its group in shared memory. This approach does not require dedication of a separate thread for communication. The pseudo-algorithm can be formulated as follows:

```
1:  Initialization;
2:  For i = 1; i <= IT do:
    /* IT is the number of iterations */
        For Ant = 1; Ant <= N / NoT + 1 do:
        /* NoT is the number of threads */
            Create Savings based Ant solution;
        Select elitist Ants;
        Download received solution if exists;
        Update and spread solution if better solution is found;
        Update pheromone matrix if needed;
        Update Savings list if needed;
    End do i;
3:  Finalization;
```

The speedup is defined for measurement of parallelization quality by the following formula:

$$S_p = T_1/T_p \qquad (8)$$

where p is the number of processors, T_1 is the execution time of the sequential algorithm and T_p is the execution time of the parallel algorithm with p processors. According to dual-core architecture used in our experiments we are calculating with number of used cores instead of processors. Similarly the efficiency is a performance metric defined by the following formula:

$$E_p = S_p/p = T_1/pT_p \qquad (9)$$

This value is typically between zero and one and estimates how well-utilized in solving the problem are processors, compared to how much effort is wasted in synchronization and communication.

4. Computational results

In our experiments we used the cluster consisting of 72 SUN X4100 nodes with two 64-bit dual core processors, each. Therefore we could use at most 4 threads per node working over the common shared memory. Reported results are average values gained over independent 15 runs for all instances. The number of customers is denoted as N and configurations mentioned below are used. Even we experienced better solution quality with different configurations, we used the same parameters settings as proposed in [8] and used in [4] to keep results comparable. We used N artificial ants for each instance, $\alpha = \beta = 5$ and $\sigma = 6$ elitist ants, the evaporation rate $\rho = 0.95$, and the neighborhood size $\lfloor N/2 \rfloor$. We ran the algorithm for 2N iterations for both instances. The algorithm did not send whole pheromone matrix between cores. Only the best σ solutions were chosen, compared with the best solutions found so far and spread between nodes every time better solution was found.

In Table 1 we can see that the time spent by communication does not increase linearly with using more cluster nodes. We can see that efficiency decreases with the increasing number of threads. The achieved efficiency for 32 cores for instance C5 is 0.59 and 0.73 for instance G19. This value is better than published in [4], where the gained efficiency is 0.37 and 0.39, respectively. So we can conclude that asynchronous communications are more suitable for the ACO as synchronous, especially for larger instances. Reduction of speedup of instance C5 against G19 is caused by the fact that time required for creating solution is smaller. Therefore, the ratio between communication and calculation is higher. Efficiency greater than 1 on the C5, G18 and G20 instances is caused by caching effect, where several threads run on separate cores of one processor. We can see that solution quality is decreasing with more threads; this is about 4% on the C15 and 3% on the G19 when 32 threads are used.

5. Conclusions

We presented parallel POSIX Threads based implementation of the ACO method using asynchronous cooperative approach for solving the VRP. We measured its speedup and efficiency in comparison with synchronous approach published in [4]. We showed that asynchronous communication increases efficiency of the ACO algorithm.

In our future work we would like to apply the presented asynchronous approach to the mixed, multi-colony ACO parallelization with focus on increasing quality of the solution. We would like to test the algorithm with different configurations. We also plan to test the dependency of the instance size on efficiency and solution quality in parallel asynchronous ACO algorithms.

Table 1

Calculated average results according to the number of threads of each measured instance, where V denotes solution quality, t_c denotes the time spent by communication and synchronization calculated per node, t_r denotes the overall execution time including communication.

Instance	Threads	Nodes	V	t_c [ms]	t_r [s]	Speed	Efficiency
C5 (n=199)	1	1	1377.68	7.96	440.70	1.000	1.000
	2	1	1376.14	16.78	219.11	2.011	1.006
	4	1	1384.03	34.06	113.03	3.899	0.975
	8	2	1396.19	36,52	61.73	7.139	0.892
	16	4	1418.81	36.43	35.48	12.422	0.776
	32	8	1435.79	36.88	23.40	18.834	0.589
G18 (n=300)	1	1	1097.49	12.49	2307.50	1.000	1.000
	2	1	1097.52	24.90	1089.14	2.119	1.059
	4	1	1101.99	62.03	563.80	4.093	1.023
	8	2	1110.80	62.39	289.89	7.960	0.995
	16	4	1123.84	56.31	149.53	15.432	0.964
	32	8	1135.54	52.85	83.65	27.585	0.862
G19 (n=360)	1	1	1498.71	16.77	5825.99	1.000	1.000
	2	1	1501.60	38.45	2929.19	1.989	0.994
	4	1	1495.88	80.55	1470.60	3.962	0.990
	8	2	1509.38	86.44	790.26	7.372	0.922
	16	4	1521.93	86.81	420.89	13.842	0.865
	32	8	1546.23	85.74	250.26	23.280	0.728
G20 (n=420)	1	1	2013.44	18.67	12421.2	1.000	1.000
	2	1	2008.52	43.71	6093.24	2.039	1.019
	4	1	2006.99	98.14	3066.11	4.051	1.013
	8	2	2026.06	101.67	1540.71	8.062	1.008
	16	4	2049.92	91.86	814.73	15.246	0.953
	32	8	2074.81	87.87	436.65	28.447	0.889

Acknowledgement

The authors would like to thank Prof. Dr. Siegfried Benkner, Institute of Scientific Computing, University of Vienna, for the possibility to use the parallel computer cluster LUNA for running their experiments.

References

[1] DANTZIG, G., RAMSER, R.: *The truck dispatching problem.* Management Science 6, pp. 80–91, 1959.
[2] CHRISTOFIDES, N., MINGOZZI, A., TOTH, P.: *The Vehicle Routing Problem.* Combinatorial Optimization, Wiley, Chichester, 1979.
[3] DORIGO, M., GAMBARDELLA, L. M.: *Ant Colony System: A cooperative learning approach to the Travelling Salesman Problem.* IEEE Transactions on Evolutionary Computation, 1(1), pp. 53–66, 1997.
[4] DOERNER, K. F., HARTL, R. F., BENKNER, S., LUCKA, M.: *Cooperative Savings based Ant Colony Optimization – Multiple Search and Decomposition Approaches.* Parallel Processing Letters, 16(3), pp. 351–369, 2006.
[5] CRAINIC, T. G.: *Parallel Solution Methods for Vehicle Routing Problems.* CIRRELT- 2007-28, 2007.
[6] GOLDEN, B. L., WASIL, E. A., KELLEY, J. P., CHAO, K. M.: *The impact of metaheuristics on solving the vehicle routing problem: algorithms, problem sets, and computational results.* Fleet management and logistics, Norwell: Kluwer, 1998.
[7] LUCKA, M., PIECKA, S.: *Parallel POSIX Threads Based Ant Colony Optimization using asynchronous Communications.* Proceeding of 8th International Conference APLIMAT 2009, pp. 613–620, 2009.
[8] REIMANN, M., DOERNER, K. F., HARTL, R. F.: *D-Ants. Savings Based Ants divide and conquer the vehicle routing problem.* Computers & Operations Research, 31 (4) (2004), 563–591.
[9] ANDREWS, G. R.: *Foundations of Multithreaded, Parallel, and Distributed Programming.* Addison-Wesley, Reading, MA, 2000. http://www.cs.arizona.edu/people/greg/mpdbook.
[10] SCHROEDER-PREIKSCHAT, W.: *The logical design of parallel operating systems.* Prentice-Hall, Englewood Cliffs, NJ, pp. 29, 1994.

Josef Vican – Marian Sykora *

DESIGN OF COMPRESSION MEMBERS FOR DURABILITY

General principles of structural elements design for durability and their application in the case of compression members are presented in this paper. The degradation effects of material deterioration due to corrosion cause the loss of member resistance, but the limit states method does not directly reflects design and verification allowing for failures caused by material deterioration. The time-dependent loss of compression member resistance was computed for specific corrosion models to express the time-dependent approximation function taking into account effects of corrosion degradation. Then, the design member resistance could be obtained using buckling member resistance multiplied by a time-dependent function considering corrosion degradation effects within the member lifetime.

Keywords: durability, prediction of member resistance, time-dependent corrosion effects.

1. Introduction

Present standard methods of the reliability theory are based on verification of the structural reliability from the viewpoint of limit states caused by the permanent, variable and accidental actions. Structural failures and damages caused by environmental actions in the form of various material degradations due to aggressive environment and its changes within bridge lifetime are taking into account indirectly or neither in member design. Therefore, durability of steel structures especially bridges is ensured by choice of an appropriate material resistant against the aggressive environment, using appropriate and verified structural details ensuring to minimize environment influences and enabling inspections and maintenance, by structural redundancy and by the choice of an appropriate corrosion protection system.

The general concept of the structural member design for durability is given in standard [1], but the form is useless for practical applications. The reason is, on one hand, the shortage of information about transfer mechanism by which the environmental influences are transferred into agents acting on the structure to cause damages like corrosion. On the other hand, the processes have stochastic character depending on many random variable parameters, so that the mathematic interpretation for structural design is very complicated.

Generally, two concepts of structural verification for durability should be used. The first approach is based on the concept of the bridge design structural service life t_s, where the probability shall be verified according to formulae

$$P_f = P(t_S \leq t_D) \leq P_{fd}, \tag{1}$$

where t_D is the bridge design lifetime, P_f is the probability of failure and P_{fd} is the target probability of failure. The second approach is based on the limit state concept design considering effects of environmental actions in the form

$$P_f(t) = P\{R(t) - E(t) \leq 0\} \leq P_{fd}, \tag{2}$$

where $E(t)$ represents the random variable time-dependent action effects and $R(t)$ is the random variable time-dependent structural member resistance. $E(t)$ and $R(t)$ are functions of random variables X_i, which are also time-dependent functions in relation with the models of material degradations.

Because of random variable character of many parameters entering corrosion process, the probability theory and mathematic statistics should be the most appropriate approaches to describe it. At present, several corrosion models are known, which could be generally used for analysis of the corrosion influence on the structural member reliability. Review of the most frequent corrosion models describing loss of material in time is presented in Table 1.

2. Time-dependent buckling member resistance

Taking into account the buckling effects and the effects of variable randomness which are influencing the resistance of compression member, the buckling member resistance can be expressed in the form (3), that is a function of many material and geometric random variables

$$R_c = \chi \cdot f_y \cdot A_n \cdot \varphi_A \cdot \prod_{i=b}^{n} \varphi_i, \tag{3}$$

* **Josef Vican, Marian Sykora**
Department of Structures and Bridges, Faculty of Civil Engineering, University of Zilina, Slovakia, E-mail: vican@fstav.uniza.sk

Probabilistic corrosion models — Table 1

Author	Mean	Standard deviation	Distribution
Southwell-Melchers [2]	$0.084t^{0.823}$ [mm]	$0.056t^{0.823}$[mm]	Normal
Frangopol [3]	$0.03207t^{0.5}$ [mm]	$0.00289t^{0.045}$[mm]	Normal
Qin-Cui [4]	$1.67[1 - \exp(-t/9.15)^{1.97}]$ [mm]	$0.0674[1 - \exp(-t/0.181)^{0.0294}]$ [mm]	Normal
Guedes Soares [5]	$d_{corr}(t) = 1.5(1 - e^{(-t/10)})$ where t is time in year[mm]		–

where χ is the actual value of reduction factor covering buckling effects,
f_y is the actual value of the steel yield strength f_y,
A_n is the nominal value of the member cross-sectional area,
φ_A is the ratio of actual value to nominal value of a member cross-sectional area,
$\varphi_i = F_i/F_{ni}$ is ratio of actual value to nominal value of i – partial effect on member resistance or material.

Following the Ayrton – Perry's approach, the buckling resistance of the member with initial local bow having shape of the half-wave sinusoidal imperfection and the amplitude of e_0 can be given as

$$R_c = \left\{\frac{1}{2}\left[f_y + \left(1 + e_0\frac{A}{W}\right)\cdot\sigma_{cr}\right] - \left\{\frac{1}{4}\left[f_y + \left(1 + e_0\frac{A}{W}\right)\cdot\sigma_{cr}\right]^2 - f_y\cdot\sigma_{cr}\right\}^{0.5}\right\}\cdot A_n\cdot\varphi_A\,, \tag{4}$$

where σ_{cr} is the Euler's buckling stress $\sigma_{cr} = \pi^2 E/\lambda^2$,
R_c is the buckling resistance of a compression member,
η is the Perry's factor, defined as $\eta = e_0/r$,
r is the core abscissa, $r = W/A$,
λ is the member slenderness,
E is the Young's modulus of elasticity,
e_0 is the amplitude of equivalent geometric imperfection (amplitude of the half-wave sinusoidal imperfection of member centre line).

Taking into account the random variables influences on member resistance the equation can be rewritten into following formula according to [8]

$$R_c = \left\{\frac{1}{2}\Psi - \left[\frac{1}{4}\Psi^2 - \frac{\varphi_{fy}\,\varphi_E\,\varphi_I}{\varphi_A\cdot\bar{\lambda}_n^2}\right]^{0.5}\right\}\cdot\varphi_A\cdot A_n\cdot f_{yn}\,, \tag{5}$$

where $$\Psi = \varphi_{fy} + \left(1 + e_{0k}\cdot\frac{An}{Wn}\cdot\varphi_{e0}\cdot\frac{\varphi_A}{\varphi_W}\right)\frac{\varphi_E\cdot\varphi_1}{\varphi_A\cdot\bar{\lambda}_n^2}\,, \tag{6}$$

$$\lambda_n = \sqrt{\frac{f_{yn}L_{crn}^2 A_n}{\pi^2 E_n I_n}}\,, \tag{7}$$

and the basic random variables are for the Ayrton – Perry's approach introduced in the non dimensional forms as a ratio φ_i, when the actual value is divided by nominal value.

$$\varphi_{fy} = \frac{f_y}{f_{yn}};\ \varphi_{e0} = \frac{e_0}{e_{0n}};\ \varphi_E = \frac{E}{E_n};\ \varphi_A = \frac{A}{A_n};\quad \varphi_W = \frac{W}{W_n};\ \varphi_I = \frac{I}{I_n}. \tag{8}$$

In the case of an exact solution, the influence of both geometric and structural imperfections should be taken into account. Because of difficulties to obtain their values (especially structural imperfections), the equivalent geometric imperfection e_0 allowing for the initial member bow and also the structural imperfections is more often to use. The characteristic value of this imperfection amplitude is expressed by the following equation

$$e_{0,k} = e_{0,n} = \alpha(\bar{\lambda} - 0.2)W_{el}/A\,, \tag{9}$$

where α is the imperfection factor for relevant buckling curve,
$\bar{\lambda} = (f_y/\sigma_{cr})^{0.5}$ is the non-dimensional slenderness.

The equation (9) can be rewritten using Ayrton – Perry's formula as follows

$$e_{0,k} = \frac{(1-\chi)\cdot(1-\bar{\lambda}^2\chi)}{\chi}\frac{W_{el}}{A} = (1-\chi)\cdot(1-\bar{\lambda}^2\chi)\frac{M_{el,k}}{\chi\cdot N_{pl,k}} = (1-\chi)\cdot\left(1-\frac{N_{pl,k}}{N_{cr}}\chi\right)\frac{M_{el,k}}{\chi\cdot N_{pl,k}}. \tag{10}$$

Then, for the design value of the amplitude of the equivalent geometric imperfection, the following equation is valid

$$e_{0,d} = e_{0,k}\cdot\frac{1-\chi N_{pl,k}/(\gamma_{M1}\cdot N_{cr})}{1-\chi N_{pl,k}/N_{cr}} = e_{0,k}\cdot\frac{1-\bar{\lambda}^2\chi/\gamma_{M1}}{1-\bar{\lambda}^2\chi}. \tag{11}$$

The mean value and the standard deviation of the equivalent imperfection amplitude e_0 needed for further probabilistic analysis, can be obtained assuming the normal distribution function using solution of equation system

$$\begin{aligned} e_{0,k} &= \mu_{e0} + \beta_{5\%}\cdot\sigma_{e0} = \mu_{e0} + 1.645\cdot\sigma_{e0} \\ e_{0,d} &= \mu_{e0} + \beta_d\cdot\alpha_{R(E)}\cdot\sigma_{e0} = \mu_{e0} + 3.8\cdot 0.8\cdot\sigma_{e0} \end{aligned}\,. \tag{12}$$

3. Parametric study of the time-dependent buckling member resistance considering corrosion degradation

To determine effects of the corrosion losses on the buckling member resistance, the parametric study of I – cross-sectional member was realized (the results of previous parametric studies for 6 beams realized for bending moment resistance are presented in [9]). All cross-sectional dimensions b_f, t_f, h_w, t_w as well as the yield strength f_y were considered as random variables according to above mentioned studies for cross-section no. 2 which consists of flanges with nominal dimensions 150 × 16 mm and web with parameters of 268 × 6 mm. Standard deviations of the considered cross-sections were determined using the standard for tolerances of web heights and flange widths [6] and according to standard for tolerances of web and flange thicknesses [7] (class A). They were obtained based on the assumption of the normally distributed random variable standard tolerance a and that 95% of random variable realizations are in interval (so called 2σ rule)

$$\langle m_x - a; m_x + a \rangle . \tag{13}$$

To determine the buckling resistance of a compression member respecting the time-dependent degradation effects caused by corrosion of its flanges, the above mentioned approach was used. Numerical analysis of the determining time-dependent buckling resistance taking into account the randomness of the cross-sectional dimensions and yield strength was realized using software MATLAB. This parametric study was calculated for the non-dimensional slenderness $\lambda_y = \lambda_z = 0.25; 0.50; 0.75; 1.00; 1.25$ and 1.50. The values of random variables were generated by means of Latin Hypercube Sampling (LHS) for 10 000 samples considering normal distribution. For all the prediction models of corrosion losses (see Table 1), the time-dependent development of buckling resistance within the design working life of 100 years was computed. The corrosion losses d_{corr} described by the previously mentioned probabilistic models were considered as reduction of flange thickness given by formula $t_{f,red} = t_f - d_{corr}$, where t_f is the flange thickness without corrosion effects.

Degradation function $F(t)$ was derived as the ratio of time-dependent resistance $R(t)$ of degraded member and resistance R_0 at the beginning of bridge member exploitation.

$$F(t) = R(t)/R_0 . \tag{14}$$

For further general use, the degradation function $F(t)$ has been approximated by means of regression techniques based on the function given by formula (15) using MATLAB software environment.

$$F(t) = a \cdot e^{-b.t} + c . \tag{15}$$

Mean values of computed buckling resistances taking into account the random variables effects can be seen in Fig. 1. In Fig. 2, the mean values of compression member buckling resistances in dependence on non-dimensional slenderness $\bar{\lambda}_y$ and $\bar{\lambda}_z$ for relevant corrosion models are shown.

The strong correlation between fitted function and actual degradation function expressed by means of regression coefficient R-square was obtained. From this study follows, that value ***b*** in formula

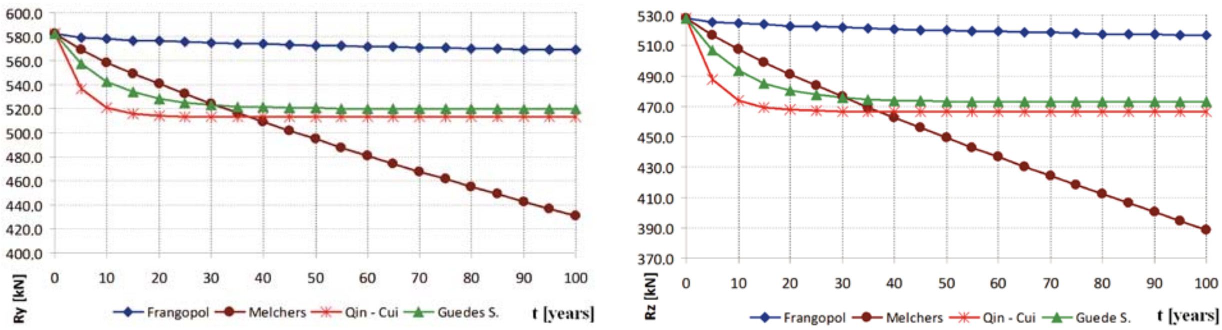


Fig. 1 Time-dependent development of mean values of buckling resistances for non-dimensional slenderness $\bar{\lambda}$ = 1.50 (R_y on the left, R_z on the right)

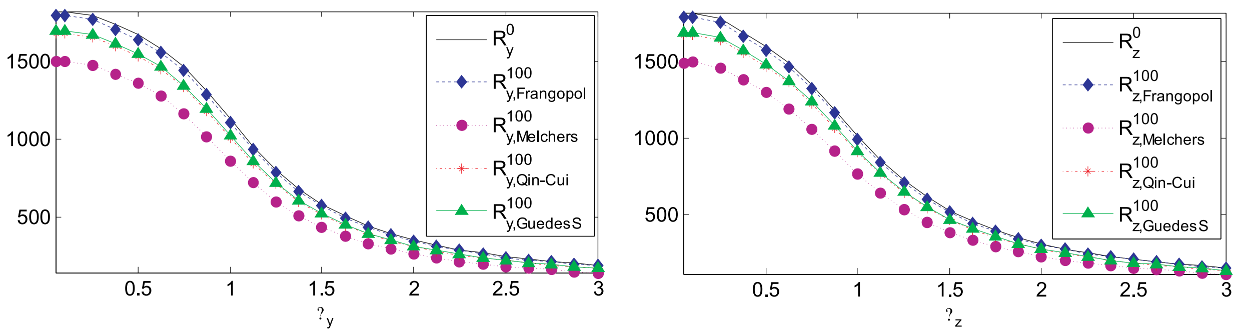


Fig. 2 The buckling resistances Ry and Rz in dependence on non-dimensional slenderness $\bar{\lambda}$ for cross-section no. 2 (in zero time and after 100 years)

(15) remains for relevant corrosion model unified and constants ***a***, ***c*** are only changing. This fact leads to an idea to apply the relevant corrosion model in dependence on environmental condition using known "degradation constant ***b***" and the "shape constants ***a***, ***c***" should be computed in dependence on cross-sectional characteristics. For practical use, the degradation constant ***b*** is supposed to represent time-dependent corrosion progress depending on the relevant corrosion prediction model and constants ***a***, ***c*** are defining the cross-sectional characteristics.

There was also observed in the study that the values of all approximation constants are identical for both relevant buckling modes y or z and, therefore, the buckling mode about axis y has only been investigated in following studies.

For better understanding of this problem and due to effort to get a general solution, the second parametric study for 2000 I- beams with random cross-sectional characteristics was realized allowing for maximum non-dimensional slenderness $\bar{\lambda} \leq 3.0$. This study brought the knowledge that for compression members the degradation constant ***b*** has the same value as was derived for members subjected to bending [9]. The possibility to use the equation for determining the shape constant ***a*** in the form of (16) derived for bending members was also proved but with the limited validation for compression members having non-dimensional slenderness $\bar{\lambda} \leq 0.2$.

$$a_{\bar{\lambda} \leq 0.2} = 2b_f/A \cdot \alpha_{corr.model}\,, \qquad (16)$$

where b_f is the flange width of I cross-section,
A is the cross-sectional area,
$\alpha_{corr.model}$ is the corrosion constant for relevant corrosion model given in Table 2.

For member with non-dimensional slenderness $\bar{\lambda} > 0.2$, the determination of shape constant a is more complicated because the value of this constant varies in dependence on the actual non-dimensional slenderness. The random variable cross-sectional characteristics and the time are other complicating factors. Therefore, the sensitivity analysis for realized second parametric study of 2000 I- beams was worked up. The results of this analysis are shown in Fig. 3.

Using results of sensitivity analysis, the numerically determined dependence of shape constant ***a*** on the non-dimensional slenderness $\bar{\lambda}$ was obtained as a ratio of shape constant $\boldsymbol{a}(\bar{\lambda})$ for arbitrary relative slenderness and value of shape constant $\boldsymbol{a}(\bar{\lambda} = 0)$ for compression members with zero non-dimensional slenderness. The final dependence determined for 2000 I- beams and for corrosion model according to Frangopol is shown in Fig. 4.

In the frame of exact solution the function describing the dependence of shape constant ***a*** on non-dimensional slenderness $\bar{\lambda}$

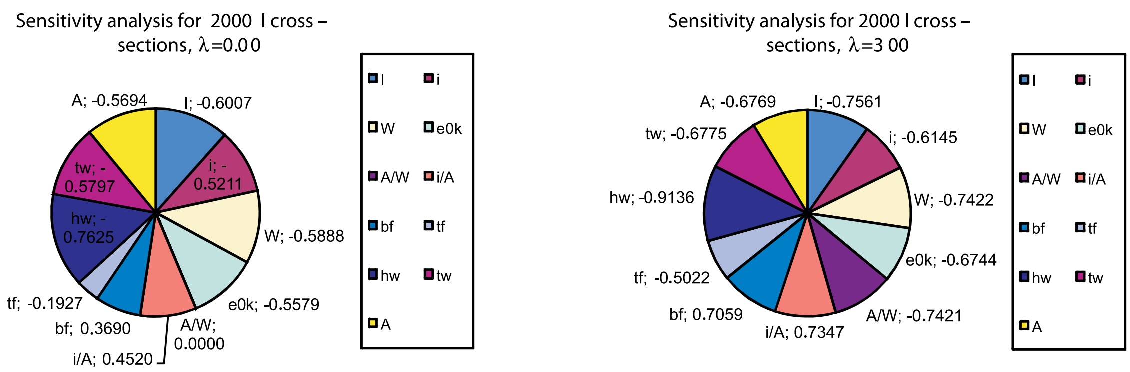


Fig. 3 Sensitivity analysis for 2000 I cross-sections in dependence on non-dimensional slenderness $\bar{\lambda}_y$

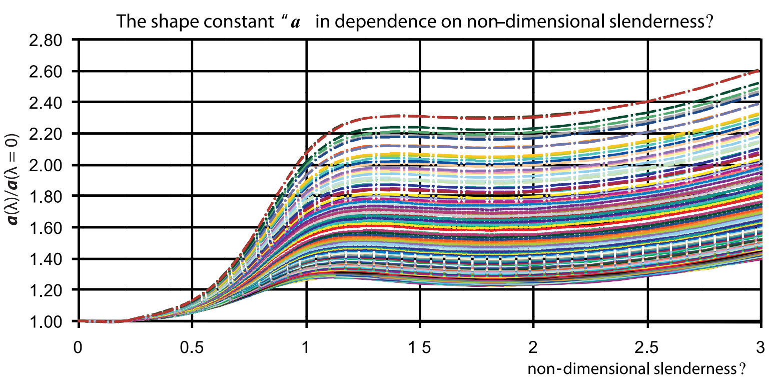


Fig. 4 The shape constant a in dependence on non-dimensional slenderness $\bar{\lambda}$

was not found due to very complicated multi-dimensional problem depending on 4 variables requiring to work with large number of data. To simplify the problem, such a case was verified, where the buckling resistance is computed in every year and for all non-dimensional slenderness range using the value of constant ***a*** valid for the case of pure compression member, i. e. $\bar{\lambda} \leq 0.2$. The results of this simplification were compared to the results obtained by the numerical calculations and the analysis results showed that the progress of ratios of buckling resistance mean values is almost linear in time. Nevertheless, the deviation between exact and simplified solution is increasing in dependence on time. But for common corrosion model, the deviation increase does not exceed the value of 4%. Providing that this deviation increase could be neglected the resistance ratio in time zero can be used, which is identical for all corrosion models, and the equation for a correction factor $\Delta(\lambda)$ would be obtained in the following form

$$\Delta(\lambda) = 1.023 + 0.015 \cdot \cos(2.55\bar{\lambda}) - 0.07 \cdot \sin(2.55\bar{\lambda}). \tag{17}$$

Finally, the time-dependent buckling resistance of the deteriorated compression member at the time *t* could be obtained using the following formulae

$$N_{Rd}(t)_{\bar{\lambda} \leq 0.2} = N_{Rd,t=0} \cdot F(t) = A \cdot f_y \cdot \left[1 + a(e^{-bt} - 1)\right], \tag{18}$$

$$N_{Rd}(t) = N_{Rd,t=0} \cdot F(t) = A \cdot \chi \cdot f_y / \Delta(\lambda) \cdot \left[1 + a(e^{-bt} - 1)\right], \tag{19}$$

$$a = 2 \cdot b_f / A \cdot \alpha_{corr.model}, \tag{20}$$

where the value of the constant ***a*** is calculated according to (20) while for constants ***b***, α_{corr}. model the values in accordance with table 2 are valid.

Values of corrosion constants. Table 2

Corrosion model	Constant	
	b	$\alpha_{corr.\ model}$
Frangopol	0.0194	3.278E-04
Melchers	0.0055	8.460E-03
Qin - Cui	0.2158	1.655E-03
Guedes Soares	0.1003	1.490E-03

4. Conclusion

The member resistance is significantly influenced due to degradation effects in the form of corrosion and therefore the corrosion effects should be taken into account for structural design. This paper presents simplified approach to the determination of the time-dependent member resistance. Nevertheless, the above described approach for design of compression members for durability is not worked up for standard using up to now. The more complex stochastic analysis comparing obtained results with experimental ones should be realized.

Acknowledgement

The paper presents results of the research activities supported by the Slovak Research and Development Agency under the contract No. APVV-20-010005 and by the Slovak Grant Agency, grant No. 1/0311/09.

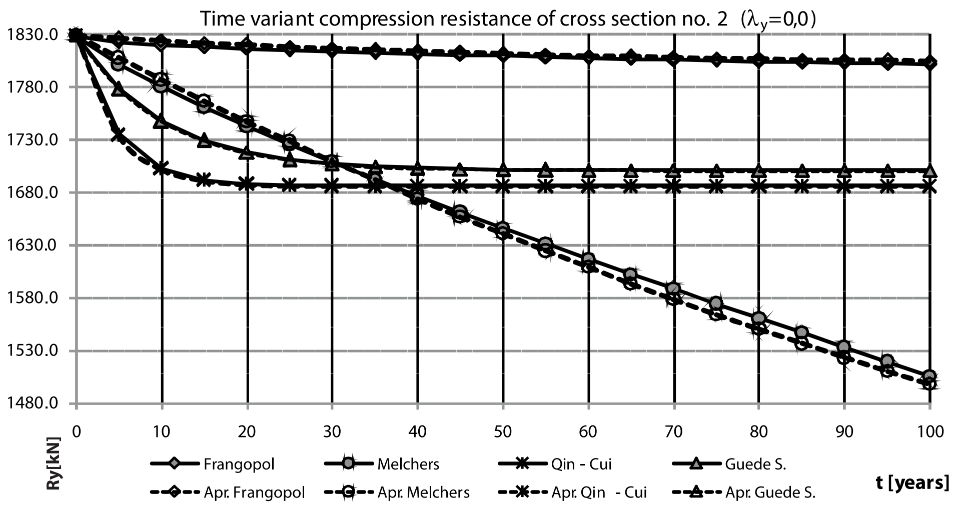


Fig. 5 Comparison of the numerically and according to (18) determined time-dependent member buckling resistance

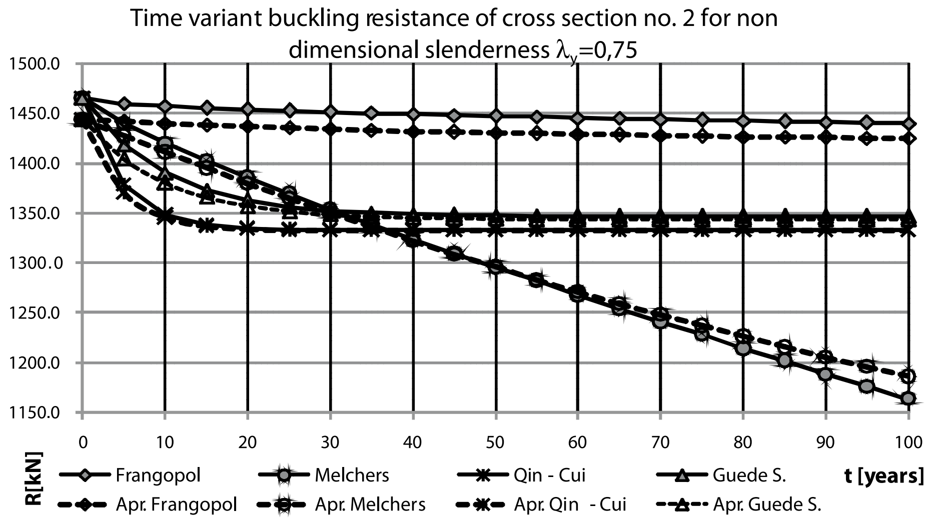


Fig. 6 Comparison of the numerically and according to (19) determined time-dependent member buckling resistance

References

[1] ISO DIS 13823: *General Principles on the Design of structures for Durability.* ISO/TC 98/SC 2.

[2] MELCHERS, R. E.: *Corrosion Uncertainty Modelling for Steel Structure.* Journal of Constructional Steel Research, Elsevier, 1999.

[3] AKGUL, F., FRANGOPOL, D. M.: *Lifetime Performance Analysis of Existing Steel Girder Bridge Superstructures.* Journal of Structural Engineering v. 130, No. 12, pp. 1875-1888.

[4] QIN, S., CUI, W.: *Effect of Corrosion Models on the Time-dependent Reliability of Steel Plated Elements.* Marine Structures, Elsevier, 2003

[5] GUEDES-SOARES, C., GARBATOV, Y.: *Reliability of Maintained, Corrosion Protected Plates Subjected to Non-linear Corrosion and Compressive Loads.* Marine Structures, Elsevier, 1999.

[6] STN P ENV 1090-1: *Execution of Steel Structures. Part 1 – General Rules and Rules for Buildings,* SUTN, Bratislava, 2001.

[7] STN EN 10029+AC 1998: *Tolerances on Dimensions, Shape and Mass for Hot Rolled Steel Plates 3 mm Thick or Above.* SUTN, Bratislava, 1998.

[8] PALES, D., SADOVSKY, Z.: *On Reliability Verification and Column Buckling.* Building Research Journal, vol. 44, 3/1996

[9] VICAN, J., SYKORA, M.: *Design of Bridge Structural Members for Durability.* In: International Conference VSU' 2008, Sofia 2008, pp. 111–116.

[10] VICAN, J., SYKORA, M.: *Design of Compression Members for Durability (in Slovak).* In: International Conference 70 Years of FCE STU, Bratislava, 2008.

Zdena Kralova – Martin Krajcovic *

VARIANCE IN A SECOND LANGUAGE PRONUNCIATION QUALITY

The study assessed the relation between the overall degree of perceived foreign accent in non-natives' English speech and a variety of factors. Five native speakers of English were asked to auditorily evaluate the samples of free English speech produced by each of 99 non-native talkers using the equal-appearing interval scale. The 5-point scale was used to rate each of the variables for each non-native speaker and the interclass correlation coefficients were calculated to identify the most important predictors of foreign accent.

Key words: second language acquisition – pronunciation – factors.

1. Introduction

This paper reports research which attempts to identify factors which explain the variance in the second language (L2) phonic competence. The authors investigate the influence and interrelatedness of certain intralingual and paralingual variables, usually considered significant in the foreign language acquisition. The main focus of the study is the level of phonology, i. e. the acquisition of the phonological subsystem of a foreign language. The data collected by empirical research presented in this paper show that, although the influence of some of these variables is doubtless and remarkable, it is not always possible to single out one of them and to disregard the rest of them. Neither is it always possible to say which of these influences will prevail under some circumstances. The acquisition of the phonological system is an immensely complex process, influenced by numerous factors simultaneously. We tried to "capture" the possible influence of several potentially relevant factors, usually intuitively recognised, by professionals and laymen alike, as highly influential.

For the last 40 years, beginning with the classic article of Asher and García [1], the variety of variables influencing the acquisition of the phonological system of a foreign language has been investigated in a large number of experimental studies. Although considerable individual variation can be observed in adults' mastery of L2 production and perception, the L2 literature, as well as our common experience as language users, generally indicates that the most adult L2 learners will permanently speak L2 with a foreign accent. One interpretation of this observation is, in the scientific study of the foreign accent, that success in acquiring the phonetics and phonology of a L2 is dependent on a number of factors which influence the performance of individual L2 users. Identifying such factors may be important for the teaching and learning of foreign languages [2]. The studies dealing with the issue published so far differ greatly in terms of the languages, subjects, methods or procedures examined. These differences appear to be responsible for the often divergent results the studies have yielded. The relative importance of many factors is uncertain because many variables relating to subject characteristics tend to be confounded and because of lack of adequate experimental control in some studies.

2. L2 pronunciation research

The acquisition of the phonetic/phonological components in the second language is a complex and dynamic process which is influenced by the context and conditions in which the language is learned. The overall quality of L2 pronunciation is likely to vary as a function of the characteristics of the subjects examined. The subjects examined in previous studies differed in a number of potentially important ways. Most of the previous studies examined English as the target L2 being learned. The native languages spoken by the subjects were far more diverse. The nonnative subjects mostly differed in L2 experience, the age at which they started learning L2, the length of residence in the L2-speaking country, the degree of motivation to speak a L2 and many other variables, [e. g., 3 – 12]. The fact that the subjects examined in previous research differ along the dimensions mentioned above often makes direct comparisons across studies problematic. However, this should not lead one to conclude that the degree of L2 pronunciation accuracy cannot be scaled reliably and validly.

The studies of the overall degree of L2 foreign accent also differ in terms of the techniques used to elicit nonative speech samples. In most studies, subjects have been asked to read words, sentences or paragraphs [e. g., 1, 4 and 8]. In a number of studies, subjects have also been asked to produce samples of free (i. e., extemporaneous) L2 speech [e. g., 13 – 16]. And finally, there

* **Zdena Kralova[1], Martin Krajcovic[2]**
[1] Department of English Language and Literature, Faculty of Science, University of Zilina, Slovakia, E-mail: zdena.kralova@fpv.uniza.sk;
[2] Department of Industrial Engineering, Faculty of Mechanical Engineering, University of Zilina, Slovakia

have been studies in which subjects were asked to repeat speech materials after hearing a native speaker model in a direct repetition technique [9] or a delayed repetition technique [17]. Some researchers have used more than one elicitation technique [e. g., 13, 15, 18].

In most instances, analyses were carried out to determine if the degree of L2 foreign accent varied as a function of type of elicitation. Both [13] and [15] reported that read speech was judged to be more strongly foreign accented than extemporaneous speech, probably due to differences in the reading ability. The samples of the extemporaneous speech may be problematic, too. This is because they may contain morphosyntactic and lexical errors influencing the ratings. In addition, subjects may be able to avoid difficult L2 sounds or sound sequences. [12] states that more reliable measures may be obtained if raters evaluate a fixed set of materials. However, we consider free speech samples more appropriate for this kind of research as the free speech imitates real communication acts when the speaker concentrates on the content more than on the form of the speech.

A control group of native speakers was recruited in most of the studies cited so far in addition to groups of nonnative subjects. The numbers of raters used in various L2 pronunciation studies differed greatly, ranging from just one [19] to 85 [20]. It is not known at present how many raters are needed to provide a reliable estimate. Unfortunately, not all the studies have included such a control group [21, 22]. This may lead to several problems – for example, it remains uncertain how native speakers would have performed under the specific circumstances of a particular experiment.

One important methodological question pertains to the characteristics of the listeners. In some studies, naive raters were recruited to evaluate speech samples [e. g., 1, 9]. In other studies, "expert" raters such as linguists or teachers participated [e. g., 16, 23]. [15] reported that experienced raters generally perceived a higher degree of L2 foreign accent in nonnative speech than experienced raters, while [8] found no significant differences between experienced raters and inexperienced raters. In Slovakia it seems rather a problem to recruit a homogeneous group of English native speakers who would be willing to participate in an experiment (often a time consuming one) and who would not be trained or experienced EFL teachers.

The listeners who evaluate the L2 speech usually use a rating scale to indicate the quality of L2 pronunciation they perceive in a speech sample. No standard scale for measuring the accuracy of L2 pronunciation has been developed so far. The equal-appearing interval (EIA) scales differ in resolution. A 5-point scale has been used most commonly [e. g., 8, 13, 14]. [9, 24] employed a continuous scale to evaluate foreign accent. [24] carried out a research to determine whether the foreign accent is a metathetic continuum (a continuum that can be divided into equal intervals ranging from high to low) or a prothetic continuum (a continuum that is not amenable to linear partitioning). They indicated that foreign accentedness is a metathetic continuum, which means that it is appropriate to use an EAI scale, and found that a 9-point (or 11-point) scale should be used to rate L2 speech samples for the degree of the foreign accent.

3. Determinants of L2 pronunciation

The above mentioned differences between studies in design and methodology, have led researchers to draw rather conflicting conclusions about the influence of certain factors on L2 pronunciation. The factors that have received the most attention in literature are undoubtedly the age of L2 learning and the length of stay in a L2-speaking environment. Unfortunately, hardly any study in the existing literature [26] has examined the changes in the degree of L2 pronunciation accuracy in a longitudinal design. Affective and social factors as well as individual aptitude have been seen as possible predictors of the second language phonetic performance [3]. The ability to mimic native speakers and the contact with native speakers have also been reported as having a significant influence on the second language pronunciation [15, 27, 28]. Specific phonetic training has generally been found to be positively associated with the phonetic development in L2. Other factors such as attitudinal and motivational variables have sometimes – though not always – proved to be influential [15].

It seems reasonable to assume that the accuracy with which nonnative speakers pronounce a L2 is, to some extent, dependent on their L1 phonic system. Unfortunately, only a few studies have examined the effect of the L1 background on the degree of L2 foreign accent [e.g., 16, 27, 28]. The authors identified the L1 background as the most important predictor of the foreign accent degree. A few studies have examined the effects of the amount of L1 use on the degree of L2 foreign accent [17, 29]. The authors conclude that the L1 use patterns exerted a significant negative influence on the degree of L2 foreign accent. However, the subjects in those studies were not matched for other variables, the relative importance of the L1 background therefore remains uncertain.

The language-learning motivation can no longer be treated as a single construct, and it needs to be examined in broad terms including many components. These components include cognitive, affective, and behavioral aspects [30]. [30] identifies and describes the motivational components of learning and success in learning EFL employing statistical methods – the factor analysis data reduction technique using the SPSS program called FACTOR (SPSS inc., 2004). He distinguishes the intrinsic/integrative and extrinsic/instrumental orientation subscales of motives, the latter (esp. the desire to have a higher social status and to get a better job) being found to be a critical component in the L2 learning motivation. Findings from this study revealed that the lower the English proficiency level of students, the lower the degree of their intrinsic motivation.

A thorough literature review showed that motivation has some general effect on L2 pronunciation. [9] identified factors designated "integrative motivation" and "concern for L2 pronunciation" as significant predictors of the L2 pronunciation quality.

However, the results obtained so far suggest that the strength of concern for the L2 pronunciation accuracy does not automatically lead to perfect L2 speech. [31] hypothesized that a strong correlation will be found between the variable "professional motivation" and the level of L2 pronunciation, but none of his subjects (highly motivated native English teachers of German) received ratings "native-like pronunciation." [13] and [15] found no evidence that motivation affects the degree of L2 foreign accent. It is important to note that motivation has not been quantified very precisely so far, though different procedures have been used to measure motivational variables. In most cases, the subjects have been asked to rate the importance of good L2 pronunciation for them [e. g., 27, 31]. It is difficult to assess the exact correlation of motivation and L2 pronunciation as many other concomitant variables may be responsible for the results obtained.

Much of the research to date has demonstrated that the age of arrival in a society where the target language is used is the most important factor explaining the variance in the second language phonological competence [e. g., 9, 12, 13, 15, 31]. These results have been used as evidence to support the critical period hypothesis [32, 33]. However, most of these studies concentrated only on immigrant communities. The previous L2 pronunciation research has shown that early learners speak L2 with a lower degree of a foreign accent than late learners. However, no study has yet provided convincing evidence for the claim that L2 speech will automatically be accent-free if it is learned before the critical period (CP). According to the CP hypothesis, the complete mastery of a L2 is no longer possible if learning begins after the end of the CP [6, 33]. A number of researchers have suggested that there may be several CPs, each affecting different linguistic abilities and the first ability to be lost would be the one needed to develop a native-like pronunciation of a L2 [23, 34].

CP effects have usually been attributed to an age-related loss of neural plasticity or to some sort of neurofunctional reorganization that occurs during the development [e. g., 33 – 36]. A few studies have pointed to what now seems to be a paradoxical effect, that is, better performance by late learners of a L2 [19, 37]. [34] concluded that late learners might have an initial avantage over early learners in L2 pronunciation ability, albeit an advantage that is only temporary.

It appears that factors other than the age of learning also have an influence on the quality of L2 pronunciation, for example, the chronological age or amount of L1 and L2 use, etc. In many studies, immigrants' speech has been studied and their age of the first exposure to L2 has been indexed as the age at which the subjects first arrived in a L2-speaking country [1, 6, 13, 17, 21]. The results of these L2 foreign accent studies do, in fact, support the view that the earlier in life one learns a L2, the better it will be pronounced. The finding that there is a linear relationship between the age of learning and the degree of L2 foreign accent led both [13] and [34] to suggest that there is a "sensitive" rather than a "critical" period for L2 learning.

Different suggestions have been made as to when the putative critical or sensitive period for L2 speech learning ends: [6] suggested the age of 15 years; [38] 12 years; and [34] 6 years of age. The last one has been supported by the results of a number of studies [e.g., 9, 21]. However, a few studies have also shown that the age of L2 learning lower than 6 years does not automatically lead to accent-free L2 speech [e.g., 11, 15] and a number of studies have provided evidence that some L2 learners with age of L2 learning of over 12 years may avoid speaking their L2 with a detectable foreign accent [8, 31].

Not every study has shown a significant effect of the length of stay in a L2-speaking environment on the level of L2 pronunciation. In many studies this variable has been operationalized as the "length of residence", which specifies the number of years spent in a community where L2 is the predominant language. In those studies where a stay (or often residence) effect was found, the stay was a less important predictor of the degree of L2 foreign accent than the age [e.g., 24]. Studies that have reported an influence of the stay on the L2 speech include [1, 17, 28]. However, there are also several studies that have not found a stay effect [13, 14, 22]. The discrepancies across the studies may be due, in part, to the fact that stay-factor only provides a rough index of the overall L2 experience.

The influence of the language use on L2 pronunciation was first examined by [27] and [28], who asked learners of English to estimate how much time they had spent speaking English with native speakers. In their study of immigrants a composite variable combining years of residence in the USA and a number of months of cohabitation with native speakers was identified as one of the most important predictors of the degree of L2 foreign accent. Three other studies showed no apparent effect of speaking with L2 native speakers on the L2 pronunciation accuracy [15, 22, 24]. [11] claimed that those learners who continue to speak their L1 frequently have significantly stronger foreign accents in L2 than individuals speaking their L1 infrequently do.

Previous research has provided divergent findings concerning the influence of gender on the degree of L2 foreign accent. Some authors [21] identified gender as a significant predictor of the degree of L2 foreign accent. However, most studies have not identified gender as a significant predictor of the degree of L2 foreign accent [27, 24]. In summary, the results obtained for gender do not lead to any strong conclusions. Some studies reported a significant influence of gender, whereas others did not. The majority of studies that identified gender as a significant predictor of degree of L2 foreign accent found that females usually received higher ratings than males. However, two studies [1, 9] suggested that the effect of gender may vary as a function of age of L2 learning and the amount of L2 experience.

4. Methodology

Ninety-nine (74 female and 25 male) subjects examined in our study are Slovak first-year university students enrolled in the English Language and Literature course. They were approximately at an intermediate level of English proficiency, aged between 18 and 20 years and had typically started learning English at elementary school with a focus on grammar-based instruction. The majority have never lived in an English-speaking country.

The 5 native English control subjects (3 American – 2 male, 1 female; 2 British – 1 male, 1 female) were asked to auditorily evaluate their pronunciation (English phonic competence – EPC) produced by each of the 99 talkers. They were asked to rate each talker by a mark on a scale from 1 (very poor English pronunciation) to 5 (excellent English pronunciation). The raters were English native speakers more or less experienced in ELT who had lived in Slovakia for several months/years.

The pronunciation samples (a free, extemporaneous talk in English) were recorded and other information was elicited from the participants by the way of tests and questionnaire (Table 2). The L2 background questionnaire asked about personal details, experience and language learning attitudes and beliefs.

Relying on the data most frequently presented in research papers and theoretical discussions, we aimed primarily at investigating the possible influence of factors upon the pronunciation accuracy observed in the interlanguage of learners of English as a foreign language (Table 1). The pronunciation task was recorded on a recorder with a condense microphone. The collection took place in a classroom at the University of Zilina, Slovakia and time allotment was 2 minutes for each student. The data were collected by means of a questionnaire and the tests filled out by each subject after their production of English. The questionnaire contained 18 questions relevant for English as a foreign language. Some of these questions provided facts, some other provided the subject's subjective judgement or opinion. The questionnaires required 30 minutes to complete and the testing time varied according to the variable tested (Table 2).

Native speakers were asked to auditorily evaluate the samples of free English speech produced by each of the non-native talkers using the equal-appearing interval 5-point scale. The 5-point scale was used to rate each of the variables for each non-native speaker (Table 3) and the interclass correlation coefficients (Table 4) were calculated to identify the most important predictors of foreign accent. An average rating was obtained for each speaker and the variable EPC was computed by averaging across each rater's score.

To assess their relative contribution to EPC, the variables were submitted to a simple correlation. Correlation coefficients (r) were calculated for each factor by applying the scoring coefficients generated by the principal component analysis to standardized values for subjects' responses to the questionnaire and test items. The variables were correlated with the total pronunciation rating and were entered into statistical analysis (Table 5).

5. Results

Preliminary analysis revealed that much the same EPC ratings were provided by the five raters. The mean for all the native speaker evaluations was 2.93 and there was a standard deviation of 0.73. Initially, the nine measured variables were correlated with the total pronunciation ratings (0.2787 on the 0.05 level). These produced four significant correlations (r) as shown in Table 4:

PsF3 *(contact with native speakers of English)* $r = 0.3928$
IF1 *(English grammatical competence)* $r = 0.3723$
PpF2 *(self-evaluation of English pronunciation)* $r = 0.3013$
PpF1 *(verbal intelligence in Slovak)* $r = 0.2970$

The correlation between EPC and the factors is rather weak. The strongest correlation is between EPC – PsF3 and EPC – IF1 which accounted for 15.6% and 13.9% variance in the pronunciation ratings obtained by native speakers of English – the coefficient of determination (Table 5). The coefficient of determination (Fig. 1) reveals the percentage of EPC variability which can be explained by the linear relationship with a factor.

Classification of factors Table 1

FACTORS OF ENGLISH PHONIC COMPETENCE (EPC)	
I. Intralingual factors (IF)	**II. Paralingual factors (PF)**
IF1 – English grammatical competence	***Psycholingual factors (PpF)***
IF2 – English metaphonetic competence	*PpF1 – verbal intelligence in Slovak*
IF3 – other foreign languages acquired	*PpF2 – self-evaluation of English pronunciation*
	PpF3 – motivation for English language acquisition
	Sociolingual factors (PsF)
	PsF1 – age of English language learning
	PsF2 – length of stay in an English speaking environment
	PsF3 – contact with native speakers of English

Intralingual factors – characteristics related to a language system

Paralingual factors – temporal characteristics partly related to a language system

Testing methods Table 2

FACTOR	Abbr.	Testing method
English grammatical competence	IF1	Quick Placement Test Trial Version 2A (University of Cambridge)
English metaphonetic competence	IF2	English Phonetics and Phonology Test (Z. Kralova)
other foreign languages acquired	IF3	Language Background Questionnaire (Z. Kralova)
verbal intelligence in Slovak	PpF1	Psychological Clinics Standard (Slovak Republic)
self-evaluation of English pronunciation	PpF2	Language Background Questionnaire (Z. Kralova)
motivation for English language acquisition	PpF3	Language Background Questionnaire (Z. Kralova)
age of English language learning	PsF1	Language Background Questionnaire (Z. Kralova)
length of stay in an English speaking environment	PsF2	Language Background Questionnaire (Z. Kralova)
contact with native speakers of English	PsF3	Language Background Questionnaire (Z. Kralova)

Evaluation scale Table 3

EVALUATION SCALE					
	5	**4**	**3**	**2**	**1**
EPC	excellent	very good	good	average	poor
IF1	70 – 57	56 – 43	42 – 29	28 – 15	14 – 0
IF2	40 – 33	32 – 25	24 – 17	16 – 9	8 – 0
IF3	more languages	1 Germanic	1 non-Slavic/non-Germanic	1 Slavic	none
PpF1	≥ 145	144 – 131	130 – 116	115 – 86	85 – 71
PpF2	excellent	very good	good	average	poor
PpF3	combined	internal	external	indefinite	none
PsF1	1 – 6 years	7 – 10 years	11 – 14 years	15 – 18 years	≥ 19 years
PsF2	year/s	month/s	week/s	day/s	none
PsF3	daily	weekly	monthly	yearly	none

Correlation of EPC and IF/PF Table 4

Student	**EPC**	**Intralingual factors**			**Paralingual factors**					
					Psycholingual factors			***Sociolingual factors***		
		IF1	**IF2**	**IF3**	**PpF1**	**PpF2**	**PpF3**	**PsF1**	**PsF2**	**PsF3**
1	**4**	3	3	4	5	3	3	4	4	4
2	**4**	4	4	4	2	4	3	4	1	5
3	**2**	3	2	4	2	2	4	4	1	1
4	**3**	4	3	3	4	3	3	3	1	2
5	**3**	3	3	4	5	3	3	2	1	3
6	**1**	3	2	4	3	3	3	3	1	1
7	**3**	3	3	4	4	3	4	4	1	1
8	**3**	4	3	5	2	4	3	3	1	5
9	**2**	3	3	4	4	3	3	3	1	2
10	**3**	3	2	4	4	2	3	4	5	4
11	**3**	4	3	4	5	4	3	4	1	1
12	**4**	4	3	4	4	4	3	4	3	5
13	**2**	2	2	1	3	3	2	3	1	1
14	**2**	3	2	4	4	3	3	3	1	3
15	**2**	3	2	3	3	3	3	4	1	1
16	**3**	2	3	4	3	4	3	4	1	3
17	**2**	3	2	4	3	3	4	4	1	1
18	**3**	3	3	3	2	3	3	3	4	4

Student	EPC	Intralingual factors			Paralingual factors					
					Psycholingual factors			*Sociolingual factors*		
		IF1	IF2	IF3	PpF1	PpF2	PpF3	PsF1	PsF2	PsF3
19	**2**	3	3	2	3	3	4	4	1	1
20	**3**	3	3	1	4	4	2	4	1	1
21	**2**	4	3	2	2	2	4	3	5	4
22	**3**	3	2	4	4	3	5	4	1	1
23	**3**	3	3	5	3	4	4	3	1	3
24	**3**	3	3	3	4	3	4	4	1	1
25	**2**	3	2	4	3	4	2	3	1	1
26	**4**	3	3	3	4	3	2	3	4	4
27	**3**	4	3	4	4	3	3	4	1	1
28	**3**	3	3	2	4	3	3	4	4	4
29	**3**	3	3	4	4	3	4	4	5	5
30	**2**	4	3	5	3	2	2	3	1	4
31	**3**	4	2	4	4	5	4	3	1	1
32	**3**	3	3	2	3	3	3	4	1	1
33	**2**	3	2	3	4	3	3	3	3	2
34	**3**	3	3	4	3	4	4	4	1	1
35	**5**	5	2	5	4	4	3	4	1	3
36	**1**	3	2	4	1	4	3	3	1	1
37	**2**	3	3	5	4	3	4	4	1	3
38	**2**	3	3	1	2	3	3	3	1	1
39	**3**	3	2	4	5	4	3	4	4	4
40	**2**	3	2	4	5	4	3	4	1	1
41	**3**	3	3	2	2	3	2	4	4	4
42	**2**	3	3	4	3	3	3	4	1	1
43	**4**	3	2	4	4	4	4	4	5	5
44	**4**	3	3	4	5	4	3	1	2	2
45	**2**	3	2	4	4	3	3	4	1	1
46	**3**	3	2	4	5	3	2	3	4	4
47	**3**	4	5	4	5	3	4	3	4	3
48	**3**	3	3	4	3	3	5	2	1	1
49	**3**	3	3	3	4	3	2	4	1	1
50	**4**	4	4	4	4	3	3	4	1	1
51	**3**	3	2	4	4	3	4	3	1	3
52	**3**	4	3	1	4	4	1	3	1	1
53	**3**	3	2	3	2	2	3	4	1	1
54	**5**	4	3	4	4	4	4	4	5	5
55	**3**	3	3	4	4	3	3	4	1	1
56	**2**	3	3	4	3	3	4	3	1	1
57	**3**	4	2	4	4	3	3	4	1	1
58	**3**	3	3	4	4	3	3	2	1	2
59	**3**	4	3	3	4	3	4	4	4	3
60	**3**	3	3	4	4	2	5	4	1	3
61	**3**	3	3	4	5	3	3	4	1	1
62	**3**	4	3	4	4	3	2	4	1	1
63	**4**	4	3	4	3	3	4	4	1	2
64	**3**	3	2	4	4	3	3	4	1	1
65	**2**	3	2	5	4	3	3	3	1	1
66	**3**	4	3	5	3	3	4	4	1	2
67	**3**	2	2	4	4	2	4	4	1	1
68	**3**	4	3	3	4	3	3	4	3	1
69	**3**	3	2	3	4	3	3	4	1	3

Student	EPC	Intralingual factors			Paralingual factors					
					Psycholingual factors			*Sociolingual factors*		
		IF1	IF2	IF3	PpF1	PpF2	PpF3	PsF1	PsF2	PsF3
70	**2**	3	3	4	4	3	4	4	1	1
71	**3**	3	3	3	4	4	3	4	4	4
72	**4**	3	3	1	4	3	4	4	5	4
73	**3**	3	2	4	3	2	3	4	1	3
74	**3**	3	3	4	5	3	3	4	1	1
75	**4**	3	4	4	4	4	4	3	1	1
76	**3**	3	3	5	3	2	4	3	1	2
77	**3**	3	3	4	4	3	2	4	3	3
78	**3**	4	2	4	4	4	3	2	1	2
79	**4**	3	3	5	5	4	3	4	2	4
80	**3**	3	3	3	4	3	3	4	1	1
81	**4**	3	3	3	5	3	4	4	1	3
82	**3**	4	2	1	5	4	4	4	1	1
83	**3**	3	2	1	4	2	3	3	1	1
84	**3**	3	2	4	3	3	3	4	4	3
85	**3**	4	2	4	3	3	3	3	1	1
86	**3**	3	2	2	3	2	3	4	1	2
87	**4**	3	3	5	3	3	5	4	1	1
88	**2**	2	3	5	5	2	3	2	1	1
89	**4**	4	2	4	3	4	3	2	1	1
90	**2**	3	2	3	1	2	1	4	1	1
91	**3**	3	3	3	4	4	1	3	4	3
92	**3**	4	3	3	4	4	5	3	4	3
93	**3**	3	2	4	3	4	4	4	1	1
94	**2**	2	2	4	4	3	3	4	1	1
95	**4**	4	2	5	4	4	4	4	1	3
96	**3**	3	3	1	3	3	3	3	5	5
97	**3**	3	3	5	4	4	4	2	1	1
98	**3**	3	2	5	5	3	4	1	1	3
99	**3**	3	3	4	3	4	3	4	4	3
Mean	2.93	3.22	2.67	3.61	3.61	3.18	3.24	3.51	1.81	2.20
r	-	**0.3723**	0.2761	0.0815	**0.2970**	**0.3013**	0.1475	0.1072	0.2559	**0.3928**

Statistical analysis Table 5

	EPC	IF1	IF2	IF3	PpF1	PpF2	PpF3	PsF1	PsF2	PsF3
Mean	2.93	3.22	2.67	3.61	3.61	3.18	3.24	3.51	1.81	2.20
Modus	3	3	3	4	4	3	3	4	1	1
Median	3	3	3	4	4	3	3	4	1	2
Standard deviation	0.7282	0.5422	0.5860	1.0619	0.9408	0.6569	0.8176	0.7159	1.3902	1.3482
Coefficient of variation	24.86%	16.83%	21.98%	29.45%	26.09%	20.65%	25.22%	20.43%	76.89%	61.23%
Covariance		0.1470	0.1178	0.0630	0.2044	0.1441	0.0878	0.0559	0.2591	0.3880
Correlation coefficient		0.3723	0.2761	0.0815	0.2970	0.3013	0.1475	0.1072	0.2559	0.3928
Coefficient of determination		**0.1386**	0.0762	0.0066	0.0890	0.0908	0.0217	0.0114	0.0655	**0.1561**

6. Conclusion

The study assessed the relation between the overall degree of the perceived foreign accent in non-natives' English speech and a variety of intralingual and paralingual factors claimed to affect the L2 pronunciation. The empirical data clearly indicated that the so-called foreign accent phenomenon is a highly complex one. Namely, it is not always possible to detect a clear correlation

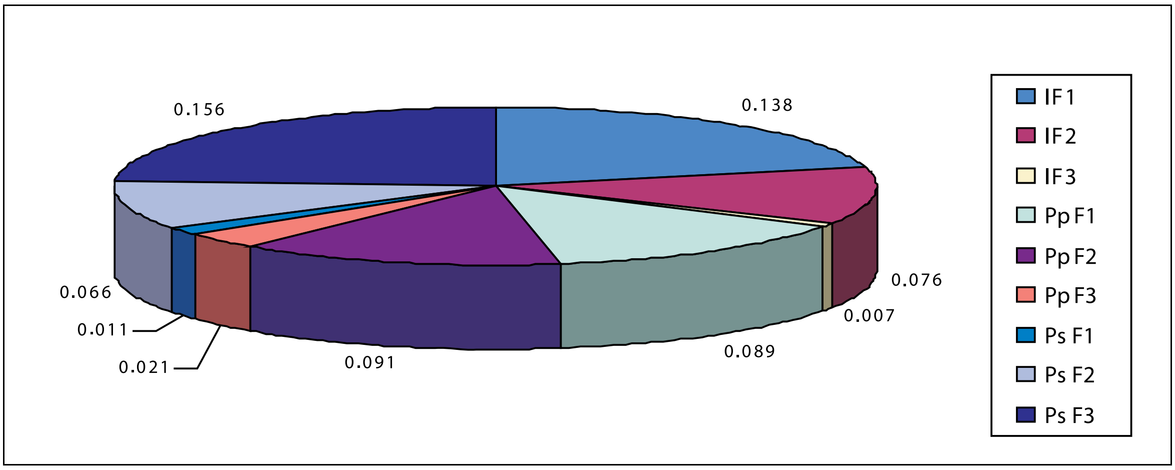


Fig. 1 Coefficients of determination

between the observed level of phonic performance of a foreign language learner and any of the factors usually considered crucial for the reduction of a foreign accent. Thus it is not always possible to isolate those variables which should be considered the most important, nor is it always possible to pin down the exact way in which a particular variable exerts its influence on the process of L2 acquisition. The analysis showed that the mutual influence of a number of factors is so complex that the influence of each and every one of them cannot be singled out and analysed independently, without taking into consideration all the others.

The results show that none of the factors proved crucial for the differences in the foreign language pronunciation rating and no statistically significant correlation could be established between the most of the variables. A statistically significant correlation (0.2787 on the 0.05 level) is an argument for the conclusion that only four of the analysed variables may be considered significant in the foreign language acquisition:

PsF3 (contact with native speakers of English),
IF1 (English grammatical competence),
PpF2 (self-evaluation of English pronunciation),
PpF1 (verbal intelligence in Slovak)

Note, however, that longitudinal research is needed to determine more precisely at what point in L2 learning, if any, the individual factors cease to have an ameliorative effect on the pronunciation of L2. Although it is not always possible to discern the exact way in which some of these variables exert the influence on the linguistic behaviour of the learner, they must always be taken into account when analysing the process of language acquisition at all linguistic levels, but at the level of phonetics in particular. The identification of aspects that influence the overall L2 pronunciation accuracy may be important for teaching and learning second languages.

References

[1] ASHER, J. J., GARCIA, R. (1969):. *The Optimal Age to Learn a Second Language.* The Modern Language Journal, n. 53, pp. 334-341.
[2] HLAVNOVA, A. (1999): *Language as Means of Communication.* Communications - Scientific Letters of University of Zilina, n. 1, pp. 90-93.
[3] FLEGE, J. E. (1987): *A Critical Period for Learning to Pronounce Foreign Languages.* Applied Linguistics, n. 8, pp. 162-177.
[4] FLEGE, J. E. (1988): *Factors Affecting Degree of Perceived Foreign Accent in English Sentences.* Journal of the Acoustical Society of America, n. 84, pp. 70-79.
[5] SINGLETON, D. (1989): *Language Acquistion: the Age Factor. Multilingual Matters*. Clevedon: Avon.
[6] PATKOWSKI, M. S. (1990): *Age and Accent in a Second Language. Reply to James Emil Flege.* Applied Linguistics, n. 11, pp. 73-89.
[7] FLEGE, J. E. (1991): *Perception and Production: the Relevance of Phonetic Input to L2 Phonological Learning.* In: Crosscurrents in Second Language Acquisition. T. Heubner - C. A. Ferguson (eds.), Amsterdam.
[8] BONGAERTS, T., VAN SUMMEREN, C., PLANKEN, B., SCHILS, E. (1997): *Age and Ultimate Attainment in the Pronunciation of a Foreign Language.* Studies in Second Language Acquisition, n. 19, pp. 447-465.
[9] FLEGE, J. E., MUNRO, M. J., MACKAY, I. R. A. (1995): *Factors Affecting Strength of Perceived Foreign Accent in a Second Language.* Journal of the Acoustical Society of America, n. 97, pp. 3125-3134.
[10] BIALYSTOCK, E. (1997): *The Structure of Age: in Search of Barriers to SLA.* Second Language Research, n. 2, pp. 116-137.

[11] FLEGE, J. E., FRIEDA, E. M., NOZAWA, T. (1997): *Amount of Native-language (L1) Use Affects the Pronunciation of an L2.* Journal of Phonetics, n. 25, pp. 169-186.
[12] PISKE, T., MACKAY, I. R. A., FLEGE, J. E. (2001): *Factors Affecting Degree of Foreign Accent in an L2: a Review.* Journal of Phonetics, n. 29, pp. 191-215.
[13] OYAMA, S. (1976): *The Sensitive Period for the Acquisition of a Nonnative Phonological System.* Journal of Psycholinguistic Research, n. 5, pp. 261-283.
[14] PIPER, T.: CANSIN, D. (1988): *Factors Influencing the Foreign Accent.* The Canadian Modern Language Review, n. 44, pp. 334-342.
[15] THOMPSON, I. (1991): *Foreign Accents Revisited: the English Pronunciation of Russian Immigrants.* Language Learning, n. 41, pp. 177-204.
[16] KRALOVA, Z. (2005): *Slovak-English Phonic Interference (in Slovak).* Zilina, Fakulta prirodnych vied Zilinskej univerzity v Ziline.
[17] FLEGE, J. E., YENI-KOMSHIAN, G., LIU, H. (1999): *Age Constraints on Second Language Acquisition.* Journal of Memory and Language, n. 41, pp. 78-104.
[18] MARKHAM, D. (1997): *Phonetic Imitation, Accent, and the Learner.* Lund: Lund University Press.
[19] SNOW, C. E., HOEFNAGEL-HÖHLE, M. (1977): *Age Differences in the Pronunciation of Foreign Sounds.* Language and Speech, n. 20, pp. 357-365.
[20] NEUFELD, G. G. (1979): *Towards a Theory of Language Learning Ability.* Language Learning, n. 29, pp. 227-241.
[21] TAHTA, S., WOOD, M., LOEWENTHAL, K. (1981): *Foreign Accents: Factors Relating to Transfer of Accent from the First Language to a Second Language.* Language and Speech, n. 24, pp. 265-272.
[22] ELLIOTT, R. E. (1995): *Field Independence/Dependence, Hemispheric Specialization, and Attitude in Relation to Pronunciation Accuracy in Spanish as a Foreign Language.* The Modern Language Journal, n. 79, pp. 356-371.
[23] FATHMAN, A. (1975): *The Relationship Between Age and Second Language Productive Ability.* Language Learning, n. 25, pp. 245-253.
[24] FLEGE, J. E., FLETCHER, K. L. (1992): *Talker and Listener Effects on Degree of Perceived Foreign Accent.* Journal of the Acoustical Society of America, n. 91, pp. 370-389.
[25] SOUTHWOOD, M. H., FLEGE, J. E. (1999). *Scaling Foreign Accent: Direct Magnitude Estimation Versus Interval Scaling.* Clinical Linguistics and Phonetics, n. 13, pp. 335-449.
[26] MEADOR, D., FLEGE, J. E., MACKAY, I. R. A. (2000): *Factors Affecting the Recognition of Words in a Second Language.* Bilingualism: Language and Cognition, n. 3, pp. 55-67.
[27] SUTER, R. W. (1976): *Predictors of pronunciation accuracy in second language learning.* Language Learning, n. 26, pp. 233-253.
[28] PURCELL, E. T., SUTER, R. W. (1980): *Predictors of Pronunciation Accuracy: A Reexamination.* Language Learning, n. 30, pp. 271-187.
[29] GUION, S. G., FLEGE, J. E., LOFTIN, J. D. (2000): *The Effect of L1 Use on Pronunciation of Quichua-Spanish Bilinguals.* Journal of Phonetics, n. 28, pp. 27-42.
[30] SEMMAR, Y. (2006): *An Exploratory Study of Motivational Variables in a Foreign Language Learning Context.* Journal of Language and Learning, n. 1, pp. 118-132.
[31] MOYER, A. (1999): *Ultimate Attainment in L2 Phonology.* Studies in Second Language Acquisition, n. 21, pp. 81-108.
[32] PENFIELD, W., ROBERTS, L. (1959): *Speech and Brain Mechanisms.* Princeton: Princeton University Press.
[33] LENNEBERG, E. H. (1967): *The Biological Foundations of Language.* New York: Wiley.
[34] LONG, M. H. (1990): *Maturational Constraints on Language Development.* Studies in Second Language Acquisition, n. 12, pp. 251-285.
[35] LAMENDELLA, J. T. (1977): *General Principles of Neurofunctional Organization and their Manifestation in Primary and Nonprimary Language Acquisition.* Language Learning, n. 27, pp. 155-196.
[36] NEVILLE, H. J., MILLS, D. L., LAWSON, D. S. (1992): *Fractionating Language: Different Neural Subsystems with Different Sensitive Periods.* Cerebral Cortex, n. 2, pp. 244-258.
[37] OLSON, L. L., SAMUELS, S. J. (1973): *The Relationship Between Age and Accuracy of Foreign Language Pronunciation.* Journal of Educational Research, n. 66, pp. 263-268.
[38] SCOVEL, T. (1988): *A Time to Speak: a Psycholinguistic Inquiry into the Critical Period from Human Speech.* New York: Newbury House/Harper and Row.

Peter Brida – Jozef Benikovsky *

RADIO MAP FRAMEWORK FOR GSM POSITIONING

This paper introduces radio map framework for creation of GSM (Global System for Mobile communications) network radio maps. These radio maps are structured in the way that they can be utilized later on as a data source for fingerprinting-based localization methods. Attributes of fingerprinting localization method as well as particular radio signal parameters are explained. The framework handles data measurement, data transfer and data visualization scenarios. Data are basically various radio signal parameters coupled with position coordinates. These data make up a database, which allows centralized storage and opens a path to service-oriented localization architecture in the future. Visualization provides means to verify radio map quality, distribution and density in perceptible and user friendly way as well as means to present and publish achieved results. Entire system is aimed to provide usability by multiple users and further optimization. Centralization allows easy implementation of changes and system maintenance at one place.

Keywords: Localization, fingerprinting, radio map, GSM, received signal strength.

1. Introduction

There was undoubtedly great progress made in the area of localization systems and services over last few years and current market provides numerous navigation and localization devices. Location based services attract more subscribers every year (see [1]) and have great potential. However, there are also drawbacks of this technological advance. Growing number of standards, technical recommendations and resource exhaustion (e.g. radio wave frequency spectrum) reduces possibilities for development of new and independent technologies.

Commonly used localization methods are based on observation of miscellaneous signal parameters. Methods used in cellular networks are Cell Identification (Cell ID), Received Signal Strength (RSS), Angle of Arrival (AoA), Time of Arrival (ToA) and Time Difference of Arrival (TDoA) [2],[3], [4] and [5]. The research of many specialists is focused on RSS method [6], [7], [8], [9] and [10], which can be then processed using trilateration technique [11], [12] or Fingerprinting technique [7], [8] and [13]. Fingerprinting seems to be most accurate and affordable technique as it is suitable for Non Line-of-Sight (NLoS) environments as well as it is more immune to multipath than trilateration.

This work is aimed to show solution that uses fingerprinting algorithm and existing technologies and combines them into *radio map framework* for localization in GSM network. The framework is basically a set of measurement algorithms, communication interfaces and database of various signal parameters usable by fingerprinting method to provide localization itself. Fingerprinting method is briefly described in Chapter 1.1 below. Framework architecture is described in Chapter 2.

1.1 Fingerprinting

Localization methods such as RSS, AoA, ToA and their mutual combinations [14] suffer from inaccuracy caused by estimation of distance between Mobile Station (MS) and Base Stations (BSs) influenced by radio channel effects such as multipath propagation or delay spread.

The fingerprinting method even benefits from aforementioned radio channel properties. It is based on unique "fingerprints" of radio signal properties measured at certain spots. Accuracy of the method in radio networks is determined by two factors [13]. Firstly, signal properties vary very much at relatively small area. For instance, in few meters range, signal from a BS can attenuate, lose or be replaced with stronger one. Secondly, these signals are relatively constant in time. It allows data gathering and their use in future. These fingerprints form the radio map shown in Fig. 1 and described below.

A disadvantage of the method is sensitivity for environment changes – movement of pedestrians and cars, construction and demolition of buildings or weather conditions – they altogether affect signal properties. It is necessary to update the map, but basically, buildings and walls affect the signal most of all and therefore update is not needed very often.

Fingerprinting method consists of two steps. At first, radio map for particular area is created. It is basically a database of spots with known position (coordinates) coupled with various signal properties, e.g. received signal level (RxLev), signal angles or propagation time. This step is called offline phase. The *offline* phase is to be facilitated by the radio map framework presented in this paper.

* **Peter Brida, Jozef Benikovsky**
Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina, Slovakia, E-mail: peter.brida@fel.uniza.sk

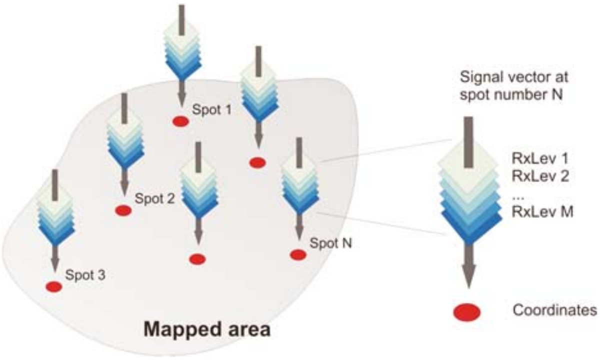


Fig. 1 Radio map for fingerprinting using RxLev

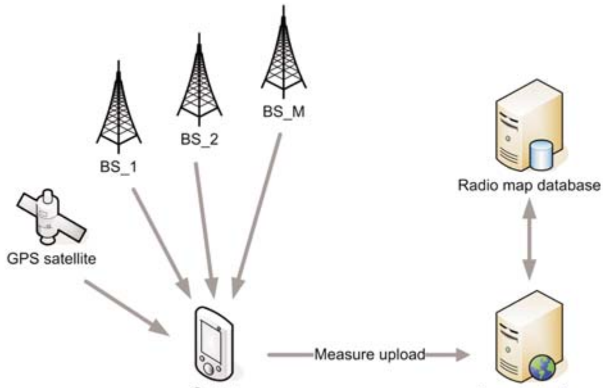


Fig. 2 Gauge subsystem

After the radio map is created, second phase can take place. MS measures signal properties at unknown spot. Then the radio map is searched to find a best match from existing spots. This step is called online phase. The *online* phase is not part of this paper.

2. Radio map framework architecture

The radio map framework consists of two parts – gauge subsystem and radio map itself [15]. Gauge subsystem is used to measure necessary signal parameters, such as identifiers of neighbour base stations coupled with received signal strength. Furthermore it measures Global Positioning System (GPS) coordinates and maps them to signal parameters. This is later user to convert signal information back into geographical coordinates. Finally, it allows transferring this data to radio map.

Radio map is the centralized storage of all measured data and gathers all essential information for fingerprinting online phase. It is actually the database, which can be accessed and queried by all clients or users.

2.1 Gauge subsystem

Fig. 2 shows components that make up gauge subsystem. *BS_1*, *BS_2* and *BS_M* in Fig. 2 represent GSM base transceiver stations. *GPS satellite* represents GPS network utilized to read "precise" position for radio map spots. *Gauge* is a device used to measure GSM as well as GPS signals, interpret them and transfer (upload) measurement results to *communication server*. The *communication server* then calls *radio map database* interfaces to append new results to database and gets notified if there were any errors.

Measurements were performed using pocket computer HP iPAQ hw6515d with Microsoft® Windows CE® 4.21 and Microsoft® .NET Compact Framework 2.0 installed. This device is able to measure GPS position as well as to utilize Subscriber Identity Module Toolkit (SIM TK) to measure GSM signal parameters. Radio map database is stored in Oracle® Express 10.0 database, which is free to use but has limited capabilities. This allows data access using standardized Structured Query Language (SQL).

2.2 Radio map

Radio map consists of spots. Every spot represents point in real world, where measurement was performed. Tab. 1 shows a snip of the radio map database.

Spots are uniquely identified by **spot_id**, which is its database identifier and is not related to GSM or GPS network. Position of the spots is defined by their position coordinates, which are **latitude** (in degrees), **longitude** (in degrees) and **altitude** (in meters). As shown in Tab. 1, columns with same spot_id have same position coordinates. There can be up to 7 measurements of GSM signal parameters performed for every spot, one for serving base station and one for up to six neighbour base stations. There are 3 measurements for spot 28914 and 4 measurements for spot 28915 shown in Tab. 1. Parameters measured from GSM network are:

- **RxLev** – received signal strength (in dBm) of BS under measure,
- **CI** – cell identity, which is unique identifier of BS in GSM network. It is not always available, therefore BCCH and BSIC are measured as well,
- **BCCH** – Broadcast Control Channel number,
- **BSIC** – Base Station Identification Code. Together with BCCH uniquely identifies BS in GSM network.

As shown in Tab. 1, spots with SPOT_ID equal to 28914 and 28915 are close together (according to latitude and longitude row). It is clear that at both spots, signal from base station with CID = 23342 is received with same RxLev. However the RxLev from the subsequent base stations are different. Furthermore, number of base stations differs as well. That shows the power of fingerprinting – even for near spots, signal parameters can change significantly and localization can be performed.

Radio map database example from Zilina-Banova, Slovakia, GSM operator Orange Slovensko. Table 1

SPOT_ID	28914	28914	28914	28915	28915	28915	28915
RxLev	−73	−67	−78	−79	−65	−73	−78
CID	20872	24343	24342	22401	39893	20872	24342
BCCH	3	13	8	10	20	3	8
BSIC	29	27	27	24	26	29	27
Latitude	49,20206	49,20206	49,20206	49,20204	49,20204	49,20204	49,20204
Longitude	18,72063	18,72063	18,72063	18,72065	18,72065	18,72065	18,72065
Altitude	390	390	390	391	391	391	391

2.3 Radio map visualization

The radio map visualization provides inevitable means to get an overview of data in radio map. It is generated by web server and Google Maps™ API using connection into radio map database.

It allows displaying of all spots from radio database at once as shown in Fig. 3. Spots are shown as black dots.

Fig. 3 Radio map database with all spots

Besides, it allows picking one BS and displaying all spots where signal of this particular BS was available as shown in Fig. 4.

3. Future works

Future works can be focused on radio map extrapolation, which would allow estimation of radio signal properties at unknown spots as well as comprehensive visualization options. This would also allow better accuracy in areas of radio map with weaker or missing signal information.

The development of fingerprinting online phase component would enable radio map framework to become usable service-oriented solution for localization in GSM networks.

Fig. 4 Radio map database filtered for BS with CID=4F81, BSIC=29 and BCCH=15

For positioning in areas like buildings, where GPS signal is not available in general, extension of the framework could be proposed. It would allow measuring signal parameters without need of GPS signal. This would have allowed indoor and outdoor positioning using only one device, if localization service had been implemented.

4. Conclusion

Proposed solution shows combination of existing technologies to create robust framework for further localization development. It provides means to create radio map database of any area covered by GSM signal with no dependency on certain mobile network operator. By use of Oracle® Express database and corresponding SQL support, it gives a hand to advanced data research. Data visualization on the other hand provides means for user to easily understand radio map and get an overview of its current state. The core of the proposed solution is "open system" and it can be utilized for various location based services, e.g. alternative positioning solution in intelligent transport system for monitoring vehicles in dense urban areas where GPS does not reliably operate.

Acknowledgement

The work on this paper was partially supported by the grant VEGA 1/4065/07 of Scientific Grant Agency of the Slovak Republic.

References

[1] *Location Based Services Forecast from ABI Research, cit.* [2008-11-12], available on internet: http://www.twine.com/item/11djpjdj2-172/location-based-services-forecast-from-abi-reserach

[2] WANG X., WANG Z., O'DEA B.: *A TOA-based location algorithm reducing the errors due to non-line-of-sight (NLOS) propagation.* IEEE Transactions on Vehicular Technology, 2003, Vol. 52, No. 1, pp. 112–116.

[3] NEMEC Z., BEZOUSEK P., *The Time Difference of Arrival Estimation of Wi-Fi Signals.* Radioengineering, Vol. 17, No. 4, pp. 51–54, ISSN: 1210-2512, 2008.

[4] DELIGIANNIS N., KOTSOPOULOS S., *Mobile Positioning Based on Existing Signalling Messaging in GSM Networks.* In Proceedings of 3rd International Mobile Multimedia Communications Conference (MSAN), 27–29 August 2007, Greece: Nafpaktos.

[5] BRIDA P., CEPEL P., DUHA J., *Mobile Positioning in Next Generation Networks*. (Chapter XI), in Kotsopoulos, S. & Ioannou, K. (Ed.), Handbook of Research on Heterogeneous Next Generation Networking: Innovations and Platforms (pp. 223–252). New York, Hershey: IGI Global (Information science reference), 10/2008. ISBN 978-1-60566-108-7.

[6] BRIDA P.,CEPEL P., DUHA J., *"A Novel Adaptive Algorithm for RSS Positioning in GSM Networks"*, in CSNDSP 2006 PROCEEDINGS, ISBN 960-89282-0-6. Patras, Greece, 19–21 July, 2006, pp. 748–751.

[7] ANNE K. R., KYAMAKYA K., ERBAS F., TAKENGA C., CHEDJOU J. C., *GSM RSSI-based Positioning Using Extended Kalman Filter for Training ANN*. IEEE Vehicular Technology Conference No. 60, Los Angeles, USA, 2004, Vol. 7, pp. 4141–4145, ISBN: 0-7803-8521-7.

[8] LAITINEN H., NORDSTROM T., LAHTEENMAKI J., *Location of GSM Terminals Using a Database of Signal Strength Measurements.* in URSI XXV, National Convention on Radio Science, Helsinki, September 2000.

[9] TAKENGA C., KYAMAKYA K., QUAN W., *On the Accuracy Improvement Issues in GSM Location Fingerprinting.* Vehicular Technology Conference, 2006, VTC-2006 Fall, ISBN: 1-4244-0062-7, Hannover 25-28 Sept. 2006, pp. 1–5.

[10] NERGUIZIAN C., DESPINS C., *AFFES S., Indoor Geolocation with Received Signal Strength Fingerprinting Technique and Neural Networks.* J. N. de Souza et al. (Eds.): ICT 2004, LNCS 3124, 2004, pp. 866–875, ISBN: 978-3-540-22571-3.

[11] M. HATA, T. NAGATSU, *Mobile Location Using Signal Strength Measurements in a Cellular System.* IEEE Trans. On Vehicular Technology, May 1980, Vol. VT-29, pp. 245–252.

[12] WONG C. L. C., LEE M. C., CHAN R. K. W., *GSM-based mobile positioning using WAP.* Wireless Communications and Networking Conference. WCNC. 2000 IEEE, September 2000, Vol. 2, pp. 874–878.

[13] OTSASON, V.: *Accurate Indoor Localization Using Wide GSM Fingerprinting,* Master's thesis, Tartu, 2005.

[14] BENIKOVSKY, J.: *Creation of System for Localization in Real GSM Network (in Slovak)*, Master's thesis, Zilina, 2008.

[15] BENIKOVSKY, J., BRIDA, P.: *Positioning System Based on Fingerprinting Method for Mobile Cellular Networks.* The Mediterranean Journal of Computers and Networks, Vol. 5, 2/2009, pp. 59–67, ISSN: 1744-2397.

Ivan Rados – Ladislav Schwartz *

ANALYSIS OF IMPACT OF THE DIFFERENT NODE ARCHITECTURES AND WAVELENGTH CHANNEL PROTECTION ON THE AVAILABILITY WDM RING NETWORKS

*This paper investigates the availability of WDM (**W**avelength **D**ivision **M**ultiplying) ring networks. An analysis of different nodes, composed of passive and active components, is then carried out, as well as their availability. The availability for optical link, and especially the Span Protection and Route Diversity, is analysed as well. The wavelength channel protection principles being used in WDM ring networks are particularly explicated: OCh (Optical Channel) dedicated and OCh shared protection. The availability models for a WDM ring using these types of protection are given as well as the impact of availability for nodes and optical links on the entire ring availability. Data on the intensity of failures and mean time to repair are taken from different literature.*

Key words: WDM, wavelength channel, availability, protection.

1. Introduction

The availability of services is an important issue in today's optical networks. The increase in data transmission and development of Internet traffic require the high capacity transmission systems. Thanks to the development of fotonic technologies, systems are designed based on Wavelength Division Multiplexing (WDM) having high transmission speeds by a single fiber. In those networks, the service interrupt for whatever reason, caused by equipment failure, or by human factor, can cause unavailability of communication services, as well as heavy losses to the users and network operators. Therefore, the application of protection in such systems has important role for the network operators [1]. Contracts between the operators and their users are always made on the Service Level Agreement (SLA) which is very strict, as far as availability and service quality are concerned. The possibility to ensure high availability is the key one for the operators to maintain the existing and to aquire new users; this mostly reflecting in the price of their network.

The operators may largely invest to equip their network with the high quality of hardware [2]. However, the quality improvement of these components beyond the present level for commercial systems is hard to achieve, and it is beyond their control. More acceptable way to ensure high availability is the use of different protection strategies. Although there are more protection strategies we shall concentrate in this work at the wavelength channel protection of WDM which is very similar to the path protection of SDH (Synchronous Digital Hierarchy) technology [3].

In this work we are going to analyze the availability between two WDM ring terminals using the wavelength channel protection:

- OCh Dedicated protection,
- OCh Shared protection.

This work is finding out what is the impact of a failure on optical cables and equipment on the availability of connection. The models and analysis of different node structures are given as well as their impact on availability and the availability of node depending on the manner it is being used in the network (add, drop, pass-through). In this work we focus on the wavelength channel protection strategy, which consists of ensuring the protection channel for each working channel on the end to end basis.

2. Availability theory

Availability A in some system is the probability that it will be functional in a given time, or it is the time relation in which the system is correct in relation to the total time. Unavailability U of a system is the probability complementary to availability. At reporting about system performances, the unavailability is often expressed as a Mean Down Time (MDT) in minutes of the year or:

$$MDT = \frac{MTTR}{MTTR + MTTF} \qquad (1)$$

Where *MTTF* is Mean Time to Failure, and *MTTR* is Mean Time to Repair [4]. In this work we suppose that *MTTF* for the wavelength channel is constant and independent on component

* **Ivan Rados[1], Ladislav Schwartz[2]**
[1] Department of Transmission Systems HT d.d., Mostar, Bosnia and Herzegovina,
[2] Faculty of Electrical Engineering, University of Zilina, Slovakia, E-mail ladislav.schwartz@fel.uniza.sk,

ageing. $MTTF = 1/\lambda$, where λ is the intensity of failure usually expressed in FIT-a (Failure in Time), $1FIT = 1$ failure in hours.

The basic transport entity of WDM systems is a wavelength channel. This is the simplex type of connection with a standard of value between two nodes (for example, 2.5 Gb/s). The wavelength channel in an unprotected system (serial structure) generally passes through nodes which can be optical add/drop multiplexers, optical switches and optical links. The availability of a serial structure equals to the availability product of individual components, namely:

$$A_s = \prod_{i=1}^{n} A_i \qquad (2)$$

This structure is unable to restore the communication between two nodes in the case of a failure on any of components or the connection is functional if and only if all components of the structure are correct. A substantial availability improvement can be obtained by the application of a parallel structure, which means that in addition to the working path there is a protection one, too. It is certainly important that these two paths are completely independent or that they do not have common components [5]. If we characterize the availabilities of working and protection path as the two completely independent elements (w and p), then the availability of such a structure is the union of two non-disjunctive events or the events which do not exclude each other because there is no complementary variable.

$$A_p = P(w \cup p) \qquad (3)$$

In order that these two events be mutually exclusive it is necessary to transform this union of non-disjunctive events into a union of disjunctive events in which the complementary variables show up. By a simple analysis using De Morgan's laws we obtain that the availability is

$$A = P\left|w \cup (\overline{w} \cap p)\right| \qquad (4)$$

By using the distributive law with regard to the union and section

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \qquad (5)$$

we obtain that

$$A_p = P(w \cup \overline{w}) \cap (w \cup p) \qquad (6)$$

Since $w \cup \overline{w} = 1$ (full set) we can write

$$A_p = P\left[1 \cap (w \cup p)\right] \qquad (7)$$

and to the identity $1 \cap p = p$, so is

$$A_p = P(w \cup p) \qquad (8)$$

And this is actually the availability of the parallel structure. The connection availability for a wavelength channel is [7].

$$A_p = A_k = A_w + A_p - A_w A_p \qquad (9)$$

The connection between two nodes in a parallel structure will be incorrect if, and only if there is a failure on the working and protection path. Of course, the connection can not be established if a failure occurs in terminal nodes because they are the common components of the working and protection path.

3. Availability of the link and different node architectures

Since the nodes are generally connected by optical links, irrespective of their architecture, in this section we shall analyze the availability for links and different node architectures.

3.1 Link availability

In order to determine the wavelength channel availability between two nodes we shall introduce some terms relative to the availability of the node and the optical link connecting two nodes [6].

- a_{OL_i}, availability of i-th link on working path,
- $-a_{n_j}$, availability of j-the node which belongs to the working path.

An unprotected link consists of the optical fiber, optical booster (a_{BOA}), amplifier (a_{OA}), line amplifier (a_{LOA}) and optical pre-amplifier (a_{OPA}). In order that the connection will be functional, all components optical link must be correct. The most often cause of an optical link failure is the breakdown of optical fiber. Since, in the case of the cable breakdown, mostly all the fibers break, we shall suppose that the failures of fiber and cable are fully dependent. Therefore, instead of the availability for fiber, we shall take the availability for cable (a_{OC}).

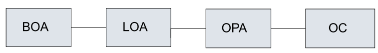


Fig. 1 Unprotected link structure

The availability of the link without the protection equals to the availability product of individual components.

$$a_{OLNP_i} = a_{BOA} \times a_{LOA} \times a_{OPA} \times a_{OC} \qquad (10)$$

In the second instance shown in figure 2, two nodes are connected by two fibers belonging to the same cable.

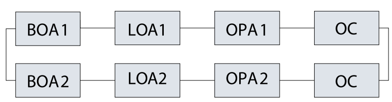


Fig. 2 Span protected link

The availability for this structure (Span Protection-SP) equals to the parallel addition of the individual branch availabilities. So it can be written:

$$\begin{aligned} a_{OL.,SP} &= a_w \cup a_p = a_w + a_p - a_w \cap a_p \\ &= a_{BOA1}a_{LOA1_1}a_{POA1}a_{OC} + a_{BOA2}a_{LOA2}a_{OPA2}a_{OC} - \\ &- (a_{BOA1}a_{BOA2}a_{LOA1}a_{LOA2}a_{OPA1}a_{OPA2}a_{OC}a_{OC}) \end{aligned} \quad (11)$$

Although in the above equation's small brackets we have a product of two equal members $a_{OC} \times a_{OC}$, only one is present in the formula for availability a_{OC} because the cause of node failure is the same so we get:

$$\begin{aligned} a_{OL.,SP} &= a_{BOA1}a_{LOA1_1}a_{POA1}a_{OC} + a_{BOA2}a_{LOA2}a_{OPA2}a_{OC} - \\ &- (a_{BOA1}a_{BOA2}a_{LOA1}a_{LOA2}a_{OPA1}a_{OPA2}a_{OC}) \end{aligned} \quad (12)$$

In this type of protection it is actually the protection in the case of a failure on some of the amplifiers but not the optical fibers.

In the third type of protection (route diversity-RD), the link structure is same as in span protection, in that the nodes are connected by optical fibers belonging to completely separated cables, as shown in figure 3.

$$\begin{aligned} a_{OL.,RD} &= a_{BOA1}a_{LOA1_1}a_{POA1}a_{OC1} + a_{BOA2}a_{LOA2}a_{OPA2}a_{OC2} - \\ &- (a_{BOA1}a_{BOA2}a_{LOA1}a_{LOA2}a_{OPA1}a_{OPA2}a_{OC1}a_{OC2}) \end{aligned} \quad (13)$$

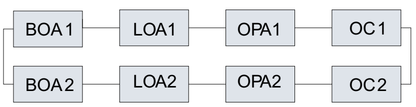


Fig. 3 Completely separated links

This type of protection protects the connection in the case of a failure on transmission equipment as well as optical cables.

3.2 Availability for different node structure

Nodes in which a wavelength channel is being dropped and added are called termination nodes. Nodes through which a wavelength channel runs from "west" to "east" and are located between terminal nodes are called pass-through nodes. Since the wavelength channel passes through the different components inside these two types of nodes their availability is different.

- $-a_{n_jt}$, terminal node is a working one
- $-a_{n_jp}$, pass-through node is working one

If all nodes of the same type are the same we have:
$a_{n_jt} = a_{nt}, \forall j,$
$a_{n_jp} = a_{np}, \forall j.$

3.2.1 Node with passive components

If passive components are used, the node structure in the 1+1 wavelength channel protection are show in figure 4

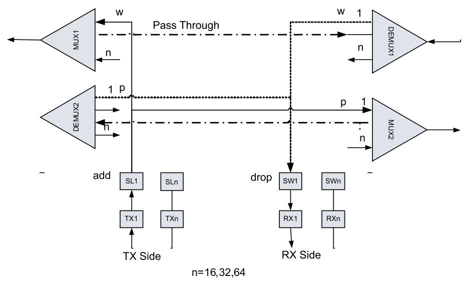


Fig. 4 Node structure for 1+1 protection

For the instance of a terminal node (bidirectional communication), the wavelength channel passes through the transmitter, divider, multiplexers, demultiplexers, switch and receiver, so that the terminal node availability for one wavelength channel is

$$\begin{aligned} a_{nt,1+1(1)} &= a_{TX1} \times a_{SPL1} \times (a_{MUX1} \cup a_{MUX2}) \times \\ &(a_{DMUX1} \cup a_{DMUX2}) \times a_{SW1} \times a_{RX1} \end{aligned} \quad (14)$$

In pass through nodes, a wavelength channel passes through the demultiplexers and multiplexers so the availability for pass-through nodes is:

$$a_{np,1+1(1)} = a_{MUX1} \times a_{DMUX1} \times a_{MUX2} \times a_{DMUX2} \quad (15)$$

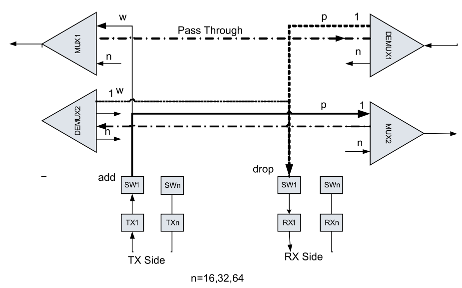


Fig. 5 Node structure for 1:1 protection

In 1:1 protection, if passive components are being used, the wavelength channel in terminal nodes passes through the transmitter, switch, multiplexers, demultiplexers, switch and receiver so that the terminal node availability for one wavelength channel and 1:1 is

$$\begin{aligned} a_{nt,1:1(1)} &= a_{TX1} \times a_{SW1} \times (a_{MUX1} \cup a_{MUX2}) \times \\ &(a_{DMUX1} \cup a_{DMUX2}) \times a_{SW1} \times a_{RX1} \end{aligned} \quad (16)$$

The availability for pass through nodes is same as for the 1+1 protection. The availability for all wavelength channels (n) inside a terminal node should be

$$a_{nt,1+1(n)} = a''_{TX1} \times a''_{SPL1} \times (a_{MUX1} \cup a_{MUX2}) \times (a_{DMUX1} \cup a_{DMUX2}) \times a''_{SW1} \times a''_{RX1} \quad (17)$$

3.2.2 Node with active components

An optical switch with add/drop possibilities is used as the active component. Figure 6 represents an optical add/drop multiplexer (OADM) using such an optical switch.

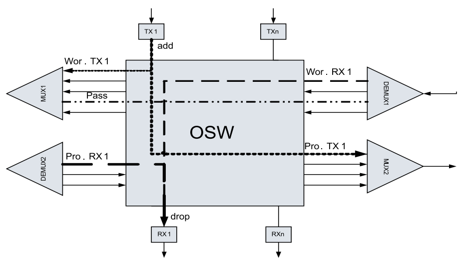


Fig. 6 Node structures with an active optical switch

In the case that the node is used as terminal one, the wavelength channel passes through the transmitter, optical switch, multiplexers, demultiplexers and receiver so the availability for one wavelength channel is

$$a_{nt,(1)} = a_{TX1} \times a_{OSW} \times (a_{MUX1} \cup a_{MUX2}) \times (a_{DMUX1} \cup a_{DMUX2}) \times a_{RX1} \quad (18)$$

The availability for all the wavelength channels in the terminal node should be

$$a_{nt,(n)} = a''_{TX1} \times a_{OSW} \times (a_{MUX1} \cup a_{MUX2}) \times (a_{DMUX1} \cup a_{DMUX2}) \times a''_{RX1} \quad (19)$$

If the node is a pass-through one wavelength channel passes through the multiplexers and optical switch so the availability is

$$a_{NP} = a_{MUX1} \times a_{DMUX1} \times a_{MUX2} \times a_{DMUX2} \times a_{OSW} \quad (20)$$

The availability of terminal nodes is same for the 1+1 and 1:1 protection because the optical switch can act both as the divider and switch.

4. Availability for WDM ring

Since the installation of more SDH line systems between two nodes is very expensive, as the capacities of optical cables exhaust considerably, the need for high transmission capacity system requiring only two fibers, has arisen. Such are WDM systems based on wavelength multiplexing which uses the wavelength channel protection. In this section we shall analyse the availability for WDM ring which uses the wavelength channel protection. In this type of protection the switching is carried out on a wavelength channel level so that the protection capacity is one wavelength. There are two types of protection: OCh dedicated protection ring (OCh DPRing) and OCh shared protection ring (OCh SPRing).

4.1 OCh dedicated protection ring

This type of protection requires two fibers in a ring. Each wavelength channel is being routed on the working path along one side of the ring, and the corresponding dedicated wavelength channel, along the opposite side. Bidirectional wavelength requirements are supported by two wavelength channels; one in each direction. Two types of dedicated protection are possible: 1+1 and 1:1.

In the ring network which uses the 1+1 wavelength channel protection the wavelength channel in the source node is being duplicated and concurrently delivered in both directions of the ring; working and protection [11].

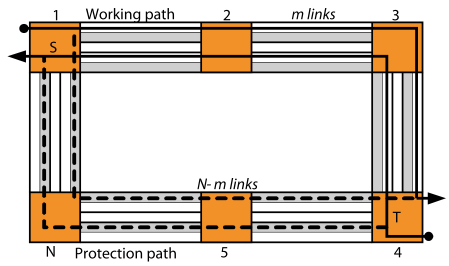


Fig. 7 1+1 wavelength channel protection under conditions without a failure

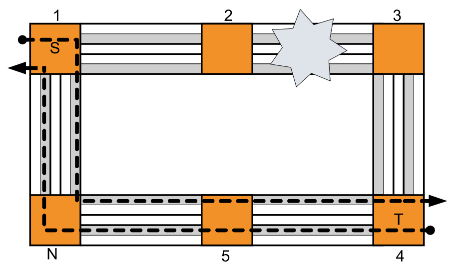


Fig. 8 1+1 wavelength channel protection in the case of a failure on working path

Under ordinary conditions in the terminal node, the receivers gets two signal copies (with a different delay) and chooses the best one. In the case of a failure on the working path, the receiver chooses the signal that it gets from the protection path. This is so called single ended protection because the switching is carried out only on one (receiving) side. It is important that the working and protection path do not have common components in order that one failure would not cause total communication break down, and this means that the component failures on the working and protection paths are fully independent.

For each wavelength channel of the working path, a corresponding wavelength channel on the protection path is reserved and therefore, we are talking about an 1+1 wavelength channel protection [8].

Since the wavelength channels on the protection path are persevered we are talking about so called dedicated protection in which no active switch is required. The wavelength channels on the working and protection paths have the same wavelength but that is not a rule. If different wavelengths are being used, the wavelength converters not having big impact on WDM ring availability must be applied.

Let the network generally consist of N nodes and N links connecting those nodes [9]. If a wavelength channel on the working path P0 passes through the *m* number of optical links between terminal nodes, the availability for the working path is equal to the availability product for optical links and nodes through which this wavelength channel passes [10].

$$\begin{aligned} a_{st}(P_0) &= \prod_{i,j\in P_0} a_{OL_i} \times a_{n_j} = \prod_{i\in P_0} a_{OL_i} \times \prod_{j\in P_0} a_{n_j} \\ &= (a_{nt})^2 \times (a_{np})^{m-1} \times \left[\prod_{i\in P_0}^{m} a_{OL_i}\right] \end{aligned} \tag{21}$$

In the case of a failure on the working path, the wavelength channel passes the $N-m$ number of optical links and the $N-m-1$ of nodes on the protection path P1.

The availability for the protection path is [10].

$$\begin{aligned} a_{st}(P_1) &= \prod_{i,j\in P_1} a_{OL_i} \times a_{n_j} = \prod_{i\in P_1} a_{OL_i} \times \prod_{j\in P_1} a_{n_j} \\ &= (a_{nt})^2 \times (a_{np})^{N-m-1} \times \left[\prod_{i\in P_1}^{m} a_{OL_i}\right] \end{aligned} \tag{22}$$

The availability for the wavelength channel between the s and t nodes is completely determined by these two paths so that the availability for the wavelength channel in the case of 1+1 protection is calculated as the availability of two branches which are fully independent.

$$A_{st}(a) = a_{st}(P_0) + a_{st}(P_1) - \left[a_{st}(P_0) \times a_{st}(P_1)\right] \tag{23}$$

$$\begin{aligned} A_{st}(a) &= (a_{nt})^2 \times (a_{np})^{m-1} \times \left[\prod_{i\in P_0}^{m} a_{OL_i}\right] + \\ &+ (a_{nt})^2 \times (a_{np})^{N-m-1} \times \left[\prod_{i\in P_1}^{N-m} a_{OL_i}\right] - \\ &- \left\{ \begin{matrix} (a_{nt})^2 \times (a_{np})^{m-1} \times \left[\prod_{i\in P_0}^{m} a_{OL_i}\right] (a_{nt})^2 \times \\ \times (a_{np})^{N-m-1} \times \left[\prod_{i\in P_1}^{N-m} a_{OL_i}\right] \end{matrix} \right\} \end{aligned} \tag{24}$$

where a determines the availability for optical links and nodes. Although in the equation brackets we have the product of the same two members $(a_{nt})^2 \times (a_{nt})^2$ only one is taken in the expression for availability $(a_{nt})^2$ because the cause of node failure is the same one.

Therefore we have

$$A_{st}(a) = (a_{nt})^2 \times \left\{ \begin{matrix} (a_{np})^{m-1} \times \left[\prod_{i\in P_0}^{m} a_{OL_i}\right] + \\ + (a_{np})^{N-m-1} \times \left[\prod_{i\in P_1}^{N-m} a_{OL_i}\right] - \\ - (a_{np})^{N-2} \times \left[\prod_{i\in P_0, P_1}^{N} a_{OL_i}\right] \end{matrix} \right\} \tag{25}$$

If we assume that optical links have the same length, their availability is same, i.e

$a_{OL_i} = a_{OL}, \forall\, i.$

In this case, the availability between the s and t nodes is

$$A_{st}(a) = (a_{nt})^2 \times \left[\begin{matrix} (a_{np})^{m-1} \times (a_{OL})^{m} + \\ + (a_{np})^{N-m-1} \times (a_{OL})^{N-m} - \\ - (a_{np})^{N-2} \times (a_{OL})^{N} \end{matrix} \right] \tag{26}$$

In 1:1 protection, the wavelength channel on the front side is not permanently duplicated so that the switching is performed dual-ended which requires a protocol to coordinate it. From the availability aspect, the same availability expression applies as for the 1+1 protection. The advantage of 1:1 protection to the 1+1 one is that the auxiliary capacity, under conditions without a failure, can be used for the subordinated traffic transmission.

In the case of a failure, the auxiliary capacity is used as the protection one for the superordinated traffic. A disadvantage of 1:1 protection is the complexity because the switching for different OADMs must be synchronous.

4.2 OCh shared protection ring

This is the unique type of protection in the ring WDM networks having no equivalent in the SDH ring networks. This protection also uses 2 fibers for connecting the ring nodes. It is actually very similar to an 1:1 protection because it does not have prereserved protection wavelength channels. The difference is that half of the wavelength channels on each fiber is reserved for working traffic and the other half for protection traffic, while in an 1:1 protection all the wavelength channels on a fiber are reserved for the protection, and in a situation without a failure they can be used for the subordinated traffic. Working channels in one fiber are protected by working channels in another fiber. Protection wavelength channels pass along the ring in the opposite direction than the working channels do. Two directions of the bidirectional wavelength requirement are routed on the same side of the ring in opposite fibers. The same wavelength can be reused for some other non-overlapping requirement between different nodes.

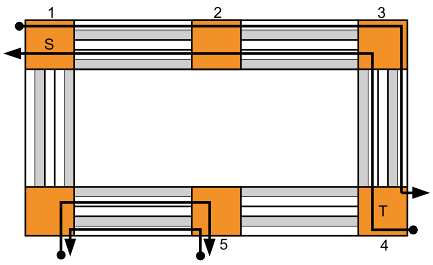


Fig. 9 OCh shared protection under conditions without a failure

When a failure occurs, the damaged wavelength channels are switched in terminal OADMs to another side of the ring and then they use the protection wavelength channels in that fiber. Bidirectional traffic requirements will be routed by using different wavelengths for both directions; otherwise, the wavelength conversion is needed for switching the traffic from working wavelength channel to the protection channel in the opposite direction. In this type of protection there is no prededicated wavelength channels for each wavelength channel of the working path, but there are resources for the damaged wavelength channels. Protection paths will be formed only after a failure on the working path.

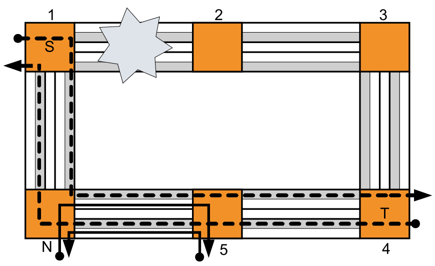


Fig. 10 OCh shared protection under conditions with a failure

When a failure damages all the links between two adjacent OADMs in the ring, the protection path will be established between the source and destination OADM. When a failure occurs, the switching action is required in the source and destination OADM. However, the switching must be carried out for each damaged wavelength. From the availability aspect, the same expression applies for this type of protection as for the 1:1 dedicated protection, in that case the capacity in the ring is more efficient.

5. Numerical results

As shown before, the same expression applies for the availability of OCh dedicated and OCh shared protection. The only difference is in the nodes structure. For availability calculation we shall use this expression by changing the nodes structure and using different optical link protections: **S**pan **P**rotection and **R**oute **D**iversity. In order to calculate the availability for a WDM system one needs to know the nodes and optical links availability. For availability calculation, in general, one needs to know the intensity of failures for the individual components. The data are shown in Table 1, and are taken from different sources [9].

NOTE: w is the number of wavelength channels

Availability data for optical components Tab. 1

Component/Device	Symbol	Failure Rate(FIT)
Booster Amplifier	BOA	3200
Line Amplifier	LOA	3200
Pre - Amplifier	POA	3200
Multiplexer	MUX	25xW
Demultiplexer	DEMUX	25xW
Optical Switch	OSW	1000
Fix Transmitter	TRX	186
Tunable Transmitter	TX	745
Fix Receiver	RX	70
Tunable Receiver	RCX	470
Switch	SW	50
Spliiter	SPL	50
Cable (per km)	OC	100

Unavailability and MDT for an optical link $\lambda = 100$ FIT/km Tab. 2

	$U \times 10^{-6}$	MDT (min/year)
MTTR=12		
Span Protection	96.01	50.51
Route Diversity	0.04	0.02
MTTR=21		
Span Protection	168.01	88.30
Route Diversity	0.13	0.07

Optical link unavailability results from the above table are obtained on the assumption that the link length is 80 km and MTTR=12 hours and 21 hours. As it can be seen from Table 2, the unavailability for optical link, as expected, is substantially higher for Span Protection than it is for Route Diversity. Namely, in SP, fibers of the both paths (working and protection) belong to the same cable, so that the same cause of failure actually breaks down the entire communication between the nodes and, therefore, traffic as well. It is also noticeable that the increase in time to repair considerably increases the unavailability. Unavailability for optical nodes with passive and active components was calculated for the instances of 16 and 64 wavelengths, at a time to repair of 4 and 6 hours, which is real time often being taken for the device repair.

As it can be seen from tables 3 and 4, the unavailability for terminal nodes, irrespective of active and passive components, is almost independent on the number of wavelengths. This is logical because the multiplexers are, in fact, in parallel connection for both types of protection (SP and RD).

Unavailability and MDT (min/year) for a node with passive components — Table 3

	$U \times 10^{-6}$		MDT	
MTTR=4 h	λ = 16	λ = 64	λ = 16	λ = 64
Terminal	1.42	1.42	0.75	0.75
Pass-through	6.40	25.60	3.36	13.46
MTTR=6 h				
Terminal	2.13	2.13	1.12	1.12
Pass-through	9.60	38.42	5.04	20.18

Unavailability and MDT (min/year) for a node with active components — Table 4

	$U \times 10^{-6}$		MDT	
MTTR=4 h	λ = 16	λ = 64	λ = 16	λ = 64
Terminal	5.02	5.02	2.64	2.64
Pass-through	10.40	29.60	5.47	15.56
MTTR=6 h				
Terminal	7.53	7.53	3.96	3.96
Pass-through	15.60	44.40	8.20	23.34

In pass-through nodes, the increase in the number of wavelengths has an essential impact on the unavailability increase. The reason is a serial connection of the multiplexers (for PC and MTTR=4 hours this increase is about 4 times; for AC it is about 2.84 times).

The reason for this different increase is in the fact that in pass-through nodes of AC structure there is an optical switch in serial connection. It can also be seen that the availability for nodes (both terminals and pass-through) with active components, in general, is higher than it is for the same nodes with passive components (for example, for MTTR=6 hours, terminal node with active components has 3.5 times higher unavailability than the same node with PC does). The reason is in the fact that an active optical switch has a few times higher probability of failures than the passive switch and divider do. It is also noticeable that the increase in the device time to repair increases unavailability, irrespective of whether the AC or PC components are concerned, for about 1.5 times regardless of the number of wavelengths.

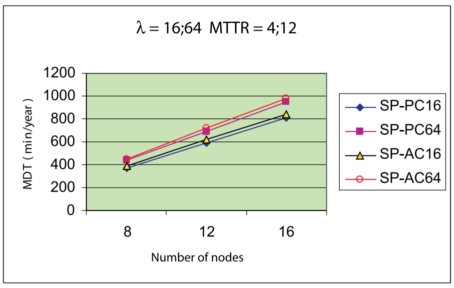


Fig. 13 MDT for Span Protection for MTTR=4; 12 h

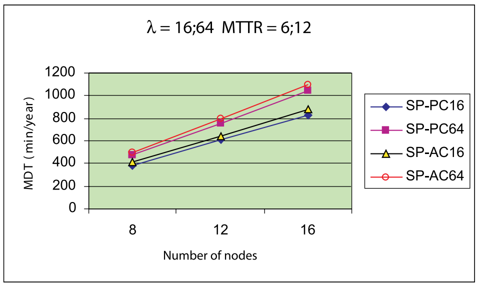


Fig. 14 MDT for Span Protection for MTTR=6; 12 h

We assumed that the ring would consist of 8, 12 and 16 nodes; 16 and 64 wavelengths, and that the distance between optical amplifiers and nodes is 80 km. We also analyzed the impact of the reduction in optical link and device time to repair on system availability. It is worst of node that in figures below, during the time to repair (MTTR), the first number refers to the device repair while the other is for the link repair.

In figures 13 and 14 are shown the unavailability and/or MDT for Span Protection with passive and active components, 16 and 64 wavelengths, for various device and optical link times to repair. It is noticeable that the unavailability is almost independent on the number of wavelengths, because the multiplexer in the terminal node is in parallel connection.

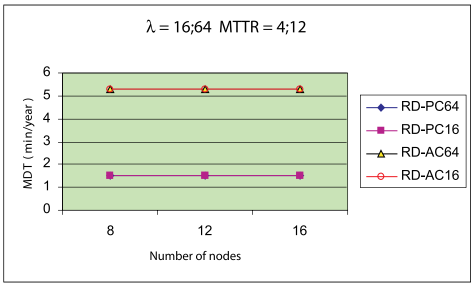


Fig. 15 MDT for Route Diversity for MTTR=4; 12 h

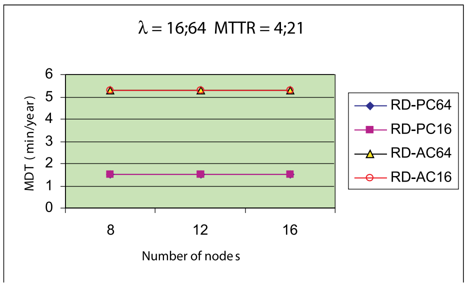


Fig. 16 MDT for Route Diversity for MTTR=4; 21h

A somewhat higher unavailability is also noticeable if active components are concerned, because there is an optical switch having a significant intensity of failures. As expected too, the unavailability increases by increasing the number of nodes and total distance between terminal nodes, because the impact of optical link on the entire unavailability is considerably higher than it is the impact of nodes in this type protection (for example, for λ = 16, N = 12 the link impact on the entire unavailability for nodes is from 90.25 % to 9.75 %). In SP, this is an expected result because both the working and the protection path are connected to optical fibers belonging to the same cable. The unavailability increase if optical link time to repair MTTR=21, is substantially higher then the MTTR increase for the device. This is expectable because of the significant impact of optical link unavailability on total unavailability.

Figures 15, 16 and 17 show the WDM ring unavailability for 8, 12 and 16 nodes, which use a Route Diversity protection. It is noticeable that the unavailability for the equal number of nodes and wavelengths is a few dozen times lower than is in SP. The reason is that the working and protection path have no common components (complete separation), so that the breakage of working path optical link does not have the impact in the communication between terminal nodes.

In RD the impact of nodes on total unavailability is dominant, as expected, because the optical links are completely separated and in parallel connection (for example, for λ = 16, N = 8 the link impact on the entire unavailability for nodes is from 3.55 % to 96.45 %). It is also noticeable that the number of wavelengths does not have a substantial impact on the entire unavailability either, not it has the number of nodes, and the distance between terminal nodes. Once again, the reason for this is the complete separation of the working and protection path.

It can be seen that the increase in time to repair for nodes has a bigger impact on the entire unavailability than it has the increase in time to repair for optical link (MTTR increase from 4 to 6 hours does increase the unavailability for about 66 % while a MTTR increase of link from 12 to 21 hours does not increase the unavailability at all). The reason for this is again the entire paths independence which means that the one path breakage does not shut down the communication between the nodes.

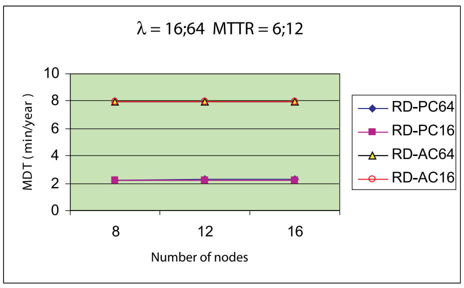


Fig. 17 MDT for Route Diversity for MTTR= 6; 12h

In figures 18 through 21 it is shown the SP and RD unavailability depending on the number of wavelengths and nodes for different times to repair for device and optical link.

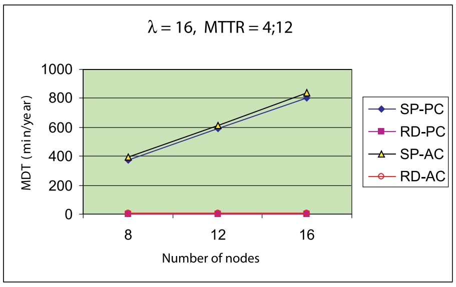


Fig. 18 MDT for MTTR= 4;12h and λ = 16

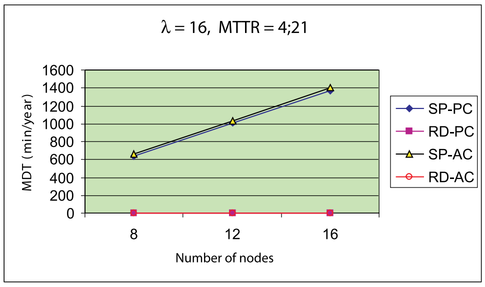


Fig. 19 MDT for MTTR=4; 21 h and λ = 16

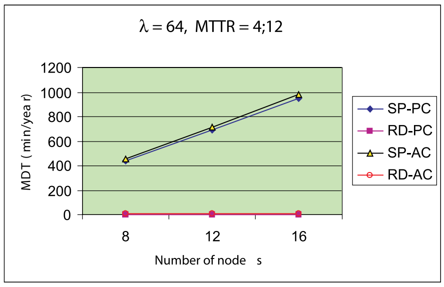


Fig. 20 MDT for MTTR=4; 12 h and λ = 64

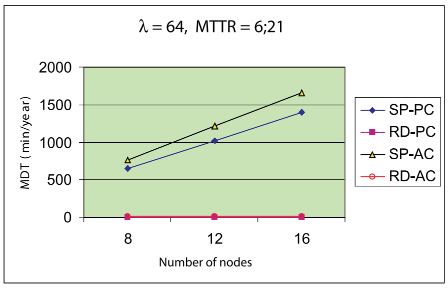


Fig. 21 MDT for MTTR=6; 21 h and λ = 64

RD, as seen, has a few dozen times lesser unavailability than SP has, for the reasons stated previously. It is also noticeable that in SP, the unavailability increases with the number of nodes, and link lengths, while in RD, that dependence is very small, almost insignificant, because the optical links parallel structure is concerned where a failure in one link does not have the impact on another, so that the communication between the nodes remains retrieved.

The highest unavailability, as understandable has SP where MTTR=6; 21 because the optical link impact in this type protection is dominant on the total unavailability. The reason for this it that for optical cable breakage the communication between the terminal nodes is being lost for as long as the very same has not been repaired, because both the working and protection are connected to the fibers of same cable (for N = 16 an increase in time to repair from 4 to 12 to 6 and 21 increases the SP unavailability for 59 %).

Of course, in a WDM ring, irrespective whether SP and RD, the terminal nodes have the biggest impact on the total unavailability because their failure leads to the total communication shut down between those nodes.

6. Conclusion

This paper investigates the impact of optical links and node architectures on the availability of connections in WDM ring networks. This work's result analysis shows that Span Protection has substantially higher unavailability than Route Diversity because both directions are connected to the optical links whose fibers belong to the same cable. In RD protection the working and protection paths are completely independent and so such a system is only unavailable in the case of a simultaneous failure on both paths (working and protection) and in the case of a failure in terminal nodes.

By the analysis it is also obtained that the numbers of wavelengths, as well as nodes build-up of active and passive components, do not have a big impact on WDM ring availability, irrespective of the type of protection. It is also established that the impact of optical link time to repair is considerable, while in RD more considerable is the impact of device repair, for the reason that the RD nodes availability has a dominant impact on total unavailability. A WDM ring availability improvement can be achieved by the reduction of time to repair both the optical links and devices. That requires a better maintenance and team organization.

However, a considerable availability improvement can be achieved by the ring networks construction whose working and protection paths have no common components, and they are fully separated, thus offering high availability services to the users.

List of abbreviations

BOA	Booster Optical Amplifier
DPRing	Dedication Protection Ring
LOA	Link Optical Amplifier
MTTF	Mean Time to Failure
MTTR	Mean Time to Repair
OC	Optical Cable
OCh	Optical Channel
OCh DPRing	Optical Channel Dedication Protection Ring
OCh SPRing	Optical Channel Share Protection Ring
OPA	Optical per Amplifier
RX	Receiver
SPL	Splitter
SPRing	Share Protection Ring
SW	Switch
TX	Transmitter
1FIT	1 failure in hours

List of symbols

a_{BOA}	availability of optical booster amplifier
a_{DMUX}	availability of demultiplexer
A_k	Availability of connection
a_{LOA}	availability of optical line amplifier
a_{MUX}	availability of multiplexer
a_{n_j}	availability of j-the node which belongs to the working path
a_{n_jp}	availability of pass-through node
$a_{np,1+1(1)}$	availability of pass through node for 1+1 protection
a_{n_jt}	availability of terminal node
$a_{nt,1+1(1)}$	availability of termination node for 1+1 protection (one wavelength channel)
$a_{nt,1:1(1)}$	availability of termination node for 1:1 protection (one wavelength channel)
$a_{nt,1+1(n)}$	availability of termination node for 1+1 protection (for n wavelength channel)
a_{OC}	availability for optical cable
a_{OL_i}	availability of i-th link on working path
a_{OLNP_i}	availability of optical link, no protection
$a_{OL,RD}$	availability of optical link, route diversity
$a_{OL,SP}$	availability of optical link, span protection
a_{OPA}	availability of optical pre-amplifier
A_p	Availability of a parallel structure
a_{RX}	availability of receiver
A_s	Availability of a serial structure
a_{SPL1}	availability of splliter
$A_{st}(a)$	availability for the wavelength channel between the *s* and *t* nodes
$A_{st}(P_0)$	availability of working path between *s* and *t* nodes
$A_{st}(P_1)$	availability of protection path between *s* and *t* nodes
a_{SW}	availability of switch
a_{TX1}	availability of transmitter

References

[1] WILLEMS G., ARIJS P., PARYS V. W., DEMEESTER P.: *Capacity vs. Availability Trade-offs in Mesh-Restorable WDM Networks*, Design of Reliable Communication Networks (DRCN), Proc., Budapest, Hungary, Oct. 2001.

[2] SCHUPKE A. D.: *Reliability Models of WDM Self-Healing Rings, Second International Workshop on the Design of Reliable Communication Networks (DRCN),* Munich, Germany, April 9–12, 2000.
[3] JURDANA, I., MIKAC, B.: *An Availability Analysis of Optical Cables,* Proceedings of WAON'98, pp. 153–160,
[4] INKRET, R.: *Shaping of Entirely Optical Transmission Network with Wave Multiplex/masters Degree Thesis,* Faculty of Electrical engineering and computing/IT; University in Zagreb, 1998,
[5] INKRET, R.: *Availability Modelling of Multi-Service All-Optical Transmission Network,* Doctoral Thesis, University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, 2004,
[6] ARCI, D., MAIER, G., PATTAVINA, A., PETECCHI, D., TORNATORE, M.: *Availability Models for Protection Techniques in WDM Networks,* Design of Reliable Communication Networks (DRCN) 2003, Banff, Alberta, Canada, October 19–22, 2003, pp. 158–166,
[7] WILSON, M., R.: *The Quantitative Impact of Survivable Network Architecture on Service Availability*, IEEE Communications Magazine, May 1998, pp. 122–127,
[8] INKRET, R., MIKAC, B.: *Availability Analysis of Different WDM Network Protection Scenarios,* Proceedings of WAON'98, pp. 121–128,
[9] RADOS, I.: *Influence of Common Path on Availability of the Ring Network,* Journal of Telecommunication and Information Technology, National Institute of Telecommunication, Poland, No. 3/2004.
[10] TOMSU, P., SCHMUTZER, C.: *Next Generation Optical Networks,* PrenticeHall PTR Upper Saddle River, NJ, 2002.
[11] GROVER, W,. D.: *Mesh-Based Survivable Networks, Options and Strategies for Optical,* MPLS, SONET and ATM Networks, Prentice Hall PTR 2004.

Andrej Chu *

ANT COLONY OPTIMIZATION METHOD AND SPLIT-DELIVERY VEHICLE ROUTING PROBLEM

This paper deals with a split delivery vehicle routing problem, which is a modification of a vehicle routing problem. It consists in delivery routes optimization in communications network containing initial city of all routes and a given number of places, which is necessary to include in delivery routes, where a customer can be served by more than one vehicle. The objective is to find a set of vehicle routes that serve all the customers and the total distance traveled is minimized. The split delivery vehicle routing problem is NP hard, therefore we present a solution approach by three heuristics, and a metaheuristics called Ant colony optimization (ACO).

Keywords: split delivery vehicle routing problem, integer programming, ant colony optimization, metaheuristics

1. Introduction

Vehicle routing problem (VRP) is a classical problem in operations research. It consists in delivery routes optimization in communications network containing depot of all routes and a given number of cities, which is necessary to include in delivery routes. In VRP a set of customers needs to be served and a fleet of capacitated vehicles is available to do so. The objective is the minimization of costs, which usually means minimizing the total distance traveled. In most VRPs it is assumed that the demand of a customer is given and is less than or equal to the capacity of a vehicle and that each customer has to be served by exactly one vehicle, i.e., there is a single-visit assumption. The condition is that the sum of demands of the cities on the route should be less than or equal to the capacity of vehicle. The vehicle capacity is the limitary factor in this problem.

It is obvious that when a customer's demand exceeds the vehicle capacity it is necessary to visit that customer more than once. Even when all customer demands are less than or equal to the vehicle capacity, it may be beneficial to use more than one vehicle to serve a customer. In the split delivery vehicle routing problem (SDVRP) the single-visit assumption is relaxed and each customer may be served by more than one vehicle.

The SDVRP is, similarly like the traveling salesman problem (which is reduced on this vehicle routing problem in the case of big enough capacity of a vehicle V), an NP hard problem.

Let us set the mathematical model SDVRP, which is based on Miller-Tucker-Zemlin formulation of the traveling salesman problem.

The binary variable x_{ij}^k equals 1, if the edge (i,j) is included in the solution, so it means that the vehicle goes from the city i to the city j, otherwise this variable's value is equal to zero. Each customer has a demand q_i, which can be less than, equal, or greater than the vehicle capacity V. The variable q_i^k represents the quantity delivered to the i-th customer on k-th the route.

The distance between the cities i and j is c_{ij}. Each vehicle has capacity V and has to start and finish its tour at the depot. A customer may be visited more than once.

The objective function (1) represents the sum of all edges distances in the solution, hence the sum of all routes length of the solution. The equation (2) assures that the k-th route leaves the depot just once. The equation (3) assures that only one edge comes out from city i on the k-th route. Similarly, the (4) sets that number of edges coming from the city j on the k-th route is the same as number of edges coming into the city (0 or 1). The equations (3) and (4) are not applied on the city 1, because from and into the city 1 there are as many edges as many routes there are. The inequality (5) defines variable u_i, which represents the demand on the route k from the city 1 to the city i. This condition also has the anti-cycling effect – it prevents creating the sub-cycles in the solution. The condition (7) means that demand on the route k does not exceed the vehicle capacity V. Constraints (8) ensure that the demand q_i of customer i is completely satisfied. Constraints (9) impose that a delivery to a customer i on route k can only take place if route k is selected and the total quantity delivered on a selected route cannot exceed the vehicle capacity V.

The SDVRP model is (1)–(11):

* **Andrej Chu**
Faculty of Informatics and Statistics, University of Economics, Czech Republic, E-mail: andrej.chu@vse.cz

$$\min \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_{ij}^{k} \quad (1)$$

$$\sum_{j=2}^{n} x_{1j}^{k} \le 1,\ k = 1, 2, ..., K \quad (2)$$

$$\sum_{j=2}^{n} x_{ij}^{k} \le 1,\ i = 2, ..., n;\ k = 1, 2, ..., K \quad (3)$$

$$\sum_{i=1}^{n} x_{ij}^{k} = \sum_{i=1}^{n} x_{ji}^{k},\ j = 2, 3, ..., n;\ k = 1, 2, ..., K \quad (4)$$

$$u_i^k + q_j^k - V\left(1 - x_{ij}^k\right) \le u_j^k,\ \ i = 1, 2, ..., n;$$

$$j = 2, 3, ..., n;\ i \ne j;\ k = 1, 2, ..., K \quad (5)$$

$$u_1^k = 0,\ k = 1, 2, ..., K \quad (6)$$

$$q_i^k \le u_i^k \le V,\ \ i = 2, 3, ..., n;\ k = 1, 2, ..., K \quad (7)$$

$$\sum_{k=1}^{K} q_i^k = q_i,\ \ i = 2, 3, ..., n \quad (8)$$

$$0 \le q_i^k \le q_i \sum_{j=1}^{n} x_{ij}^k,\ \ i = 2, 3, ..., n;\ k = 1, 2, ..., K \quad (9)$$

$$x_{ii}^k = 0,\ \ i = 1, 2, ..., n;\ k = 1, 2, ..., K \quad (10)$$

$$x_{ij}^k \in \{0, 1\},\ \ i, j = 1, 2, ..., n;\ k = 1, 2, ..., K \quad (11)$$

2. Heuristic methods

As the split delivery vehicle routing problem is NP hard, for the large problem the solution of the model cannot be obtained for acceptable computer time consumption. It will be useful to propose the heuristic methods for the large scale problem solution. The modified heuristic methods for the traveling salesman problem respective the vehicle routing problem will be used. There are the following heuristics: nearest neighborhood method, insert method and savings method. All the methods are illustrated on a numerical experiment. The differences of the results obtained by those methods are shown on the case study.

It is assumed that the distance matrix C is symmetric, not negative. Let us denote M the set of cities that was not included in any route. In the beginning of the method the set M equals $\{2, 3, ..., n\}$. The heuristic method ends when the set M is empty. The route will be denoted as $tr = (tr(1), tr(2), ..., tr(m))$, where $tr(1) = tr(m) = 1$.

2.1 The nearest neighborhood method

By this method the following steps are executed until the set M is empty:

Step 1. Let us denote the city with the lowest distance c_{1i} as k and let it form the route $tr(1) = 1$, $tr(2) = k$, $tr(3) = 1$, let it set $m = 3$. If $q_k \le V$, delete the city k from the set M and set $\overline{V} = V - q_k$, and continue to step 2. Otherwise the city is not deleted from the set M, and city has the demand $p_k = q_k - V$, the route is closed and the method continues by the step 1 (starting the new route).

Step 2. If the set M is empty, the method stops. Let us find the city k from M which minimizes the distance $c_{tr(m-1).k}$ and for which the route enlargement tr by the city k by inserting it after the city $tr(m-1)$ satisfies the condition $q_k \le \overline{V}$. If $q_k \le \overline{V}$, the route is closed and the method follows by the step 1 (starting the new route), otherwise step 3.

Step 3. Let us enlarge the route tr by the city k by inserting this city after the city $tr(m-1)$, then increment m by 1. If $q_k \le \overline{V}$ delete the city k from the set M, set free capacity $\overline{V} = \overline{V} - q_k$ and continue by step 2, otherwise set $q_k = q_k - \overline{V}$ and continue to step 1.

2.2 The insertion method

The following steps are done until the set M is not empty.

Step 1. Let us denote the city k with the highest distance c_{1i} and put the route $tr(1) = 1$, $tr(2) = k$, $tr(3) = 1$, and $m = 3$. The city k is deleted from the set M. If the set M is empty the method ends.

Step 2. Find city k from the set M by following these conditions:

- it minimizes number $d = c_{tr(i),k} + c_{k,tr(i+1)} - c_{tr(i),tr(i+1)}$ for all $i = 1, 2, ..., m-1$ and $k \in M$,
- the route is enlarged by the city k by inserting this city between the city $tr(i)$ and $tr(i+1)$, where i minimizes the value d. If $q_k \le V$, the city k is deleted from the set M and $\overline{V} = \overline{V} - q_k$, go to the step 2. Otherwise the city is not deleted from the set M, and city has the demand $q_k = q_k - V$, the route is closed and the method continues by the step 1 (starting the new route).
- enlarge the route tr by the city k by inserting this city between the city $tr(i)$ and $tr(i+1)$, increase m by 1. If the set M is empty the method ends, otherwise it follows the step 2.

2.3 The savings methods

The following steps are executed until the set M is empty:

Step 1. If the set M contains only the city k, the route $tr(1) = 1$, $tr(2) = k$, $tr(3) = 1$ is formed, let $m = 3$. If $q_k \le V$ then the k will be deleted from the set M and the method ends. If not, k is not deleted from M, $q_k = q_k - V$, repeat step 1.

Step 2. Let us find the pair of cities from M in the form (k,l), that maximizes savings $s_{ij} = c_{1i} + c_{j1} - c_{ij}$. If this pair (k,l) does not exist, form $(1,k,1)$, $(1,l,1)$, stop.

Step 3. Put the route $tr(1) = 1$, $tr(2) = k$, $tr(3) = i$, $tr(4) = 1$, $m = 4$. If $q_k + q_k - V$, then k and l will be deleted from M and $V' = V - (q_k + q_l)$, step 4.

Step 4. Find i from the set M that maximizes s_{ik}, or j from M that maximizes s_{ij}. If $s_{ik} \geq s_{ij}$ (or j does not exist), the city i is inserted into the route before the city k and if $q_i > V'$ then the city i is deleted from the set M. Let s be increased by one and k equal i. If , the city i is not deleted from M and $q_i = q_i - V'$. The route is closed.

Find j that maximizes s_{ij} when inserting j into the route after the city i. If $s_{ik} \geq s_{ij}$ (or i does not exist), the city j is inserted into route after the city i and if $q_j > V'$, city j is deleted from M. Let m be increased by one. If $q_j > V'$, the city i is not deleted from M and $p_j = q_j - V'$. The route is closed.

If neither i nor j does exist, the route ends and method continues by the step 1.

2.4 ACO metaheuristics

Ant colony optimization (ACO) is a metaheuristics which has been invented in quite recent time. The idea was first introduced by Marco Dorigo 15 years ago. Since then it has been applied on many known problems and significant progress has been made in optimizing this metaheuristics by improving its attributes, finding right parameters values and especially combining it with another metaheuristics, e. g. local search (Dorigo & Schützle, 2004).

The main idea of this metaheuristics lies in simulating the behavior of natural systems, in this case ant colonies. The ants usually search for the food for survival by exploring their environment. The most important part of this process is pheromone deposition, which every ant does when he moves outside the anthill. The more ants move on one path, the more pheromones they deposit and attract other ants to use the path. They also deposit pheromones when moving from the food source to the anthill, so the richer the food source is, the more ants are attracted to it. This is achieved also by depositing more pheromone when returning from richer food sources.

This idea has been transformed into ant colony optimization by transforming the food-search problem into mathematical problems, mostly represented by graphs, and by defining artificial ants. Artificial ants are quite similar to the real ants. The difference is that they usually move on the edges of the graphs and construct feasible solutions of a problem. They are also often enhanced by memory and other features needed for the solution construction. When constructing the solution they follow the pheromone and heuristic information.

The main framework of the algorithm is as follows:

0. Initialization – computing the initial pheromone and heuristic information.
1. Solution construction – m solutions are built by m ants.
2. Pheromone evaporation – pheromone laid on the paths evaporates every iteration with specified rate.
3. Pheromone deposition – ants with w best solutions deposit pheromone on their paths.

Steps 1-3 are repeated until the stop condition is fulfilled. The stop condition can be set as a fixed number of iterations or as the solution improvement no longer occurs after a predefined number of iterations.

Our paper reports about the attempt to apply the ACO metaheuristics on SDVRP, which has been never done before.

Solution construction is done by an artificial ant by sequential selection of the nodes and collection of the goods in the graph until its capacity is full. Thus, it creates the node strings representing the solution. To accomplish this, we need first to define the heuristic information η_{ij}:

$$\eta_{ij} = d_{i1} + d_{1j} - d_{ij} \text{ for } i, j \in M; i \neq j \tag{11}$$

This heuristic information is the same as the one used in the C-W savings method introduced before. We also need the pheromone information τ_{ij}, which is initially set to a value corresponding approximately to the value of pheromone, which will be laid by the ants in one iteration (see further). In every node i, the ant computes the probability of the next move for each node:

$$p_{ij} = \frac{[\tau_{ij}]^\alpha [\eta_{ij}]^\beta}{\sum_{k \in N'} [\tau_{ij}]^\alpha [\eta_{ij}]^\beta} \text{ for } j \in M' \tag{12}$$

where M' is a set of the nodes with unsatisfied demands in this solution so far. Parameters α and β were set to 1 and 2 respectively, as they had given the best experiments results. After the probability computation the ant randomly chooses (with respect to the given probabilities) the next node. This procedure is repeated until all demands are satisfied.

After the solution is constructed, pheromone evaporation takes place. Pheromone evaporation greatly helps with the speed of solution convergence. It is described by the following formula:

$$\tau_{ij} = (1 - \rho).\tau_{ij} \text{ for } i, j \in M; i \neq j \tag{13}$$

where ρ is the pheromone evaporation rate and has been set to 0.01 in the experiments (1% of the pheromone evaporates every iteration).

The last step of an iteration is pheromone deposition. The ant lays pheromone on every edge its solution consists of. The amount of pheromone laid (added to each edge) is computed as

$$\Delta\tau = 1/T \tag{14}$$

where T is the total length of the constructed solution. This allows the ants with better solutions to lay more pheromone than those with worse solutions.

2.4.1 Computational complexity analysis

Computational complexity of the algorithm implementing metaheuristic ant colony optimization method can be expressed from the computational complexities of the partial steps.

When constructing the solution, each ant in certain node has to make decision for the move to the next node. Considering n as a number of nodes, this activity requires computing the probability for each node by (12). This requires n steps and the whole activity is repeated as many times as many vertex walks there are in the graph. The number of these walks depends on an instance's values and may vary from n to the sum of all demands of all customers (that is the case when the vehicle's capacity V is set to 1 which is highly unlikely). In real data instances it is possible to presume that there will be no need to make more than $2n$ walks, which means that the number of split deliveries will not exceed double of the number of nodes (that is when each node is split once in average). With this presumption we can expect that the number of operations will be a quadratic function with n as argument. That means that this step will run with quadratic amortized computational complexity. The asymptotic computational complexity, however, is pseudopolynomial and depends on the value of all demands and vehicle's capacity.

Pheromone laying requires walking over the found solution (or the path rings of the solution) and updating the pheromone values of the edges included in the solution. Using the same presumptions as in the previous paragraph, we can except this operation to require number of steps from n to sum of all demands of all customers. Therefore, we can assume this operation to have a linear amortized computational complexity, but again, the asymptotic complexity is pseudopolynomial.

All in all, the single iteration of one ant takes $O(n^2)$ operations in average. Having m ants, we have the computational complexity of single iteration of the whole colony $O(m.n^2)$.

Pheromone evaporations takes n^2 steps but it is conducted only once per a colony's iteration, so it has negligible effect on the overall complexity and so has the initialization phase of the algorithm.

3. Numerical experiments

Let us have 20 cities from Slovakia and Czech Republic. Bratislava (Slovakia) is a depot of all routes. The minimal distances between the cities or between the cities and the depot are known and the transportation is done in the public road network. The modified heuristic methods were applied on 10 studies with different sets of cities. The demands of all customers are given by the size of the cities and they are the same for each study. In the first experiment the capacity of a vehicle is constant.

For the ACO metaheuristics the number of ants was set to 100, but only the half of them (those with the best solutions) in each iteration were allowed to lay the pheromone. The number of iteration was fixed to 1000. These settings were found as best in the number of experiments.

After application of modified heuristic methods on different studies of SDVRP we found out that we can achieve the best result by the savings method, when considering only the heuristic methods. The nearest neighborhood method is not suitable for SDVRP, see the result table 1.

As for the ACO metaheuristics, we can see that even though it uses the same heuristic information as the savings method does, it generates better solutions in most of the input examples. In addition, there is a lot of room for improving this implementation, while it does not use any other optimization of the generated solutions, which may include local search metaheuristics or improving the strings by removing unnecessary splits at the end of some strings. Hence, the ACO metaheuristics seems to be really a promising method not only for the SDVRP to solve. It has a potential to become very efficient method to solve NP-hard problems and has good possibilities to be extended as a distributed algorithm for parallel systems.

Table 1

	Number of cities	routes	Nearest neighb. method (km)	Savings method (km)	Insertion method (km)	ACO method (km)
1	11	4	2555	1944	2245	1962
2	10	4	3063	2520	2457	2427
3	13	5	3804	3596	3471	3559
4	11	4	2688	2398	2419	2402
5	15	5	3463	3199	3150	3044
6	14	6	4521	3558	3865	3531
7	17	6	4123	3746	3911	3750
8	18	7	5163	4058	4476	4040
9	19	7	4692	4320	4538	4236
10	20	7	4667	4456	4725	4417

In the second experiment the capacity of a vehicle is not constant. We studied dependence of a total traveled distance on the capacity of a vehicle. We applied the three modified heuristic methods on more than 4000 SDVRPs with a different capacity of a vehicle and we marked down their dependence between the total traveled distance and the capacity of vehicle. The representative graph 1 represents application of a modified savings method on the SDVRP. This analysis can be used operatively, when the decision of the used vehicles capacity is to be taken.

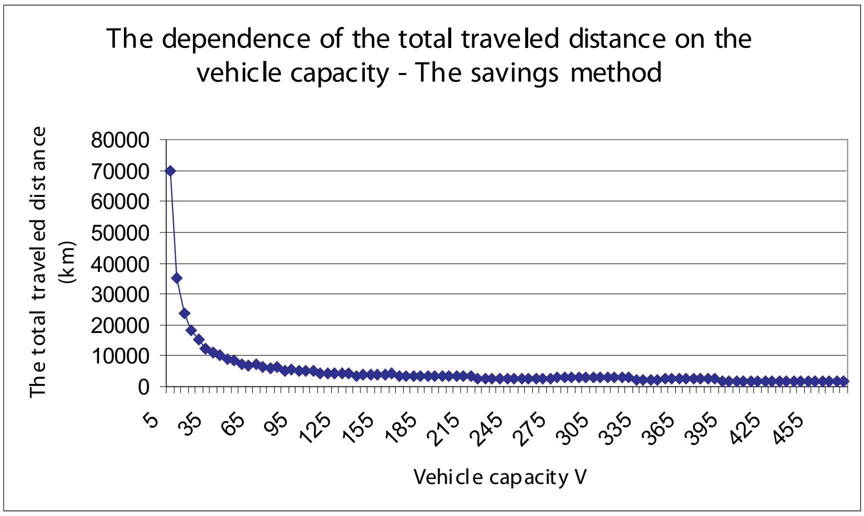


Fig. 1 Results of the second experiment

References

[1] BODIN, L., GOLDEN, B.: *Classification in Vehicle Routing and Scheduling.* Networks, Vol. 11, 1981
[2] DANTZIG, G.B.; RAMSER, J. H.: *The Truck Dispatching Problem.* Management Science 6 (1): 80–91, 1959
[3] DORIGO, M., SCHÜTZLE, T.: *Ant Colony Optimization.* MIT Press, 2004, ISBN 0-262-04219-3
[4] DROR, M., TRUDEAU, P.: *Split Delivery Routing.* Naval Reasearch Logistics, 1990
[5] EVANS J. R., MINIEKA E.: *Optimization Algorithms for Networks and Graphs.* Marcel Dekker, Inc., New York, 1992
[6] FABRY, J.: *Dynamic Round and Cartage Problems (in Czech),* Dissertation thesis. Praha: VSE-FIS, 2006
[7] GUTIN, G., PUNNEN, A. P.: *The traveling salesman problem and its variations.* Kluver 2002. ISBN 1-4020-0664-0
[8] JANACEK, J.: *Mathematical Programming (in Czech),* EDIS ZU, Zilina, 1999, ISBN 80-7100-573-8
[9] JANACEK, J.: *Optimalization in Transport Networks (in Czech)*, EDIS ZU, Zilina, 2003, ISBN 80-8070-031-1
[10] PELIKAN, J.: *Discreet Models in Operating Research (in Czech),* Professional Publishing, 2001, ISBN 80-86419-17-7.

Alan Vasko *

ANALYSIS OF FACTORS INFLUENCING MICROSTRUCTURE AND MECHANICAL PROPERTIES OF AUSTEMPERED DUCTILE IRON

The paper deals with some factors influencing microstructure and mechanical properties of austempered ductile iron (ADI). Final structure and properties of ADI are obtained by exactly controlled process of heat treatment of nodular cast iron. The influence of conditions of isothermal heat treatment on microstructure and mechanical properties of austempered ductile iron, especially different temperature of isothermal transformation of austenite and different holding time at this temperature, is shown in the paper.

1. Introduction

Austempered ductile iron (ADI) is new, perspective construction material with an excellent combination of strength, plasticity and toughness. Special properties of ADI are given by unique structure of matrix which is created by acicular ferrite and retained austenite. Technical literature often describes this matrix as bainite (although it does not contain carbides). The matrix is also referred to as ausferrite microstructure (high carbon austenite + ferrite). The structure of ADI is obtained by exactly controlled process of heat treatment of nodular cast iron (Fig. 1) [1,2].

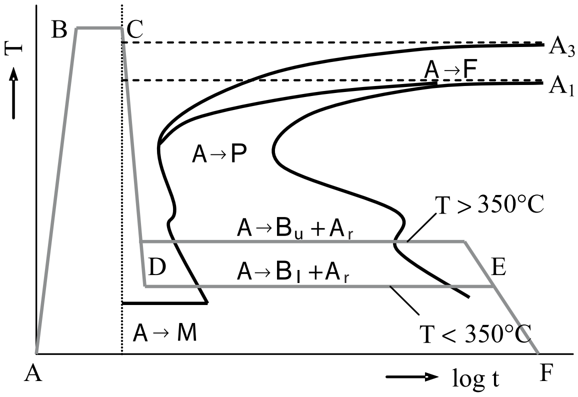


Fig. 1 Process of isothermal heat treatment in the diagram of isothermal transformation of austenite

The isothermal heat treatment consists of the following stages:
- heating to the austenitization temperature (AB);
- holding time at the austenitization temperature (BC);
- quick cooling to the temperature of isothermal transformation of austenite (CD) so that no other transformation of austenite is carried out before reaching the temperature of isothermal transformation;
- holding time at this temperature (DE) until austenite is transformed into bainite;
- cooling to the ambient temperature (EF) which is usually realized slowly in order to prevent formation of stress [3].

The final structure and properties of austempered ductile iron are dependent on the following metallurgical factors:
- *microstructure of initial nodular cast iron* – nodular cast iron has to have perfectly-nodular graphite with the nodularity higher than 80 % and the count of graphitic nodules higher than 150 mm^{-2} and it usually has fine-grained pearlite-ferritic matrix. The structure has to be without free carbide, segregation, porosity and other casting defects;
- *basic chemical composition* – the content of carbon and silicon is dependent on the wallthickness of cast (the content of carbon ranges from 3.3 % for the wallthickness of 100 mm up to 3.7 % for the wallthickness of 5 to 25 mm, the content of silicon is from 1.9 % for the wallthickness of 100 mm up to 2.4 % for the wallthickness of 5 to 25 mm), the content of manganese should be lower than 0.2 % because it causes the occurrence of martensite at grains boundaries;
- *austenitization temperature and holding time at this temperature* – they shall insure homogeneous austenitic matrix. The austenitization temperature is chosen in the range from 820 to 950 °C in dependence on the content of carbon. Higher temperature causes higher hardness, lower temperature causes higher tensile strength and good elongation. Holding time at the austenitization temperature is usually from 1 to 3 hours and it depends on an initial structure of nodular cast iron. High content of ferrite in matrix before the heat treatment lengthens the holding time; pearlitic matrix needs shorter holding time than ferritic;

* **Alan Vasko**
Department of Material Engineering, Faculty of Mechanical Engineering, University of Zilina, Slovakia, E-mail: alan.vasko@fstroj.uniza.sk

- *cooling rate to the temperature of isothermal transformation of austenite* – a cooling rate has to be quick enough to prevent the occurrence of undesirable ferrite and pearlite;
- *temperature of isothermal transformation of austenite and holding time at this temperature* – they both determine the final structure and properties of ADI in a decisive way. The temperature of isothermal transformation is usually within 250 and 450 °C. At higher temperatures (350 to 450 °C) it is possible to obtain ADI with lower strength and hardness but higher elongation and toughness and better fatigue characteristics. At lower temperatures (250 to 350 °C) it is possible to get ADI with higher strength, hardness and abrasion resistance but lower elongation and toughness. The holding time at temperature of isothermal transformation of austenite is within 0.5 and 4 hours when austenite is transformed into upper or lower bainite;
- *alloying elements* – they have strong influence on quality and properties of ADI. Some of them (for example copper, molybdenum or nickel) decrease a critical cooling rate by moving the diagram of isothermal transformation of austenite to the right [4,5].

2. Experimental material and methods

The influence of conditions of isothermal heat treatment on microstructure and mechanical properties of austempered ductile iron was searched on several sets of specimens which were different in temperature of isothermal transformation of austenite and holding time at this temperature.

Experimental material was cast in the electric induction furnace ISTOL. The basic charge was formed by pig iron, steel scrap and additives for the control of chemical composition. FeSiMg7 modifier was used for modification and FeSi75 inoculant was used for inoculation.

Ferrite-pearlitic nodular cast iron was used as basic material for isothermal heat treatment. Chemical composition of the basic material is presented in Tab. 1.

The austenitization temperature was 920 °C and the holding time at this temperature was 30 minutes. The isothermal transformation of austenite was realized in AS 140 salt bath at the temperatures 420, 380, 320 a 250 °C and the holding time at this temperature was from 30 to 300 minutes (by 30 min.) [6,7].

Chemical composition of basic material Table 1

Element	C	Si	Mn	S	P	Cu	Ni
content (weight %)	2.78	4.69	0.49	0.017	0.050	0.92	1.10

The metallographic analysis of specimens of basic material (after casting) and specimens after isothermal heat treatment was made by the light metallographic microscope Neophot 32. The microstructure was evaluated by STN EN ISO 945 (STN 42 0461) and by image analysis (using Lucia software) [8, 9].

The tensile test was made by STN EN 10002-1 by means of the testing equipment ZDM 30 with loading range $F = 0$ to 50 kN. The Rockwell hardness test was made by STN EN ISO 6508-1 by means of the testing equipment LECO LR-3E with a diamond cone forced into specimens under the load $F = 1\,471$ N (150 kp). The measured values of Rockwell hardness were converted to Brinell hardness by STN EN ISO 18265.

The fatigue tests were made by STN 42 0362 at high-frequency sinusoidal cyclic push-pull loading (frequency $f \approx 20$ kHz, load ratio $R = -1$, temperature $T = 20 \pm 5$ °C) using the ultrasonic testing equipment KAUP-ZU and the testing procedures.

The mechanical tests were realized on specimens made from basic material (after casting) and material after isothermal heat treatment.

3. Experimental results

3.1 Metallographic analysis

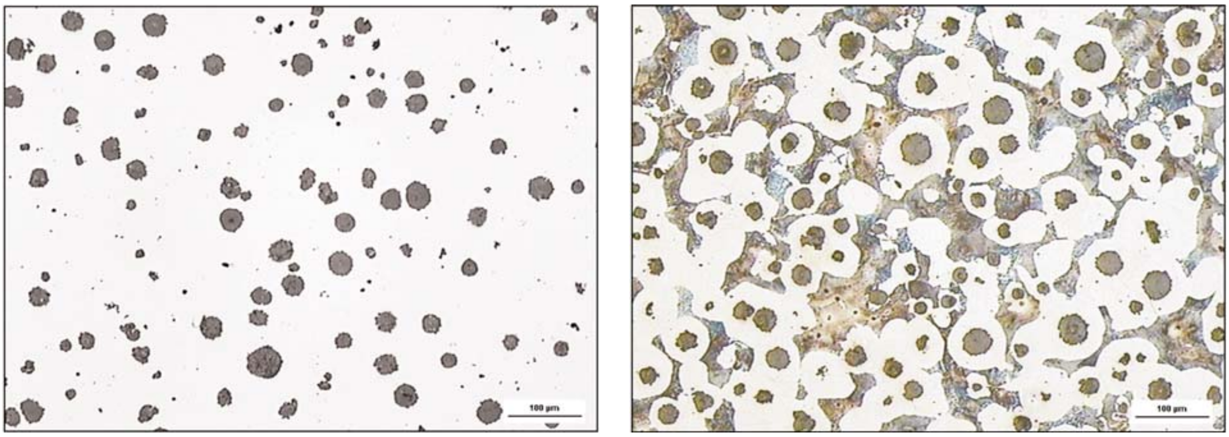
From the microstructural point of view the basic material (after casting) is ferrite-pearlitic nodular cast iron (Fig. 2) with 57% content of ferrite in a matrix, the size of graphite within 15 and 60 μm and count of graphitic nodules 205 mm^{-2}. Graphite occurs only in a perfectly-nodular (80%) and imperfectly-nodular (20%) shape.

After isothermal heat treatment ADI was obtained. The specimens after isothermal heat treatment with the temperature of isothermal transformation of austenite 420 and 380 °C have a matrix created by upper bainite and retained austenite (Fig. 3). The specimens after isothermal heat treatment with the temperature of isothermal transformation of austenite 320 and 250 °C have a matrix created by lower bainite and retained austenite (Fig. 4). The content of retained austenite is slightly decreased with increasing holding time in all sets of specimens. The shape, size and count of graphitic nodules in the specimens after isothermal heat treatment are not changed in comparison with the specimen of basic material (after casting).

3.2 Mechanical properties

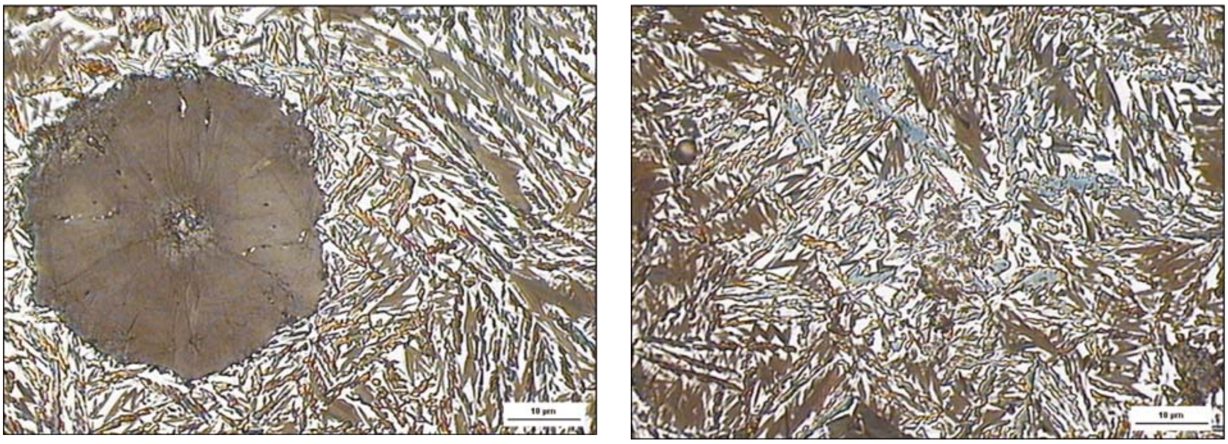
The changes in microstructure of specimens after isothermal heat treatment caused a change in mechanical properties.

Tensile strength of the basic material (ferrite-pearlitic nodular cast iron) is 711 MPa, hardness of the basic material is 250 HB. The isothermal heat treatment induced considerable improvement of tensile strength and hardness in comparison with the basic material. The tensile strength and hardness of the specimens after isothermal heat treatment are increased with a decreasing temperature of isothermal transformation of austenite (Tab. 2). The hardness is decreased with an increasing holding time at the temperature of isothermal transformation of austenite in all sets of specimens (Fig. 5).



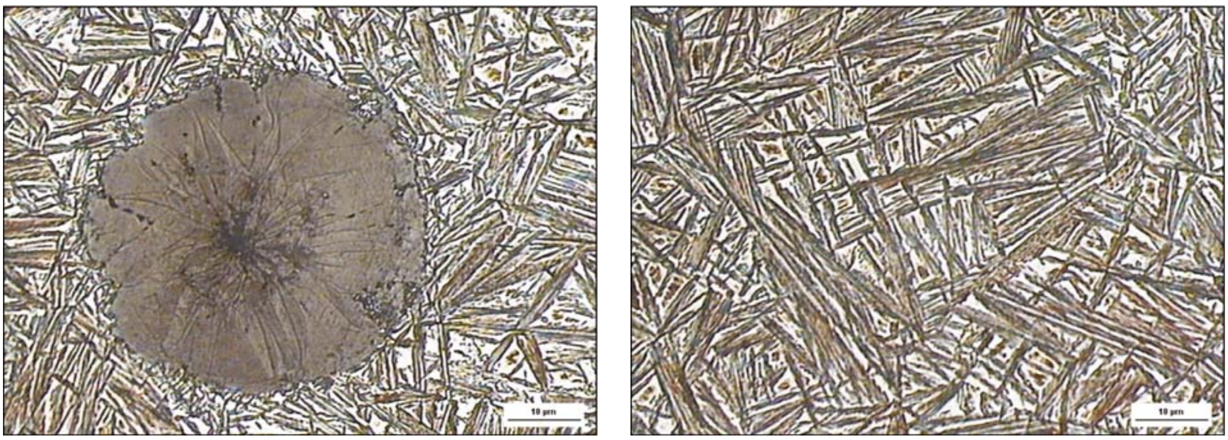
a) non-etched b) etched by 3 % Nital + pikric acid

Fig. 2 Microstructure of basic material (after casting) – ferrite-pearlitic nodular cast iron



a) etched by 3 % Nital b) etched by Klemm I

Fig. 3 Microstructure of specimens after isothermal heat treatment – 380°C/ 60', ADI – matrix created by upper bainite and retained austenite



a) etched by 3 % Nital b) etched by Klemm I

Fig. 4 Microstructure of specimens after isothermal heat treatment – 320°C/ 60', ADI – matrix created by lower bainite and retained austenite

Fatigue strength of the basic material is about 390 MPa. The isothermal heat treatment induced a decrease of fatigue properties in comparison with the basic material. Fatigue endurance of the specimens after isothermal heat treatment is decreased with a decreasing temperature of isothermal transformation of austenite.

Mechanical properties of chosen specimens Table 2

Specimen	Matrix	R_m (MPa)	HB	σ_c (MPa)
basic material	ferrite + pearlite	711	250	390
420 °C/60'	upper bainite + retained austenite	980	474	378
380 °C/60'		1040	495	361
320 °C/60'	lower bainite + retained austenite	1164	492	328
250 °C/60'		1551	530	276

The increase of tensile strength and hardness with a decreasing temperature of isothermal transformation of austenite is caused by a change of matrix from upper bainite to lower bainite. The decrease of fatigue strength with a decreasing temperature of isothermal transformation of austenite is due to the same change of microstructure.

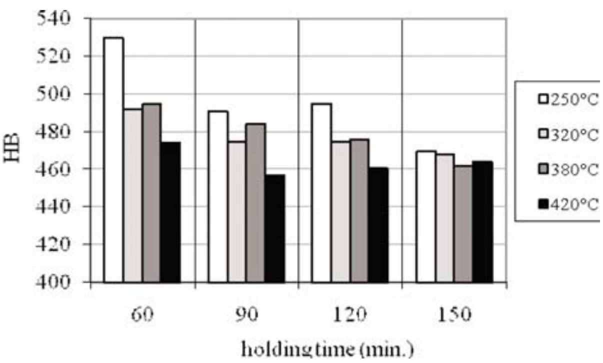


Fig. 5 Influence of temperature of isothermal transformation of austenite and holding time at this temperature on Brinell hardness

The measured values of tensile strength, hardness and fatigue strength correspond to results of similar experiments abroad, for example [10,11].

4. Conclusion

Final microstructure and mechanical properties of casts from ADI are markedly dependent on the temperature of isothermal transformation of austenite and holding time at this temperature. Their influence on the microstructure and mechanical properties of ADI can be summarized in following points:

- the specimens with higher temperature of isothermal transformation of austenite have the matrix created by upper bainite and retained austenite and the specimens with lower temperature of isothermal transformation of austenite have the matrix created by lower bainite and retained austenite;
- the content of retained austenite is decreased with increasing holding time at the temperature of isothermal transformation;
- the shape, size and count of graphitic nodules are not changed in dependence on the temperature of isothermal transformation of austenite and in dependence on the holding time at this temperature;
- the tensile strength and hardness of the specimens are increased with decreasing temperature of isothermal transformation of austenite, but the elongation and toughness is decreased;
- the hardness of the specimens is decreased with increasing holding time at the temperature of isothermal transformation;
- the fatigue strength is decreased with decreasing temperature of isothermal transformation of austenite.

The choice of temperature of isothermal transformation of austenite and holding time at this temperature depends on a required combination of mechanical properties.

Acknowledgement

This work has been supported by the Cultural and Educational Grant Agency of the Ministry of Education of the Slovak Republic, grant No. 3/6078/08 *Creation of a laboratory and textbooks for the teaching of the subject "Properties and use of materials".*

References

[1] DORAZIL, E.: *Austempered Ductile Iron - ADI (in Czech)*, Slevarenstvi, Vol. 41, 4/1993, p. 181-183.
[2] SKOCOVSKY, P., BOKUVKA, O., KONECNA, R., TILLOVA, E.: *Material Science for Specializations of Mechanical Engineering (in Slovak)*, EDIS Zilina, 2006.
[3] SKOCOVSKY, P., PODRABSKY, T.: *Graphitic Cast Irons (in Slovak),* EDIS Zilina, 2005.
[4] GEDEONOVA, Z., JELC, I.: *Metallurgy of Cast Irons (in Slovak),* HF TU Kosice, 2000.
[5] VECHET, S., KOHOUT, J., KLAKURKOVA, L.: *Fatigue Properties of Austempered Ductile Iron in Dependence of Transformation Temperature,* Materials Science (Medziagotyra), Vol. 14, 4/2008, p. 324-327.
[6] VASKO, A., KOPAS, P.: *Influence of Conditions of Isothermal Heat Treatment on Microstructure and Mechanical Properties of Austempered Ductile Iron (ADI)*, In: TRANSCOM 2007, Section 6, Zilina, 2007, p. 255-258.

[7] KOPAS, P., HANZLIKOVA, K., ZAPLETAL, J., VASKO, M.: *Influence of Heat Treatment and Chemical Composition on Mechanical Properties of ADI (in Slovak)*, In: Nekonvencne technologie 2006, Zilina, 2006, p. 1/18-9/18.
[8] SKOCOVSKY, P., VASKO, A.: Q*uantitative Evaluation of Structure of Cast Irons (in Slovak),* EDIS Zilina, 2007.
[9] BELAN, J.: *Quantitative Metallography of Mechanical Worked Ni-base Superalloys (in Slovak),* Acta Mechanica Slovaca, Vol. 12, No. 3-A, 2008, p. 33–38.
[10] SALMAN, S. – FINDIK, F. – TOPUZ, P.: *Effects of Various Austempering Temperatures on Fatigue Properties in Ductile Iron,* Materials and Design, Vol. 28, 7/2007, p. 2210–2214.
[11] KIM, Y. J. – SHIN, H. – PARK, H. – LIM, J. D.: *Investigation into Mechanical Properties of Austempered Ductile Cast Iron (ADI) in Accordance with Austempering Temperature,* Materials Letters, Vol. 62, 3/2008, p. 357–360.

Lubomir Miklanek *

PROPERTIES OF STRAIN-GAUGE LOAD CELLS APPLIED FOR STATOR REACTION MEASUREMENT

This contribution deals with experience with the measurement of the stator reaction of the dynamometer using load cells. Originally, the stator reaction had been measured using a balance mechanism. Within the framework of the author's research, it was necessary to improve the accuracy of torque measurement of the dynamometer. To accomplish this requirement, the old balance mechanism was replaced by a strain-gauge load cell. Two various types of the load cells were applied in order to arrange the measuring range to the expecting range of values of the torque. During the calibration, hysteresis in voltage output of the both applied load cells was observed. The maximal value of the hysteresis was about 0.5 % of the measured range of both of the load cells. Moreover, a certain temporary-plasticity of both of the load cells occurred.

Keywords: Strain-gauge, load cell, hysteresis, dynamometer, torque.

1. Introduction

An older type of eddy current dynamometer [1] was applied in laboratory for torque measurement. Originally, torque of the engine, as a stator reaction, was measured using a balance mechanism. A precise measurement of the torque of the internal combustion engine (ICE) is very important in the author's research especially in case of the low engine load. Moreover, precise measurement of the torque enables the measurement with a constant value of the tested engine torque (Mt).

As the increase of the measurement accuracy had been required, the balance mechanism was replaced by a commercial strain-gauge load cell, see Fig. 1. A measuring amplifier [2] was used both as a source for Wheatstone's bridge excitation voltage and as an output signal conditioning. Load cell output signal was converted by the measuring amplifier to the unified voltage output signal (±10V). This voltage output depends on the load (measured torque) and can be evaluated either by the voltmeter or by the appropriate inputs of the Data AcQuisition System (DAQ).

Note that all the measured data are acquired during the measurement using a DAQ system that was developed and commonly used in the author's laboratory as a standard part of the experimental equipment. More information on the test bed has been presented in [1] and [3].

Moreover, to protect the applied load cell against the excessive values of load, a protection (Force Stop) was designed [4] and installed near the load cell, as illustrated in Fig. 1. The force stop was designed in such a way that the activating of the force stop is adjustable. More about the load cell protection will be described bellow.

The next paragraph will describe the performed calibration measurements of both of the load cells and their detected properties.

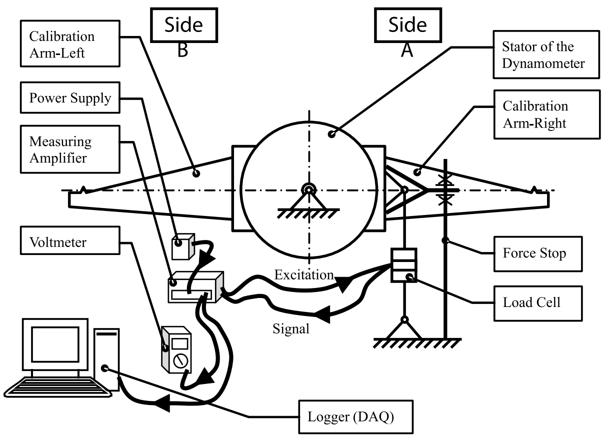


Fig. 1 Scheme of the measurement of the stator reaction using the load cell.

2. Performed Calibration Measurements

Within the framework of the author's research a question of enhanced accuracy torque (Mt) measurement was solved. It con-

* **Lubomir Miklanek**
Josef Bozek Research Center, Faculty of Mechanical Engineering, Czech Technical University in Prague, Czech Republic,
E-mail: lubomir.miklanek@fs.cvut.cz

cerns the test bed with the S 781 NG engine (it is four-cylinder internal combustion, spark-ignition engine made by manufacturer Škoda with max. value of Mt = 90 Nm). To solve this task, the load cell (further marked as *load cell-1*) with these parameters was applied: (Max. loading (Rated load): 490 N; Gauge-factor: 1.8553 mV/V; Input impedance: 415 Ω; Material: anodized aluminum; 6-wire technique with two sensor lines), [5].

The properties of the applied load cell were detected by the calibration measurements. The calibration of the load cell was performed for both sides A and B (both torque direction), see Fig. 1. The measured patterns are presented in Fig. 2 and Fig. 3. Used marking in Figures: *Mtcalibration* is calibration torque; *Mt* is measured torque using the load cell.

There was also a requirement to increase the accuracy of higher torque measurement of other engine. The engine should be a mass-produced four-cylinder 1.9-liter light-duty diesel engine, turbocharged with the air inter-cooler. Its nominal power was of 66 kW at 4000 rpm and peak torque of 202 N.m at 1900 rpm. In order to solve this task, the load cell (further marked as *load cell-2*) with these parameters was applied: (Max. loading (Rated load): 1962 N; Gauge-factor: 2.0011 mV/V; Input impedance: 395 Ω; Material: stainless steel; 6-wire technique with two sensor lines), [6].

The properties of this load cell-2 were detected by the calibration as in case of the load cell1. The calibration was performed for both sides A and B (both torque direction) as well. The measured patterns are presented in Fig. 4 and Fig. 5. The marking used in these Figures is the same as in Figs. 2 and 3.

2.1 Calibration of the Load Cell-1 on Side B

The calibration began on side B (tension). The load cell-1 was under no loading a few hours before the calibration process.

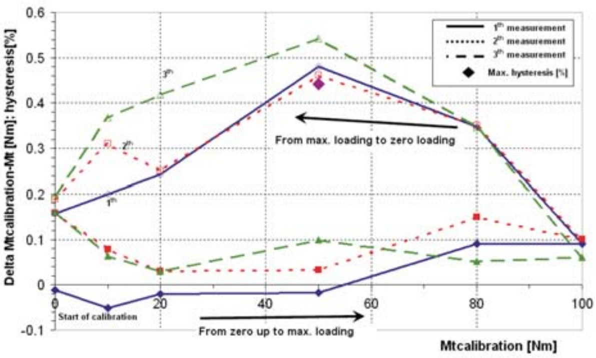


Fig. 2 Measured patterns during the calibration of the applied load cell-1 on side B

The calibration was performed from the zero loading up to the maximal loading and back to the zero loading. This calibration process was repeated three times consecutively, see Fig. 2. As it is seen in Fig. 2, hysteresis occurred between both the increasing and decreasing loading. Its maximal value was about 0.45 % of the measured range during the calibration.

Moreover, certain temporary-plasticity occurred. The term "temporary-plasticity" will be explained in the paragraph below. This observed temporary-plasticity caused a certain offset in the measuring amplifier voltage output especially at the zero loading of the load cell, as it is seen in Fig. 2. Its maximal value was about 0.17 % of the measured range at the zero loading.

2.2 Calibration of the Load Cell-1 on Side A

After the calibration on side B, the calibration on side A (compression) was performed. The measured patterns are presented in Fig. 3. As in the previous case, all three calibration measurements were performed consecutively. The hysteresis with its maximal value of about 0.66 % was occurred.

However, the observed temporary-plasticity from the previous calibration on side B at the zero loading was changed to the direction of load, see Fig. 3. As it can be seen from Fig. 3, the change of temporary-plasticity direction was made just after the first calibration measurement, which is very interesting. It means the maximal value of the temporary-plasticity was doubled (about 2x 0.17 % of the measured range) at the end of the first calibration measurement.

Observed temporary-plasticity was not changed significantly during the second and third calibration measurement and the value was about 0.17 % at the zero loading, see Fig. 3.

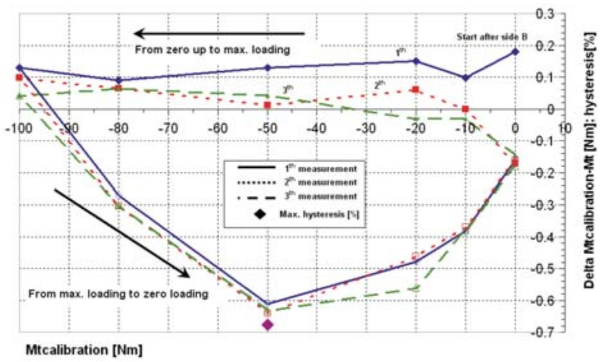


Fig. 3 Measured patterns during the calibration of the applied load cell-1 on side A

2.3 Calibration of the Load Cell-2 on the Side B

As mentioned before, for the higher value of the torque measurement, the load-cell-2 was applied. The calibration began on side B (tension). On the contrary to the calibration of the load cell-1, the load cell-2 was under loading a few minutes before the calibration process.

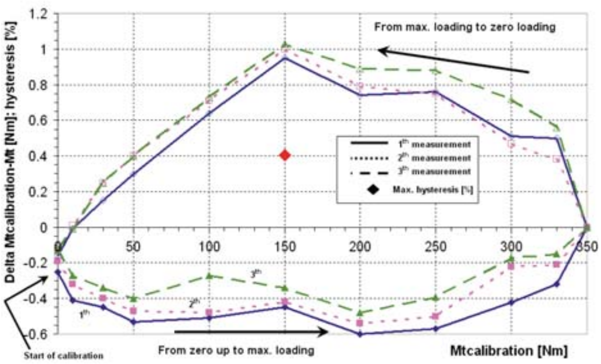


Fig. 4 Measured patterns during the calibration of the applied load cell-2 on side B

As previously, the calibration was performed from the zero loading up to the maximal loading (tension) and back to the zero loading. This calibration process was repeated three times consecutively, see Fig. 4. As it is seen in Fig. 4, hysteresis occurred between both the increasing and decreasing loading. Its maximal value was below 0.5 % (0.4 %) of the measured range during the calibration.

Probably due to the fact that the load cell-2 was loaded before the calibration in the direction of calibration loading, the temporary-plasticity did not occur, see Fig.4.

2.4 Calibration of the Load Cell-2 on Side A

After the calibration on side B, the calibration on side A (compression) was performed. The measured patterns are presented in Fig. 5. As in the previous case on side B, all three calibration measurements were performed consecutively. The hysteresis with its maximal value of about 0.44 % (below 0.5 % again) occurred between both the increasing and decreasing loading.

Moreover, the temporary-plasticity at the zero loading occurred as can be seen in Fig. 5. Observed temporary-plasticity at the zero loading was changed to the direction of load. The change of temporary-plasticity direction was made just after the first calibration measurement, similarly to the load cell-1 in Fig. 3.

Based on the patterns in Fig. 3 (load cell-1) the temporary-plasticity is at the end of the first calibration measurement two times more. It means the value on the Y-axis (Delta Mtcalibration-Mt) at the zero loading of the load cell-2 should be about "–0.5 Nm" without any influence of the plasticity. Upon this presumption, the maximal value of the temporary-plasticity should be about 0.086 % of the measured range. It is about two times less than in case of the load cell-1.

In Figs. 3 and 5 it can be seen that the properties of both load cells were quasi turned over by a change of the direction of the load. After finishing the first calibration, the properties of the load cells change only negligibly, which is interesting.

3. Temporary-plasticity elimination

As mentioned before, certain plasticity was observed after the calibration of both load cells, see Figs. 2, 3 and 5. Nevertheless, it was observed that the mentioned plasticity (on sides A and B) disappeared by itself after a few hours of the zero loading of the load cell under constant temperature conditions. Time record of the partial elimination of plasticity (about 66 % after 4.5 h) is shown in Fig. 6. Overall elimination of plasticity takes about 7 to 8 hours.

So-called Shape Memory Alloys (SMA) have similar properties (return to the original shape after being deformed), [7], [8]. These alloys allow a so-called pseudo-plasticity phenomenon. It means that alloy can be deformed by mechanical stress under normal temperature conditions. After finishing the mechanical stress, the deformation of alloy is fixed. Nevertheless, the alloy returns to the original shape (before the deformation) by heating the alloy on the appropriate temperature (about 400 – 500 °C).

However, the measured properties of the load cell, shown in Fig. 6 (elimination of plasticity under almost constant temperature

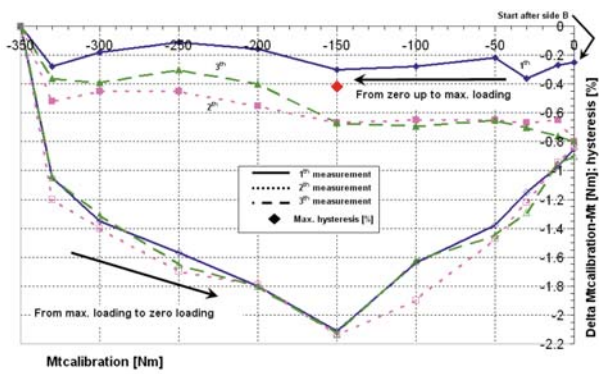


Fig. 5 Measured patterns during the calibration of the applied load cell-2 on side A

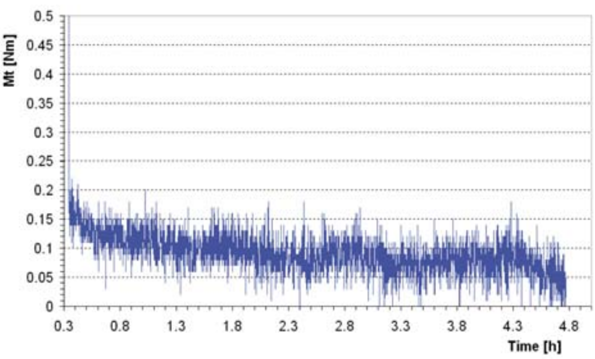


Fig. 6. Time-record of the partial elimination of temporary-plasticity by itself after a few hours under zero loading.

conditions after a few hours) seem not to be properties like pseudo-plasticity.

Based on the above described fact and assumption that the material of the load cell is not SMA, the author gives a name to this phenomenon: Temporary-plasticity. Please note, the author is no expert in metallurgy.

4. Load Cell Protection

Protection of the applied load cell is possible using a purpose-designed force stop, see Fig. 1. The design of the force stop makes an adjustment of its activation possible. When the loading of the tension-compression load cell is less than the preset value, the force stop is not active. However, when the loading of the load cell is larger than the preset value, the force stop is active. Once the force stop is activating, the load is split into both the load cell and the force stop. This splitting causes the change in measured patterns of loading of an applied load cell, see Fig. 7. In this Figure, the measured patterns of the load cell-1 are presented before and after the force stop activating.

It is seen that after the force stop activating, the loading of the load cell increases only slightly compared to the loading increasing without the force stop. Note, the maximum safe static overload of this type of load cell is about 150 % of the rated load. The force stop should be activated at latest on the rated load.

The patterns of load cell-2 before and after the force stop activating are very similar.

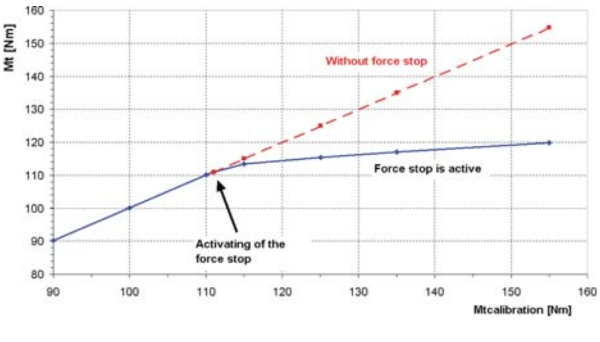


Fig. 7 Measured patterns of loading of the load cell-1 before and after force stop activating

5. Conclusion

Properties of the applied load cell (Max. loading (Rated load): 490 N; Gauge-factor: 1.8553 mV/V; Input impedance: 415 Ω; Material: anodized aluminum; 6-wire technique with two sensor lines) was observed during the calibration in both directions of stress – tensile stress (side B) and compression stress (side A). During the calibration, the hysteresis in voltage output signal was observed (about 0.45 % of the measured range in direction of tensile stress and about 0.66 % in direction of compression stress). Moreover, certain temporary-plasticity of the load cell (about 0.17 % of the measured range) occurred. This temporary-plasticity caused the small offset in output voltage especially in the zero loading.

Moreover, the other load cell (Max. loading (Rated load): 1962 N; Gauge-factor: 2.0011 mV/V; Material: stainless steel; 6-wire technique with two sensor lines) was calibrated for another engine torque measurement in both directions of stress – tensile stress (side B) and compression stress (side A).

Surprisingly, observed hysteresis in voltage output signal of the load cell-2 was about 0.4 % of the measured range as well, although the load cell-2 was made from different material as the load cell-1.

However, the observed temporary-plasticity of the load cell-2 was almost on the half value compared to the load cell-1. It must be noted that the rated load of the load cell-2 is four-times larger than the rated load of the load cell-1.

Based on the measured patterns in Figs. 3 and 5, the properties of both load cells are changed by a change of direction of the load. Moreover, the properties of the load cells are almost constant after the first loading (from the zero loading up to the maximal loading and back to the zero loading) in case of a change in the direction of the load.

In the author's opinion, these observed properties of the applied load cells have no significant effect on the required accuracy of measurements compared to the measured ranges.

Acknowledgement

This contribution has been elaborated with the support of the project 1M6840770002 – Josef Božek Research Center of Engine and Automotive Technology. The author gratefully acknowledges this support.

References

[1] MIKLANEK, L. *Utilization of Automation Test Process in Transient Engine Operating State.* Mecca, 3/2003, vol. I, p. 42–48, ISSN 1214-0821.
[2] HOTTINGER BALDWIN MESSTECHNIK GMBH: *Clip Electronic.* Operating Manual, 2008.
[3] TAKATS, M.: *In-cylinder Pressure Recording and Data Acquisition System.* Mecca, 4/2003, vol. 1, p. 29-35, ISSN 1214-0821, 2003.
[4] MIKLANEK, L.: *Strain-gauge Load Cell. Patent application,* No. 2007-433, Industrial Property Office, Prague, 2007.
[5] TEDEA-HUNTLEIGH ELECTRONICS. CO.LTD: *Load Cell Test Data Sheet-Model 614, 50kg.* 2008.
[6] TEDEA-HUNTLEIGH ELECTRONICS. CO.LTD: *Load Cell Test Data Sheet-Model 616, 200kg.* 2007.
[7] http://en.wikipedia.org/wiki/Shape_memory_alloy
[8] http://cs.wikipedia.org/wiki/Slitiny_s_tvarovou_pam%C4%9Bt%C3%AD.

Katarzyna Nosal *

HOW TO CHANGE THE TRAVELLERS' MOBILITY BEHAVIOURS? - EXAMPLES OF MOBILITY PLANS

The paper presents a mobility management concept as well as gives examples of mobility plans which are fundamental instruments of mobility management. A mobility plan includes activities for the promotion of sustainable mobility modes: bikes, public transport, walking trips and new means of private car usage (carpooling, carsharing systems). It also consists of solutions which realize an idea of sustainable transportation. Mobility plan activities allow to fulfill travellers' mobility needs and shape their mobility behaviours.

Keywords: mobility management, mobility plan, reduction of traffic congestion, travel behaviours.

1. Introduction

Nowadays cities are affected by an increasing number of cars. Car traffic leads to considerable problems related with congestion, parking, accidents as well as environmental pollution. Thus a very important issue is the change of the people's mobility behaviours towards more sustainable transport means: public transport, bikes, walking trips or shared car usage – carpooling and carsharing systems.

We can shape travellers' attitudes and behaviours using the mobility management concept. Mobility management is an approach to the passenger transport, oriented on promotion of the sustainable mobility modes and on management of the demand for car usage [1].

Change of mobility attitudes and behaviours is a very long and not easy process. There is a necessity to take into consideration specific needs and expectations of different users. That is why the mobility management concept consists of a range of services and instruments corresponding to individual clients' needs that can be flexibly adapted to various conditions and expectations of target groups. The core of mobility management is made by "soft" measures like:

- Information and consulting (providing travelers with information about sustainable mobility modes, analyzing the present transport situation, looking for solutions, evaluating alternatives and recommendation of the best solution, e.g. comparison of travel time and costs of different transport modes).
- Transport services and products (this category includes not only the standard services and products as tickets, maps with city transport network but also innovative services and products: loyalty programs, public transport tickets which provide entrance to cultural or sports events etc.).
- Products sale and reservation, e.g. sale of the public transport tickets or rail ticket reservation in the area of the institution.
- Organization and coordination of services and solutions (e.g. organization of the carpooling or carsharing systems in institutions, coordination of mobility services for the handicaps).
- Educational and marketing activities (sustainable mobility trainings, seminars, happenings, marketing campaigns etc.).

"Soft" measures usually enhance the effectiveness of the "hard" measures related with transport infrastructure development (e.g. new tram lines or new bike paths construction). Activities within mobility management "soft" measures do not require large financial investments – in comparison to the "hard" measures.

2. Mobility plan – fundamental instrument of mobility management

The fundamental instrument of the mobility management concept is a mobility plan (also: travel plan, trip reduction plan, green commuter plan, mobility actions, and mobility programme). The mobility plan integrates the above mentioned "soft' measures in one package of the activities. The mobility plan includes strategies for promotion of the idea of sustainable mobility and also consists of solutions and measures which realize the idea of sustainable mobility (in this case "to realize the idea of sustainable mobility" means: to create good conditions for trips made by sustainable mobility means") [2].

General aims of the mobility plans can be defined as:

- Change of travel behaviours of target groups towards pro – ecological transport modes (public transport, bikes, walking trips, new forms of private car usage – carpooling, carsharing systems).

* **Katarzyna Nosal**
Chair of Transport Systems, Faculty of Civil Engineering, Cracow University of Technology, Krakow, Poland, E-mail: nosal@transys.wil.pk.edu.pl

- Improvement in accessibility to the institutions or to the areas.
- Reduction of parking needs.
- Reduction of traffic congestion.

Mobility plans are implemented for these places which generate and attract large traffic flows as e.g.: distinguished city areas (city centers, housing estates, industrial areas), big institutions and companies (municipalities, schools, universities, hospitals, shopping centers etc) [3]. This instrument of mobility management is appropriate to temporary events – trade fairs, concerts, sports matches – the organization of these events has a significant influence on the urban transport system and results often in paralysis of the city. Mobility plan solutions help to reduce the negative effects of temporary events organization.

Implementation of the mobility plans requires appropriate methodology, coordinate educational and informational activities as well as conscious participation of target groups (travellers, employees, employers, students, clients etc.). Very important issues are: monitoring of all activities, ability to drawn a conclusion from realization of activities and evaluation of their results.

The stages of the mobility plan implementation are the following [4]:

1. Definition of the mobility plan aims.
2. Determination of target area and target groups.
3. Analysis of the present situation (surveys about mobility behaviours and preferences, inventories of solutions and connections within PT, bikes, private transport and walking trips, analysis of parking situation etc.).
4. Definition of necessary activities to implementation (conceptual designs, marketing and educational activities etc.).
5. Implementation of proposed activities and solutions.
6. Evaluation of implemented activities.

As it has been mentioned, the mobility plan is a powerful instrument which can be used in the process of changing travellers' mobility behaviours. Two examples of the implemented mobility plans are presented below to prove it.

3. An integrated mobility plan for the Cracow University of Technology (2005 – 2008)

The first mobility plan in Poland was implemented by the Cracow University of Technology (Chair of Transport Systems) in the framework of the EU CIVITAS CARAVEL project [5]. It aimed at the change of the University employees' and students' mobility behaviours towards sustainable transport modes.

The University was affected by an increasing number of cars in the city – more employees and students used private cars to get to the University campuses which are located in different city areas. Alternative modes of transport are considered as less up-market and insecure or simply did not exist.

The first phase of the mobility plan realization was an analysis of situation before the measure implementation (2005, 2006). Mobility plan performers carried out surveys about employees' and students' travel behaviours and preferences as well as inventories concerning trip conditions to and from University campuses realized by public transport, bikechrbtys and cars. Results of the analysis allowed to formulate main activities within realization of the mobility plan.

In order to reach measure objectives, the following activities have been implemented [6] [7]:

3.1 Carpooling system at Cracow University of Technology

Carpooling is a more efficient way of private car usage – when more people use one vehicle during travelling to work, learning or leisure places, thereby economizing on fuel and reducing parking needs. The carpooling system called "Let's drive together" ("Jedzmy razem") has been implemented. This system is directed to employees and students and it helps them to contact each other and find a driver or passenger. Exchange of information is done by a special website "Let's drive together".

3.2 New information website www.info-komunikacja.one.pl with data concerning sustainable mobility means

The new information website "Info.Komunikacja" was established in order to provide information about public transport modes (bus and tram connections, schedules, ticket prices etc.), bikes (bike paths map, location of bike racks, bike rentals in Krakow etc.) and carpooling system (information concerning carpooling website, benefits from travelling in carpooling etc.). Employees and students could also use the travel planner, which enable them to precise plan of trips by public transport or car. After determination of a transport means, origin and destination places, the user gets trip route printed in the map and the travel time.

3.3 Mobility Consultant post at the University

A mobility consultant is a person who gives employees and students advice and information about travelling, and has influence on mobility behaviours. A mobility consultant gives information concerning bus, tram, rail and air transport connections and tariffs, car usage costs, bike paths in a city, carpooling system, development plans of city transport infrastructure etc.

3.4 New bike policy

Before implementation of the mobility plan there was insufficient amount of bike racks in the area of University campuses. As a result of the analysis of the biking situation 23 new bike racks have been newly installed and the amount of bike parking places has grown by 120. Also the information boards with bike racks location were set up in the of University campus area.

3.5 New car parking policy

Within new car parking policy the cost of parking in the area of the University was doubled in 2007 in order to encourage parking users to use public transport and bikes and to reduce the parking places shortage. According to the declarations in survey 2008.5% of employees and 9% of extramural students have changed means of transport.

3.6 Series of seminars and workshops

During various seminars the sustainable mobility means and Caravel project were presented. An excellent idea was also a case study in the form of discussion with students concerning ideas of PT marketing actions. During workshops students also had to create concepts of mobility plans, goods distribution, demand responsive transport (a form of bus transport with flexible routing and scheduling, operating between pick-up and drop-off locations according to passengers needs).

3.7 Marketing actions for promotion of sustainable transport modes

In the mobility plan framework many marketing actions were realized, e.tg. brochures, leaflets, papers, posters and happenings - where the use of bikes, public transport and carpooling system was promoted. Fig. 1. presents one of these happenings – bike happening. It was a very successful action. During this event "The Book of University Bikers" was established as a base of people who travel by bike to the University and in their free time. At the same occasion, students could have their bikes inspected at free bike maintenance points and could get prizes in competitions. To avoid burglary the police marked the bicycles. It is worth mentioning that rector of the University also participated in the Ceremonial Bike Rides.

In 2008, thanks to the mobility plan activities, the share of one–person car trips to the University campuses decreased: – staff: from 45% to 41% and extramural students: from 50% to 30%. It can be explained as a change from car trips (as a driver) to carpooling trips. The share of carpooling trips increased - staff: from 1% to 5%, full-time students: from 0% do 7%, and extramural students: from 1% to 17%. Due to the increase in carpooling trips the need for parking at the University area was reduced. The number of staff travelling to the University by bikes was doubled [4].

4. Integrated mobility actions for the San Martino Hospital in Genoa (2005–2008)

The measure called "Integrated mobility actions for the San Martino Hospital in Genoa" is another successful mobility plan implemented within the EU CIVITAS CARAVEL project.

San Martino Hospital in Genoa is one of the biggest in Italy. More than 4.500 employees work in this hospital that covers more than 30.000 m^2 with an internal road network of about 20 km. It is located along one of the main street connecting the eastern part of the city and the centre, where the traffic flow in the morning peak hour is about 4.000 vehicles. Every day the hospital generates and attracts a large number of employees, patients, visitors and medical students. Moreover, due to the big pressure of vehicles on the area, the problem of lack of parking places is relevant [8].

An objective of the mobility plan for the hospital was to regulate all the public and private traffic generated/attracted by the hospital and by the surrounding areas in a more efficient way.

In order to achieve the objective following activities were implemented:

- Development of The Home Work Trip Plan (HWTP) for the hospital staff with a particular focus on the testing of a carpooling system.
- Establishment of the Infomobility Platform for employees and visitors with data about traffic and mobility in the city. The Infomobility Platform allows users to obtain real time transport information from www.mobilitypoint.it and to use the carpooling system website. Two kiosks with the access to the internet

Fig. 1 Bike happening, April, 2007

Fig. 2 One of the two kiosks with access to the Infomobility Platform

and to the Infomobility Platform were installed in crucial locations inside the hospital.
- Increase of bus connections to the several departments inside the hospital.
- Extension of the public transport system (additional bus lines) inside of the hospital area for people coming early in the morning and late in the evening.
- Use of electric vehicles for goods distribution inside the hospital area.
- Dissemination of the mobility plan and sustainable mobility modes through: training, workshop, announcements on the notice boards inside hospital buildings, in the intranet and special bus (Infomobility Bus) parked outside the hospital days before the workshop. An Infomobility Bus is a kind of mobility point where one can find information about sustainable mobility means as well as place where s/he can be given suggestions and ideas concerning the city transport organization [9].

Fig. 3 Infomobility Bus [9]

Implementation of the mobility plan for the San Martino hospital allowed to obtain the following results (2008) [10]:
- The traffic flows in the street nearby the hospital decreased (in the west direction, in morning peak hours by 31%).
- The share of car trips to the hospital decreased from 42% to 39%.
- The share of walking trips increased from 5.5% do 11.5%.

5. Conclusion

The change of travellers' behaviours is a very difficult process and its effects are usually not noticeable right after the implementation of appropriate instruments. But it is possible to meet traveller's mobility needs and to encourage them to travel in a more sustainable way as it was presented above. Activities realized within the mentioned mobility plans allowed to shape the mobility needs and attitudes of the target groups. These two examples have shown that very important issues are:
- Demonstration of more efficient and integrated modes of travelling.
- Influence on users mobility via development of bike and public transport infrastructure, improvement in services, information and consulting.
- Promotion of the sustainable transport via marketing and educational actions.

References

[1] *Definition and Categorisation of Mobility Management Measures.* Materials of the MAX project: Successful Travel Awareness Campaigns and Mobility Management Strategies, 2007, www.epomm.org/docs/MAX_Defn_catgsn_MM_measures_Final.doc
[2] www.epomm.org – *European Platform on Mobility Management*
[3] www.mobilitymanagement.be – *Toolbox for Mobility Management Measures in Companies*
[4] NOSAL, K.: *Integrated Mobility Management,* In proc. of conference Innowacyjne rozwiazania transportowe – doswiadczenia z realizacji projektu CIVITAS CARAVEL, Krakow, 2009
[5] www.civitas-initiative.org – CIVITAS – Cleaner and Better Transport in Cities
[6] NOSAL, K.: *An Integrated Mobility Plan for the Cracow University of Technology*, Proc. of conference Wspomaganie decyzji w projektowaniu i zarządzaniu transportem, SITK, Poznan, 2007
[7] NOSAL, K.: *An Integrated Mobility Plan for the Cracow University of Technology*, Proc. of: I International Interdisciplinary Technical Conference of Young Scientists, Poznan, 2008
[8] Caravel Inception Report, 11.8: *Integrated mobility actions for the San Martino Hospital*, Genoa 2005
[9] PIREDDU, P.: *Mobility Marketing & Mobility Forum, the experience of Genoa,* Proc. of conference CIVITAS CARAVEL Summer University Burgos, Burgos, 2007
[10] Caravel evaluation report, 11.8: *Integrated mobility actions for the San Martino Hospital,* Genoa, 2008.

Viliam Lendel *

PROBLEMS AND KEY FACTORS OF CRM IMPLEMENTATION

CRM value chain identifies five key steps in the development and implementation of a CRM strategy: customer portfolio analysis, customer intimacy, network development, value proposition development, and manage customer life cycle. These five primary stages of the CRM value chain represent three main sequential phases of CRM strategy: analysis, resource development and implementation. This article deals with CRM value chain and diagnostics of the level of Slovak companies in the CRM area.

Keywords: CRM, CRM Value Chain, implementation, customer.

1. Introduction

In developed countries the change of business processes are distinguished in the orientation to a customer. Enterprises base their actions in the market not on "suspicions" or "experience", but on knowledge that is acquired by analysing customer's data. However knowledge-based activity of an enterprise is possible only when having processed the data on their basis motivated decisions to find, attract and keep customers are taken. This explains why at present it has particularly become fashionable to speak about customer relationship management (CRM) and CRM value chain.

CRM has a lot of definitions. CRM is everything what is related to satisfaction of customer's needs. Dick Lee states that CRM is perceived as "chain reaction", which is caused by new strategic initiatives of communication with a customer when high level of information technologies development and constant customer's need to get better quality of service are achieved [7].

Buttle understands customer relationship management as simply a bridge between marketing and information technologies. CRM decisions impact on marketing, operations, sales, customer service, human resources and information technologies [1].

CRM represents a business strategy designed to reduce costs and increase profitability by solidifying customer loyalty. CRM brings together information from all data sources within an organization to give one, view of each customer in real time. There are three key elements to a successful CRM initiative: people, process, and technology [5].

An interesting view on the term CRM has been brought by Payne. He understands customer relationship management as a strategic approach concerned with creating improved shareholder value through the development of appropriate relationships with key customers and customer segments. In his opinion CRM unites the potential of information technologies and relationship marketing strategies to deliver profitable, long-term relationships. CRM provides enhanced opportunities to use data and information both to understand customers and implement relationship marketing strategies better. This requires a cross-functional integration of people, operations, processes and marketing capabilities that is enabled through information technology and applications [8].

The rapid convergence of CRM and e-commerce has changed the face of enterprise customer relationship management and marketing. Today we can observe a boom of new services on the Internet. They have become very popular in a short time and they are used by millions of people all over the world [6].

2. The CRM Value Chain

The CRM value chain is a proven model which businesses can follow when developing and implementing their CRM strategies (Fig. 1). It has been five years in development and has been piloted in a number of business-to-business and business-to-consumer settings, with both large companies and small, medium enterprises: IT, software, telecoms, financial services, retail, media, manufacturing, and construction. The model is grounded on strong theoretical principles and the practical requirements of business [1].

The ultimate purpose of the CRM value chain process is to ensure that the company builds long-term mutually-beneficial relationships with its strategically – significant customers. Not all customers are strategically significant [1].

* **Viliam Lendel**
Department of Management Theories, Faculty of Management Science and Informatics, University of Zilina, Slovakia,
E-mail: viliam.lendel@fri.uniza.sk

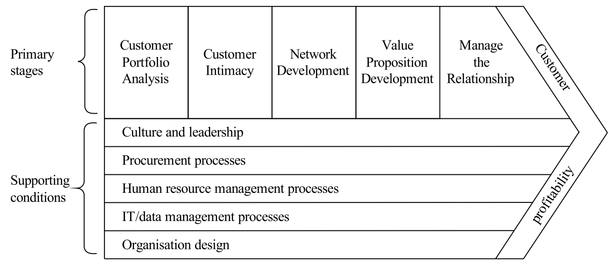


Fig. 1 The CRM value chain [1]

The CRM value chain has the five steps (Tab. 1). There are customer portfolio analysis (CPA), customer intimacy, network development, value proposition development and managing the relationship. At each stage of the value chain there are concepts, tools and processes to help create and implement successful strategy.

The CPA step analyses the customer base to identify customers to target with different value propositions [1]. CPA involves using customer and market data to decide which customers to serve. The companies are looking for groups of customers. They want to segment customer base. They want to determine which groups of customers are most profitable. Which customers spend the most on their goods and services? The result that the companies are looking for is their target customer base. They rate and segment their clients into groups that are most desirable to do business with because they meet their criteria for what a desirable customer are [2].

The second step involves the business in getting to know the selected customers as segments or individuals and building a customer data-base which is accessible to all those, whose decisions or activities impact upon customer attitude and behaviour [1]. This stage looks at the individual customer. Having found the segment(s) that the companies want to pursue, they want to get to know the people in that segment very well. They want to know them better than their competition knows them [2]. Customer intimacy involves getting to understand customers and their requirements.

Step three involves building a strong network of relationships with employees, suppliers, partners and investors who understand the requirements of the chosen customers [1].

Step four involves developing, with the network's compliance, propositions which create value jointly for the customer and company [2]. Network development and value proposition development are focused on building or acquiring resources to create and deliver value to customers [3].

The fifth and final stage is to manage the customer relationship. Managing the customer relationship is about implementing CRM by acquiring and retaining customers, and developing their value.

These steps are iterative and reflexive. They are iterative in the sense that the five-step process is repetitive and continuous. It is not a one-time process that leads to a strategy that is serviceable for ever. For example, in a dynamic environment in which competitors keep improving their value proposition it is important to review periodically which customers to serve, what to serve them and how to deliver the value [3]. Data are used across all 5 stages of the CRM Value Chain.

3. Situation in Slovak enterprises

From March 2007 to February 2009 I carried out the research specialized in diagnostics of the level of Slovak companies in the CRM area. For better understanding the value of the research, I addressed medium and large businesses. 230 top managers of Slovak medium (79 %) and large (21 %) businesses participated in the research.

Search subject has been companies acting in all branches of national industry on area of Žilina self-administrative region. Specifically it is about companies, relegated as medium and large companies on the base of employee count by Statistical Office of the Slovak Republic. *Target group* of the research are *companies*. For these companies it is essential to fulfill below written criteria of assortment to be sorted as target group:

Five stages of the CRM Value Chain [4] Table 1

Customer Portfolio Analysis	Customer Intimacy	Network Development	Value Proposition Development	Manage The Relationship
– Market segmentation – Sales forecasting – ABC – Life-time value – Customer analysis toolkit	– Customer database development – Internal data – Data enhancement – Data warehousing – Data mining – Benchmarking – Privacy – Database technology and software	– Network management – Internal buy-in – External network – Network position – E-commerce – EDI/Extranets/ Portals	– Sources of customer value – Customer experience – Process reengineering – People issues – Technology enablement	– Customer acquisition – Customer retention – Customer development – Organisation design – Metrics

- Acting on area of Žilina self-administrative region,
- Employees count higher than 50.

On the base of these criteria it can be said, that target group consists of medium and large companies, acting on area of Žilina self-administrative region. Object of the research (final respondents) are managers from middle or top management in these companies.

Representative technique has been chosen as sample selection method. To be specific, t*echnique of base selection* has been used. This technique uses *full-range searching*.

Sample size represents 210 of respondents (medium and large companies) by required 95 % interval of reliability and maximal admissible fault 5 % [9]. Actual count – 230 respondents says that sample of asked companies may be considered as representative.

Data gathering was running by two main ways – by **personal questioning** and by **electronic questionnaire**. Telephone contact or e-mail communication was made before personal questioning. Internet environment was also used for the data gathering. Electronic questionnaire was made through PHP and was placed on internet site of the faculty: http://fria.fri.uniza.sk/~lendel/dotaznik.php. By personal questioning it was gathered 121 questionnaires, which represents 53 % of all questionnaires. Electronic questionnaire was filled by 109 managers from medium and large companies.

The purpose of the research was to find and analyse the current level of CRM area on the base of identifying main factors that affect the level of using CRM information system and process of implementation in the company. The current situation of CRM application based on results of the research is presented in Figure 2.

Almost one fourth of respondents (Fig. 2) did not deal with this problem. In the phase of study is 10 percent of respondents, 7 percent is in decision-making phase of CRM application importance for the company. 11 percent of respondents implements CRM in the company's practice. Almost half respondents (49 percent) said that CRM is in full operation in the company.

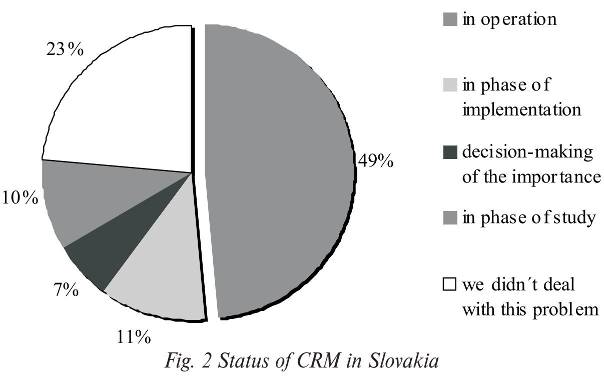


Fig. 2 Status of CRM in Slovakia
Source: Own research

70 percent of respondents said that conception is the most important phase of CRM implementation process (Fig. 3). This phase informs about necessity of exactly defined criteria and conditions. 4 percent said that phase of selection, but on the other hand 17 percent said that phase of implementation is the most important. Main target of this phase is to successfully adapt the software and organizational structure. The phase of implementation finishes with testing and system realisation. 9 percent said about phase of realisation that it's the most important. For successful CRM information system implementation, it is necessary to have skilled employees. Assurance of regular communication is the most important in this phase.CRM implementation and CRM realization phase contains processes, which are essential for proofing and bringing the new CRM system to full operation in company. CRM is implemented into current company structure, where the software proofing is made and testing operation is running. After that the software is presented to company managers and eventual suggestions are being processed.

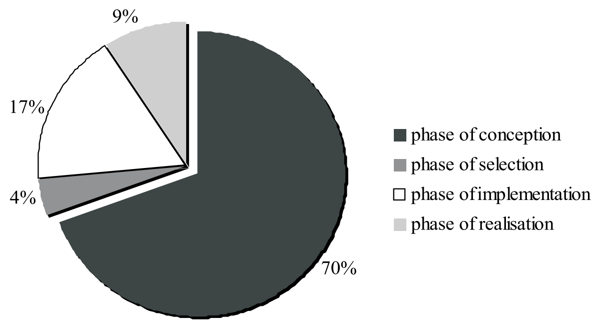


Fig. 3 Most important phase of process of CRM implementation in company's opinion
Source: Own research

The top managers identified the key problem areas of CRM implementation in the company. They selected the following problems:
- Change of customer demands.
- Loss of coordination (reason: very long process of implementation).
- Mismatched training.
- Permanent distrust of new technology.
- Mismatched definition of requisites before implementation.
- Insufficient consulting before installation.
- Insufficient details about processes and information flow.
- Low level of staff motivation.
- Insufficient trust between management and staff.

Top managers identified the following key problem areas of CRM implementation in the company:
- Low level of staff motivation (47.8 %).
- Insufficient details about processes and information flow (31.3 %).
- Mismatched definition of requisites before implementation (34.4 %).
- Permanent distrust of new technology (30.4 %).

As much as 54 % asked companies consider application of CRM into company as continuous process. 12 % of respondents quoted that CRM implementation lasted more than 12 months. 14 % of asked companies say that process lasted from 8 to 12 months, other 11 % claim that duration was from 4 to 8 months and about 10 % of respondents said the whole implementation process lasted less than 4 months.

Besides building the CRM in the company, 24 % of asked companies used outsourcing services. 51 % of respondents implemented CRM using own sources.

4. Key elements influencing the CRM building in company

Part of theoretical outputs of problem solution is also attempt to feature knowledge in the form of *hexagonal stellar model* (Fig. 4), which contains key elements influencing the CRM building in company. It consists of two triangles, each with elements representing main factors that influence CRM building in company:
- People.
- Processes.
- Technologies.
- Strategy.
- Organizational structure.
- Management.

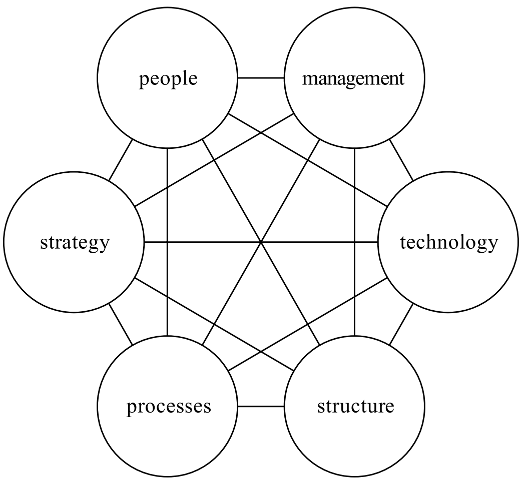


Fig. 4 Hexagonal stellar model
Source: Own elaboration

Customer relationship management is in the sense of hexagonal stellar model understood as a compact system which includes vision, strategy, values of company culture and policy, company processes, their sources, goals and metrics with direct relation to customer.

The first triangle contains the key elements that form the main proportion (base) of the CRM. They are:
- People.
- Processes.
- Technologies.

Progress in the *information technologies* area provided the companies with methods of collection, saving, analyzing and sharing information about customers, which rapidly increased their ability to react on each customer's needs, to attract new and maintain current customers. Other important and indispensable element in the triangle are *processes*. All business processes in company should be oriented on customer. The most important element within the triangle are *people*. Success rate of building long-term and mutually profitable relations with customers will depend on performance and approach of people.

The second triangle contains elements that are important from the view of securing CRM function in company and are essential for building the CRM model. They create environment, in which the CRM can be built, and also resources and ways to achieve it. These elements are as follows:
- Strategy.
- Organizational structure.
- Management.

Support of *top management* and involvement of all employees is essential for successful CRM building in company. Creating the *strategy* is also important, same as in each key step in company run. Strategy with solid rules, but able to modify on the base of specific company conditions. These facts and changes have to reflect also in a new *organizational structure* of the company.

5. Conclusion

At present most organisations recognise evident benefit of CRM and almost every enterprise either uses certain CRM technologies, supporting its business, or evaluates specific benefit of the CRM technology and plans its future realisation.

For CRM to be truly effective, an organization must convince its staff that change is good and that CRM will benefit them. Then it must analyze its business processes to decide which need to be reengineered. Next is to decide what kind of customer information is relevant and how it will be used. Responsible team or top management must choose the right CRM solution. This process, depending upon the size of the company and the breadth of data, can take anywhere from a few weeks to a year or more.

Summarising viewpoints of various authors, it would be possible to highlight that in order to successfully implement CRM, it is necessary to balance and integrate technologies, processes and people. These elements are closely related to enterprise's strategy, processes of technologies, and processes of integration of overlapping functions as well as orientation to customers.

References

[1] BUTTLE, F. *The CRM Value Chain,* Macquarie University, Sydney, April 2000, Available on Internet: www.wtcbrescia.it/upload/0-FButtle_CRMvalchain.pdf
[2] STEVENSON, J. *The CRM Value Chain.* 2007, Available on Internet: http://it.toolbox.com/blogs/customer-centric/the-crm-value-chain-20548
[3] *CRM Value Chain.* Available on Internet, http://www.studymarketing.org/articles/Managing_CRM/CRM_Value_Chain.html
[4] *Customer Relationship Management, Concepts and Tools. Chapter 2,* The Customer Relationship Management Value Chain. Available on internet: www.sm.au.edu/uploadfiles/1181004054_Chapter_2_the_crm_value_chain.ppt
[5] *What is CRM?* Available on internet: http://www.destinationcrm.com/Articles/CRM-News/Daily-News/What-Is-CRM-46033.aspx
[6] KUBINA, M.: *Nontraditional tools of marketing on the Internet:* Journal of Information, Control and Management systems, Faculty of Management Science and Informatics, University of Zilina, 2008. Vol. 6, 2/2008. ISSN 1336-171
[7] LEE, D. *Why climb the CRM mountain?* The Customer Relationship Primer, 2nd ed., 2001, Available on Internet: http://sirnet.metamatrix.se/material/SIRNET_bakgrundsmaterial/CRM_0105.pdf
[8] PAYNE A.: *Handbook of CRM: Achieving Excellence in Customer Management,* Butterworth-Heinemann, 2005. ISBN 978-07506-6437-0, 2005.
[9] http://www.raosoft.com/samplesize.html

COMMUNICATIONS

Marija Kopf – Petr Cenek *

A CONTRIBUTION TO THE TRAFFIC STATE ESTIMATION BY MEANS OF IMAGE PROCESSING

The success of traffic simulations depends largely on the simulation model validity and on accuracy of input data. For input data acquisition, video cameras are used to survey the traffic at junctions or at other places along the road to collect video data. The velocity, distance between the vehicles, acceleration and other relevant parameters can be extracted from the collected data. The estimated values are used for calibration of the simulation model. The data that has to be collected for calibration or validation of the simulation system, differ from country to country and reflects the driver behavior. For that reason, the data collection must be performed and the simulation system calibrated, each time when modeling a different traffic area. The more accurate the obtained parameters, the more accurate the traffic simulation would be. The image processing methods and the expected accuracy of traffic state estimation, such as vehicle position, speed and acceleration, is discussed in the paper and applied to the evaluation of driver behavior necessary for the calibration of the microscopic simulation systems.

Keywords: microscopic systems, transportation simulation, error estimation, image processing

1. Introduction

There are many areas of traffic applications where the estimation of traffic states is necessary. A variety of methods can be applied depending on the environment, i.e. the objects in the scene and many other factors. This paper concentrates on the estimation of traffic states necessary for calibration of the microscopic simulation models. Such models describe vehicle dynamics in detail and can be used to evaluate future impacts of projected new road constructions or changes in traffic control. Furthermore, the traffic states of interest are position of the vehicle, its velocity and acceleration. Also in this subfield, there are many possibilities how to obtain the mentioned traffic state values. The important characteristic of the estimated traffic states is the acquired data precision. There are not many known results on defining the methodology which investigates all possible error sources during the process of traffic state evaluation. The main goal of this paper is to contribute to the methodology for the error estimation. The traffic applications that can profit from this work are the traffic simulation systems which have to be calibrated and this calibration has to be validated.

The second chapter describes the car following model. This model is used by the majority of the traffic simulation systems. The third chapter describes the common framework structure for vision surveillance systems which is used to collect the data from the scene. The error sources are identified and the proposal how to calculate and validate them. The chapters four and five describe the calculation of the position and velocity and the calibration of the system, respectively. The final chapter is the conclusion and the outline of the future work.

2. Car Following Models

Microscopic traffic flow models simulate the behavior of single traffic participants. The dynamic variables of the model for the simulated vehicles are the position and the velocity of the single vehicle. There are many models used in traffic simulation which basically belong to two major groups: car-following models and cellular automation models. For this work the model of interest was the car-following model.

Car following models are also known as time-continuous models, since the transportation flow is modeled as a continuous not a discrete process.

The car following process is depicted in Fig. 1, [1]. The vehicles are considered inertial systems. All vehicles are modeled with the same average characteristics: the average length L_V, the average speed v and the average distance L from the previous vehicle. The parameter L_S is the minimum safety distance between the vehicles. All vehicles are modeled with the same average parameters.

The basic characteristics of the transportation flow are:
v – Speed [km/h],
K – Density [veh/km],
Q – Traffic volume (intensity) [veh/h].

* **Marija Kopf[1], Petr Cenek[2]**
[1]Department of Transportation Networks, Faculty of Management Science and Informatics, University of Zilina, Slovakia,
E-mail: mkopf@mkopf.de

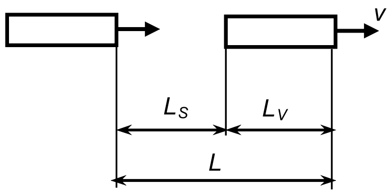


Fig. 1 The Car Following Process

The density stands for an average number of vehicles on a unit road length. It can be estimated as

$$K = \frac{1}{L} \tag{1}$$

The traffic volume is defined as an average number of vehicles that pass a certain point on a road during a unit time interval. The following equation gives the basic characteristics of a steady flow

$$Q = K * v \tag{2}$$

The calibration of the microscopic systems is a challenging task. The simulation system must be accurate in order to be applicable for solving the transportation problems.

After the calibration parameters are chosen, the simulation model must be validated for its accuracy. The formal representation of how the model corresponds to reality is as follows:

$$PR[|reality - simulation\_prediction| < \delta] > \alpha \tag{3}$$

The symbol δ represents the difference between the simulation model and reality, α is the probability that the acceptable difference δ is met. The metrics suitable for measuring the difference between the reality and the predicted results obtained from the simulation.

The inputs for the simulators are of various forms and origin. Geometric inputs, like lane width, junctions etc, are available through documented map sources. Others are obtained by calibrating the system, i.e. using the field data. The field data can be classified in the following categories.

- Parameters that can be directly estimated from field data (vehicle mix, arrival rates, turning percentages)
- Parameters not directly measurable (drivers aggressiveness)
- Tuning parameters that are not real but are required for the model (free flow speed, lost time)

The traffic states can be determined by the statistical analysis of traffic flow data observed from surveillance videos at two freeway locations in the network. The vehicle characteristics and performance data included vehicle length, maximum speed, maximum acceleration and deceleration rates, etc.

3. Vision Surveillance Framework

This chapter explains the framework of such a system and shows the layer where this work aims to search for the error sources.

Nearly every visual surveillance system starts with motion detection. Motion detection aims at segmenting regions corresponding to moving objects from the rest of an image. Subsequent processes such as tracking and behavior recognition are greatly dependent on it. The process of motion detection usually involves environment modeling, motion segmentation, and object classification, which intersect each other during processing.

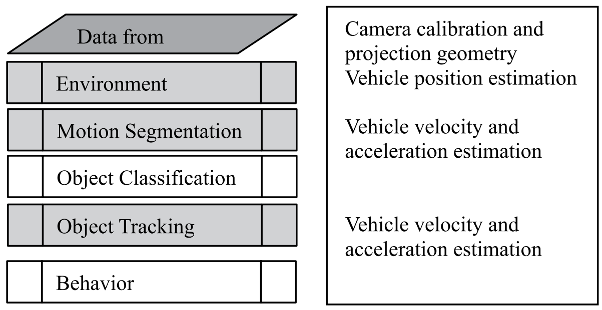


Fig. 2 Vision Surveillance Framework

After the motion detection, surveillance systems generally track moving objects from one frame to another in an image sequence. The tracking algorithms usually have considerable intersection with motion detection during processing. Tracking over time typically involves matching objects in consecutive frames using features such as points, lines or blobs. Tracking methods are divided into four major categories: region-based tracking, active-contour-based tracking, feature based tracking, and model-based tracking.

Behavior understanding involves the analysis and recognition of motion patterns, and the production of high-level description of actions and interactions.

4. Position and Velocity Estimation

The described framework shows the sequence of steps used for the traffic surveillance. For each step there is a possibility to choose between different methods, depending on the application of interest. Each chosen method in each framework step is an error source. For estimating traffic states in this work, the following steps depicted in Fig. 2 will be considered:

- Data from Camera
- Environment Modeling
- Motion Segmentation
- Object Tracking

Possible errors when collecting data from camera are related to the camera calibration and projection geometry. Errors due to environment modeling relate to the precise estimation of the

Fig. 3 Position Estimation: Determining the vanishing point and using landmarks

vehicle position. Possible errors during motion segmentation and object tracking are related to the estimation of velocity and acceleration. The problems to solve are how to estimate the real dimensions from a video picture, depending on the camera position and the road position in the video image. It is possible to use the GPS to measure the position of chosen landmarks. The accuracy of GPS measurements of absolute positions is in the range of 0.01m to 15m. The theoretical analysis is done to predict the precision and errors which can be expected for each of the parameters.

The second error source is the quantization both in time and in space. The traffic states are derived from the sequence of video images. The position of the vehicle is evaluated in each video frame. These values give information about the velocity and acceleration. The video sequence is discrete in time, since it consists of individual frames, and in space, since the image consists of individual pixels. It will be evaluated what filtering can be used and the influence of the sampling frequency on the accuracy of results.

The frequency of frames is an important parameter for the calculation of velocity and acceleration. The influence of the frequency on the velocity and acceleration error will be evaluated. It is obvious, that the higher velocities alse demand the higher frequency of the video frames. It will be considered what minimum frame frequency is necessary for evaluation of vehicle velocity and acceleration depending on the range of these traffic states. The theoretical analysis will be done to determine, depending on the measured velocity, what frequency rates are sufficient so that the increase of frame frequency above this value results in no significant decrease in error. In case, that the vehicles overlap, there is a problem of identifying the seperate vehicles. A vehicle either remains unidentified, or it is not identified in certain frames. This problem can be significantly reduced with the increase of the frame frequency, since the vehicles, if not identified in all frames, would be identified in some of them and by carefully choosing the algorithms for vehicle tracking.

5. Calibration of the Microscopic Systems

The statistical data obtained from the surveillance systems are used as input of the model generator of a microscopic simulation system. How to obtain the vehicle count, velocity and acceleration was described in previous chapters. The velocity of an observed vehicle is measured at the specific points in the scene to enable modeling of the driver behavior.

The figure below shows one measurement of successive velocities at the junction to model the behavior of drivers when turning to the right.

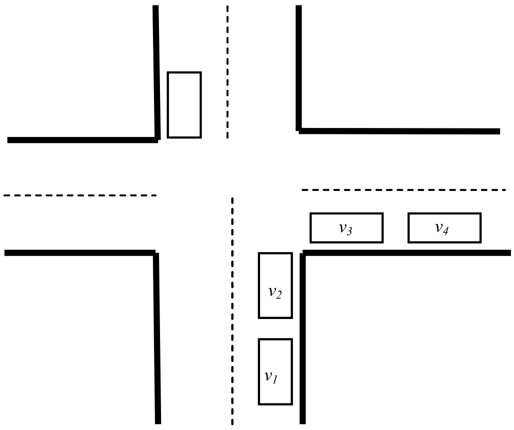


Fig. 4 Velocity estimation at various positions in the scene

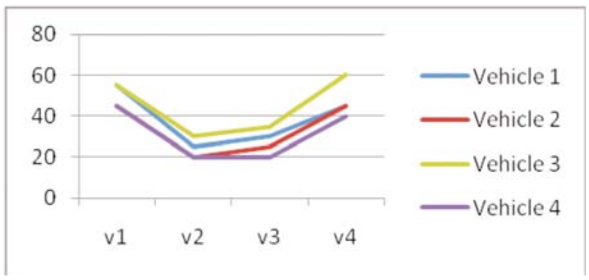


Fig. 5 Velocity profiles of various driver behaviors

6. Conclusion

The further work will concentrate on the implementation steps. The implementation of the image processing methods as a software program should enable practical validation of the proposed method improvements. The video data collected with cameras will be processed. The calculated traffic states are going to be used as the calibration parameters of the simulated model. The simulation model at the University of Zilina will be used. The model shall be calibrated with the gained traffic states describing the driver behavior. The simulation will deliver the simulated traffic sequences which can be used to verify the results. Based on these results, the final evaluation of results and conclusion will be made. The statistical evaluation of the results will be also performed.

References

[1] CENEK, P.: *Models and Optimization in Transports and Logistics.* University of Zilina, 2008

[2] PURSULA, M.: *Simulation of Traffic Systems – An Overview.* Journal of Geographic Information and Decision Analysis, vol.3, 1/1999, pp.1–8

[3] BAYARRI, B.: *Assessing Uncertainties in Traffic Simulation: A Key Component in Model Calibration and Validation.* TRB 83rd Annual Meeting, pp. 32–40, January 2004

[4] SEUL, M.: *Practical Algorithms for Image Analysis.* Cambridge University Press, 2000, ISBN-10: 052188411X

[5] JIMINEZ, T.: *A Road Traffic Simulator: Car-Following and Lane- Changing.* Proceedings of the 14^{th} European Simulation Multiconference on Simulation and Modelling, 2000

[6] KUMAR, P.: *Framework for Real-Time Behavior Interpretation From Traffic Video.* IEEE Transactions on Intelligent Transportation Systems, Vol. 6, 1/2005

[7] GRAMMATIKOPOULOS, L., KARRAS, G., PETSA, E.: *Automatic Estimation of Vehicle Speed from Uncalibrated Video Sequences, International Symposium on Modern Technologies,* Education and Professional Practice in Geodesy and Related Fields, Sofia 2005

[8] ZHANG, Z.X. , LI, M., HUANG, K.Q. , TAN, T. N.: *Robust automated ground plane rectification based on moving vehicles for traffic scene surveillance,* 15^{th} IEEE International Conference on Image Processing, 2008

[9] BOSE, B., GRIMSON, E.: *Ground Plane Rectification by Tracking Moving Objects,* Proc. of Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2003

[10] H. RAKHA, B. HELLINGA, M. VAN AERDE, W. PEREZ: S*ystematic Verification, Validation and Calibration of Traffic Simulation Models,* Annual Meeting, TRB, Washington, DC, 1996

[11] MASOUD, O., ROGERS, S., PAPANIKOLOPOULOS, N.: *Monitoring Weaving Sections*, Report no. CTS 01-06, University of Minnesota, USA, October 2001

[12] GONZALES, R., R. WOODS, R., EDDINS, S.: *Digital Image Processing,* 2^{nd} Edition, Prentice Hall; 2nd edition (January 15, 2002), ISBN-10: 0201180758

[13] SEUL, M. et al.: *Practical Algorithms for Image Analysis,* Cambridge University Press, 2000, ISBN-10: 0521660653.

Peter Tarabek *

GENERAL VECTORIZATION OF LINE OBJECTS IN DRAWN MAPS

Drawn maps consist of multiple object types. The most important are line objects which represent infrastructure. Attributes of these objects are essential for many tasks but in raster format they provide only low level information. Vectorization must be used to obtain vector data. In this paper general vectorization process consisting of five stages is proposed. For these stages short discussion and basic recommendations are given and some proper methods are presented.

Keywords: vectorization, skeletonization, skeleton, thinning, raster-to-vector

1. Introduction

Extraction of graphical information from raster maps is a complex problem. Vector data which represent the necessary information on networks and need less space for their storage are used in many tasks and are essential for GIS and CAD applications. Drawn maps consist mainly of line objects whose length is much larger than the thickness. These objects may represent road infrastructure but the raster data provide only low level information. Vectorization process needs to be used to transform this data to vector form. Conservation of connectivity and shape of the line objects are two most important conditions for vectorization process.

2. Vectorization process

There are three main approaches which can be used in vectorization process: sparse-pixel vectorization (SPV), matching opposite contours (MOC) and skeletonization.

SPV [1, 2] is fastest from the mentioned approaches because it processes only small portion of pixels in image. The main disadvantage of SPV methods is that they not preserve connectivity if the parameters are not set up correctly. MOC methods [2, 3] don't visit all pixels and work either directly on the contours or on a polygonal approximation of it. However hard-to-master thresholds and heuristics need to be used for complex drawings [2].

Skeletonization methods compute medial axis called skeleton. There are many papers about different skeletonization techniques [4, 5]. Some of these techniques can directly produce vectors. The most common are thinning techniques [6, 7, 8] which remove outer pixels layer by layer in iterative process until one pixel thick skeleton is produced. Thinning conserve connectivity and shape of the objects and is used in the proposed vectorization process. On the other hand thinning is sensitive to noise and sometimes produces distortion in junction points.

The process of vectorization proposed in this paper consists of 5 stages (Fig. 1) and it is focused on line objects. Because of huge variety of possible representations of individual objects, there is no specific vectorization process which can deal with all drawn maps. When dealing with line objects this situation is easier. These objects are represented in a very similar manner so if the right methods are applied for each stage this process can be used for a wide range of different maps.

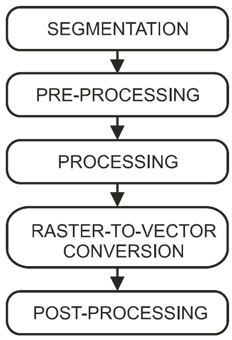


Fig. 1 Stages of the proposed general vectorization process

2.1 Segmentation

Besides line objects, maps usually contain other objects like text, symbols and regions. In this stage line objects are separated from others. The result is a binary image where all line objects

* **Peter Tarabek**
Department of Transportation Networks, Faculty of Management Science and Informatics, University of Zilina, Slovakia,
E-mail: Peter.Tarabek@fri.uniza.sk

should be represented as black pixels and all other objects including the background as white pixels. This reduces amount of data and it also speeds up and simplifies future processing.

The separation of line objects is accomplished by thresholding. A general condition for threshold can be defined as follows:

$$g(x,y) = \begin{cases} 0, \text{ if } f(x,y) < T \\ 1, \text{ } if \text{ } f(x,y) \geq T \end{cases} \quad (1)$$

where $g(x,y)$ is the value of pixel with coordinates x, y in a resulting binary image, $f(x,y)$ is the original value of pixel and T is a threshold value.

Several threshold values are usually used to correctly separate objects in image. There are many local and global threshold techniques which can be used. Although a correct automatic set up of threshold parameters for all kinds of maps is not possible some type of automation is achievable [9, 10, 11].

2.2 Pre-processing

In pre-processing, a binary image is processed to remove imperfections and to amplify desired features of line objects. A binary image should be improved according to weaknesses of techniques used in a processing step to prevent future errors. Imperfections in a binary image are caused by a bad condition of paper maps, process of scanning, usage of segmentation methods, complexity of map and variety of possible representations. In Fig. 2 thinning was used to create a skeleton from an inaccurate input. The acquired skeleton differs from the desired skeleton because thinning is sensitive to imperfections in the original image.

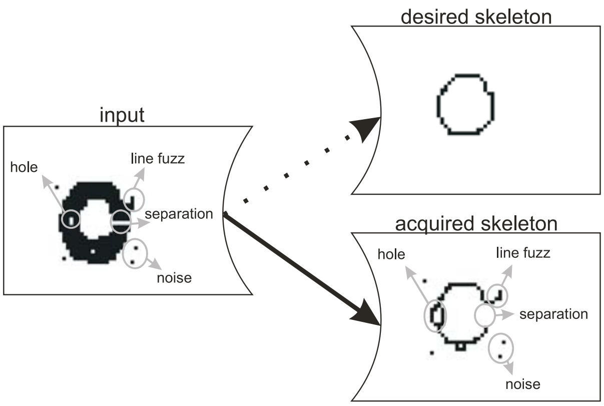


Fig. 2 Desired and acquired skeleton of input image

Accurate pre-processing for thinning should fulfill these tasks:
- remove isolated small objects
- reduce boundary noise (contour intrusions and protrusions)
- fill small holes in objects
- connect disconnected objects

Binary morphology operators opening and closing provide very good results for this kind of tasks. Operations opening and closing can be combined to remove all the mentioned imperfections. Order of operations and number of repetitions are the most important decisions which need to be made in order to obtain accurate results [12].

2.3 Processing

In this step the binary image is prepared for raster-to-vector conversion. Features like thickness of objects, information about their contours and number of components can be extracted and some level of feed back is possible by measuring the quality of result [13] and analyzing extracted features.

Thinning which produces a skeleton is used in the proposed vectorization process. The skeleton is ideal for line objects because it is represented by a set of one pixel thick lines which are a natural representation of line objects such as roads. Thinning should fulfill these requirements:
- Skeleton should be one pixel thick
- Connectivity should be preserved
- Shape and position of the junction points should be preserved
- Skeleton should lie in the middle of a shape (medial axis)
- Skeleton should be immune to noise (especially to boundary noise)
- Excessive erosion should be prevented (length of lines and curves should be preserved)

As noted previously, thinning is sensitive to noise and to other imperfections like holes in objects. These problems can be solved by binary morphology operators recommended in the previous chapter. Another often mentioned disadvantage is a low computational speed which can be improved by using the contour approach to thinning proposed in [14].

2.4 Raster-to-vector conversion

Conversion to vector form for one pixel thick skeleton can be performed in two main steps: nodes and edges recognition. Usually the local approach is used. In this approach 3×3 neighborhood of each pixel is inspected. In this case candidates for nodes can be recognized by having a connectivity number (CN) $>$ 2 (number of black to white transitions in 3×3 neighborhood). Nodes are selected from these candidates based on additional rules. A special case of nodes are end points which can be recognized by having CN = 1. In the next step edges are recognized. Edge pixels have CN = 2 so they can be easily traced. This approach produces excessive nodes and edges and need further processing to yield accurate results.

Another possibility is to form clusters of candidates for nodes. For each candidate the priority number based on the number of 8-neighbors (N8) candidates and the number of 4-neighbors (N4) candidates is computed. For each cluster, the candidate with

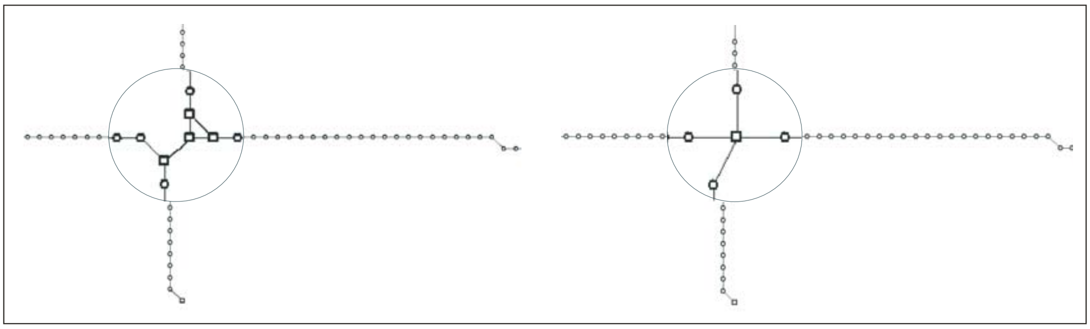


Fig. 3 Local approach (left) and cluster approach (right)

largest priority is selected for a node. This approach is used in the proposed vectorization process. The difference between the local approach and cluster approach is shown in Fig.3.

2.5 Post-processing

After the raster-to-vector conversion is done vector data usually contain a large number of vertices that can be reduced by some kind of polygonal approximation [15, 16]. Also straight lines and arcs can be recognized in this phase [17]. Vector data can be used to perform additional processing including pruning, removing incorrectly separated objects, improving quality of junction points and recognizing attributes such as length, width and color of the edges.

3. Conclusion

The general vectorization process for line objects in drawn maps was discussed in this paper. For each of its steps a short discussion of the problem was given and suitable methods were recommended. The vectorization process is based on thinning which conserves connectivity and shape of line objects. The operator is required only at the beginning of the process to set up parameters. Processing and raster-to-vector conversion are fully automatic and do not require any settings.

The results of the proposed vectorization process are shown in Figs. 4 and 5. The processing time for the map in Fig. 4 with dimensions 985×1114 pixels was less than 2 seconds. For the map with dimensions 600x396 pixels shown in Fig. 5, processing time was less than 0.1 second.

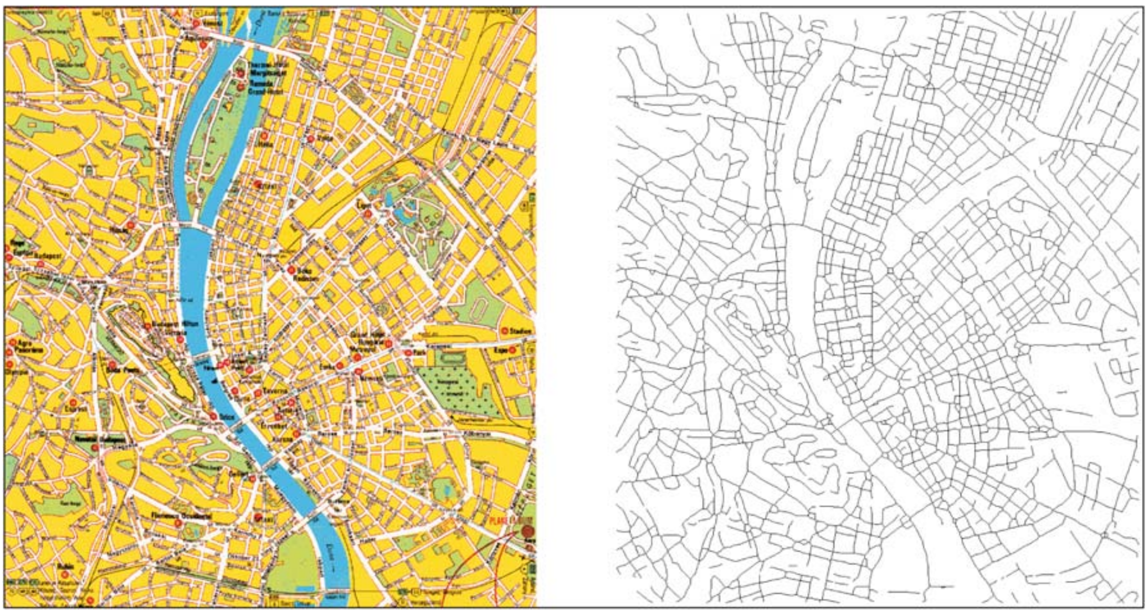


Fig. 4 Result of proposed vectorization process

Fig. 5 Result of proposed vectorization process

References

[1] DORI, D., LIU, W.: *Sparse Pixel Vectorization: An Algorithm and Its Performance Evaluation,* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, 1999, ISSN 0162-8828, pp. 202–215.

[2] TOMBRE, K., AH-SOON, C., DOSCH, P., MASINI, G., TABBONE, S.: *Stable and Robust Vectorization: How to Make the Right Choices,* Graphics Recognition-Recent Advances, 2000, ISBN 978-3-540-41222-9, pp. 3–18.

[3] HORI, O., TANIGAWA, S.: *Raster-to-Vector Conversion by Line Fitting Based on Contours and Skeletons,* Proceedings of the Second International Conference on Document Analysis and Recognition, 1993, pp. 353–358.

[4] ZOU, J. J., HONG, Y.: *Vectorization of Cartoon Drawings,* Selected papers from the Pan-Sydney workshop on Visualisation - Volume 2, 2000, ISSN 1445-1336, pp. 77–78.

[5] SUZUKI, S., UEDA, N.: *Robust Vectorization Using Graph-Based Thinning And Reliability-Based Line Approximation,* Proceedings CVPR '91, 1991, ISBN 0-8186-2148-6, pp. 494–500.

[6] BERNARD, T. M., MANZANERA, A.: *Improved Low Complexity Fully Parallel Thinning Algorithm,* In Proceedings of the 10th international Conference on Image Analysis and Processing, 1999, ISBN 0-7695-0040-4.

[7] ZHANG, T. Y., SUEN, C. Y.: *A fast parallel algorithm for thinning digital patterns,* Commun. ACM, 27(3), 1984, ISSN 0001-0782, pp. 236–239.

[8] LAM, L., LEE, S., SUEN, C. Y.: *Thinning Methodologies-A Comprehensive Survey,* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, No. 9, 1992, ISSN 0162-8828, pp. 869–885.

[9] BOBIN, J., STARCK, J.-L., FADILI, J., MOUDDEN, Y., DONOHO, D.L.: *Morphological Component Analysis: an adaptative thresholding strategy,* IEEE Transactions on Image Processing, Vol. 16, No. 11, 2007, ISSN 1057-7149, pp. 2675–2681.

[10] OTSU, N.: *A threshold selection method from gray-level histogram,* IEEE Transactions on System, Man and Cybernetics, Vol. 9, No. 1, 1979, ISSN 0018-9472, pp. 62–66.

[11] LIAO, P. S., CHEN, T. S., CHUNG, P. C.: *A fast Algorithm for Multilevel Thresholding,* Journal of Information Science and Engineering 17, 2001, pp. 713–727.

[12] TARABEK, P.: *Morphology image pre-processing for thinning algorithms,* Journal of Information, Control and Management Systems, Vol. 5, No.1, 2007, ISSN 1336-1716, pp. 131–138.

[13] TARABEK, P.: *Performance measurements of thinning algorithms,* Journal of Information, Control and Management Systems, Vol. 6, No. 2, 2008, ISSN 1336-1716, pp. 125–132.

[14] TARABEK, P.: *A Contour Approach to Thinning Algorithms,* Communications – Scientific Letters of the University of Zilina, Vol. 11, 3/2009, ISSN 1335-4205, pp. 43–48.

[15] KOLESNIKOV, A., FRANTI, P.: *Data Reduction of Large Vector Graphics,* Research Report, A-2003-2, Department of Computer Science, University of Joensuu, Finland, 2003, ISBN 952-458-311-9.

[16] SARFRAZ, M.: *Optimal Curve Fitting to Digital Data,* International Journal of WSCG, Vol. 11, No. 1, 2003, ISSN 1213-6972.

[17] DOSCH, P., MASINI, G., TOMBRE, K.: *Improving Arc Detection in Graphics Recognition,* Proc. 15th Int'l Conf. Pattern Recognition (ICPR '00), Vol. 2, 2000, ISBN 0-7695-0750-6, pp. 243–246.

COMMUNICATIONS – Scientific Letters of the University of Zilina Writer's Guidelines

1. Submissions for publication must be unpublished and not be a multiple submission.
2. Manuscripts written **in English language** must include **abstract** also written in English. The submission should not exceed **10 pages** with figures and tables (format A4, Times Roman size 12). The **abstract** should not exceed 10 lines.
3. Submissions should be sent: **by e-mail** (as attachment in application MS WORD) to one of the following addresses: *komunikacie@uniza.sk* or *holesa@uniza.sk* or *vrablova@uniza.sk* or *polednak@fsi.uniza.sk* **with a hard copy** (to be assessed by the editorial board) **or on a CD** with a hard copy to the following address: Zilinska univerzita, OVaV, Univerzitná 1, SK-010 26 Zilina, Slovakia.
4. Abbreviations, which are not common, must be used in full when mentioned for the first time.
5. Figures, graphs and diagrams, if not processed by Microsoft WORD, must be sent in electronic form (as GIF, JPG, TIFF, BMP files) or drawn in contrast on white paper, one copy enclosed. Photographs for publication must be either contrastive or on a slide.
6. References are to be marked either in the text or as footnotes numbered respectively. Numbers must be in square brackets. The list of references should follow the paper (according to **ISO 690**).
7. The author's exact **mailing address of the organisation where the author works, full names, e-mail address or fax or telephone number,** must be enclosed.
8. The editorial board will assess the submission in its following session. In the case that the article is accepted for future volumes, the board submits the manuscript to the editors for review and language correction. After reviewing and incorporating the editor's remarks, the final draft (before printing) will be sent to authors for final review and adjustment.
9. The deadlines for submissions are as follows: September 30, December 31, March 31 and June 30.

COMMUNICATIONS

SCIENTIFIC LETTERS OF THE UNIVERSITY OF ZILINA
VOLUME 11

Address of the editorial office:
Zilinská univerzita
Office for Science and Research
(OVaV)
Univerzitna 1
SK 010 26 Zilina
Slovakia
E-mail: komunikacie@nic.uniza.sk,
polednak@fsi.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