JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

Preface

This issue contains many interesting papers. The starting one, *Decision Support under Risk by Optimization of Scenario Importance: Weighted OWA Aggregations*, by Włodzimierz Ogryczak and Tomasz Śliwiński, addresses an important problem of evaluation of decision outcomes under several scenarios to form an overall objective functions; this is a basic problem in decision support under uncertainty. The proposed approach is to use a fuzzy operator defined as the so-called weighted OWA (WOWA) aggregation. The WOWA aggregation, similar to the classical ordered weighted averaging (OWA), uses the preferential weights assigned to the ordered values (i.e., to the worst value, the second worst and so on) rather than to the specific criteria. This makes it possible to model diverse preferences with respect to the risk. Simultaneously, importance weighting of scenarios can be introduced. In this paper, solution procedures are analyzed for optimization problems with the WOWA objective functions related to decisions under risk. Linear programming formulations are introduced for optimization of the WOWA objective representing risk averse preferences. Their computational efficiency is demonstrated.

The next paper, *Path Diversity Protection in Two-Layer Networks*, by Mateusz Dzida, Tomasz Śliwiński, Michał Zagożdżon, Włodzimierz Ogryczak, and Michał Pióro, addresses an optimization problem related to issues of future network architecture, namely, dimensioning links in a resilient two-layer network. A particular version of the problem which assumes that links of the upper layer are supported by unique paths in the lower layer is considered. Two mixed-integer programming formulations of this problem are presented and discussed. Direct resolving of these formulations requires pre-selection of "good" candidate paths in the upper layer of the network. Thus, the paper presents an alternative approach which is based on decomposing the resolution process into two phases, resolved iteratively. The first phase subproblem is related to designing lower layer path flows that provide the capacities for the logical links of the upper layer. The second phase is related to designing the flow patterns in the upper layer with protection assured through diversity of paths. In this phase we take into account the failures of the logical links that result from the failures of the lower layer links (so called *shared risk link groups*).

The third paper, *Hierarchical Multiobjective Routing in MPLS Networks with Two Service Classes – A Meta-Heuristic Solution*, by Rita Girão-Silva, Jośe Craveirinha, and João Clímaco, begins by reviewing a two-level hierarchical multicriteria routing model for MPLS networks with two service classes (QoS and best effort services) and alternative routing, as well as the foundations of a heuristic resolution approach, previously proposed by the authors.

Afterwards a new approach is described, of metaheuristic nature, based on the introduction of simulated annealing and tabu search techniques in the structure of the dedicated heuristic. The application of the developed procedures to a benchmarking case study shows that, in certain initial conditions, this approach provides improvements in the final results, especially in more "difficult" situations detected through sensitivity analysis.

The fourth paper, *Propagation Path Loss Modeling in Container Terminal Environment*, by Ryszard J. Katulski, Jacek Stefański, and Jarosław Sadowski, describes a novel method of path loss modeling for radio communication channels in container port area. Multivariate empirical model is presented, based on multidimensional regression analysis of real path loss measurements from container terminal environment. The measurement instruments used in propagation studies in port area are also described.

The fifth paper, *Model for Balancing Aggregated Communication Bandwidth Resources*, by Piotr Pałka, Kamil Kołtyś, Eugeniusz Toczyłowski, and Izabela Żółtowska, presents a multicommodity bandwidth exchange model BACBR (balancing aggregated communication bandwidth resources) for the purpose of aggregating similar offers in bandwidth market auctions and market clearing. In this model offers submitted to sell (or buy) the same, similar, or equivalent network resources (or demands for end-to-end connections) are aggregated into single commodities. BACBR model is based on an earlier balancing communication bandwidth trade (BCBT) model. It requires much less variables and constraints then original BCBT, however, the outcomes need to be disaggregated. The general model for disaggregation is also given in the paper.

The sixth paper, *Incorporating Customer Preference Information into the Forecasting of Service Sales*, by Piotr Rzepakowski, describes the phenomenon of customers preference change when they are getting more familiar with services or being motivated to change their buying habits. Different sources of motivation induce customers to change their behavior: an advertisement, a leader in a reference group, satisfaction from services usage and other experiences, but usually those reasons are unknown. Nevertheless, people vary in susceptibility to suggestions and innovations, and also in preference structure change dynamics. Historical information about the preference structure gives additional information about uncertainty in forecasting activity. In this work the conjoint analysis method was used to find customer preference structure and to improve a prediction accuracy of telecommunication services usage. The results have shown that prediction accuracy increases about by one percent point, what results in a 20 percent increase after using proposed algorithm modification.

The seventh paper, *Multiobjective Approach to Localization in Wireless Sensor Networks*, by Michał Marks and Ewa Niewiadomska-Szynkiewicz, describes a complex problem of wireless sensor network localization that can be solved using diverse types of methods and algorithms. There are several criteria which are essential when we consider wireless sensor networks. The objective is to determine accurate estimates of nodes location under the constraints for hardware cost, energy consumption and computation capabilities. In this paper the application of stochastic optimization for performing localization of nodes is discussed. Two phase scheme is described that uses a combination of the trilateration method, along with the simulated annealing optimization algorithm. Two variants of proposed technique are discussed, i.e., centralized and distributed. The attention is paid to the convergence of proposed algorithm for different network topologies and trade-off between its efficiency and localization accuracy.

The eighth paper, *Comparative Study of Wireless Sensor Networks Energy-Efficient Topologies and Power Save Protocols*, by Ewa Niewiadomska-Szynkiewicz, Piotr Kwaśniewski, and Izabela Windyga, describes ad hoc networks that are the ultimate technology in wireless communication that allow network nodes to communicate without the need for a fixed infrastructure. The paper addresses issues associated with control of data transmission in wireless sensor networks (WSN) – a popular type of ad hoc networks with stationary nodes. Since the WSN nodes are typically battery equipped, the primary design goal is to optimize the amount of energy used for transmission. The energy conservation techniques and algorithms for computing the optimal transmitting ranges in order to generate a network with desired properties while reducing sensors energy consumption are discussed and compared through simulations. A new clustering based approach is described that utilizes the periodical coordination to reduce the overall energy usage by the network.

The ninth paper, *Parallel and Distributed Simulation of Ad Hoc Networks*, by Andrzej Sikora and Ewa Niewiadomska-Szynkiewicz, describes advances in modeling and simulation as traditional methods used to evaluate wireless network design. This paper addresses issues associated with the application of parallel discrete event simulation to mobile ad hoc networks design and analysis. The basic characteristics and major issues pertaining to ad hoc networks modeling and simulation are introduced. The focus is on wireless transmission and

mobility models. Particular attention is paid to the MobASim system, a Java-based software environment for parallel and distributed simulation of mobile ad hoc networks. The design, performance and possible applications of presented simulation software are described.

The tenth paper, *The Non-Didactic Aspects of e-Learning Quality*, by Ewa Stemposz, Andrzej Jodłowski, and Alina Stasiecka, presents results of research on the quality of e-learning transcending the classical didactic point of view. It illustrates a discussion of criteria, measures or metrics developed on the basis of statistical analysis of data gathered from e-learners that evaluated the quality exploited e-learning applications and systems. The main contribution of the paper is the proposal for the quality metrics with the features concerning e-learning platforms in the technological and human aspects.

The eleventh paper, *Heuristic Analysis of Transport System Efficiency Based on Movement of Mobile Network Users*, by Grzegorz Sabak, describes results of introductory research focused on possibility to use location data available in a mobile network for the analysis of transport system status and efficiency. The details of a system capable of detecting abnormal traffic situation (accidents, heavy congestion) are described. This system (called VASTAR) uses a neural network to learn and store certain characteristic of the analyzed part of a road system. Based on a measured divergence from normal characteristic, a notification about non-typical situation is triggered. The results of a computational experiment using real-world location data and simulation of abnormal situation are provided. The proposed system can be a relatively low cost way to improve competitiveness of a mobile network operator by allowing him to offer new type of informational service. It could also aid municipal authorities by providing support for decisions regarding road traffic control and management and be used by emergency services as a monitoring an alarming tool for detecting abnormal road traffic situations when other means of observation are unavailable.

The twelfth paper, *Analytical Modeling of the WCDMA Interface with Packet Scheduling*, by Maciej Stasiak, Piotr Zwierzykowski, and Janusz Wiewióra, presents the application of a new analytical model of the full-availability group carrying amixture of different multi-rate traffic classes with compression property for modeling the WCDMA radio interface with packet scheduling. The proposed model can be directly used for modeling of the WCDMA interface in the UMTS network servicing different traffic classes. The described model can be applied for a validation of the efficiency of the WCDMA interface measured by the blocking probability and the average carried traffic for particular traffic classes.

The thirteenth paper, *Perspective for Using the Optical Frequency Standards in Realization of the Second*, by Karol Radecki, concerns an alternative definition of the second. The second is currently defined by the microwave transition in cesium atoms. Optical clocks offer the prospects of stabilities and reproducibilities that exceed those of cesium. This paper reviews the progress in frequency standards based on optical transitions, recommended by International Committee for Weights and Measures, as a secondary representation of the second. The operation of these standards is briefly described and factors affecting stability and accuracy of these and some new optical clocks are discussed.

The fourteenth paper, *PHY Abstraction Methods for OFDM and NOFDM Systems*, by Adrian Kliks, Andreas Zalonis, Ioannis Dagres, Andreas Polydoros, and Hanna Bogucka, presents diverse PHY abstraction methods for both orthogonal and non-orthogonal systems are presented, which allow to predict the coded block error rate (BLER) across the subcarriers transmitting this FEC-coded block for any given channel realization. First the efficiency of the selected methods is investigated and proved by the means of computer simulations carried out in orthogonal muticarrier scenario. Presented results are followed by the generalization and theoretical extension of these methods for non-orthogonal systems.

The fifteenth paper, *Technical and Regulatory Issues of Emergency Call Handling*, by Wojciech Michalski, presents selected technical and regulatory aspects of emergency call handling in communication between citizens and authorities in case of distress. Among the most important technical aspects of emergency call handling are recognition and treatment of emergency call by originating network, routing of such call to the appropriate public safety answering point (PSAP), delivering call-related information to the PSAP as well as architecture and organization of PSAPs. From the legal point of view, of importance are the obligations for the Member States and stakeholders involved in E112 project included in the EU directives, actions of European Commission related to providing access to the location information as well as obligations concerning emergency call handling included in Polish national law.

We wish the Readers interesting reading.

Andrzej P. Wierzbicki
Guest Editor

Paper

Decision Support under Risk by Optimization of Scenario Importance Weighted OWA Aggregations

Włodzimierz Ogryczak and Tomasz Śliwiński

Abstract—The problem of evaluation outcomes under several scenarios to form overall objective functions is of considerable importance in decision support under uncertainty. The fuzzy operator defined as the so-called weighted OWA (WOWA) aggregation offers a well-suited approach to this problem. The WOWA aggregation, similar to the classical ordered weighted averaging (OWA), uses the preferential weights assigned to the ordered values (i.e., to the worst value, the second worst and so on) rather than to the specific criteria. This allows one to model various preferences with respect to the risk. Simultaneously, importance weighting of scenarios can be introduced. In this paper we analyze solution procedures for optimization problems with the WOWA objective functions related to decisions under risk. Linear programming formulations are introduced for optimization of the WOWA objective representing risk averse preferences. Their computational efficiency is demonstrated.

Keywords*— *aggregation methods, decisions under risk, OWA, scenarios, WOWA.

1. Introduction

In decision problems under uncertainty, we consider, the decision is based on the maximization of a scalar (real valued) outcome. The final outcome is uncertain and only its realizations under various scenarios are known. Exactly, for each scenario S_i ($i \in I = \{1,2,\ldots,m\}$) the corresponding outcome realization is given as a function of the decision variables $y_i = f_i(\mathbf{x})$. We are interested in larger outcomes under each scenario. Hence, the decision under uncertainty can be considered a multiple criteria optimization problem:

$$\max \{ \ (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})) \ : \ \mathbf{x} \in \mathscr{F} \ \}, \qquad (1)$$

where $\mathbf{x}$ denotes a vector of decision variables to be selected within the feasible set $\mathscr{F} \subset R^q$ and $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ is a vector function that maps the feasible set $\mathscr{F}$ into the criterion space R^m. The feasible set $\mathscr{F}$ is usually defined by some constraints. The elements of the criterion space we refer to as achievement vectors. An achievement vector $\mathbf{y} \in Y$ is attainable if it expresses outcomes of a feasible solution $\mathbf{x} \in \mathscr{F}$ ($\mathbf{y} = \mathbf{f}(\mathbf{x})$). The set of all the attainable achievement vectors is denoted by A, i.e., $A = \{\mathbf{y} = \mathbf{f}(\mathbf{x}) \ : \ \text{for some } \mathbf{x} \in \mathscr{F}\}$.

From the perspective of decisions under uncertainty, model (1) only specifies that we are interested in maximization of all objective functions f_i for $i \in I$. In order to make it operational, one needs to assume some solution concept specifying what it means to maximize multiple objective functions. The solution concepts are defined by aggregation functions $a : R^m \to R$. Thus the multiple criteria problem (1) is replaced with the (scalar) maximization problem:

$$\max \{a(\mathbf{f}(\mathbf{x})) \ : \ \mathbf{x} \in \mathscr{F}\}.$$

The most commonly used aggregation is based on the weighted mean where positive importance weights p_i ($i = 1,\ldots,m$) are allocated to several scenarios:

$$A_{\mathbf{p}}(\mathbf{y}) = \sum_{i=1}^{m} y_i p_i . \qquad (2)$$

The weights are typically normalized to the total 1 ($\sum_{i=1}^{m} p_i = 1$) with possible interpretation as scenarios (subjective) probabilities. The weighted mean enables to define the importance of scenarios but it does not allow one to model the decision maker's preferences regarding the distribution of outcomes. The latter is crucial when aggregating various realizations of the same (uncertain) outcome under several scenarios and one needs to model risk averse preferences [1].

The preference weights can be effectively introduced within the fuzzy optimization methodology with the so-called ordered weighted averaging (OWA) aggregation [2]. In the OWA aggregation the weights are assigned to the ordered values (i.e., to the largest value, the second largest and so on) rather than to the specific criteria. This guarantees a possibility to model various preferences with respect to the risk. Since its introduction, the OWA aggregation has been successfully applied to many fields of decision making [3]–[6]. The weighting of the ordered outcome values causes that the OWA optimization problem is nonlinear even for linear programming (LP) formulation of the original constraints and criteria. Yager [7] has shown that the OWA optimization can be converted into a mixed integer programming problem. We have shown [8], [9] that the OWA optimization with monotonic weights can be formed as a standard linear program of higher dimension.

The OWA operator allows one to model various aggregation functions from the maximum through the arithmetic mean to the minimum. Thus, it enables modeling of various preferences from the optimistic to the pessimistic one. On the other hand, the OWA does not allow one to allocate any importance weights to specific scenarios. Actually, the weighted mean (2) cannot be expressed in terms of

the OWA aggregations. Several attempts have been made to incorporate importance weighting into the OWA operator [10], [11]. Finally, Torra [12] has incorporated importance weighting into the OWA operator within the weighted OWA (WOWA) aggregation introduced as a particular case of Choquet integral using a distorted probability as the measure. The WOWA averaging is defined by two weighting vectors: the preferential weights $\mathbf{w}$ and the importance weights $\mathbf{p}$. It covers both the weighted means (defined with $\mathbf{p}$) and the OWA averages (defined with $\mathbf{w}$) as special cases. Actually, the WOWA average becomes the weighted mean in the case of equal all the preference weights and it is reduced to the standard OWA average for equal all the importance weights. Since its introduction, the WOWA operator has been successfully applied to many fields of decision making [13], [14] including metadata aggregation problems [15], [16].

In this paper we analyze solution procedures for optimization problems with the WOWA objective functions modeling decisions under risk. A linear programming formulations are introduced for optimization of the WOWA objective with increasing preferential weights thus representing risk averse preferences. The paper is organized as follows. In Section 2 we introduce formally the WOWA operator and derive some alternative computational formula based on direct application of the preferential weights to the conditional means according to the importance weights. Further, in Section 3, we analyze the orness/andness properties of the WOWA operator with monotonic preferential weights and the corresponding risk profiles. In Section 4 we introduce the LP formulations for maximization of the WOWA aggregation with increasing weights. Finally, in Section 5 we demonstrate computational efficiency of the introduced models.

2. The WOWA Aggregation

Let $\mathbf{w} = (w_1, \ldots, w_m)$ be a weighting vector of dimension m such that $w_i \geq 0$ for $i = 1, \ldots, m$ and $\sum_{i=1}^{m} w_i = 1$. The corresponding OWA aggregation of outcomes $\mathbf{y} = (y_1, \ldots, y_m)$ can be mathematically formalized as follows [2]. First, we introduce the ordering map $\Theta : R^m \to R^m$ such that $\Theta(\mathbf{y}) = (\theta_1(\mathbf{y}), \theta_2(\mathbf{y}), \ldots, \theta_m(\mathbf{y}))$, where $\theta_1(\mathbf{y}) \geq \theta_2(\mathbf{y}) \geq \cdots \geq \theta_m(\mathbf{y})$ and there exists a permutation τ of set I such that $\theta_i(\mathbf{y}) = y_{\tau(i)}$ for $i = 1, \ldots, m$. Further, we apply the weighted sum aggregation to ordered achievement vectors $\Theta(\mathbf{y})$, i.e., the OWA aggregation is defined as follows:

$$A_{\mathbf{w}}(\mathbf{y}) = \sum_{i=1}^{m} w_i \theta_i(\mathbf{y}), \tag{3}$$

where $w_i \geq 0$ for $i = 1, \ldots, m$ are normalized weights ($\sum_{i=1}^{m} w_i = 1$). The OWA aggregation (3) allows one to model various aggregation functions from the maximum ($w_1 = 1$, $w_i = 0$ for $i = 2, \ldots, m$) through the arithmetic mean ($w_i = 1/m$ for $i = 1, \ldots, m$) to the minimum ($w_m = 1$, $w_i = 0$ for $i = 1, \ldots, m-1$).

Now, let again $\mathbf{w} = (w_1, \ldots, w_m)$ be an m-dimensional vector of preferential weights $w_i \geq 0$ for $i = 1, \ldots, m$ and $\sum_{i=1}^{m} w_i = 1$. Additionally, let $\mathbf{p} = (p_1, \ldots, p_m)$ be an m-dimensional vector of importance weights such that $p_i \geq 0$ for $i = 1, \ldots, m$ and $\sum_{i=1}^{m} p_i = 1$. The corresponding weighted OWA aggregation of vector $\mathbf{y} = (y_1, \ldots, y_m)$ is defined [12] as follows:

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{y}) = \sum_{i=1}^{m} \omega_i \theta_i(\mathbf{y}) \tag{4}$$

with the weights ω_i defined as

$$\omega_i = w^*\Big(\sum_{k \leq i} p_{\tau(k)}\Big) - w^*\Big(\sum_{k<i} p_{\tau(k)}\Big), \tag{5}$$

where w^* is an increasing function interpolating points $(\frac{i}{m}, \sum_{k \leq i} w_k)$ together with the point (0.0) and τ representing the ordering permutation for $\mathbf{y}$ (i.e., $y_{\tau(i)} = \theta_i(\mathbf{y})$). Moreover, function w^* is required to be a straight line when the point can be interpolated in this way. For our purpose of decision support under risk we will focus on the linear interpolation thus leading to the piecewise function w^*.

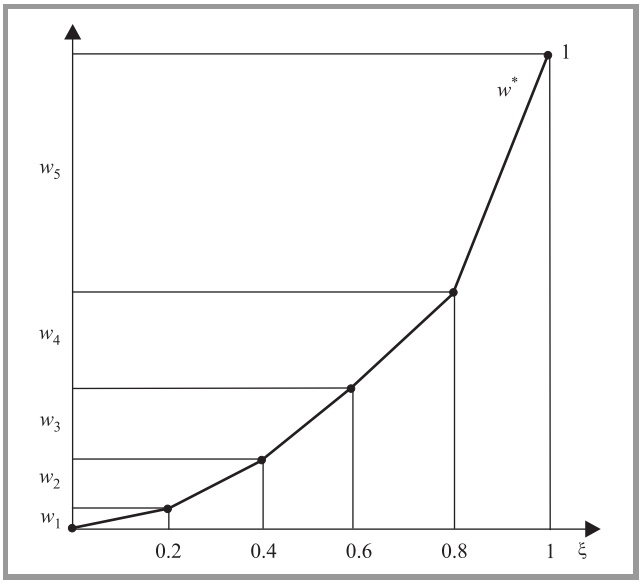


Fig. 1. Function w^* for $\mathbf{w} = (0.05, 0.1, 0.15, 0.2, 0.5)$.

To illustrate the WOWA average let us consider two outcome vectors $\mathbf{y}' = (3,1,2,4,5)$ and $\mathbf{y}'' = (1,1,2,6,4)$, where individual outcomes correspond to five scenarios. While introducing preferential weights $\mathbf{w} = (0.05, 0.1, 0.15, 0.2, 0.5)$ one may calculate the OWA averages: $A_{\mathbf{w}}(\mathbf{y}') = 0.05 \cdot 5 + 0.1 \cdot 4 + 0.15 \cdot 3 + 0.2 \cdot 2 + 0.5 \cdot 1 = 2$ and $A_{\mathbf{w}}(\mathbf{y}'') = 0.05 \cdot 6 + 0.1 \cdot 4 + 0.15 \cdot 2 + 0.2 \cdot 1 + 0.5 \cdot 1 = 1.7$. Further, let us introduce importance weights $\mathbf{p} = (0.1, 0.1, 0.2, 0.5, 0.1)$ which means that results under the third scenario are 2 times more important then those under scenario 1, 2 or 5, while the results under scenario 4 are even 5 times more important. To take into account the importance weights in

the WOWA aggregation (4) we introduce the following piecewise linear function (cf. Fig. 1):

$$w^*(\xi) = \begin{cases} 0.05\xi/0.2, & 0 \le \xi \le 0.2 \\ 0.05 + 0.10(\xi - 0.2)/0.2, & 0.2 < \xi \le 0.4 \\ 0.15 + 0.15(\xi - 0.4)/0.2, & 0.4 < \xi \le 0.6 \\ 0.3 + 0.2(\xi - 0.6)/0.2, & 0.6 < \xi \le 0.8 \\ 0.5 + 0.5(\xi - 0.8)/0.2, & 0.8 < \xi \le 1.0 \end{cases}$$

and calculate weights ω_i according to formula (4) as w^* increments corresponding to importance weights of the ordered outcomes, as illustrated in Fig. 2. In particular, one get $\omega_1 = w^*(p_5) = 0.025$ and $\omega_2 = w^*(p_5 + p_4) - w^*(p_5) = 0.275$ for vector $\mathbf{y}'$ while $\omega_1 = w^*(p_4) = 0.225$ and $\omega_2 = w^*(p_4 + p_5) - w^*(p_4) = 0.075$ for vector $\mathbf{y}''$. Finally,

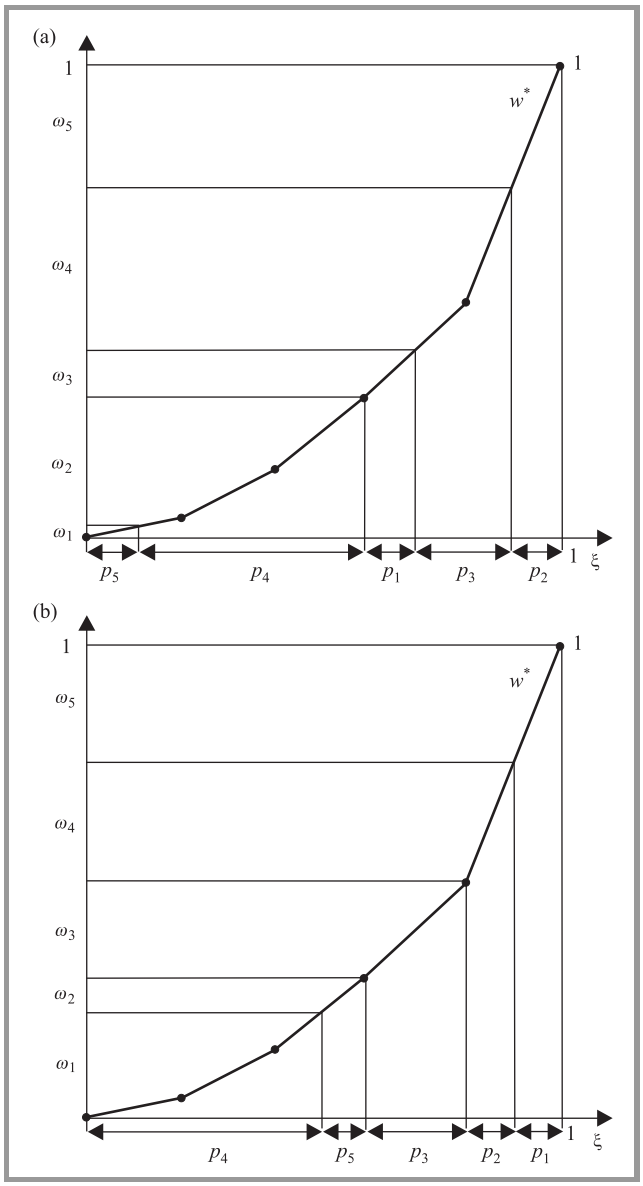


Fig. 2. Definition of weights ω_i for WOWA formula (4) for $\mathbf{w} = (0.05, 0.1, 0.15, 0.2, 0.5)$ and $\mathbf{p} = (0.1, 0.1, 0.2, 0.5, 0.1)$: (a) vector $\mathbf{y}' = (3,1,2,4,5)$; (b) vector $\mathbf{y}'' = (1,1,2,6,4)$.

$A_{\mathbf{w},\mathbf{p}}(\mathbf{y}') = 0.025 \cdot 5 + 0.275 \cdot 4 + 0.1 \cdot 3 + 0.35 \cdot 2 + 0.25 \cdot 1 = 2.475$ and $A_{\mathbf{w},\mathbf{p}}(\mathbf{y}'') = 0.225 \cdot 6 + 0.075 \cdot 4 + 0.2 \cdot 2 + 0.25 \cdot 1 + 0.25 \cdot 1 = 2.55$.

Note that one may alternatively compute the WOWA values by using the importance weights to replicate corresponding scenarios and calculate then OWA aggregations. In the case of our importance weights $\mathbf{p}$ we need to consider five copies of scenario 4 and two copies of scenario 3 thus generating corresponding vectors $\tilde{\mathbf{y}}' = (3,1,2,2,4,4,4,4,4,5)$ and $\tilde{\mathbf{y}}'' = (1,1,2,2,6,6,6,6,6,4)$ of ten equally important outcomes. Original five preferential weights must be then applied respectively to the average of the two largest outcomes, the average of the next two largest outcomes, etc. Indeed, we get $A_{\mathbf{w},\mathbf{p}}(\mathbf{y}') = 0.05 \cdot 4.5 + 0.1 \cdot 4 + 0.15 \cdot 4 + 0.2 \cdot 2.5 + 0.5 \cdot 1.5 = 2.475$ and $A_{\mathbf{w},\mathbf{p}}(\mathbf{y}'') = 0.05 \cdot 6 + 0.1 \cdot 6 + 0.15 \cdot 5 + 0.2 \cdot 2 + 0.5 \cdot 1 = 2.55$. We will further formalize this approach and take its advantages to build the LP computational models.

Function w^* can be defined by its generation function g with the formula $w^*(\alpha) = \int_0^\alpha g(\xi)\, d\xi$. Introducing breakpoints $\alpha_i = \sum_{k \le i} p_{\tau(k)}$ and $\alpha_0 = 0$ we get

$$\omega_i = \int_0^{\alpha_i} g(\xi)\, d\xi - \int_0^{\alpha_{i-1}} g(\xi)\, d\xi = \int_{\alpha_{i-1}}^{\alpha_i} g(\xi)\, d\xi$$

and finally [17], [18]:

$$\begin{aligned} A_{\mathbf{w},\mathbf{p}}(\mathbf{y}) &= \sum_{i=1}^{m} \theta_i(\mathbf{y}) \int_{\alpha_{i-1}}^{\alpha_i} g(\xi)\, d\xi \\ &= \int_0^1 g(\xi) \overline{F}_{\mathbf{y}}^{(-1)}(\xi)\, d\xi\,, \end{aligned} \tag{6}$$

where $\overline{F}_{\mathbf{y}}^{(-1)}$ is the stepwise function $\overline{F}_{\mathbf{y}}^{(-1)}(\xi) = \theta_i(\mathbf{y})$ for $\alpha_{i-1} < \xi \le \alpha_i$. It can also be mathematically formalized as follows. First, we introduce the right-continuous cumulative distribution function (cdf):

$$F_{\mathbf{y}}(d) = \sum_{i=1}^{m} p_i \delta_i(d)\,, \tag{7}$$

where

$$\delta_i(d) = \begin{cases} 1 & \text{if } y_i \le d \\ 0 & \text{otherwise} \end{cases}$$

which for any real (outcome) value d provides the measure of outcomes smaller or equal to d. Next, we introduce the quantile function $F_{\mathbf{y}}^{(-1)} = \inf\{\eta : F_{\mathbf{y}}(\eta) \ge \xi\}$ for $0 < \xi \le 1$ as the left-continuous inverse of the cumulative distribution function $F_{\mathbf{y}}$, and finally $\overline{F}_{\mathbf{y}}^{(-1)}(\xi) = F_{\mathbf{y}}^{(-1)}(1 - \xi)$.

Formula (6) provides the most general expression of the WOWA aggregation allowing for expansion to continuous case. The original definition of WOWA allows one to build various interpolation functions w^* [19] thus to use different generation functions g in formula (6). Let us focus our analysis on the the piecewise linear interpolation function w^*. It is the simplest form of the interpolation function. Note, however, that the piecewise linear functions may be built with various number of breakpoints, not necessarily m. Thus, any nonlinear function can be well

approximated by a piecewise linear function with appropriate number of breakpoints. Therefore, we will consider weights vectors $\mathbf{w}$ of dimension n not necessarily equal to m. Any such piecewise linear interpolation function w^* can be expressed with the stepwise generation function:

$$g(\xi) = nw_k \text{ for } (k-1)/n < \xi \le k/n, \quad k = 1,\dots,n. \quad (8)$$

This leads us to the following specification of formula (6):

$$\begin{aligned} A_{\mathbf{w},\mathbf{p}}(\mathbf{y}) &= \sum_{k=1}^{n} w_k n \int_{(k-1)/n}^{k/n} \overline{F}_{\mathbf{y}}^{(-1)}(\xi)\, d\xi \\ &= \sum_{k=1}^{n} w_k n \int_{(k-1)/n}^{k/n} F_{\mathbf{y}}^{(-1)}(1-\xi)\, d\xi\,. \end{aligned} \quad (9)$$

Note that $n\int_{(k-1)/n}^{k/n} \overline{F}_{\mathbf{y}}^{(-1)}(\xi)\, d\xi$ represents the average within the kth portion of $1/n$ largest outcomes, the corresponding conditional mean [20], [21]. Hence, formula (9) defines WOWA aggregations with preferential weights $\mathbf{w}$ as the corresponding OWA aggregation but applied to the conditional means calculated according to the importance weights $\mathbf{p}$ instead of the original outcomes. Figure 3 illustrates application of formula (9) to computation of the WOWA aggregations for vectors from Fig. 2.

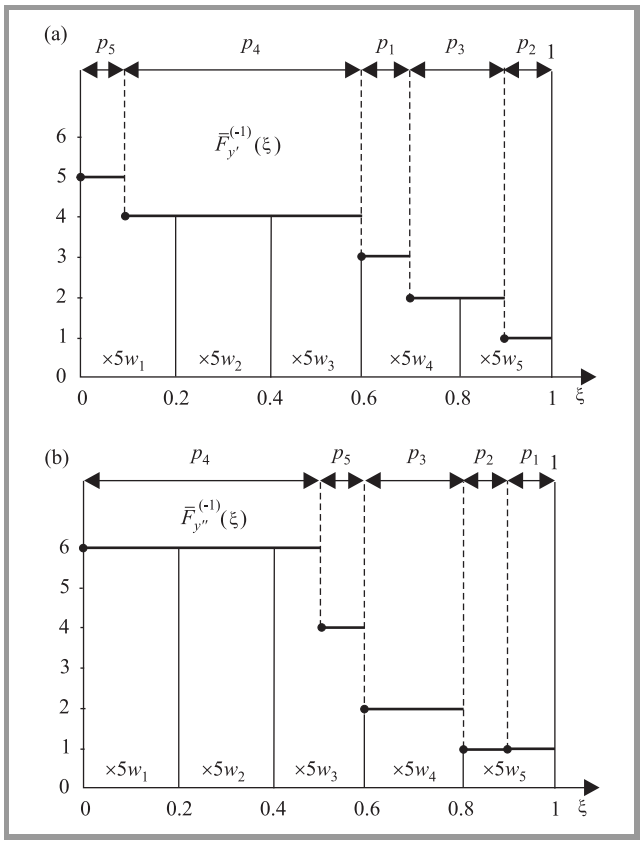


Fig. 3. Formula (9) applied to WOWA calculations for $\mathbf{p} = (0.1, 0.1, 0.2, 0.5, 0.1)$: (a) vector $\mathbf{y}' = (3,1,2,4,5)$; (b) vector $\mathbf{y}'' = (1,1,2,6,4)$.

We will treat formula (9) as a formal definition of the WOWA aggregation of m-dimensional outcomes $\mathbf{y}$ defined by the m-dimensional importance weights $\mathbf{p}$ and the n-dimensional preferential weights $\mathbf{w}$. Formula (9) may be reformulated to use the tail averages:

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{y}) = \sum_{k=1}^{n} nw_k \left(L\Big(\mathbf{y},\mathbf{p},1-\frac{k-1}{n}\Big) - L\Big(\mathbf{y},\mathbf{p},1-\frac{k}{n}\Big)\right) \quad (10)$$

with $L(\mathbf{y},\mathbf{p},\xi)$ defined by left-tail integrating of $F_{\mathbf{y}}^{(-1)}$, i.e., $L(\mathbf{y},\mathbf{p},0) = 0$,

$$L(\mathbf{y},\mathbf{p},\xi) = \int_0^{\xi} F_{\mathbf{y}}^{(-1)}(\alpha) d\alpha \quad \text{for } 0 < \xi \le 1 \quad (11)$$

and $L(\mathbf{y},\mathbf{p},1) = A_{\mathbf{p}}(\mathbf{y})$ thus representing the weighted average. Finally,

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{y}) = \sum_{k=1}^{n} w'_k L\Big(\mathbf{y},\mathbf{p},\frac{k}{n}\Big) \quad (12)$$

with weights

$$\begin{aligned} w'_k &= n(w_{n-k+1} - w_{n-k}) \text{ for } k = 1,\dots,n-1 \\ w'_n &= nw_1\,. \end{aligned} \quad (13)$$

Graphs of functions $L(\mathbf{y},\mathbf{p},\xi)$ (with respect to ξ) take the form of convex piecewise linear curves, the so-called absolute Lorenz curves [22] connected to the relation of the second order stochastic dominance (SSD). Therefore, formula (12) relates the WOWA average to the SSD consistent risk measures based on the tail means [23] provided that the importance weights are treated as scenario probabilities.

3. The Orness and Risk Preferences

The OWA aggregation may model various preferences from the optimistic (max) to the pessimistic (min). Yager [2] introduced a well appealing concept of the orness measure to characterize the OWA operators. The degree of orness associated with the OWA operator $A_{\mathbf{w}}(\mathbf{y})$ is defined as

$$\text{orness}(\mathbf{w}) = \sum_{i=1}^{m} \frac{m-i}{m-1} w_i\,. \quad (14)$$

For the max aggregation representing the fuzzy OR operator with weights $\mathbf{w} = (1,0,\dots,0)$ one gets $\text{orness}(\mathbf{w}) = 1$ while for the min aggregation representing the fuzzy AND operator with weights $\mathbf{w} = (0,\dots,0,1)$ one has $\text{orness}(\mathbf{w}) = 0$. For the average (arithmetic mean) one gets $\text{orness}((1/m,1/m,\dots,1/m)) = 1/2$. Actually, one may consider a complementary measure of andness defined as $\text{andness}(\mathbf{w}) = 1 - \text{orness}(\mathbf{w})$. OWA aggregations with orness greater or equal 1/2 are considered or-like whereas the aggregations with orness smaller or equal 1/2 are treated as and-like. The former correspond to rather optimistic preferences while the latter represents rather pessimistic preferences.

The OWA aggregations with monotonic weights are either or-like or and-like. Exactly, decreasing weights $w_1 \ge w_2 \ge \dots \ge w_m$ define an or-like OWA operator, while increas-

ing weights $w_1 \le w_2 \le \ldots \le w_m$ define an and-like OWA operator. Actually, the orness and the andness properties of the OWA operators with monotonic weights are total in the sense that they remain valid for any subaggregations defined by subsequences of their weights. Namely, for any $2 \le k \le m$ one gets

$$\sum_{j=1}^{k} \frac{k-j}{k-1} w_{i_j} \ge \frac{1}{2} \quad \text{and} \quad \sum_{j=1}^{k} \frac{k-j}{k-1} w_{i_j} \le \frac{1}{2}$$

for the OWA operators with decreasing or increasing weights, respectively. Moreover, the weights monotonicity is necessary to achieve the above total orness and andness properties. Therefore, we will refer to the OWA aggregation with decreasing weights as the totally or-like OWA operator, and to the OWA aggregation with increasing weights as the totally and-like OWA operator.

Yager [24] proposed to define the OWA weighting vectors via the regular increasing monotone (RIM) quantifiers, which provide a dimension independent description of the aggregation. A fuzzy subset Q of the real line is called a RIM quantifier if Q is (weakly) increasing with $Q(0) = 0$ and $Q(1) = 1$. The OWA weights can be defined with a RIM quantifier Q as $w_i = Q(i/m) - Q((i-1)/m)$ and the orness measure can be extended to a RIM quantifier (according to $m \to \infty$) as follows [24]:

$$\text{orness}(Q) = \int_0^1 Q(\alpha)\ d\alpha\,. \tag{15}$$

Thus, the orness of a RIM quantifier is equal to the area under it. The measure takes the values between 0 (achieved for $Q(1) = 1$ and $Q(\alpha) = 0$ for all other α) and 1 (achieved for $Q(0) = 1$ and $Q(\alpha) = 0$ for all other α). In particular, $\text{orness}(Q) = 1/2$ for $Q(\alpha) = \alpha$ which is generated by equal weights $w_k = 1/n$. Formula (15) allows one to define the orness of the WOWA aggregation (4) which can be viewed with the RIM quantifier $Q(\alpha) = w^*(\alpha)$ [25]. Let us consider piecewise linear function $Q = w^*$ defined by weights vectors $\mathbf{w}$ of dimension n according to the stepwise generation function (8). One may easily notice that decreasing weights $w_1 \ge w_2 \ge \ldots \ge w_n$ generate a strictly increasing concave curve $Q(\alpha) \ge \alpha$ thus guaranteeing the or-likeness of the WOWA operator. Similarly, increasing weights $w_1 \le w_2 \le \ldots \le w_n$ generate a strictly increasing convex curve $Q(\alpha) \le \alpha$ thus guaranteeing the and-likeness of the WOWA operator. Actually, the monotonic weights generate the totally or-like and and-like operators, respectively, in the sense that

$$\int_0^1 \frac{Q(a+\alpha(b-a)) - Q(a)}{Q(b) - Q(a)} d\alpha \ge \frac{1}{2} \tag{16}$$

or

$$\int_0^1 \frac{Q(a+\alpha(b-a)) - Q(a)}{Q(b) - Q(a)} d\alpha \le \frac{1}{2} \tag{17}$$

for the WOWA operators with decreasing or increasing weights, respectively.

Actually, the absolute Lorenz curve represent a dual characterization of the second stochastic dominance relation [22] which is the most general mathematical model of the risk averse preferences in decisions under risk [26]. Formula (12) represents the WOWA aggregation with increasing preferential weights as the weighted (positive) combination of n tail averages. Therefore, the WOWA objective functions with increasing preferential weights are SSD consistent and they represent the risk averse aggregations of outcomes under several scenarios. Moreover, such WOWA averages may be interpreted as the dual utility criteria within the theory developed by Yaari [27] which was recently reintroduced [28] in a simplified form of the spectral risk measures $\int_0^1 \phi(\xi) F_{\mathbf{y}}^{(-1)}(\xi) d\xi$, where decreasing (nonincreasing) distortion function ϕ represents risk averse preferences. Indeed, according to (6),

$$A_{\mathbf{w},\mathbf{p}}(\mathbf{y}) = \int_0^1 g(\xi) \overline{F}_{\mathbf{y}}^{(-1)}(\xi)\ d\xi = \int_0^1 g(1-\xi) F_{\mathbf{y}}^{(-1)}(\xi)\ d\xi$$

thus representing a spectral risk measure with distortion function $\phi(\xi) = g(1-\xi)$, nonincreasing for the increasing weights w_k. Similarly, the generalized WOWA can be expressed with $\phi(\xi) = g_\beta(1-\xi)$ nonincreasing for the relatively increasing weights w_k. As pointed out by Acerbi [28], the subjective risk aversion of a decision maker can be encoded in a function $\phi(\xi)$ defined for all possible $\xi \in (0,1]$ and one cannot see any arbitrary choice of function $\phi(\xi)$. The WOWA aggregations allows one to seek an appropriate function defined by a few preferential weights and possibly breakpoints (for the generalized WOWA).

4. Linear Programming Models

Consider maximization of a risk averse WOWA aggregation defined by increasing weights $w_1 \le w_2 \le \ldots \le w_n$

$$\max\{A_{\mathbf{w},\mathbf{p}}(\mathbf{y})\ :\ \mathbf{y} = \mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in \mathscr{F}\}\,. \tag{18}$$

Due to formula (12), the problem may be expressed as

$$\max\{\sum_{k=1}^{n} w'_k L(\mathbf{y}, \mathbf{p}, \frac{k}{n})\ :\ \mathbf{y} = \mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in \mathscr{F}\}$$

with positive weights w'_k defined by (13).

According to (11), values of function $L(\mathbf{y}, \mathbf{p}, \xi)$ for any $0 \le \xi \le 1$ can be given by optimization:

$$\begin{aligned} L(\mathbf{y}, \mathbf{p}, \xi) = \min_{s_i} \{ &\sum_{i=1}^{m} y_i s_i : \\ &\sum_{i=1}^{m} s_i = \xi;\ 0 \le s_i \le p_i,\ \forall\ i\ \}\,. \end{aligned} \tag{19}$$

The above problem is an LP for a given outcome vector $\mathbf{y}$ while it becomes nonlinear for $\mathbf{y}$ being a vector of variables.

This difficulty can be overcome by taking advantage of the LP dual to (19). Introducing dual variable t corresponding to the equation $\sum_{i=1}^{m} s_i = \xi$ and variables d_i corresponding to upper bounds on s_i one gets the following LP dual expression of $L(\mathbf{y},\mathbf{p},\xi)$

$$L(\mathbf{y},\mathbf{p},\xi) = \max_{t,d_i} \{\, \xi t - \sum_{i=1}^{m} p_i d_i : \quad t - d_i \le y_i,\ d_i \ge 0 \quad \forall\ i\}. \tag{20}$$

Therefore, maximization of the WOWA aggregation (18) can be expressed as follows:

$$\begin{aligned} \max_{t_k,d_{ik},y_i,x_j} \quad & \sum_{k=1}^{n} w'_k \left[\frac{k}{n} t_k - \sum_{i=1}^{m} p_i d_{ik}\right] \\ \text{s.t.} \quad & t_k - d_{ik} \le y_i,\ d_{ik} \ge 0 \quad \forall\ i,k \\ & \mathbf{y} = \mathbf{f}(\mathbf{x}), \quad \mathbf{x} \in \mathscr{F}. \end{aligned}$$

Consider multiple criteria problems (1) with linear objective functions $f_i(\mathbf{x}) = \mathbf{c}_i\mathbf{x}$ and polyhedral feasible sets:

$$\max\{(y_1, y_2, \ldots, y_m) \ :\ \mathbf{y} = \mathbf{C}\mathbf{x}, \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \quad \mathbf{x} \stackrel{\ge}{=} \mathbf{0}\}, \tag{21}$$

where $\mathbf{C}$ is an $m \times q$ matrix (consisting of rows $\mathbf{c}_i$), $\mathbf{A}$ is a given $v \times q$ matrix and $\mathbf{b} = (b_1, \ldots, b_v)^T$ is a given RHS (right hand side) vector. For such problems, we get the following LP formulation of the WOWA maximization (18):

$$\max_{t_k,d_{ik},y_i,x_j} \quad \sum_{k=1}^{n} \frac{k}{n} w'_k t_k - \sum_{k=1}^{n}\sum_{i=1}^{m} w'_k p_i d_{ik} \tag{22}$$

$$\text{s.t.} \quad \sum_{j=1}^{q} a_{rj} x_j = b_r \qquad r = 1,\ldots,v \tag{23}$$

$$y_i - \sum_{j=1}^{q} c_{ij} x_j = 0 \qquad i = 1,\ldots,m \tag{24}$$

$$d_{ik} \ge t_k - y_i,\ d_{ik} \ge 0 \qquad i = 1,\ldots,m,\ k = 1,\ldots,n \tag{25}$$

$$x_j \ge 0 \qquad j = 1,\ldots,q \tag{26}$$

Model (22)–(26) is an LP problem with $mn + m + n + q$ variables and $mn + m + v$ constraints. Thus, for problems with not too large number of scenarios (m) and preferential weights (n) it can be solved directly. Note that WOWA model (22)–(26) differs from the analogous deviational model for the OWA optimizations [8] only due to coefficients within the objective function (22) and the possibility of different values of m and n.

The number of constraints in problem (22)–(26) is similar to the number of variables. Nevertheless, for the simplex approach it may be better to deal with the dual of (22)–(26) than with the original problem. Note that variables d_{ik} in the primal are represented with singleton columns. Hence, the corresponding rows in the dual represent only simple upper bounds.

Introducing the dual variables: u_r $(r = 1,\ldots,v)$, v_i $(i = 1,\ldots,m)$ and z_{ik} $(i = 1,\ldots,m;\ k = 1,\ldots,n)$ corresponding to the constraints (23), (24) and (25), respectively, we get the following dual:

$$\begin{aligned} \min_{z_{ik},v_i,u_r} \quad & \sum_{r=1}^{v} b_r u_r \\ \text{s.t.} \quad & \sum_{r=1}^{v} a_{rj} u_r - \sum_{i=1}^{m} c_{ij} v_i \stackrel{\ge}{=} 0 && j = 1,\ldots,q \\ & v_i - \sum_{k=1}^{n} z_{ik} \ge 0 && i = 1,\ldots,m \\ & \sum_{i=1}^{m} z_{ik} = \frac{k}{n} w'_k && k = 1,\ldots,n \\ & 0 \le z_{ik} \le p_i w'_k && i = 1,\ldots,m \\ & && k = 1,\ldots,n \end{aligned} \tag{27}$$

The dual problem (27) contains: $m + n + q$ structural constraints, $m + v$ unbounded variables and mn bounded variables. Since the average complexity of the simplex method depends on the number of constraints, the dual model (27) can be directly solved for quite large values of m and n. Moreover, the columns corresponding to mn variables z_{ik} form the transportation/assignment matrix thus allowing one to employ special techniques of the simplex SON (special ordered network) algorithm [29] for implicit handling of these variables. Such techniques increase dramatically efficiency of the simplex method but they require a special tailored implementation. We have not tested this approach within our initial computational experiments based on the use of a general purpose LP code.

5. Computational Tests

In order to analyze the computational performances of the LP model for the WOWA optimization, similarly to [8], we have solved randomly generated problems of portfolio optimization according to the (discrete) scenario analysis approach [6]. There is given a set of securities for an investment $J = \{1,2,\ldots,q\}$. We assume, as usual, that for each security $j \in J$ there is given a vector of data $(c_{ij})_{i=1,\ldots,m}$, where c_{ij} is the observed (or forecasted) rate of return of security j under scenario i (hereafter referred to as outcome). We consider discrete distributions of returns defined by the finite set $I = \{1,2,\ldots,m\}$ of scenarios with the assumption that each scenario can be assigned the importance weight p_i that can be seen as the subjective probability of the scenario. The outcome data forms an $m \times q$ matrix $\mathbf{C} = (c_{ij})_{i=1,\ldots,m;j=1,\ldots,q}$ whose columns correspond to securities while rows $\mathbf{c}_i = (c_{ij})_{j=1,2,\ldots,q}$ correspond to outcomes. Further, let $\mathbf{x} = (x_j)_{j=1,2,\ldots,q}$ denote the vector of decision variables defining a portfolio. Each variable x_j expresses the portion of the capital invested in the corresponding security. Portfolio $\mathbf{x}$ generates outcomes

$$\mathbf{y} = \mathbf{C}\mathbf{x} = (\mathbf{c}_1\mathbf{x}, \mathbf{c}_2\mathbf{x}, \ldots, \mathbf{c}_m\mathbf{x}).$$

The portfolio selection problem can be considered as an LP problem with m uniform objective functions $f_i(\mathbf{x}) = \mathbf{c}_i\mathbf{x} = \sum_{j=1}^{q} c_{ij}x_j$ to be maximized [6]:

$$\max \{\mathbf{Cx} : \sum_{j=1}^{q} x_j = 1;\ x_j \geq 0,\ j = 1,\ldots,q\}.$$

Hence, our portfolio optimization problem can be considered a special case of the multiple criteria problem and one may seek an optimal portfolio with some criteria aggregation. Note that the aggregation must take into account the importance of various scenarios thus allowing importance weights p_i to be assigned to several scenarios. Further the preferential weights w_k must be increasing to represent the risk averse preferences (more attention paid on improvement of smaller outcomes). Thus we get the WOWA maximization problem:

$$\max \{A_{\mathbf{w},\mathbf{p}}(\mathbf{Cx}) : \sum_{j=1}^{q} x_j = 1;\ x_j \geq 0,\ j = 1,\ldots,q\}. \quad (28)$$

Our computational tests were based on the randomly generated problems (28) with varying number q of securities (decision variables) and number m of scenarios. The generation procedure worked as follows. First, for each security j the maximum rate of return r_j was generated as a random number uniformly distributed in the interval $[0.05, 0.15]$. Next, this value was used to generate specific outcomes c_{ij} (the rate of return under scenarios i) as random variables uniformly distributed in the interval $[-0.75r_j, r_j]$. Further, strictly increasing and positive weights w_k were generated. The weights were not normalized which allowed us to define them by the corresponding increments $\delta_k = w_k - w_{k-1}$. The latter were generated as uniformly distributed random values in the range of 1.0 to 2.0, except from a few (5 on average) possibly larger increments ranged from 1.0 to $n/3$. Importance weights p_i were generated according to the exponential smoothing scheme, which assigns exponentially decreasing weights to older or subjectively less probable scenarios: $p_i = \alpha(1-\alpha)^{i-1}$ for $i = 1,2,\ldots,m$ and the parameter α is chosen for each test problem size separately to keep the value of p_m around 0.001.

We tested solution times for different size parameters m and q. The basic tests were performed for the standard WOWA model with $n = m$. However, we also analyzed the case of larger n for more detailed preferences modeling, as well as the case of smaller n thus representing a rough preferences model. For each number of securities q and number of criteria (scenarios) m we solved 10 randomly generated problems (28). All computations were performed on a PC with the Athlon 64, 1.8 GHz processor employing the CPLEX 9.1 package. The 600 seconds time limit was used in all the computations.

In Tables 1 and 2 we show the solution times for the primal (22)–(26) and the dual (27) forms of the computational model, being the averages of 10 randomly generated problems. Upper index in front of the time value indicates the number of tests among 10 that exceeded the time limit. The empty cell (minus sign) shows that this occurred for all 10 instances. Both forms were solved by the CPLEX code without taking advantages of the constraints structure specificity. The dual form of the model performs much better in each tested problem size. It behaves very well with increasing number of securities if the number of scenarios does not exceed 100. Similarly, the model performs very well with increasing number of scenarios if only the number of securities does not exceed 50.

Table 1
Solution times [s] for the primal model (22)–(26)

Scenarios (m)	Number of securities (q)							
	10	20	50	100	150	200	300	400
10	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50	0.8	1.0	1.4	1.7	1.7	1.7	1.6	1.7
100	21.0	29.0	34.1	41.8	51.9	70.0	95.4	86.9
150	187.0	243.6	312.9	354.2	402.7	474.8	^{1}474.9	^{6}562.2

Table 2
Solution times [s] for the dual model (27)

Scenarios (m)	Number of securities (q)							
	10	20	50	100	150	200	300	400
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1
50	0.0	0.0	0.3	0.3	0.4	0.4	0.5	0.6
100	0.4	0.6	1.5	6.7	8.4	10.2	11.6	13.4
150	1.3	2.0	3.8	24.2	49.0	59.2	62.1	62.7
200	3.0	4.1	8.7	66.6	144.8	225.0	243.0	246.9
300	9.7	14.3	31.0	^{1}291.4	^{4}491.3	–	–	–
400	22.8	34.4	82.2	^{4}344.1	^{7}555.1	–	–	–

Table 3
Solution times [s] for different numbers of preferential weights ($q = 50$)

Number of scenarios (m)	Number of preferential weights (n)									
	3	5	10	20	50	100	150	200	300	400
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1
20	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.3	0.3
50	0.0	0.0	0.0	0.0	0.3	0.2	0.4	0.6	1.1	1.6
100	0.0	0.0	0.1	0.3	1.0	1.5	1.5	2.1	3.7	5.1
150	0.0	0.1	0.2	0.5	3.7	4.5	3.8	4.6	7.7	11.0
200	0.1	0.1	0.3	1.2	8.0	11.6	7.1	8.7	13.8	23.7
300	0.1	0.3	0.7	3.5	19.6	29.1	12.2	16.9	31.0	41.2
400	0.2	0.4	1.6	6.5	36.6	48.6	16.4	28.4	45.8	82.2

Table 3 presents solution times for different numbers of the preferential weights. The number of securities equals 50. It can be noticed that increasing the number of preferential weights and thus the number of breakpoints in the interpolation function induce moderate increase in the computational complexity. On the other hand, the computational efficiency can be significantly improved by reducing the number of preferential weights to a few which can

be reasonable in non-automated decision making support systems.

6. Concluding Remarks

The WOWA aggregation [12] represents a universal tool allowing one to combinine outcomes under several scenarios to form overall objective functions taking into account both the risk aversion preferences depicted with the preferential weights allocated to ordered outcomes as well as the scenarios importance expressed with weights allocated to several scenarios. The ordering operator used to define the WOWA aggregation is, in general, hard to implement within optimization problems. We have shown that the risk averse WOWA aggregations are characterized by the increasing weights and their optimization can be modeled by introducing auxiliary linear constraints. Hence, an LP decision under risk problem with the risk averse WOWA aggregation of outcomes under several scenarios can be formed as a standard linear program. Moreover, it can be further simplified by taking advantages of the LP duality.

Our computational experiments show that the LP formulation enables to solve effectively medium size WOWA problems. Actually, the number of few hundred scenarios efficiently covered by the dual LP model in less a minute for problems with limited number of structural variables seems to be quite enough for most applications to decisions under risk. The problems have been solved directly by a general purpose LP code. Taking advantages of the constraints structure specificity may remarkably extend the solution capabilities. In particular, the simplex SON algorithm [29] may be used for exploiting the LP embedded network structure in the dual form of the model. This seems to be a very promising direction for further research.

Acknowledgment

The research was partially supported by the Polish Ministry of Science and Higher Education under grant N N516 4307 33.

References

[1] W. Ogryczak, "Multiple criteria optimization and decisions under risk", *Contr. Cyber.*, vol. 31, pp. 975–1003, 2002.

[2] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making", *IEEE Trans. Syst., Man Cyber.*, vol. 18, pp. 183–190, 1988.

[3] M. Grabisch, S. A. Orlovski, and R. R. Yager, "Fuzzy aggregation of numerical preferences", in *Fuzzy Sets in Decision Analysis, Operations Research and Statistics*. Dordrecht: Kluwer, 1999, pp. 31–68.

[4] R. R. Yager and D. P. Filev, *Essentials of Fuzzy Modeling and Control*. New York: Wiley, 1994.

[5] R. R. Yager and J. Kacprzyk, *The Ordered Weighted Averaging Operators: Theory and Applications*. Dordrecht: Kluwer, 1997.

[6] W. Ogryczak, "Multiple criteria linear programming model for portfolio selection", *Ann. Oper. Res.*, vol. 97, pp. 143–162, 2000.

[7] R. R. Yager, "Constrained OWA aggregation", *Fuzzy Sets Syst.*, vol. 81, pp. 89–101, 1996.

[8] W. Ogryczak and T. Śliwiński, "On solving linear programs with the ordered weighted averaging objective", *Eur. J. Opnl. Res.*, vol. 148, pp. 80–91, 2003.

[9] W. Ogryczak and A. Tamir, "Minimizing the sum of the k largest functions in linear time", *Inform. Proc. Let.*, vol. 85, pp. 117–122, 2003.

[10] R. R. Yager, "Including importances in OWA aggegations using fuzzy systems modeling", *IEEE Trans. Fuzzy Syst.*, vol. 6, pp. 286–294, 1998.

[11] H. L. Larsen, "Importance weighted OWA aggregation of multicriteria queries", in *Proc. North Amer. Fuzzy Inform. Proc. Soc. Conf. NAFIPS'99*, New York, USA, 1999, pp. 740–744.

[12] V. Torra, "The weighted OWA operator", *Int. J. Intell. Syst.*, vol. 12, pp. 153–166, 1997.

[13] V. Torra and Y. Narukawa, *Modeling Decisions Information Fusion and Aggregation Operators*. Berlin: Springer-Verlag, 2007.

[14] A. Valls and V. Torra, "Using classification as an aggregation tool for MCDM", *Fuzzy Sets Syst.*, vol. 115, pp. 159–168, 2000.

[15] E. Damiani, S. De Capitani di Vimercati, P. Samarati, and M. Viviani, "A WOWA-based aggregation technique on trust values connected to metadata", *Electr. Notes Theor. Comp. Sci.*, vol. 157, pp. 131–142, 2006.

[16] D. Nettleton and J. Muniz, "Processing and representation of metadata for sleep apnea diagnosis with an artificial intelligence approach", *Medic. Inform.*, vol. 63, pp. 77–89, 2001.

[17] W. Ogryczak and T. Śliwiński, "On decision support under risk by the WOWA optimization", in *Ninth European Conference on Symbolic and Quanlitative Approaches to Reasoning with Uncertainty ESQARU 2007*, *LNAI*, vol. 4724. Heidelberg: Springer, 2007, pp. 779–790.

[18] W. Ogryczak and T. Śliwiński, "On optimization of the importance weighted OWA aggregation of multiple criteria", in *International Conference Computational Science and Its Applications ICCSA 2007*, *LNCS*, vol. 4705. Heidelberg: Springer, 2007, pp. 804–817.

[19] V. Torra, "The WOWA operator and the interpolation function W*: Chen and Otto's interpolation method revisited", *Fuzzy Sets Syst.*, vol. 113, pp. 389–396, 2000.

[20] W. Ogryczak and T. Śliwiński, "On equitable approaches to resource allocation problems: the conditional minimax solution", *J. Telecommun. Inform. Technol.*, no. 3, pp. 40–48, 2002.

[21] W. Ogryczak and M. Zawadzki, "Conditional median — a parametric solution concept for location problems", *Ann. Oper. Res.*, vol. 110, pp. 167–181, 2002.

[22] W. Ogryczak and A. Ruszczyński, "Dual stochastic dominance and related mean-risk models", *SIAM J. Opt.*, vol. 13, pp. 60–78, 2002.

[23] R. Mansini, W. Ogryczak, and M. G. Speranza, "Conditional value at risk and related linear programming models for portfolio optimization", *Ann. Oper. Res.*, vol. 152, pp. 227–256, 2007.

[24] R. R. Yager, "Quantifier guided aggregation using OWA operators", *Int. J. Intell. Syst.*, vol. 11, pp. 49–73, 1988.

[25] X. Liu, "Some properties of the weighted OWA operator", *IEEE Trans. Syst. Man Cyber. B*, vol. 368, pp. 118–127, 2006.

[26] A. Müller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*. Chichester: Wiley, 2002.

[27] M. E. Yaari, "The dual theory of choice under risk", *Econometrica*, vol. 55, pp. 95–115, 1987.

[28] C. Acerbi, "Spectral measures of risk: a coherent representation of subjective risk aversion", *J. Bank. Finan.*, vol. 26, pp. 1505–1518, 2002.

[29] F. Glover and D. Klingman, "The simplex SON method for LP/embedded network problems", *Math. Progr. Study*, vol. 15, pp. 148–176, 1981.

Włodzimierz Ogryczak is a Professor and Deputy Director for Research in the Institute of Control and Computation Engineering (ICCE) at the Warsaw University of Technology, Poland. He received both his M.Sc. (1973) and Ph.D. (1983) in mathematics from Warsaw University, and D.Sc. (1997) in computer science from Polish Academy of Sciences. His research interests are focused on models, computer solutions and interdisciplinary applications in the area of optimization and decision making with the main stress on: multiple criteria optimization and decision support, decision making under risk, location and distribution problems. He has published three books and numerous research articles in international journals.
e-mail: wogrycza@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Tomasz Śliwiński is an Assistant Professor of the Optimization and Decision Support Division in the Institute of Control and Computation Engineering (ICCE) at the Warsaw University of Technology, Poland. He received the M.Sc. and Ph.D. degrees in computer science in 1999 and 2007, respectively, both from the Warsaw University of Technology. His research interests focus on the column generation techniques and discrete optimization methods. He is an author and co-author of several research articles in international journals.
e-mail: tsliwins@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

Path Diversity Protection in Two-Layer Networks

Mateusz Dzida, Tomasz Śliwiński, Michał Zagożdżon, Włodzimierz Ogryczak, and Michał Pióro

Abstract— The paper addresses an optimization problem related to dimensioning links in a resilient two-layer network. A particular version of the problem which assumes that links of the upper layer are supported by unique paths in the lower layer is considered. Two mixed-integer programming formulations of this problem are presented and discussed. Direct resolving of these formulations requires pre-selection of "good" candidate paths in the upper layer of the network. Thus, the paper presents an alternative approach which is based on decomposing the resolution process into two phases, resolved iteratively. The first phase subproblem is related to designing lower layer path flows that provide the capacities for the logical links of the upper layer. The second phase is related to designing the flow patterns in the upper layer with protection assured through diversity of paths. In this phase we take into account the failures of the logical links that result from the failures of the lower layer links (so called *shared risk link groups*).

Keywords— link dimensioning, path diversity, resilient routing, two-layer network optimization.

1. Introduction

One of the most important internet architectures is based on IP-over-WDM (wavelength division multiplexing) networks. IP-over-WDM refers to a complex network model which uses a layered structure of resources, operated according to two distinct network protocols. Resources of a layered network form a hierarchical structure with each layer constituting a proper network.

In the considered network model the lower layer is an WDM network composed of a set of fibers connecting WDM cross-connects. The upper layer is an IP network composed of a given set of logical connections between routers (see [1] and [2]). The logical IP links are supported by paths composed of the WDM links (WDM paths).

In this paper we address an optimization problem related to dimensioning capacity of the WDM fibers. Throughout the paper we assume that each IP link is supported by an unique path in the WDM layer (uniqueness property). Dimensioning cost for a given realization of the upper layer links is calculated as a sum of capacity costs of distinct links of the lower layer. For a given set of traffic requirements (demands) the considered problem consists in determining a set of IP layer paths (IP paths) and their realizations in the WDM layer for which the total dimensioning cost is minimized.

In the paper we consider two-layer network model for which the WDM links are subject to failures. We assume that during a failure one of the WDM links becomes unavailable. Failure of an WDM link causes that all IP links associated with affected paths are unavailable too. Thus, a failure of single WDM link may cause unavailability of multiple IP links. In the literature such failure model is called shared risk resource group (see [3]).

Suppose that flows assigned to a specific traffic demand are bifurcated and diversified (recall that the uniqueness property refers only to the WDM layer). Diversification of the flows constitutes a means for protecting the IP-over-WDM network against failures. We refer to this protection type as path diversity (see [4]).

In the general two-layer network model the set of lower layer nodes is wider than the set of the upper layer nodes, i.e., only selected sites comprise either WDM cross-connects and IP routers, while the others comprise only WDM cross-connects. Note that when all sites comprise devices of both kinds the problem can be reduced to dimensioning IP links, each related to the corresponding WDM link.

In the following we present a mixed-integer programming (MIP) formulation of the problem related to dimensioning the WDM links in the resilient IP-over-WDM network. As this formulation requires identifying all possible paths in the IP layer, resolving it using general MIP solvers is not efficient. Thus, we propose a dedicated method based on decomposing the resolution process into two iteratively invoked phases.

The paper is organized as follows. In Section 2 we formulate the considered problem as a mixed-integer program. The resolving method is presented in Section 3. The numerical results illustrating the efficiency of the method are discussed in Section 4. The paper is summarized in Section 5, where the conclusions are drawn.

2. Problem Formulation

Let $\mathscr{V}$ and $\mathscr{W}$ be the set of IP nodes and WDM nodes, respectively. We define $\mathscr{E}$ and $\mathscr{F}$ as the link sets associated with upper and lower layer, respectively. The network graphs associated with network layers are denoted by $\mathscr{G}(\mathscr{V},\mathscr{E})$ – the IP layer graph and $\mathscr{H}(\mathscr{W},\mathscr{F})$ – the WDM layer graph. In the balance of this paper we assume that each origin-destination (O-D) pair, constituting the set of demands $d \in \mathscr{D}$, is associated with a specific portion of requested bandwidth h_d. By $\mathscr{P}_d$ we denote a set of candidate paths for demand d.

The IP layer links $e \in \mathscr{E}$ are supported by paths in the WDM layer. For each link $e \in \mathscr{E}$ we define a set of such

candidate paths $\mathscr{P}_e$. Throughout the paper we assume that each WDM link can be subject to a failure, and only single link $f \in \mathscr{F}$ can fail at a time. Still, a specific realization of the IP links can induce complex failures of multiple links in the IP layer.

The set of failure states is denoted by $\mathscr{S}$. Each element of $\mathscr{S}$ corresponds to a failure of a specific link $f \in \mathscr{F}$, what is determined by values of constants ρ_{fs}. Value zero of ρ_{fs} indicates that the corresponding link f is affected, and cannot be used to transit traffic in failure state s; $\rho_{fs} = 1$, otherwise.

Objective in the considered problem is to find capacities of the WDM links of the minimal cost. The unit cost of link f capacity is given by ξ_f, while the total capacity of this link is determined by variable g_f. The total link flow is calculated as a sum of particular path flows realizing the related demands; they are denoted by z_{eq}. Due to assumed uniqueness property, each candidate path is associated with a binary variable u_{eq}. Thus, the flow associated with a path, selected to support link e ($u_{eq} = 1$), must carry entire flow of this link, determined by value of variable y_e. Binary variable w_{es} constitutes a link-failure incidence factor which determines if IP link e is affected by a failure of the associated WDM path. Subsequently, binary variables r_{dps} determine if IP path p is affected by failure s due to unavailability of at least one WDM links traversed by this path. Finally, we define x_{dps} as a variable representing the traffic flow assigned to IP path p in state s. An MIP formulation of the discussed problem reads:

$$\min \sum_{f\in\mathscr{F}} \xi_f g_f, \tag{1}$$

$$\text{st.} \quad \sum_{p\in\mathscr{P}_d} x_{dps} \geq h_d \qquad d \in \mathscr{D}, s \in \mathscr{S}, \tag{2}$$

$$x_{dps} \leq x_{dp\mathscr{O}} \qquad d \in \mathscr{D}, p \in \mathscr{P}_d, s \in \mathscr{S}, \tag{3}$$

$$x_{dps} \leq (1 - r_{dps})h_d \qquad d \in \mathscr{D}, p \in \mathscr{P}_d, s \in \mathscr{S}, \tag{4}$$

$$\sum_{d\in\mathscr{D}} \sum_{p\in\mathscr{P}_{ed}} x_{dp\mathscr{O}} \leq y_e \qquad e \in \mathscr{E}, \tag{5}$$

$$|\mathscr{E}_p| r_{dps} \geq \sum_{e\in\mathscr{E}_p} w_{es} \qquad d \in \mathscr{D}, p \in \mathscr{P}_d, s \in \mathscr{S}, \tag{6}$$

$$\sum_{q\in\mathscr{Q}_e} z_{eq} = y_e \qquad e \in \mathscr{E}, \tag{7}$$

$$\sum_{e\in\mathscr{E}} \sum_{q\in\mathscr{Q}_{fe}} z_{eq} \leq \rho_{fs} g_f \qquad f \in \mathscr{F}, s \in \mathscr{S}, \tag{8}$$

$$\sum_{q\in\mathscr{Q}_e} u_{eq} \leq 1 \qquad e \in \mathscr{E}, \tag{9}$$

$$z_{eq} \leq u_{eq} M \qquad e \in \mathscr{E}, q \in \mathscr{P}_e \tag{10}$$

$$\sum_{q\in\mathscr{P}_e} \sum_{f\in\mathscr{F}_q} \rho_{fs} u_{eq} \leq |\mathscr{F}| w_{es} \qquad e \in \mathscr{E}, s \in \mathscr{S}. \tag{11}$$

System of three constraints (2)–(4) assures that at least h_d amount of bandwidth survives in each situation $s \in \mathscr{S}$ (according to the principles of the path diversity protection model). Equivalently, this can be expressed by the following nonlinear inequality:

$$\sum_{p\in\mathscr{P}_d} r_{dps} x_{dp\mathscr{O}} \geq h_d \quad d \in \mathscr{D}, s \in \mathscr{S}. \tag{12}$$

Constraint (5) is a capacity constraint. Constraint (6) indicates if certain path p is affected by specific failure s ($r_{dps} = 1$), i.e., if at least one of the IP links supporting path p is failed. The appropriate values of variables w_{es} are induced by constraint (11) which forces $w_{es} = 1$ when at least one of the links of the path supporting link e fails in state s (i.e., $\rho_{fs} u_{eq} = 1$).

The presented mathematical model corresponds to a classical two-layer dimensioning problem where the capacities of the IP links (given by y_e, $e \in \mathscr{E}$) are supported by the WDM path flows (according to constraint (7)). Constraint (9) is used to assure the uniqueness property. Constraint (8) is a capacity constraint of the WDM links. The considered objective is minimizing the total capacity installation cost associated with these links.

Below, we present an optimization model which does not require the predefined lists of candidate paths $\mathscr{Q}_e$, $e \in \mathscr{E}$. Instead, the so called node-link [2] notation of multicommodity flow optimization is used [2]. The node-link notation implicitly take into account all possible paths. Let $f \in \delta^+(v)$ and $f \in \delta^-(v)$ be the sets of links outgoing from and incoming to node $v \in \mathscr{W}$, respectively, Δ_{ve} be a constant which is equal to 1 if v is the starting node of e, to -1 if v is terminating node of e, and to 0, otherwise, and variables z_{fe} denote the WDM flows associated with the paths supporting the IP links. Accordingly, we define u_{fe} as a binary variable determining if link f is contained in the path supporting e:

$$\min \sum_{f\in\mathscr{F}} \xi_f g_f, \tag{13}$$

$$\text{st.} \quad (2) - (6),$$

$$\sum_{f\in\delta+(v)} z_{fe} - \sum_{f\in\delta-(v)} z_{fe} = \Delta_{ve} y_e, \qquad v \in \mathscr{V}, e \in \mathscr{E}, \tag{14}$$

$$\sum_{f\in\delta+(v)} u_{fe} \leq 1 \quad v \in \mathscr{V}, e \in \mathscr{E}, \tag{15}$$

$$z_{fe} \leq M u_{fe} \quad e \in \mathscr{E}, f \in \mathscr{F}, \tag{16}$$

$$\sum_{e\in\mathscr{E},} z_{fe} \leq g_f \quad f \in \mathscr{F}. \tag{17}$$

Constraint (14) expresses a flow conservation principle which is characteristic of the node-link notation. The uniqueness property is assured by integrality of u_{fe} and constraint (15). Due to (16) the flows of selected links are non-negative (M is maximal capacity of an IP link). Finally, inequality (17) constitutes a capacity constraint related to the WDM links.

Observe that values of vector $w = (w_{es} : e \in \mathscr{E}, s \in \mathscr{S})$ can be calculated as products of ρ_{fs} for the set of links $f \in \mathscr{F}$ determining the realization of specific link e.

3. Resolution Approach

To the best of our knowledge the problem related to the path diversity protection can be formulated only using the link-path notation (as far as linear constraints and continuous variables are considered). Formulation (2) is such a formulation using variables assigned to each of the candidate paths in the IP layer. Each of these variables represents a portion of the related demand. To effectively use this formulation one has to determine proper set of the candidate paths. Because the number of possible candidate paths grows exponentially with the number of nodes in the network graph, resolving a link-path formulation involving all candidate paths is inefficient. Thus, we develop a method which decomposes the resolving process into two subsequently invoked phases. The proposed method allows to identify the set of the necessary candidate paths in the IP layer using the column generation technique (see [2]), and to design the IP links realizations by resolving appropriate MIP.

The approach is based on the assumption that realizations of the IP links are known during the first phase. It means that values of variables r_{dps} are fixed and given. Thus, technique called column generation (see [2]) can be used to resolve the problem related to designing capacities of the IP links. The problem can be defined by the system of constraints (5), (12) and objective function (18):

$$\min \sum_{e \in \mathscr{E}} \zeta_e y_e, \tag{18}$$

where ζ_e is a specific capacity unit cost associated with current realization of the links in the lower layer, i.e., each ζ_e is calculated as a sum of ξ_f along paths supporting link e. The problem, referred to as master problem in the context of path generation, is denoted by **M**.

In the column (path) generation algorithm [5] not all the columns of the constraints matrix are stored. Instead, only a subset of the variables (columns) that can be seen as an approximation (restriction) of the original problem is kept. The column generation algorithm iteratively modifies the subset of variables by introducing new variables in a way that improves the current optimal solution. At the end, the set contains all the variables (paths) necessary to construct the overall optimal solution which can use all possible paths in the graph.

Let $(\lambda_d^s)^*$ $(d \in \mathscr{D}, s \in \mathscr{S})$ be current optimal dual variables associated with constraint (12) and $\Lambda_d^* = \sum_{s \in \mathscr{S}} (\lambda_d^s)^*$ $(d \in \mathscr{D})$ be current optimal auxiliary dual variables. At each iteration we are interested in generating path p for which the reduced price $\sum_{e \in p} \xi_e + \sum_{s \in \bar{\mathscr{S}}_p} (\lambda_d^s)^* - \Lambda_d^*$ has the smallest and negative value, as we can expect this will improve the current optimal solution to the greatest possible extent.

The pricing problem stated above can be approached in a way described in [6]. The basic idea is to compute the dual length $\langle p \rangle = \sum_{e \in p} \xi_e + \sum_{s \in \bar{\mathscr{S}}_p} (\lambda_d^s)^*$ of each path p, skiping, however, the computations for many paths for which apply some domination rules as proposed in [6]. The set of all non-dominated paths can be generated by means of a label-setting algorithm for shortest-path problems with resource constraints (SPPRC) [7].

As an extension of the SPPRC algorithm, we have also introduced path length limitation – an important contribution to the reduction of the size of the set of non-dominated paths. The extension is based on the observation that excessively long paths are useless as they cannot improve the current solution, or the solutions they represent are known to be worse than some already known solutions. For example, a simple path length restriction may be expressed as follows: $\sum_{e \in p} \xi_e < \Lambda_d^*$. Also, knowledge of some path p' representing a feasible solution can help to tighten the path length restriction. In such a case we are only interested in finding a path p satisfying $\sum_{e \in p} \xi_e < \sum_{e \in p'} \xi_e + \sum_{s \in \bar{\mathscr{S}}_{p'}} (\lambda_d^s)^*$. Applying path length limitation results in significant reduction of the pricing time.

Let $y^* = (y_e : e \in \mathscr{E})$ be the vector of link capacities of the IP links, obtained as the optimal solution of **M**. In the proposed approach y^* is an input for the second phase which adjusts the realizations of the IP links to better fit conditions of the IP layer. The resulting vector of the link-failure incidence factors $w = (w_{es} : e \in \mathscr{E}, s \in \mathscr{S})$ determines new values of $r = (r_{dps} : d \in \mathscr{D}, p \in \mathscr{P}_d, s \in \mathscr{S})$ for the next iteration of the procedure. The subproblem of the second phase, given by the system of constraints (14)–(17) and objective function (19), is denoted by **R**:

$$\min \sum_{e \in \mathscr{E}} \sum_{f \in \mathscr{F}} u_{fe}. \tag{19}$$

The general idea of the proposed approach is presented by Algorithm 1. Note that **M** can be formulated equivalently as a system of constraints (2)–(4) with objective function (18). Let $\lambda = (\lambda_{dps} : d \in \mathscr{D}, p \in \mathscr{P}_d, s \in \mathscr{S})$, $\beta = (\beta_{dps} : d \in \mathscr{D}, p \in \mathscr{P}_d, s \in \mathscr{S})$, and $\gamma = (\gamma_{dps} : d \in \mathscr{D}, p \in \mathscr{P}_d, s \in \mathscr{S})$ be the vectors

Algorithm 1: The decomposed iteractive procedure

Step 1: For each link $e \in \mathscr{E}$ find the cheapest realization with respect to link costs ξ_f. Denote the obtained vector of upper link capacities by y^0, link realization by u^0, and link-failure incidence by r^0.

Step 2: For fixed $r = r^0$ solve **M** using the path generation, and denote the obtained capacity vector by y^0.

Step 3: Put $y^0 \equiv 0$. Solve **R** for fixed vector $y = y^0$, and denote the obtained link realization by u^0. For each $e \in \mathscr{E}$ calculate a vector of link-failure incidence vector r^0. If stopping criterion is not met, return to Step 1.

of Lagrangean multipliers associated with constraints (2), (3), and (4), respectively. The dual corresponding to the considered formulation of **M** reads:

$$\max \sum_{d\in\mathscr{D}}\sum_{s\in\mathscr{S}} h_d\lambda_{ds} - \sum_{d\in\mathscr{D}}\sum_{p\in\mathscr{P}_d}\sum_{s\in\mathscr{S}}(1-r^*_{dps})h_d\beta_{dps}, \quad (20)$$

$$\text{st.} \sum_{s\in\mathscr{S}}\alpha_{dps} \le \sum_{e\in\mathscr{E}_p}\pi_e \quad d\in\mathscr{D},\ p\in\mathscr{P}_d, \quad (21)$$

$$\lambda_{ds} \le \alpha_{dps}+\beta_{dps} \quad d\in\mathscr{D},\ p\in\mathscr{P}_d,\ s\in\mathscr{S}, \quad (22)$$

$$\pi_e \le \xi_e \quad e\in\mathscr{E}. \quad (23)$$

Let $(\lambda^0,\alpha^0,\beta^0,\pi^0)$ be an optimal solution of problem (20)–(23). Consider path p and state s for which $r^*_{dps}=1$ and $\beta^0_{dps}>0$. Suppose that the value of r^*_{dps} is decreased to 0, and problem (3) is re-optimized. It can be shown that the value of λ_{dps} can be smaller than λ^0_{dps} due to new value of β_{dps} which is equal to zero. Thus, we conclude that it can be advantageous to set r_{dps} to zero in the next step of the procedure because we can potentially decrease the optimal value of the primal objective function, i.e., decrease the dimensioning cost.

Similarly, setting r^*_{dps} to zero for path p and state s for which $\alpha^0_{dps}>0$ can also lead to decrease of the optimal value of objective function (20). Due to these observations, in the following we consider a slightly different form of the objective function of **R**. We define $\mathscr{I}_a$ and $\mathscr{I}_b$ as sets of triplets (d,p,s) for which $\alpha_{dps}>0$ and $\beta_{dps}>0$, i.e., $\mathscr{I}_a=\{(d,p,s):\ \alpha_{dps}>0\}$ and $\mathscr{I}_b=\{(d,p,s):\ \beta_{dps}>0\}$,

$$\max \sum_{(d,p,s)\in\mathscr{I}_b}\beta^0_{dps}(1-r_{dps}) + \sum_{(d,p,s)\in\mathscr{I}_a}\alpha^0_{dps}(1-r_{dps}). \quad (24)$$

The modified **R** shall also involve an appropriate set of constraints (6) related to triples contained in $\mathscr{I}_a\cup\mathscr{I}_b$. In practical implementations of the discussed approach it can be advantageous to consider a combined objective function of **R**:

$$\min\ \varepsilon\left(\sum_{e\in\mathscr{E}}\sum_{f\in\mathscr{F}}u_{fe}\right)+(1-\varepsilon) \times\left(\sum_{(d,p,s)\in\mathscr{I}_b}\beta^0_{dps}r_{dps}+\sum_{(d,p,s)\in\mathscr{I}_a}\alpha^0_{dps}r_{dps}\right), \quad (25)$$

where ε is an optimization parameter.

Notice that for given link realizations **M** can be infeasible due to empty set of allowable candidate paths, i.e., at least one path cannot be affected in any failure state. Because of that, in the following, we consider specific inequalities which can be used to exclude the infeasible link realizations from the solution space of **R**. The basic form of the inequalities refers to a cut set in graph $\mathscr{H}(\mathscr{W},\mathscr{F})$. Let $\delta(\mathscr{W}')$ be a cut set associated with subset of nodes $\mathscr{W}'$:

$$\sum_{e\in\delta(\mathscr{W}')} w_{es} \ge 1 \quad s\in\mathscr{S}. \quad (26)$$

Inequality (26) assures that at least one IP link must be available for given cut set $\delta(\mathscr{W}')$ in $\mathscr{H}(\mathscr{W},\mathscr{F})$. In particular, (26) is valid for the set of links outgoing from one specific node, i.e., $\delta^+(v)$. Still, the number of potential cut sets grows exponentially with the number of nodes. Thus, we assume that only specific cut sets, related to one, two or three nodes could be examined in the practical implementations.

It may still appear that the feasible solution space of **M** is empty for a specific realization of the IP links, and we must exclude the current solution from the feasible solution space of **R**. For this purpose we use simple inequality which excludes binary vectors related to non-feasible link realizations. Let $\mathscr{U}_0$ and $\mathscr{U}_1$ be the sets of (e,f) pairs for which u_{ef} is equal to zero and one in the excluded realization, respectively. The discussed inequality reads:

$$\sum_{(e,f)\in\mathscr{U}_0} u_{ef} + \sum_{(e,f)\in\mathscr{U}_1}(1-u_{ef}) \ge 1. \quad (27)$$

Inequalities above are introduced into the formulation of **R** each time **M** is infeasible and a cut set which assures feasibility of **M** cannot be identified.

4. Numerical Results

The main goal of our computational experiments was to assess the efficiency of the two phase algorithm considered in the paper. For this purpose we performed a series of tests which, first, could tell us how the parameter ε in function (25) influences the algorithm efficiency and, finally, how fast the value of the generated solutions improves during the method execution. Aiming at this we implemented Algorithm 1 supported by the linear solver of CPLEX 10.0 which was used to resolve problems **R** and **M**. The computations were conducted on a PC equipped with P4 Quad Core processor and 4 GB memory.

We used two network instances from Survivable Network Design Data Library: *pdh* (11 nodes, 34 links, 24 demands) and *newyork* (16 nodes, 49 links, 240 demands). The topologies of these networks defined the topologies of the lower layers. We assumed that the nodes of the lower layer were also the nodes of the upper layer, i.e., having both IP and WDM switching capabilities. The graph of the upper layer was assumed to be fully-connected.

Using the two example networks we investigated the influence of the value of parameter ε. For this purpose we run our two phase algorithm using different values of this parameter. In the computations we assumed every single link failure in the lower layer. The time limit was set to 2 hours.

In Tables 1 and 2 we present the values of the objective function (25) of the best solution found within the assumed time limit.

Table 1
Two phase algorithm: objective values for different values of ε for the *pdh* network

ε	$0.0-0.8$	0.85, 0.88	0.9	0.95	1.0
Objective	98622.7	93323.8	93217.6	93526.2	93632.4
Optimum	91937.5				

According to the results presented in Tables 1 and 2 we conclude that the objective function (25) used when resolving the lower layer problem strongly influences the efficiency of the two phase algorithm. It appeared that the approach based on weighting both components of (25) was the most efficient one. Neither dual based indicators, nor

Table 2
Two phase algorithm: objective values for different values of ε for the *newyork* network

ε	$0.0-0.9$	0.93	0.98	0.99	1.0
Objective	26012.9	25472.7	25639.7	25831.2	25474.6
Optimum	24453.7				

shortest path lengths when used as standalone ($\varepsilon = 1$ or $\varepsilon = 0$) could provide the solutions of the same quality. The optimum value of the objective function was computed as an optimal solution of the single layer path diversity design problem (taking into account every single link failure) with

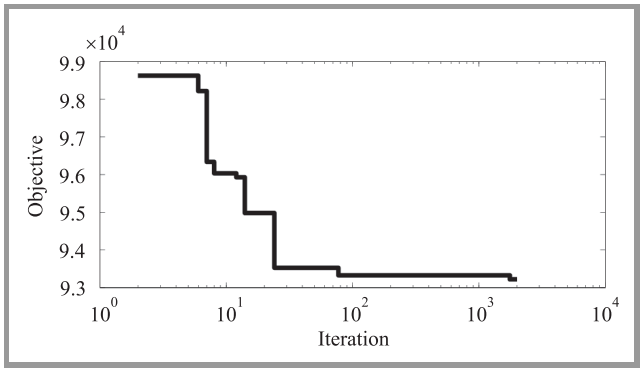


Fig. 1. Two phase algorithm: improvement of the objective function for *pdh* and $\varepsilon = 0.93$.

the network topology and the unit capacity link costs as in input networks *pdh* and *newyork*. Observe that this value of the objective function is also an optimal value of the two layer counterpart (with a full graph in the upper layer) only if all nodes of the lower layer are also the nodes in the upper layer – what is in fact true in our case. This value allowed to asses the quality of obtained solutions, which appeared to by only less than 1% worse than the optimum.

Analyzing Fig. 1, representing the algorithm convergence, we can conclude that for a proper value of ε the algorithm quickly finds a good quality solution which is only slightly improved until the algorithm terminates.

5. Concluding Remarks

In the paper we investigated an optimization problem related to dimensioning WDM links in a resilient two-layer IP-over-WDM network. In the considered problem the capacities of WDM links must be large enough to accommodate flows associated with selected realization of the IP links. Since the WDM links are subject to failures, we assumed that protection of traffic flows was assured by path diversity. In the paper we proposed a dedicated method for resolving this problem. The method was based on iterative resolving two subproblems, each related to optimizing flows in a distinct network layer. In our numerical experiments we tested the efficiency of the method for different settings of ε in objective function (25). The experiments revealed that neither dual based indicators, nor shortest path lengths when used as standalone could provide the solutions of the best quality.

Acknowledgments

The research presented in this paper has been funded by grants no. N517 397334 and N516 375736 from the Polish Ministry of Science and Higher Education. The first three authors have been additionally supported by a grant from the Faculty of Electronics and Information Technology, Warsaw University of Technology. M. Pióro has also been supported by grant no. 621-2006-5509 from Swedish Research Council.

References

[1] S. Borne, E. Gourdin, B. Liau, and A. R. Mahjoub, "Design of survivable IP-over-optical networks", in *Proc. INOC'2003 Conf.*, Paris, France, 2003, pp. 114–118.

[2] M. Pióro and D. Medhi, *Routing, Flow, and Capacity Design in Communication and Computer Networks*. San Francisco: Morgan Kaufman, 2004.

[3] D. Coudert, P. Datta, S. Perennes, H. Rivano, and M.-E. Voge, "Shared risk resource group: complexity and approximability issues", *Paral. Proces. Lett.*, vol. 17, no. 2, pp. 169–184, 2007.

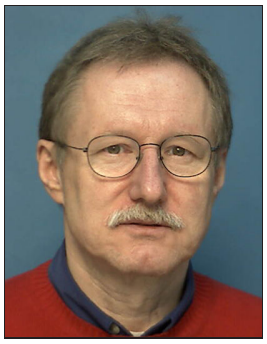
[4] M. Dzida, T. Śliwiński, M. Zagożdżon, W. Ogryczak, and M. Pióro, "Path generation for a class survivable network design problems", in *NGI 2008 Conf. Next Gener. Internet Netw.*, Cracow, Poland, 2008.

[5] G. B. Dantzig and P. Wolfe, "The decomposition algorithm for linear programming", *Oper. Res.*, vol. 8, no. 1, pp. 101–111, 1960.

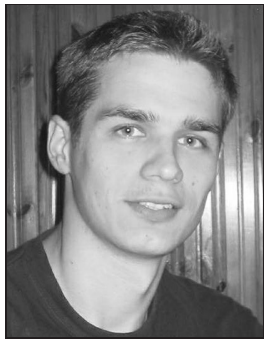
[6] T. Stidsen, B. Petersen, K. B. Rasmussen, S. Spoorendonk, M. Zachariasen, F. Rambach, and M. Kiese, "Optimal routing with single backup path protection", in *Proc. INOC 2007 Conf.*, Spa, Belgium, 2007.

[7] S. Irnich and G. Desaulniers, "Shortest path problems with resource constraints", in *Column Generation*, G. Desaulniers, J. Desrosier, and M. M. Solomon, Eds. New York: Springer, 2005, pp. 33–65.

Mateusz Dzida received the M.Sc. degree in computer science in 2003 and the Ph.D. degree in telecommunications in 2009, both from the Warsaw University of Technology, Poland. Currently he is an Assistant of the Switching and Computer Networks Division in the Institute of Telecommunications (IT) at the Warsaw University of Technology. His research interests focus on designing telecommunication networks. He is an author and co-author of several research articles in international journals.
e-mail: mdzida@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Michał Pióro received the Ph.D. degree in telecommunications in 1979 and the D.Sc. degree in 1990, both from the Warsaw University of Technology, Poland. He is a Professor and Head of Division of Computer Networks and Switching at the Institute of Telecommunications, Warsaw University of Technology, and a Full Professor at the Lund University, Sweden. In 2002 he received a Polish State Professorship. His research interests concentrate on modeling, design and performance evaluation of telecommunication systems. He is an author of four books and more than 150 technical papers presented in the telecommunication journals and conference proceedings. He has lead many research projects for telecom industry in the field of network modeling, design, and performance analysis.
e-mail: mpp@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Michał Zagożdżon received the M.Sc. degree in computer science in 2003 and the Ph.D. degree in telecommunications in 2009, both from the Warsaw University of Technology, Poland. Curretly he is an Assistant at the Switching and Computer Networks Division in the Institute of Telecommunications (IT) at the Warsaw University of Technology. His research interests focus on modeling and design of telecommunication networks. He is an author and co-author of several research articles in international journals.
e-mail: mzagozdz@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Włodzimierz Ogryczak and **Tomasz Śliwiński** – for biography, see this issue, p. 13.

Paper

Hierarchical Multiobjective Routing in MPLS Networks with Two Service Classes – A Meta-Heuristic Solution

Rita Girão-Silva, José Craveirinha, and João Clímaco

Abstract—The paper begins by reviewing a two-level hierarchical multicriteria routing model for MPLS networks with two service classes (QoS and BE services) and alternative routing, as well as the foundations of a heuristic resolution approach, previously proposed by the authors. Afterwards a new approach, of meta-heuristic nature, based on the introduction of simulated annealing and tabu search techniques, in the structure of the dedicated heuristic, is described. The application of the developed procedures to a benchmarking case study will show that, in certain initial conditions, this approach provides improvements in the final results especially in more "difficult" situations detected through sensitivity analysis.

Keywords— MPLS-Internet, multiobjective optimization, routing models, simulated annealing, tabu search.

1. Introduction and Motivation

Modern multiservice network routing functionalities have to deal with multiple and heterogeneous quality of service (QoS) requirements. This led to routing models designed to calculate and select one (or more) sequences of network resources (routes), satisfying certain QoS constraints and seeking the optimization of route related objectives. There are potential advantages in formulating important routing problems in these types of networks as multiple objective optimization problems, as these multiple objective formulations enable the trade-offs among distinct performance metrics and other network cost function(s) to be pursued in a consistent manner.

The interest in the application of multicriteria approaches to routing models in communication networks has been fostered mainly by the increasing relevance of QoS issues in the new technological platforms of multiservice networks.

An in-depth methodological analysis of issues raised by the use of multicriteria analysis in telecommunication network design and their relation with knowledge theory models is given in [1]. A review on multicriteria models in telecommunication network design problems including a section on routing models is in [2]. A recent overview on multicriteria routing models in telecommunication networks with a case study is presented in [3].

In particular, a significant number of routing models of multicriteria nature has been proposed in the context of the emergent multiprotocol label switching (MPLS) Internet networks – see [3]. This has to do mainly with the capability of implementing multiple connection-oriented services with QoS requirements. This technology is based on the introduction of label switching routers (LSRs) in the MPLS network that forward the packets (grouped in forward equivalence classes – FECs), through the so-called label switched paths (LSPs) by using a specific packet label switching technique. As a result of this and other technical capabilities of MPLS, advanced QoS-based routing mechanisms can be implemented, in particular involving "explicit routes" (i.e., routes completely determined at the originating node) for each traffic flow of a given service type.

A discussion on key methodological and modeling issues associated with route calculation and selection in MPLS networks and the proposal of a meta-model for hierarchical multiobjective network-wide routing in MPLS networks, were presented in [4]. This meta-model is associated with a network-wide multiobjective routing optimization approach of a new type. Two types of traffic flows are considered: firstly QoS type flows (first priority flows) such that, when accepted by the network, have a guaranteed QoS level, related to the required bandwidth; secondly best effort (BE) flows, that are considered in the model as second priority flows, and are carried by the network in order to obtain the best possible QoS level. The routing model incorporates an alternative routing principle: when a first choice route (corresponding to a loopless path) assigned to a given micro-flow[1], in a specific traffic flow (corresponding to a MPLS "traffic trunk") is blocked a second choice route may be attempted.

In the present model, described in detail in [5], the first priority objective functions concern network level objectives of QoS type flows, namely the total expected revenue and the maximal value of the mean blocking of all types of QoS traffic flows; the second priority objective functions are related to performance metrics for the different types of QoS services and the total expected revenue for the BE traffic flows. The traffic flows in the network are represented in an approximate stochastic form, based on the use of the concept of effective bandwidth for macro-flows and on a generalized Erlang model for estimating the blocking probabilities in the arcs, as in the model used in [6], [7].

The theoretical foundations of a specialized heuristic strategy for finding "good" compromise solutions to the very complex bi-level routing optimization problem, were also presented in [5]. In [8], a heuristic approach (HMOR-S2 – hierarchical multiobjective routing with two service classes)

[1] A micro-flow corresponds in our model to a "call", that is, a connection request with certain features.

devised to find "better" solutions to this hierarchical multiobjective routing optimization problem, was proposed and applied to a test network used in a benchmarking case study, for various traffic matrices.

This work presents a new approach, of meta-heuristic nature, that aims at finding even "better" solutions to the above hierarchical multiobjective routing optimization problem namely in very specific situations where sensitivity analysis showed that there was the potential for some improvement(s) in the first level objective functions. The basis of the approach is the following: beginning with the analytic results obtained after one run of the HMOR-S2 heuristic, a further run is executed, this time by using a new algorithm that includes a meta-heuristic strategy, namely, a simulated annealing (SA) or a tabu search (TS) strategy (see, e.g., [9], [10]).

The developed meta-heuristic procedures seek to make the most of the knowledge acquired with the problem by previous experimentation with the specialized heuristic HMOR-S2 and aim to overcome possible limitations of this heuristic detected through sensitivity analysis. We can say that the essence of the motivation underlying this work was to make the most of the previously developed substantive or core model (in the sense defined in the theory on model-based decision support [11]) on hierarchical multicriteria network-wide routing optimization, described in [4], [5], by incorporating new OR tools (namely SA and TS) in the previously developed heuristic resolution approach. That is, we tried to make the most of a synthesis of knowledge about a given automated routing decision model, acquired through theoretical analysis and extensive experimentation.

The paper is organized as follows. The two-level hierarchical multiobjective alternative routing model with two service classes is reviewed in Section 2, together with the basis of the dedicated heuristic. In Section 3, the features of the application of the two meta-heuristic techniques SA and TS, in the context of the heuristic approach, are presented. The formal description of the proposed specialized meta-heuristics applied to the routing problem are also described in Section 3. The results obtained with these procedures, by using analytic and discrete-event simulation experiments for a test network used in a benchmarking study, are revealed in Section 4. Finally, conclusions are drawn and future work is outlined in Section 5.

2. Review of the Multiobjective Routing Model

2.1. The Multiobjective Routing Model

As previously mentioned the considered model is an application of the multiobjective modeling framework for MPLS networks proposed in [4]. This framework (or "meta-model") in [4] considers hierarchical optimization with up to three optimization levels. In the first priority objective functions, global network performance metrics are considered; the second priority objective functions are concerned with performance metrics for the different types of services in the network; the third priority functions refer to performance metrics for packet streams micro-flows of the carried traffic flows and are related to average delays. Traffic flows in the network are represented in a stochastic form, considering two levels of representation: "macro" level or traffic flow level, and "micro" level (corresponding to packet streams in a traffic flow). Two classes of services are considered: QoS, that is services with guaranteed QoS levels (when accepted by the network), and BE, corresponding to traffic flows that are routed having in mind to obtain the best possible quality of service but not at the cost of deteriorating the QoS of the QoS traffic flows. This implies that QoS flows are treated as first priority traffic flows. The different service types of each class are represented through the sets $\mathscr{S}_Q$ (for QoS service types) and $\mathscr{S}_B$ (for BE service types). Note that the traffic flows of each service type $s \in \mathscr{S}_Q$ or $s \in \mathscr{S}_B$ may differ in important attributes, in particular the required bandwidth.

The model now reviewed is a simplification of the general model for QoS and BE service classes outlined in [4, Subsection 3.3], where only the macro level traffic stochastic representation was considered. In this simplification, the additional complexity which would result from the inclusion of a third optimization level in the routing model, as well as the corresponding additional computational burden associated with the stochastic model for calculating average delays, can be avoided. Therefore, the hierarchical multiobjective routing optimization model has two levels with several objective functions in each level. The first level (first priority) includes objective functions formulated at the network level for the QoS traffic, namely the expected revenue and the worst average performance among QoS services. In the second level the objective functions are concerned with average performance metrics of the QoS traffic flows associated with the different types of QoS services as well as the expected revenue of the BE traffic.

This is a network-wide[2] routing optimization approach, which takes into account the nature of the formulated objectives, enabling a full representation of the relations between the objective functions, taking into account the interactions between the multiple traffic flows associated with different services.

Also note that in this model, "fairness" objectives are explicitly considered at the two levels of optimization, in the form of min-max objectives. These objective functions seek to make the most of the proposed multiobjective formulation.

In the model the network is represented through a capacitated directed graph, where a capacity C_k is assigned to every arc (or "link") l_k, and the traffic flows are represented in a stochastic form, as shown in [4]. A traffic flow is specified by $f_s = (v_i, v_j, \overline{\gamma}_s, \overline{\eta}_s)$ for $s \in \mathscr{S} = \mathscr{S}_Q \cup \mathscr{S}_B$ and a stochastic process is assigned to it, that is in general,

[2]This means in this context that the main objective functions of a given service class depend explicitly on all traffic flows in the network.

a marked point process. The process describes the arrivals and basic requirements of micro-flows, originated at the MPLS ingress node v_i and destinated to the MPLS egress node v_j, using some LSP. The other features of the traffic flow are characterized by the vectors of "attributes" $\overline{\gamma}_s$ and $\overline{\eta}_s$, for service type s. The vector $\overline{\gamma}_s$ represents the traffic engineering attributes of flows of service type s and the vector $\overline{\eta}_s$ enables the description of mechanism(s) of admission control to all arcs l_k in the network by calls of flow f_s. In particular these attributes include information on the required *effective bandwidth* d_s and the mean duration $h(f_s)$ of each micro-flow in f_s. The use of the concept of effective bandwidth (a concept developed in [12]) in the present context (MPLS networks with explicit routes) was earlier considered by [6] and in [7], [13]. The effective bandwidth can be viewed as a stochastic measure of the utilization of network resources allowing for an approximate, although effective, representation of the effects of the variability of the rates of traffic sources of different types, as well as the effects of statistical multiplexing of different traffic flows in a network.

A teletraffic model, that underlies the routing model, enables the calculation of node to node blocking probabilities $B(f_s)$ for all flows f_s of all service types, from which the average blocking probability B_{ms}, for all traffic flows of type s, can be estimated for a given set of routes for all offered traffic flows. The maximal average blocking probability among all QoS service types, $B_{Mm|Q}$, is

$$B_{Mm|Q} = \max_{s \in \mathscr{S}_Q} \{B_{ms}\}. \tag{1}$$

This will represent the fairness objective at the network level, as a first priority objective function.

The total expected network revenues, W_Q and W_B associated with QoS and BE traffic flows, respectively, are expressed in terms of the expected revenues $w(f_s)$ per call[3] of flow f_s, and of the values of carried traffic A_s^c, for all service types:

$$W_{Q(B)} = \sum_{s \in \mathscr{S}_{Q(B)}} W_s = \sum_{s \in \mathscr{S}_{Q(B)}} A_s^c w_s.$$

The usual simplification, $w(f_s) = w_s, \forall f_s \in \mathscr{F}_s$, where $\mathscr{F}_s$ is the set of traffic flows of type s, will be considered. The total expected revenue for the traffic flows of QoS type W_Q is a first priority objective function together with the maximal blocking probability for all QoS service types, $B_{Mm|Q}$, given in Eq. (1), while the total expected revenue for the BE traffic flows, W_B, will be a second level objective function. Therefore, the routing of BE traffic, in a quasi-stationary situation, will not be made at the cost of the decrease in revenue or at the expense of an increase in the maximal blocking probability of QoS traffic flows. Nevertheless, it is important to note that while QoS and BE traffic flows are treated separately in terms of objective functions so as to take into account their different priority in the routing optimization, the interactions among all traffic flows are fully represented in the model. This is guaranteed by the used traffic modeling approach, underlying the optimization model, because the traffic model used to obtain the blocking probabilities $B(f_s)$ integrates the contributions of all traffic flows which may use every link of the network. This feature is a major difference in comparison with more common routing models that have been proposed for networks with two service classes, based on some form of decomposition of the network representation, corresponding to "virtual networks", one for each service class.

The second level of optimization includes the BE expected revenue, and $2|\mathscr{S}_Q|$ objective functions related to all QoS service types, the mean blocking probabilities for flows of type $s \in \mathscr{S}_Q$,

$$B_{ms|Q} = \frac{1}{A_s^o} \sum_{f_s \in \mathscr{F}_s} A(f_s) B(f_s),$$

where A_s^o is the total traffic offered by flows of type s and $A(f_s)$ is the mean traffic offered associated with f_s (in Erlang), and the maximal blocking probability $B_{Ms|Q}$, defined over all flows of type $s \in \mathscr{S}_Q$,

$$B_{Ms|Q} = \max_{f_s \in \mathscr{F}_s} \{B(f_s)\}.$$

This function constitutes the fairness objective defined for every service type $s \in \mathscr{S}_Q$.

Therefore the considered two-level hierarchical optimization problem for two service classes is depicted in Fig. 1.

1st level	QoS: network objectives:	$\min_{\overline{R}}\{-W_Q\}$	
		$\min_{\overline{R}}\{B_{Mm\|Q}\}$	
2nd level	QoS: service objectives:	$\min_{\overline{R}}\{B_{ms\|Q}\}$	
		$\min_{\overline{R}}\{B_{Ms\|Q}\}$	
		$\forall s \in \mathscr{S}_Q$	
	BE: network objective:	$\min_{\overline{R}}\{-W_B\}$	

subject to equations of the underlying traffic model.

Fig. 1. Problem P-M2-S2.

The decision variables $\overline{R}$ represent the network routing plans, that is, the set of all the feasible routes (i.e., node to node loopless paths) for all traffic flows. The acronym P-M2-S2 stands for "**P**roblem – **M**ultiobjective with **2** optimization hierarchical levels – with **2 S**ervice classes".

The basic teletraffic sub-model allows for the blocking probabilities B_{ks}, for micro-flows of service type s in link l_k, to be given in the form $B_{ks} = \mathscr{L}_s\left(\overline{d_k}, \overline{\rho_k}, C_k\right)$. Here $\mathscr{L}_s$ represents the basic function (implicit in the teletraffic analytical model) that expresses the marginal blocking probabilities, B_{ks}, in terms of $\overline{d_k} = (d_{k1}, \cdots, d_{k|\mathscr{S}|})$ (vector of equivalent effective bandwidths for all service types), $\overline{\rho_k} = (\rho_{k1}, \cdots, \rho_{k|\mathscr{S}|})$ (vector of reduced traffic loads ρ_{ks} offered by flows of type s to l_k) and the link capacity C_k.

This type of approximation was suggested in [6] for off-line single-objective multiservice routing optimization models

[3]The term 'call' means a node to node connection request with certain traffic engineering features.

and was also used in the multiobjective dynamic alternative routing model proposed in [7]. It enables the calculation of $\{B_{ks}\}$ through efficient numerical algorithms. We should stress that very efficient and robust approximations have to be used in a network-wide routing optimization model of the type associated with P-M2-S2, for tractability reasons.

2.2. Basis of the Heuristic Approach

The dedicated heuristic resolution approach that is the starting point for the meta-heuristics analyzed in this paper uses the theoretical foundations described by the authors in [5], which will now be reviewed.

In the hierarchical multiobjective routing problem P-M2-S2 an alternative routing principle is used. This means that the network routing plans $\overline{R} = \cup_{s=1}^{|\mathscr{S}|} R(s)$ (decision variables) for all the network services, where $R(s) = \cup_{f_s \in \mathscr{F}_s} R(f_s), s \in \mathscr{S}_Q \cup \mathscr{S}_B$ are such that $R(f_s) = (r^p(f_s)), p = 1, \cdots, M$ with $M = 2$ in our model. It is assumed that for each flow f_s the first choice route $r^1(f_s)$ will be used unless it is blocked because one of its links l_k does not have the required available bandwidth d_s (or a call is not accepted according to the probabilistic availability function ψ_{ks}). If $r^1(f_s)$ is blocked the routing method makes the current connection request attempt the second choice route $r^2(f_s)$. This request will be blocked only if $r^2(f_s)$ is also blocked. If $M > 2$, routes $r^3(f_s), \cdots, r^M(f_s)$ would be attempted in this order.

The high "complexity" of the routing problem P-M2-S2 stems from two major factors: all objective functions are strongly interdependent (via the $\{B(f_s)\}$), and all the objective function parameters and (discrete) decision variables $\overline{R}$ (network route plans) are also interdependent. Note that all these interdependencies are defined explicitly or implicitly through the underlying traffic model. Regarding computational complexity, it must be remarked that the simplest, "degenerated" single objective version of the problem, that is, concerning a model with a single objective function W_Q, one single service and no alternative routing ($M = 1$) is NP-complete in the strong sense, as shown in [14]. The addressed problem may be viewed as a bi-level, multiobjective extension of this type of problem.

Concerning the possible conflict between the objective functions in P-M2-S2, it should be observed that in many routing situations, the maximization of W_Q leads to a deterioration on some $B(f_s), s \in \mathscr{S}_Q$, for certain traffic flows $A(f_s)$ with low intensity, and this tends to increase $B_{Ms|Q}$ and, consequently, $B_{Mm|Q}$. In single-objective routing models this aspect is usually addressed by imposing upper bounds on the values $B(f_s)$. This is a major factor to justify the interest and potential advantage in using multiobjective approaches when dealing with this type of routing methods.

The resolution (in a multicriteria analysis sense) of the routing problem P-M2-S2 was earlier performed by a heuristic procedure in [8], which is briefly reviewed in this section. This heuristic is an improved version of the heuristic approach described in [5] and it is based on the recurrent calculation of solutions to a constrained bi-objective shortest path problem, formulated for every end-to-end flow f_s:

$$\text{problem } \mathscr{P}_{s2}^{(2)} : \min_{r(f_s) \in \mathscr{D}(f_s)} \left\{ m^n(r(f_s)) = \sum_{l_k \in r(f_s)} m_{ks}^n \right\}_{n=1;2} . \tag{2}$$

The path metrics m^n to be minimized are the marginal implied costs $m_{ks}^1 = c_{ks}^{Q(B)}$ (the definition of which is reviewed in the following analysis) and the marginal blocking probabilities $m_{ks}^2 = -\log(1 - B_{ks})$; $\mathscr{D}(f_s)$ is the set of all feasible loopless paths for flow f_s, which satisfy specific traffic engineering constraints (other than the effective bandwidth) for flows of type s. A typical constraint is a maximal number of arcs per path depending on the class and type of service s. The logarithmic function is just used to transform the blocking probability into an additive metric. The link cost coefficients $m_{ks}^1 = c_{ks}^{Q(B)}$ are then used in problems of form Eq. (2), when candidate solutions have to be obtained to seek the improvement of the revenue of the QoS (BE) traffic, in different steps of the heuristic procedure. According to this approach, the comparison of the efficiency of different candidate routes in the context of a multicriteria routing framework of this type should take into account both the loss probabilities experienced along the candidate routes and the knock-on effects upon the other routes in the network, effects associated with the acceptance of a call on that given route. Such effects can be measured exactly through the implied costs.

It is important to note that this auxiliary constrained bi-objective shortest path problem was used as a basis of the heuristic approach having in mind that the consideration of the metric blocking probability tends, at a network level, to minimize the maximal node-to-node blocking probabilities $B(f_s)$, while the metric implied cost tends to maximize the total average revenue W_T in a single class multiservice loss network (see [15], [16]).

Concerning the implied cost c_{ku} (resulting from the acceptance of a call of flow f_u in link l_k) this is an important mathematical concept in routing optimization in loss networks which was originally proposed by Kelly [17] for single-rate traffic networks. The definition was later extended to single route multirate traffic networks in [6], [18]. The implied cost can be viewed as the expected value of the loss of revenue in all traffic flows which may use link l_k, resulting from the acceptance of a connection request from f_u stemming from the decrease in the capacity of this link. Therefore we can say that the implied cost measures in a probabilistic manner the knock-on effects on all network routes (of all traffic flows) associated with the acceptance of a call from f_u in a link l_k. In [19], the definition of c_{ku} was adapted to multirate loss networks with alternative routing by extending the model for single-service networks given in [17]. The extension of this definition to a multi-rate network with alternative routing and two service classes was proposed in [5]. For this purpose the following definition of marginal implied costs associated with QoS (BE) traffic was put forward [5]. The *marginal*

implied cost for QoS (BE) traffic, $c_{ku}^{Q(B)}$, associated with the acceptance of a connection (or "call") of traffic f_u of any service type $u \in \mathscr{S}$ on a link l_k is defined as the expected value of the traffic loss induced on all QoS (BE) traffic flows resulting from the capacity decrease in link l_k.

In [5], a conjecture was presented, implying the marginal implied costs for QoS (BE) traffic can be obtained by solving a system of equations:

$$c_{ku}^{Q(B)} = \sum_{s \in \mathscr{S}_{Q(B)}} \frac{\zeta_{kus}}{1-B_{ks}} \left[\sum_{f_s \in \mathscr{F}_s : l_k \in r^1(f_s)} \lambda_{r^1(f_s)} \left(s_{r^1(f_s)}^{Q(B)} + \right. \right.$$
$$\left. \left. + c_{ks}^{Q(B)} \right) + \sum_{f_s \in \mathscr{F}_s : l_k \in r^2(f_s)} \lambda_{r^2(f_s)} \left(s_{r^2(f_s)}^{Q(B)} + c_{ks}^{Q(B)} \right) \right], \tag{3}$$

with

$$s_{r^2(f_s)}^{Q(B)} = w^{Q(B)}(f_s) - \sum_{l_j \in r^2(f_s)} c_{js}^{Q(B)},$$

$$s_{r^1(f_s)}^{Q(B)} = w^{Q(B)}(f_s) - \sum_{l_j \in r^1(f_s)} c_{js}^{Q(B)} - (1 - L_{r^2(f_s)}) s_{r^2(f_s)}^{Q(B)},$$

$$\zeta_{kus} = \mathscr{L}_s\left(\overline{d_k}, \overline{\rho_k}, C_k - d_{ku}\right) - \mathscr{L}_s\left(\overline{d_k}, \overline{\rho_k}, C_k\right),$$

where $s_{r^p(f_s)}^{Q(B)}$ denotes the surplus value of a call on route $r^p(f_s)$, $\lambda_{r^p(f_s)}$ is the marginal traffic carried on $r^p(f_s)$ by flow f_s, $L_{r^p(f_s)}$ represents the blocking probability for calls of f_s on route $r^p(f_s)$ ($p = 1;2$) (considering that $r^1(f_s)$ and $r^2(f_s)$ are arc-disjoint paths) and ζ_{kus} is the increase in call blocking probability for type s calls on link l_k resulting from a decrease in the capacity of l_k associated with the acceptance of a type u call. The coefficients $w^{Q(B)}(f_s)$ are the marginal expected revenues per call of f_s, such that $w^Q(f_s) + w^B(f_s) = w(f_s)$ and can be written as $w^{Q(B)}(f_s) = \alpha^{Q(B)} w(f_s)$, in terms of the coefficients $\alpha^{Q(B)} \in]0.0;1.0[$ which satisfy the normalization condition $\alpha^Q + \alpha^B = 1.0$.

A system of implicit non-linear equations can be defined in order to calculate the B_{ks} in terms of link capacities (matrix $\overline{C} = [C_k]$), the offered traffic matrix $\overline{A} = [A(f_s)]$, and the current network routing solution $\overline{R}$,

$$B_{ks} = \beta_{ks}(\overline{B}, \overline{C}, \overline{A}, \overline{R}), \tag{4}$$

with $k = 1, \cdots, |\mathscr{L}|; s = 1, \cdots, |\mathscr{S}|$ and $\overline{B} = [B_{ks}]$. Concerning the calculation of $c_{ks}^{Q(B)}$ through Eq. (3), it implies the resolution of a system of equations of the general form:

$$c_{ks}^{Q(B)} = \kappa_{ks}^{Q(B)}(\overline{c}, \overline{B}, \overline{C}, \overline{A}, \overline{R}), \tag{5}$$

where $\overline{c} = [c_{ks}^{Q(B)}]$. The numerical resolution of these two systems of equations in B_{ks} and $c_{ks}^{Q(B)}$ is performed by fixed point iterators, given the matrices $\overline{C}, \overline{A}$ and $\overline{R}$.

In the heuristic, the auxiliary constrained shortest path problem $\mathscr{P}_{s2}^{(2)}$ Eq. (2) is solved by the algorithm MMRA-S2 [5], an adaptation of a previously developed algorithmic approach, MMRA-S (modified multiobjective routing algorithm for multiservice networks), described in [7], [19]. Generally, there is no feasible solution which minimizes the two objective functions simultaneously. Hence, the resolution of this routing problem aims at finding a "best" compromise path from the set of non-dominated solutions, according to some system of preferences. In this context, path computation and selection have to be fully automated. Therefore the system of preferences is embedded in the working of the algorithm MMRA-S2. This is implemented by defining preference regions in the objective function space obtained from aspiration and reservation levels (preference thresholds) defined for the two objective functions [15], [16]. Further details on this algorithmic approach can be seen in [7].

Another important part of the addressed routing model is the underlying traffic model. This stochastic traffic model involves all the sub-models and associated numerical procedures, that are needed for obtaining all traffic related parameters, namely implied costs and blocking probabilities B_{ks} and $B(f_s)$, under certain simplifying assumptions.

A description of the traffic modeling approach used in the routing model can be seen in [4].

Now let us review the basic features of the dedicated heuristic HMOR-S2, taken as the starting point and reference procedure in the present work.

In the heuristic, a basic searching strategy is to seek for routing solutions $\overline{R}(s)$ for each service $s \in \mathscr{S}$, in order to achieve a better performance in terms of W_B, $B_{ms|Q}$ and $B_{Ms|Q}$, $s \in \mathscr{S}_Q$ while respecting the hierarchy of objective functions. This also means that network resources are left available for traffic flows of other services so that the solutions selected at each step of the procedure may improve the first priority objective functions W_Q and $B_{Mm|Q}$. The heuristic was designed in order to seek, firstly for each QoS service and starting from the services with higher effective bandwidth (considering the numbering of s, $s = 1, \cdots, |\mathscr{S}_Q|$) and, secondly, for each BE service (also beginning by the higher bandwidth services, $s = |\mathscr{S}_Q| + 1, \cdots, |\mathscr{S}|$), solutions which dominate the current one, in terms of $B_{ms|Q}$ and $B_{Ms|Q}$ for QoS services and in terms of W_B for BE services. These solutions will only be accepted if they do not lead to the worsening of any of the network functions W_Q and $B_{Mm|Q}$.

Another basic idea of the heuristic is the generation of candidate solutions $(r^1(f_s), r^2(f_s))$ for each f_s, using the mentioned algorithm MMRA-S2, and their possible selection through specific criteria, to be "tuned" throughout the execution of the heuristic. A maximal number of arcs D_s per route for each service type s is previously defined and a feasible route set $\mathscr{D}(f_s)$ is obtained for each f_s. For example, for real time QoS services, D_s is equal to the network diameter; for the non-real time QoS services, D_s is the network diameter plus 1, while for the BE services, no limits are imposed on D_s.

Note that special rules had to be constructed for the selection of candidate first choice routes $r^1(f_s)$ taking into account the network topology and the need to make a distinction between real time QoS services (typically video

and voice services) and non-real time QoS services (for example "premium data" service). These rules are described in [5].

Concerning the calculation of candidate second choice routes $r^2(f_s)$ for QoS or BE traffic, the MMRA-S2 procedure is used. Having in mind to prevent performance degradation in overload conditions, these alternative routes should be eliminated in certain conditions. This is achieved through a mechanism designated as alternative path removal (APR), an adaptation of the mechanism originally proposed in [7], [20].

The theoretical analysis of the model, confirmed by experimentation, showed that successive application of MMRA-S2 to every traffic flow does not lead to an effective resolution approach to the network routing problem P-M2-S2. This results from an instability phenomenon that arises in such path selection procedure, expressed by the fact that the route sets $\overline{R}$ often tend to oscillate between certain solutions some of which may lead to poor global network performance under the prescribed metrics.

Therefore, another core idea of the heuristic approach (similarly to multiobjective dynamic routing method for multiservice – MODR-S) [7] is the search for the subset of the path set $\overline{R}^a = \cup_{s=1}^{|\mathscr{S}|} \overline{R}^a(s) : \overline{R}^a(s) = \{(r^1(f_s), r^2(f_s)), f_s \in \mathscr{F}_s\}$ the elements of which should be possibly changed in the next route improvement cycle. Detailed analysis and extensive experimentation with the heuristic led to the proposal of a criterion for choosing candidate paths for possible routing improvement by increasing order of a function $\xi(f_s)$ of the current $(r^1(f_s), r^2(f_s))$, given in [8]. The use of this criterion considers two search cycles, where $\xi(f_s) = F_L(f_s)$ in the first cycle and $\xi(f_s) = F_C^{Q(B)}(f_s)$ in the second cycle, if the effect over QoS (BE) traffic is being considered, with

$$F_C^{Q(B)}(f_s) = (n_2 - n_1)c_1^{'Q(B)} + c_{r^1(f_s)}^{Q(B)} - c_{r^2(f_s)}^{Q(B)},$$

$$c_{r(f_s)}^{Q(B)} = \sum_{l_k \in r(f_s)} c_{ks}^{Q(B)},$$

$$c_1^{'Q(B)} = \frac{1}{n_1} \sum_{l_k \in r^1(f_s)} c_{ks}^{Q(B)} = \frac{1}{n_1} c_{r^1(f_s)}^{Q(B)},$$

$$F_L(f_s) = 1 - L_{r^1(f_s)} L_{r^2(f_s)}.$$

The aim of $F_C^{Q(B)}(f_s)$ is to give preference (concerning the potential value in changing the second choice route when seeking to improve W_Q or W_B) to the flows for which the route $r^1(f_s)$ has a low implied cost and the route $r^2(f_s)$ has a high implied cost. The factor $(n_2 - n_1)$ was introduced for normalization purposes, considering that $r^1(f_s)$ has n_1 arcs and $r^2(f_s)$ has n_2 arcs. The aim of $F_L(f_s)$ is to give preference to the choice of the flows which currently have worse end-to-end blocking probability given by $L_{r^1(f_s)} L_{r^2(f_s)}$.

Another key point tackled by the heuristic is the specification of a variable *nPaths*, which represents the number of routes with smaller values of $\xi(f_s)$ that should possibly be changed by running MMRA-S2 once again. In order to do so, the effect of each candidate route on the relevant objective functions is anticipated by solving the corresponding analytical model.

The full description and formalization of this heuristic as well as an application study are given in [8].

3. Developed Meta-Heuristics

The study of the heuristic approach HMOR-S2, the basis of which was reviewed in the previous section, was completed with a sensitivity analysis, which led to the consideration of variants of this heuristic. In the report [21], two variants to the HMOR-S2 were described, firstly the HMOR-S2$_R$ where a floating relaxation was imposed on one of the first level objective function values, and secondly the HMOR-S2$_B$ where a floating barrier was imposed on one of the first level objective function values. Extensive experimental analysis was carried out for those variants and a simulation study was also conducted. The main results of the sensitivity analysis and the SA and TS-based variants of the heuristic are now described.

3.1. Sensitivity Analysis

The purpose of the sensitivity tests applied to the HMOR-S2 heuristic was to check whether the heuristic was treating the lower level objective functions in a balanced way (that is, to check whether better values of the second level objective functions could be obtained without worsening the values of the first level objective functions) and to check whether the value of an upper level objective function could be improved at the cost of worsening the value of the other upper level objective function.

In the first set of tests, either an upper bound was imposed on one of the blocking probability functions B_{ms} or B_{Ms}, $s \in \mathscr{S}_Q$, or a lower bound was imposed on the BE traffic revenue W_B, $s \in \mathscr{S}_B$. These bounds constitute barriers, in the sense that they are more demanding than the corresponding values obtained at the end of the HMOR-S2 run.

In the second set of tests (relaxation tests), the focus was on the first level objective functions. In one of the tests, the blocking function $B_{Mm|Q}$ is no longer treated as an objective function and an upper bound on its value is imposed. This upper bound is less demanding than the corresponding value $[B_{Mm|Q}]_{\text{basis}}$ obtained at the end of the HMOR-S2 run. The purpose of this test is to check whether the QoS traffic revenue can still be improved by relaxing the value of the other main objective function. In the other test, the QoS services revenue W_Q is no longer treated as an objective function and a lower bound on its value is imposed. This lower bound is less demanding than the corresponding value $[W_Q]_{\text{basis}}$ obtained at the end of the HMOR-S2 run. The purpose of this test is to check whether the blocking function $B_{Mm|Q}$ can be improved when the value of the other objective function is relaxed.

Generally speaking, the results of the sensitivity tests for the HMOR-S2 heuristic were as expected, allowing us to assume that the heuristic is balanced in the treatment of the different objective functions. Nonetheless, there are a few results that are worth mentioning.

In the first set of tests, one or both of the upper level objective function values were worse when a barrier (i.e., a stricter value) was imposed on one of the lower level blocking probability functions or BE traffic revenue. That is, when the improvement of one of the lower level functions is imposed, the upper level objective function values tend to be worse (at least for one of those functions). There was however one situation where one of the first level objective functions improved and the other worsened. This result is not unexpected, as the two first level objective functions are conflicting in nature, but showed that there was one non-dominated solution that the basic heuristic was not able to detect so far.

In the second set of tests, in one of the sensitivity tests where the upper level objective function $B_{Mm|Q}$ ceased to be treated in the heuristic as an objective function and a relaxed upper bound was imposed on its value, a final solution with slightly better values for both $B_{Mm|Q}$ and W_Q was obtained. Therefore, in spite of allowing the value of $B_{Mm|Q}$ to increase beyond the value obtained when the basic heuristic was run, it actually diminished, and there was a slight improvement of the QoS traffic revenue. This result suggests that, in some rare cases, the heuristic is not capable of finding a solution that slightly dominates the current selected solution.

In order to try to obtain solutions with even better values for both the upper level objective functions in these very specific types of situations, new approaches were devised. These new approaches consist of the introduction of metaheuristic techniques (SA and TS) in the structure of the basic heuristic HMOR-S2.

3.2. Application of a SA Technique to the Basic Heuristic

The SA technique can be viewed as a variant of the heuristic technique of local neighbourhood search, where a subset of feasible solutions is explored in the neighbourhood of the current solution. In an optimization problem, the traditional implementations of local search always try to move towards an improvement of the objective function. However, with this type of strategy, the risk of remaining in a local optimum is high. The SA technique tries to prevent this from happening, by allowing solutions with worse values of the objective function (when compared with the value of that function in the current solution) to be taken into account. These moves towards worse solutions are done in a controlled way, and with the purpose of avoiding local minima or maxima. The probability of acceptance of a solution that is actually worse than the current solution is controlled by the variation of the objective function value and a parameter, a so-called temperature T, related to the state of the system, in particular related to the number of iterations that have occurred since the beginning of the search procedure.

A generic SA algorithm for a single objective problem, where a minimization problem is considered, with solution space S, objective function f and neighbourhood structure N, can be seen, for example, in [22].

The SA technique has been successfully used to solve many different optimization problems. This technique is easy to implement, it can be applied to a great diversity of combinatorial optimization problems and usually it allows for the calculation of adequate solutions [22]. However, in order to get good solutions, many parameters have to be carefully tuned: the cooling function $\vartheta(T)$, the neighbourhood area (based on the specific features of the problem to be solved), the probability function of acceptance of the new solution, the number of iterations *nrep* and the stopping condition. Another disadvantage, apart from the need to carefully tune the system parameters, is the execution time of the SA algorithms that tends to be very long. Experiences from many authors actually show that for a specific and well-defined problem, an algorithm specifically tailored to that problem tends to provide better results than a SA algorithm [22]. Nevertheless, many authors have applied SA techniques to telecommunication network optimization problems, such as network design and routing problems – see for instance [23]–[34].

Introduction of a SA technique in the HMOR-S2 heuristic. Many issues had to be addressed to formulate this SA-based variant, HMOR-S2$_{SA}$. Firstly the basic technique of SA had to be adapted to a hierarchical multiobjective problem. A choice was made to work only with the upper level objective functions and two different SA processes were considered simultaneously. The lower level objective functions are used as in the basic heuristic, that is, their value for the specific service under scrutiny has to improve so that the new solution may be taken into account in further steps.

Firstly, the initial temperature has to be specified. It should be high in order to guarantee that the final solution of the problem does not depend heavily on the initial solution. A high initial temperature also assures a certain diversity of solutions, which is advantageous on the initial stages of the resolution approach. Remember that the temperature decays throughout the heuristic procedure, which causes the probability of accepting new solutions that are actually worse than the current solution to diminish. This provides an intensification strategy, which should be correct for the final stages of the HMOR-S2$_{SA}$. Note that diversification-like and intensification-like strategies are already being used in the basic dedicated heuristic, HMOR-S2, as the parameter *nPaths* (that represents the number of paths that can change from the current solution to the new one) starts with a high value (that is, the new solution can be quite diverse from the current one) and decays throughout the algorithm, which means that the paths remain the same for an increasing number of origin-destination pairs. As two SA sub-algorithms are considered simultaneously, two dif-

ferent initial temperatures have to be defined, in particular, one associated with the QoS services revenue, $T_W^0 = W_Q^0 = W_Q^{initial}$, and the other associated with the blocking probability function $B_{Mm|Q}$, $T_B^0 = B_{Mm|Q}^0 = B_{Mm|Q}^{initial}$.

The features of the neighbourhood area of the current solution have to be defined. In this implementation the features of the neighbourhood change throughout the procedure. Note that this is already being made in the basic heuristic, as the portion of the state space where new feasibe solutions are sought, is defined according to the flows for which the paths may change in the current iteration. Therefore, not only the neighbourhood, where new solutions are sought, diminishes throughout the algorithm (because of the value of *nPaths*) but also it adapts to the current conditions of the resolution procedure and it is chosen in order to search for improvements in the objective function values.

The number of iterations for each temperature value also has to be determined. For higher temperatures (initial stages of the resolution procedure), *nrep* is small; for lower temperatures (final stages of the resolution procedure), *nrep* is high, so as to seek a guarantee that the neighbourhood area is thoroughly searched and no maxima (or minima) for each main objective function remain undiscovered. The value that was considered is $nrep = \left\lceil \frac{|\overline{\mathscr{F}}|+1-nPaths}{2} \right\rceil$, where $|\overline{\mathscr{F}}| = \frac{1}{|\mathscr{S}|}\sum_{s\in\mathscr{S}}|\mathscr{F}_s|$ is the average number of traffic flows per service.

The cooling mechanism has to be devised so that the temperatures do not decay too slowly or too fast. Several experiences were conducted and the cooling functions that provided the best results were $T_W^j = \left[T_W^0\left(1-\frac{j}{J}\right)\right]^a$ and $T_B^j = \left[T_B^0\left(1-\frac{j}{J}\right)\right]^a$ in iteration j, with $J = 1000;5000$ and $a = 0.1;0.01$, for the 2 simultaneous SA procedures.

The probability of accepting a new solution that is actually worse than the most adequate solution up to the current stage of the algorithm (iteration j) is $p_W^j = \exp\left(\frac{W_Q^j - \max\{W_Q\}}{T_W^j}\right)$ and $p_B^j = \exp\left(\frac{\min\{B_{Mm|Q}\} - B_{Mm|Q}^j}{T_B^j}\right)$ for the 2 simultaneous SA procedures, where $\max\{W_Q\}$ and $\min\{B_{Mm|Q}\}$ are the upper level objective function values in the most adequate solution found so far.

The stopping criterion is the same as the one used in the basic heuristic, that is, the algorithm stops when $nPaths = 0$.

The adaptation of the SA technique to the basic heuristic HMOR-S2 can be described as depicted in Fig. 2.

The complete formalization of the meta-heuristic version of HMOR-S2 using SA, HMOR-S2$_{SA}$, is in Appendix B.1 of the report [35].

Note that one of the features of a standard SA technique is the random choice of the new solution (to be taken into account at each step of the algorithm) among all the feasible solutions in the neighbourhood of the best solution found so far. However, in the adaptation of a SA-like technique to HMOR-S2, the choice of the feasible solution to be compared with the most adequate solution found so far, is done with the help of the MMRA-S2 algorithm, as in the basic heuristic. Note that the solution provided by this auxiliary algorithm is likely to produce better results than a randomly chosen solution, taking into account the foundations of the resolution procedure, given in Section 2.

I. Let the initial temperatures be $T_W^0 = W_Q^0 = W_Q^{initial}$ and $T_B^0 = B_{Mm|Q}^0 = B_{Mm|Q}^{initial}$.
II. $j = 1$
III. Define J and a.
IV. In the iteration $j \geq 1$.
  1. Let the current temperatures be $T_W^j = \left[T_W^0\left(1-\frac{j}{J}\right)\right]^a$ and $T_B^j = \left[T_B^0\left(1-\frac{j}{J}\right)\right]^a$.
  2. Cycle to be performed *nrep* times:
    (a) Calculation of a new solution, using the MMRA-S2 bi-objective algorithm.
    (b) For the new solution, let W_Q be the expected QoS service revenue and $B_{Mm|Q}$ the maximal average blocking probability for all QoS services.
    (c) Let X_W and X_B be two r.v. following a uniform distribution in]0.0;1.0[.
    (d) If $s \in \mathscr{S}_Q$, check whether ($B_{ms} \leq \min\{B_{ms}\}$ and $B_{Ms} \leq \min\{B_{Ms}\}$). If $s \in \mathscr{S}_B$, check whether ($W_B \geq \max\{W_B\}$).
      – If so:
        A. Check whether ($W_Q > \max\{W_Q\}$ and $B_{Mm|Q} < \min\{B_{Mm|Q}\}$).
        • The solution is accepted.
        B. Otherwise, check whether $X_W < \exp\left(\frac{W_Q - \max\{W_Q\}}{T_W^j}\right)$ and whether $X_B < \exp\left(\frac{\min\{B_{Mm|Q}\} - B_{Mm|Q}}{T_B^j}\right)$.
        • The solution is accepted.
        C. Otherwise, the solution is not accepted.
      Else, the solution is not accepted.
  End of the *nrep* cycle.
  3. $j \leftarrow j+1$.

The cycle ends when all the cycles "`For` (*s*)", "`For` (*ape*)" and "`For` (*nPaths*)" have been executed.

Fig. 2. The adaptation of the SA technique to the basis heuristic HMOR-S2.

Concerning the numerical complexity of this heuristic, it can be said that the instructions in the inner cycle of the procedure are executed $C_i^{\text{HMOR-S2}_{SA}} = 4|\mathscr{S}||\overline{\mathscr{F}}|^2 + 2|\mathscr{S}||\overline{\mathscr{F}}|$ times. The numerical complexity of the heuristic in terms of the number of solutions that are analyzed is $C_s^{\text{HMOR-S2}_{SA}} = \frac{|\mathscr{S}||\overline{\mathscr{F}}|}{6}\left(2|\overline{\mathscr{F}}|^2 + 9|\overline{\mathscr{F}}| + 10\right)$. For comparison, the corresponding numerical complexities of the HMOR-S2 heuristic approach (see [21]) are $C_i^{\text{HMOR-S2}} = 4|\mathscr{S}||\overline{\mathscr{F}}|$ and $C_s^{\text{HMOR-S2}} = 2|\mathscr{S}||\overline{\mathscr{F}}|(|\overline{\mathscr{F}}|+1)$. This means

that the HMOR-S2 heuristic involves a significantly lower number of calculations than HMOR-S2$_{SA}$. For further details on these calculations, see [35]. These complexity measures are an indication of the heuristic numerical complexity just at the level of the "optimization" procedures.

3.3. Application of a TS Technique to the Basic Heuristic

The TS technique is a local neighbourhood search technique applied to a dynamic neighbourhood defined in terms of the current solution and the history of the states encountered during the search up to the current instant. For example, in [10], [36] this technique is described in detail and some examples of application to different optimization problems are provided. The TS can be defined as a technique where restrictions are imposed so as to guide a search process into areas that otherwise would not be explored in the search for new solutions [10]. The restrictions are usually the exclusion of some solutions that are classified as tabu, i.e., forbidden.

The reasoning behind the TS is that the resolution of problems should include an adaptive memory and an intelligent exploration of the solution space (i.e., a guided and systematic exploration rather than a random one) [36]. An adaptive memory allows for the implementation of procedures that manage to explore the solution space in an economic and efficient way. The memory can be a short-time one and its information is used to prevent the search from remaining in a local "optimum", or it can be a long-time one and it allows for the use of intensification and diversification strategies.

A generic TS algorithm for a single objective problem, where a minimization problem is considered, with solution space S, objective function f and neighbourhood structure N, can be seen, for example, in [37].

For a successful use of the TS technique in solving many different optimization problems, many implementation choices have to be carefully made concerning key aspects: the diversification and intensification strategies, the information to be kept in memory, the neighbourhood area, the criteria to attribute a tabu status to a move (a move is a change that is imposed on a solution in order to find another different solution), the tabu tenure (i.e., the time during which a move remains tabu), the aspirational criteria and the stopping condition.

Unlike what happens in the SA technique, in the TS technique the adequate solutions are sought having in mind not only the objective function value, but also other influential factors, such as the diversification of solutions, the intensification of solutions, the aspirational criteria, the frequency of solutions and the tabu tenures.

Many authors have applied TS techniques to telecommunication network optimization problems, such as network design and routing problems – see for instance [38]–[44].

Introduction of a TS technique in the HMOR-S2 heuristic. Note that some aspects of TS-like techniques are already used in the basic heuristic. For instance, some paths are not allowed to change in certain steps (i.e., their change is tabu or forbidden). In each iteration the number of paths that can possibly change is *nPaths* and the choice of the *nPaths* flows for which the paths are liable to change is made according to the value of an auxiliary function $\xi(f_s)$ (see Subsection 2.2).

Many issues had to be addressed to formulate this variant HMOR-S2$_{TS}$. Firstly the basic technique of TS had to be adapted to a hierarchical multiobjective problem. A choice was made to focus this technique on the QoS services revenue, having in mind its central role in the system of preferences implicit in the model. In fact, given two non-dominated solutions it is usually more acceptable, from a network design point of view, to select the solution with higher QoS service revenue, at the cost of some degradation of $B_{Mm|Q}$.

The neighbourhood area where a new solution will be searched for also has to be defined. Considering a specific solution, the neighbourhood of that solution is the set of solutions that differ in the pair of routes $(r^1(f_s), r^2(f_s))$ for one flow. Therefore a move from one solution to another solution in the neighbourhood is done by choosing a new set of paths for one of the flows. The new set of paths for a flow is chosen by solving the auxiliary bi-objective shortest path problem with the MMRA-S2 algorithm. If this new set of paths for a particular flow allows for a better solution to the routing problem, then the previous set of paths for that flow becomes tabu and a move that would lead to using that previous set of paths again, is forbidden.

The tabu list is a list of moves which are tabu, so in this adaptation we consider the tabu list as a list of pairs of paths which are tabu. The maximal size of the tabu list is given by *nPaths*, which means it changes throughout the algorithm: at the beginning of the algorithm, *nPaths* is high, which means that many moves can become tabu; towards the end of the algorithm, *nPaths* decreases. New moves can be added to the tabu list and once it is full, the oldest move (at the top of the list) is withdrawn and the new move is added at the end of the list. Therefore, this list is a queue with FIFO (first-in first-out) discipline. The size of the tabu list also has an impact on the tabu tenure. Note that a tabu list is used for a specific service $s \in \mathscr{S}$ and when the algorithm proceeds to the analysis of a new service in the "services cycle" of the basic heuristic the tabu list is reinitialized.

An aspirational criterion may be defined: if the values for the upper level objective functions and for the lower level objective functions (for the service under scrutiny) of a new solution are better than the corresponding values in the most adequate solution found so far, then this new solution should always be considered as the new most adequate solution, even if it is obtained by performing a tabu move.

The information on the tabu list is kept in the memory of the resolution procedure, along with information on a vari-

I. Initialization of the frequency values $\text{freq}(f_s), \forall f_s, s \in \mathscr{S}$.
II. Cycle of services.
  1. Initialization of the tabu list, with length given by *nPaths*.
  2. Cycle in *nCycles*:
    (a) Calculation and ordering of the values of $\xi(f_s)$.
    (b) Use of MMRA-S2 to find pairs of paths for the flows f_s.
    (c) Initialization of $(W_Q(f_s) - W_Q^a) - a \cdot \text{freq}(f_s)$ for all the flows.
    (d) Cycle in *numIterations*.
      – (Search up to a maximum of *numIterations* new solutions in the neighbourhood of the current solution.)
      – Go through the ordered flows f_s according to increasing values of $\xi(f_s)$.
        A. Check whether the pair of paths proposed for the flow f_s is tabu.
        B. Keep a copy of the current pair of paths for this flow and load the new pair of paths in the solution.
        C. If the new solution is "better" than the current one (i.e., has better values for the upper level functions and for the lower level functions for the service under scrutiny).
          • If the move is tabu.
            * If the aspirational criterion is met.
              – The current solution is the most adequate up to this stage of the algorithm.
              – Increment the value of $\text{freq}(f_s)$.
            Otherwise, go back to the previous solution.
          Otherwise,
            * Increment the value of $\text{freq}(f_s)$.
            * Check whether the new solution is better than the most adequate solution up to now and if it is so, the new solution becomes the most adequate solution.
            * Add the move to the tabu list.
          • Leave the cycle of "going through the flows".
        Otherwise,
          • If the move is not tabu, keep the information on the value of $(W_Q(f_s) - W_Q^a) - a \cdot \text{freq}(f_s)$.
          • Go back to the previous solution.
        (End of the cycle of "going through the flows".)
      – If no new solution that improves the current solution was found.
        A. Choose the solution obtained with a non-tabu move, with the highest value of $(W_Q(f_s) - W_Q^a) - a \cdot \text{freq}(f_s)$.
    (End of the cycle in *numIterations*.)
  (End of the cycle in *nCycles*.)
End of the cycle of services.

Fig. 3. The adaptation of the TS technique to the basis heuristic HMOR-S2.

able $\text{freq}(f_s)$, that gives the number of times a specific flow f_s has seen its set of paths changed throughout the algorithm. This information is associated with a long-term memory. As for the solutions that are found and explored, the only information that is kept is the one concerning the most adequate solution found up to the current stage of the algorithm.

In the inner cycle of the heuristic, if new sets of paths for all the *nPaths* flows have been considered and a solution better than the current one has not been found yet, then the solution that will be used in the next stage of the algorithm will be the one originating from a non-tabu move with the highest value of $(W_Q(f_s) - W_Q^a) - a \cdot \text{freq}(f_s)$, where $W_Q(f_s)$ is the QoS services revenue value when the set of paths for flow f_s is changed, W_Q^a is the QoS services revenue value for the current solution, and a is an empirical parameter for which a value has to be chosen. The value of $(W_Q(f_s) - W_Q^a) - a \cdot \text{freq}(f_s)$ increases with the difference $(W_Q(f_s) - W_Q^a)$ (i.e., preference is given to the solutions with higher value of the QoS services revenue) and/or with lower $\text{freq}(f_s)$ (i.e., preference is given to the solutions obtained with the change of paths for a flow f_s which has not seen its paths change very often in the past stages of the algorithm). The reasoning behind this is based on a proposal in [40].

Note that this choice of solutions (with which the algorithm continues the search) tries to avoid local extremes. Instead of always proceeding with the best solution found so far, it becomes more advantageous to proceed with a solution with good value of QoS traffic revenue. The algorithm stops after a pre-defined number of iterations.

The adaptation of the TS technique to the basic heuristic HMOR-S2 can be described as depicted in Fig. 3.

The complete formalization of the TS meta-heuristic version of HMOR-S2, HMOR-S2$_{\text{TS}}$, is in Appendix B.2 of the report [35].

As for the numerical complexity of this heuristic, the instructions in the inner cycle of the procedure are executed $C_i^{\text{HMOR-S2TS}} = 4|\mathscr{S}||\overline{\mathscr{F}}|$ times and the number of solutions that are analyzed is $C_s^{\text{HMOR-S2TS}} = 2|\mathscr{S}||\overline{\mathscr{F}}|(|\overline{\mathscr{F}}|+1)$. Therefore, the numerical complexity represented by any of these measures is the same as for the HMOR-S2 heuristic (see [21]). For further details on these calculations, see also [35].

4. Experimental Results

In this section, the analytical and simulation results obtained with the HMOR-S2$_{\text{SA}}$ and the HMOR-S2$_{\text{TS}}$ heuristics in a network case study analogous to the one in [45] are presented.

4.1. Application Model

In [45] a model for traffic routing optimization and admission control in multiservice networks supporting traffic with different QoS requirements, was proposed. This

model will be used as a benchmarking study for the present work concerning upper bounds for the optimal value of the QoS traffic revenue. The objective functions to be maximized in the problem formulated in [45] are the QoS and BE flows revenues, W_Q and W_B. A bi-criteria lexicographic optimization problem was formulated, so that the improvements in W_B are to be sought under the constraint that W_Q remains with the optimal value. A two-stage heuristic procedure based on a multicommodity flow (MCF) formulation was developed to solve this problem. An admission control mechanism was applied in the first stage of the heuristic. Initially only QoS traffic in the original network $\mathscr{N}$ is taken into account and the aim is to find the optimal value of W_Q. Once this has been achieved, the BE traffic is offered to a residual network $\mathscr{N}'$, composed of arcs with the remaining capacities. In the first stage deterministic models are used in the calculation of paths, in particular mathematical programming models based on MCFs. As these models are only a rough approximation in this context and they tend to under-evaluate the blocking probabilities, Mitra and Ramakrishnan [45] propose an adaptation of the original model to obtain more "correct" models, that is models which constitute a better approximation in a stochastic traffic environment. This adaptation consists of a compensation of the required bandwidth values of the flows in the MCF model with a parameter $\alpha \geq 0.0$, so as to represent the effect of the random fluctuations of the traffic that are typical of stochastic traffic flows. The parameter α should have a high value if the need for compensation is high, due to a high variability in the point processes. The MCF-based result is mapped into the adapted model, keeping the relations between traffic intensities invariant.

Furthermore, traffic splitting was used in this traffic routing model, which means that the required bandwidth of each flow may be divided by multiple paths from source to destination, allowing for a more balanced traffic distribution in the network, hence lower blocking probabilities.

The fact that the values of W_Q obtained by this reference model provide upper bounds for the optimal value of W_Q (for the same input traffic matrix) in our model, results from the lexicographic optimization as well as the simplifications in the traffic model, the admission control and the traffic splitting mechanisms, adopted in [45].

4.2. Application of the Model to a Network Case Study

The routing model in [45] was applied to the test network depicted in Fig. 4. It has $N = 8$ nodes, with 10 pairs of nodes linked by a direct arc and a total of $|\mathscr{L}| = 20$ unidirectional arcs. The bandwidth of each arc C'_k [Mbit/s] is shown in Fig. 4. The number of channels C_k is $C_k = \left\lceil \frac{C'_k}{u_0} \right\rceil$, with basic unit capacity $u_0 = 16$ kbit/s. There are $|\mathscr{S}| = 4$ service types with the features displayed in Table 1. The values of the required effective bandwidths $d_s = \frac{d'_s}{u_0}$ [channels] $\forall s \in \mathscr{S}$ are also in the table (where d'_s is the required bandwidth in kbit/s). The expected revenue for a call of type s is assumed to be $w_s = d_s, \forall s \in \mathscr{S}$. The average duration of a type s call is h_s and D_s represents the maximum number of arcs for a type s call.

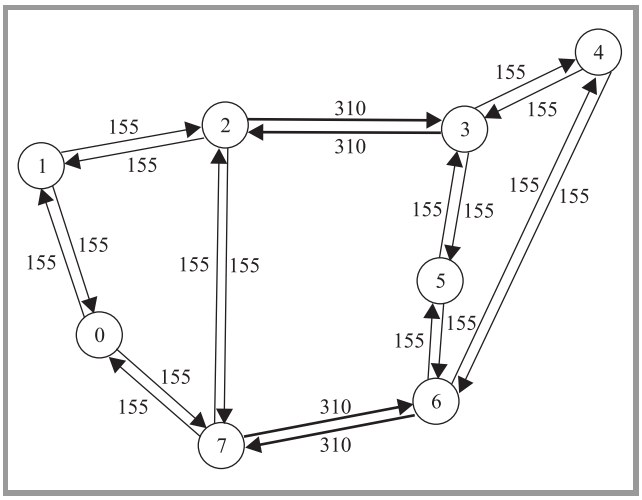


Fig. 4. Test network $\mathscr{M}$ [45], with the indication of the bandwidth of each arc C'_k [Mbit/s].

Table 1
Service features on the test network $\mathscr{M}$

Service	Class	d'_s [kbit/s]	d_s [channels]	w_s	h_s [s]	D_s [arcs]	m_s
1 – video	QoS	640	40	40	600	3	0.1
2 – premium data	QoS	384	24	24	300	4	0.25
3 – voice	QoS	16	1	1	60	3	0.4
4 – data	BE	384	24	24	300	7	0.25

A base matrix $T = [T_{ij}]$ with offered total bandwidth values from node i to node j [Mbit/s] is provided in [45]. As mentioned above, the adaptation of the MCF model to a stochastic model was based on a compensation mechanism that models the effect of random fluctuations of traffic that are typical of a stochastic traffic model. After the introduction of the compensation factor, a relation can be established between the bandwidth demand of each flow f_s for a traffic mix $T(f_s) = m_s T_{ij}$ with $m_s \in [0.0; 1.0]$ and $\sum_{s \in \mathscr{S}} m_s = 1.0$, in the MCF model and the parameters $A(f_s)$ (the mean traffic offered associated with f_s, in Erlang) and $d'_s = d_s u_0$ of the stochastic model. From [45, eq. (5.2)],

$$A(f_s) \approx \frac{T(f_s)}{d'_s} - \alpha \sqrt{\frac{T(f_s)}{d'_s}} = \frac{m_s T_{ij}}{d_s u_0} - \alpha \sqrt{\frac{m_s T_{ij}}{d_s u_0}} \text{ [Erl]}$$

if $\frac{T(f_s)}{d'_s} = \frac{m_s T_{ij}}{d_s u_0} > \alpha^2$ and both $T(f_s)$ and $A(f_s)$ are high. Otherwise,

$$A(f_s) \approx \frac{T(f_s)}{d'_s} = \frac{m_s T_{ij}}{d_s u_0} \text{ [Erl].}$$

From these data all the parameters needed by our traffic model can be obtained as shown in [5].

In this application example, results for the QoS flows revenue W_Q are presented for three values of α: $\alpha = 0.0$ corresponds to the deterministic situation; $\alpha = 0.5$ is the compensation parameter when calls arrive according to a Poisson process, service times follow an exponential distribution and the network is critically loaded; and $\alpha = 1.0$ is used for traffic flows with higher "variability".

For further details on the application of this traffic model to the network case study under analysis, see [5].

4.3. Analytical Results

In the analytical study, the meta-heuristic versions were run only once. For the routing plan obtained at the end of this single run, values for all the objective functions are computed and if the first level objective function values dominate the corresponding values for the initial solution, then this routing plan will be the final solution (Table 2).

Two different sets of tests were conducted: the (i) tests where the initial solution is the same as the one used in the basic heuristic HMOR-S2 runs, a solution which is typical of Internet routing conventional algorithms; the (f) tests where the initial solution of each meta-heuristic version is the routing plan obtained at the end of the basic heuristic runs for each specific α.

For the (i) tests, an initial solution with only one path for each flow, i.e., without an alternative path, is considered leaving it up to the heuristic to find an adequate solution with second choice paths. The initial solution is the same for all the services $s \in \mathscr{S}$ and the paths are symmetrical. The path for every flow f_s is the shortest one (that is, the one with minimum number of arcs); if there is more than one shortest path, the one with maximal bottleneck bandwidth (i.e., the minimal capacity of its arcs) is chosen; if there is more than one shortest path with equal bottleneck bandwidth, the choice is arbitrary.

As for the (f) tests, the aim is to check whether the meta-heuristic variants can improve the quality of the final solutions obtained with HMOR-S2 as an alternative to the direct use of the meta-heuristics (as in the case of the (i) tests).

The analytical results concerning W_Q were compared with results obtained with the previous heuristic HMOR-S2 [8] and with the model proposed in [45], which provides an upper bound to the objective function W_Q optimal value in P-M2-S2.

The experiences with the HMOR-S2$_{\text{SA}}$ were conducted with different temperature cooling functions and the ones that provided best results for the upper level objective functions were $T_W^j = \left[T_W^0\left(1-\frac{j}{J}\right)\right]^a$ and $T_B^j = \left[T_B^0\left(1-\frac{j}{J}\right)\right]^a$ in iteration j, with $J = 1000;5000$ and $a = 0.1;0.01$. The final results were quite similar regardless of the chosen value. An example of the results is displayed in Table 2. These results were obtained with $J = 1000$ and $a = 0.1$ in 11m30s on average in a Linux environment on a Pentium 4 processor with 3 GHz CPU and 1 GB of RAM.

The experiences with the HMOR-S2$_{\text{TS}}$ were conducted with different values for $numIterations = 10$ and a, and the ones that provided best results for the upper level objective functions were $numIterations = 10$ and $a = 20$. These results are displayed in Table 2 and they were obtained in 11m08s on average in the same computer mentioned earlier.

In Table 2, two different comparative analysis can be performed. For HMOR-S2$_{\text{SA}}$(i) and HMOR-S2$_{\text{TS}}$(i), the initial solution is the same as the one used in the corresponding basic heuristic so the table allows for a comparison of the final results obtained with HMOR-S2 and HMOR-S2$_{\text{SA}}$ or with HMOR-S2 and HMOR-S2$_{\text{TS}}$. As for the variants HMOR-S2$_{\text{SA}}$(f) and HMOR-S2$_{\text{TS}}$(f), the initial solution has the objective function values displayed in the table under HMOR-S2 (basis) so that a comparison of the initial and the final results with HMOR-S2$_{\text{SA}}$ and with HMOR-S2$_{\text{TS}}$ can be performed. If an objective function value obtained with one of the variants is the same or better than the corresponding objective function value obtained with the basic heuristic, this is indicated in bold. The table also shows the obtained results for W_Q as a percentage of the upper bound optimal values given in [45].

With the (i) version of the heuristic HMOR-S2$_{\text{SA}}$, the final results for the upper level objective functions are worse when $\alpha = 0.0$ and are the same for the other values of α. As for the (i) version of the heuristic HMOR-S2$_{\text{TS}}$, the final results for the upper level objective functions improve for $\alpha = 1.0$ but are worse for the other values of α. As these variants take longer to run than the basic heuristic and generally do not provide better results for W_Q and $B_{Mm|Q}$, when the initial solution is the same, they can not be considered a better approach for solving the routing problem. However, their use on a second stage of the resolution of the routing problem (after the basic heuristic has been used on a first stage) seems to provide interesting results. In fact, for $\alpha = 0.0;0.5$, the upper level objective function results are better with the (f) test of the heuristic HMOR-S2$_{\text{SA}}$. In particular, with the (f) application version of the heuristic HMOR-S2$_{\text{TS}}$, the upper level objective function results improve for all the values of α.

The results with $\alpha = 1.0$ for both variants are worth mentioning. After HMOR-S2$_{\text{SA}}$(f) is run, the final solution is actually the same as the initial solution. Note that the heuristics always give the initial solution as a final result if the algorithm has not succeeded in finding a better solution in terms of the objective functions W_Q and $B_{Mm|Q}$. As for HMOR-S2$_{\text{TS}}$, the values for W_Q and $B_{Mm|Q}$ in the final solution obtained with the (i) test are actually better than the ones obtained with the (f) test, although the latter are still slightly better than for the basic heuristic. This shows the dependency of the final results on the initial solution, and also shows that starting with a better solution does not necessarily lead to a better final solution.

Taking these results into account, we may conclude that a run of the basic heuristic HMOR-S2 followed by a run of the HMOR-S2$_{\text{SA}}$ variant or a run of the HMOR-S2$_{\text{TS}}$ variant may provide improved results for the routing problem

Table 2
Objective function values for the final solution for different traffic matrices

Objective functions	HMOR-S2 (basis)	HMOR-S2$_{SA}$ (i)	HMOR-S2$_{SA}$ (f)	HMOR-S2$_{TS}$ (i)	HMOR-S2$_{TS}$ (f)
	$\alpha = 0.0$				
W_Q	64731.51*	64517.97	**64795.66**$\diamond$	64619.61	**64915.35**$\star$
$B_{Mm\|Q}$	0.0898	0.107	**0.0843**	0.116	**0.0731**
$B_{m1\|Q}$	0.0898	0.107	**0.0843**	0.116	**0.0731**
$B_{m2\|Q}$	0.0199	0.0218	**0.0194**	**0.0105**	**0.0189**
$B_{m3\|Q}$	0.00216	0.00283	**0.00206**	0.00480	**0.00179**
$B_{M1\|Q}$	0.691	**0.673**	0.700	0.854	0.721
$B_{M2\|Q}$	0.0723	0.115	0.0811	**0.0434**	0.0953
$B_{M3\|Q}$	0.0287	**0.0274**	0.0295	0.0467	0.0312
W_B	17007.15	**17662.81**	**17121.51**	**17489.36**	**17163.01**
	$\alpha = 0.5$				
W_Q	60569.09†	**60569.09**	**60724.32**$\bullet$	60162.90	**60751.77**$\odot$
$B_{Mm\|Q}$	0.0424	**0.0424**	**0.0289**	0.0805	**0.0258**
$B_{m1\|Q}$	0.0424	**0.0424**	**0.0289**	0.0805	**0.0258**
$B_{m2\|Q}$	0.00534	**0.00534**	**0.00270**	0.0104	**0.00259**
$B_{m3\|Q}$	0.00119	**0.00119**	**0.000854**	0.00254	**0.000744**
$B_{M1\|Q}$	0.628	**0.628**	**0.619**	0.742	0.634
$B_{M2\|Q}$	0.0432	**0.0432**	**0.0108**	**0.0385**	**0.00769**
$B_{M3\|Q}$	0.0243	**0.0243**	**0.0237**	0.0330	0.0246
W_B	16904.99	**16904.99**	16738.50	**17664.88**	**16905.73**
	$\alpha = 1.0$				
W_Q	56100.60‡	**56100.60**	**56100.60**$\square$	**56191.34**	**56109.97**$\otimes$
$B_{Mm\|Q}$	0.0263	**0.0263**	**0.0263**	**0.0179**	**0.0252**
$B_{m1\|Q}$	0.0263	**0.0263**	**0.0263**	**0.0179**	**0.0252**
$B_{m2\|Q}$	0.00515	**0.00515**	**0.00515**	**0.00266**	**0.00494**
$B_{m3\|Q}$	0.000560	**0.000560**	**0.000560**	**0.000430**	**0.000555**
$B_{M1\|Q}$	0.544	**0.544**	**0.544**	**0.489**	0.556
$B_{M2\|Q}$	0.0185	**0.0185**	**0.0185**	**0.00955**	**0.0177**
$B_{M3\|Q}$	0.0193	**0.0193**	**0.0193**	**0.0165**	0.0200
W_B	16479.60	**16479.60**	**16479.60**	16288.89	16464.83

HMOR-S2: *) 99.35%; †) 99.57%; ‡) 99.58% of $W_Q^{\max}$ (the optimal revenue in [45]); HMOR-S2$_{SA}$(f): $\diamond$) 99.45%; $\bullet$) 99.83%; $\square$) 99.58% of $W_Q^{\max}$; HMOR-S2$_{TS}$(f): $\star$) 99.63%; $\odot$) 99.87%; $\otimes$) 99.59% of $W_Q^{\max}$.

under analysis. Finally note that the best results obtained with the meta-heuristic variants are more than 99% of the optimal value W_Q. This shows that a significant improvement on $B_{Mm|Q}$ can be obtained just with a very slight worsening on the average revenue W_Q, which gives an idea of the potential advantages of this type of multiobjective routing formulations as previously noted in [4], [5].

4.4. Simulation Results

After the analytical experiences were performed, simulation experiences, with static routing methods using the heuristics, were also carried out for the cases where more promising results were obtained. We considered that simulations with a dynamic version of the routing methods would not provide any important additional information on the quality of the variants of the heuristic. In the simulation study we used a discrete-event simulation platform developed for this type of networks, which enabled the validation of the routing model results and the evaluation of the errors intrinsic to the analytical model which provides the estimates for the objective functions.

The discrete-event stochastic simulation was applied to a static routing model, where the routing plan is the final solution obtained after the (f) test for each of the variants was run. This routing plan never changes throughout the sim-

Table 3
Average objective function values with 95% confidence intervals, for simulations with the routing plan obtained with the HMOR-S2$_{SA}$(f) and the HMOR-S2$_{TS}$(f)

Objective functions	HMOR-S2		HMOR-S2$_{SA}$(f)		HMOR-S2$_{TS}$(f)	
	analytical	static routing model	analytical	static routing model	analytical	static routing model
	Results for $\alpha = 0.0$					
W_Q	64731.51	64642.53±64.17(0.10%)	64795.66	**64704.03**±72.85(0.11%)	64915.35	**64781.55**±67.82(0.10%)
$B_{Mm\|Q}$	0.0898	*0.0887±0.00336*(3.79%)	0.0843	***0.0830***±*0.00389*(4.68%)	0.0731	**0.0749**±0.00316(4.22%)
$B_{m1\|Q}$	0.0898	*0.0887±0.00336*(3.79%)	0.0843	***0.0830***±*0.00389*(4.68%)	0.0731	**0.0749**±0.00316(4.22%)
$B_{m2\|Q}$	0.0199	0.0246±0.000647(2.63%)	0.0194	**0.0242**±0.000551(2.27%)	0.0189	**0.0243**±0.000609(2.51%)
$B_{m3\|Q}$	0.00216	0.00226±0.0000663(2.93%)	0.00206	**0.00216**±0.0000624(2.89%)	0.00179	**0.00196**±0.0000485(2.47%)
$B_{M1\|Q}$	0.691	*0.684±0.00802*(1.17%)	0.700	*0.687±0.0119*(1.74%)	0.721	*0.714±0.0180*(2.52%)
$B_{M2\|Q}$	0.0723	0.0843±0.00242(2.87%)	0.0811	0.0923±0.00377(4.09%)	0.0953	0.106±0.0107(10.05%)
$B_{M3\|Q}$	0.0287	0.0291±0.000206(0.71%)	0.0295	0.0298±0.000171(0.57%)	0.0312	0.0315±0.000236(0.75%)
W_B	17007.15	16982.33±37.02(0.22%)	17121.51	**17102.41**±40.75(0.24%)	17163.01	**17137.81**±49.80(0.29%)
	Results for $\alpha = 0.5$					
W_Q	60569.09	60491.22±50.79(0.08%)	60724.32	**60655.12**±60.57(0.10%)	60751.77	**60655.33**±57.72(0.10%)
$B_{Mm\|Q}$	0.0424	0.0460±0.00163(3.54%)	0.0289	**0.0320**±0.00162(5.08%)	0.0258	**0.0308**±0.00104(3.39%)
$B_{m1\|Q}$	0.0424	0.0460±0.00163(3.54%)	0.0289	**0.0320**±0.00162(5.08%)	0.0258	**0.0308**±0.00104(3.39%)
$B_{m2\|Q}$	0.00534	0.00809±0.000328(4.06%)	0.00270	**0.00521**±0.000329(6.32%)	0.00259	**0.00577**±0.000269(4.66%)
$B_{m3\|Q}$	0.00119	0.00126±0.0000403(3.20%)	0.000854	**0.000927**±0.0000182(1.96%)	0.000744	**0.000838**±0.0000167(2.00%)
$B_{M1\|Q}$	0.628	0.631±0.0151(2.40%)	0.619	***0.615***±*0.0210*(3.41%)	0.634	0.637±0.0157(2.46%)
$B_{M2\|Q}$	0.0432	0.0503±0.00266(5.29%)	0.0108	**0.0179**±0.00201(11.27%)	0.00769	**0.0139**±0.000742(5.35%)
$B_{M3\|Q}$	0.0243	0.0245±0.000196(0.80%)	0.0237	**0.0239**±0.000117(0.49%)	0.0246	0.0248±0.000278(1.12%)
W_B	16904.99	16899.02±38.69(0.23%)	16738.50	*16752.53±39.75*(0.24%)	16905.73	**16905.09**±39.59(0.23%)
	Results for $\alpha = 1.0$					
W_Q	56100.60	56027.72±46.92(0.08%)	56100.60	**56027.72**±46.92(0.08%)	56109.97	**56038.54**±47.33(0.08%)
$B_{Mm\|Q}$	0.0263	0.0281±0.00126(4.48%)	0.0263	**0.0281**±0.00126(4.48%)	0.0252	**0.0269**±0.00126(4.70%)
$B_{m1\|Q}$	0.0263	0.0281±0.00126(4.48%)	0.0263	**0.0281**±0.00126(4.48%)	0.0252	**0.0269**±0.00126(4.70%)
$B_{m2\|Q}$	0.00515	0.00832±0.000685(8.23%)	0.00515	**0.00832**±0.000685(8.23%)	0.00494	**0.00806**±0.000648(8.04%)
$B_{m3\|Q}$	0.000560	0.000637±0.0000154(2.42%)	0.000560	**0.000637**±0.0000154(2.42%)	0.000555	**0.000633**±0.0000168(2.65%)
$B_{M1\|Q}$	0.544	0.547±0.0281(5.13%)	0.544	**0.547**±0.0281(5.13%)	0.556	0.558±0.0192(3.45%)
$B_{M2\|Q}$	0.0185	0.0325±0.00353(10.88%)	0.0185	**0.0325**±0.00353(10.88%)	0.0177	**0.0312**±0.00331(10.60%)
$B_{M3\|Q}$	0.0193	0.0195±0.000167(0.86%)	0.0193	**0.0195**±0.000167(0.86%)	0.0200	0.0202±0.000307(1.52%)
W_B	16479.60	16453.09±17.05(0.10%)	16479.60	**16453.09**±17.05(0.10%)	16464.83	16438.45±18.54(0.11%)

ulation regardless of the random variations of traffic offered to the network. After an initialization phase that lasts for a time $t_{warm-up}$, information on the number of offered calls and effectively carried calls in the network for each flow $f_s, s \in \mathscr{S}$, is gathered, until the end of the simulation. With this information, $B(f_s), \forall s \in \mathscr{S}$ and subsequently, the values of the upper and lower level objective functions related to blocking probabilities can be estimated. As for the expected revenues, knowing the effectively carried calls in the network allows for the calculation of the carried traffic estimates and average revenues.

The results displayed in Table 3 were obtained with a total simulated time $t_{total} = 48$ h and a warm-up time $t_{warm-up} = 8$ h. It took almost 2 h to get these results in the same computer mentioned earlier.

As the results for the (i) version in Table 2 show, only the final solution for the TS-like variant and $\alpha = 1.0$ is better (in terms of the upper level objective function values) than the corresponding final solution for HMOR-S2.

In Table 3, the analytical values of each objective function are displayed, together with the simulation results (average value $\pm$ half length of a 95% confidence interval, computed by the independent replications method [46]) for these functions. If the statistical estimate of an objective function value obtained with one of the variants is the same or better than the corresponding value obtained with the basic heuristic, this is indicated in bold. Furthermore, if some simulation result is better than the corresponding analytical value, this is indicated in italic. The revenue values have 2 decimal places and the blocking probability values have 3 significant figures.

In most cases, the analytical results are outside the 95% confidence interval of the static routing model simulation results, but they are of similar magnitude. The analytical results tend to be better than the corresponding static routing model simulation results, especially in situations of lower traffic loads (which correspond to higher values of α in this routing problem application example). In fact, only for the HMOR-S2$_{\mathrm{SA}}$(f) heuristic with $\alpha = 0.0$ did we get a result where an upper level objective function analytical value was in the corresponding confidence interval and had a value worse than the corresponding static routing model simulation result. These differences between the simulation and analytic results are mainly due to the inaccuracies intrinsic to the analytic/numerical resolution, particularly those associated with the simplifications of the traffic model, and the associated error propagation. As the overflow traffic is treated as Poisson traffic, the analytical model is actually a simplification which tends to underestimate the blocking probabilities in the network (and to overestimate the revenues). The errors that result from this simplification propagate throughout the complex and lengthy numerical calculations associated with the resolution, for a great number of times, of the large systems of implicit non-linear equations (4) and (5). Further simplifications were assumed in the stochastic model for the traffic in the links: a superposition of independent Poisson flows and independent occupations of the links. A more accurate and realistic representation of the traffic flows would allow for better estimates of the blocking probabilities (see for example the numerical algorithms proposed in [47] where the representation of the traffic flows is based on their means and variance values). Nonetheless, the approximations in our model can be considered appropriate in this context for practical reasons. In fact, if more complex models were used to represent the traffic and to calculate the blockings in overflow conditions, the computational burden would be too heavy since the analytical model has to be numerically solved many times during the execution of the heuristic and the routing method would be intractable. It is important to note that, concerning accuracy, the focus is on the relative value of the results of the traffic model rather than on the absolute accuracy of such values, since the aim of the routing optimization procedure is just the comparison of routing solutions, in terms of the values of the objective functions.

The results displayed in the table for the upper level objective functions obtained with the two variants are close, but for the TS-like variant are slightly better than with the SA-like variant. Therefore, the HMOR-S2$_{\mathrm{TS}}$ heuristic may be considered more adequate to the resolution of the very complex routing problem P-M2-S2. A comparison of the results obtained with both variants shows that the analytical and simulation results are coherent, in the sense that whenever the analytical value of an objective function is better for the TS-like variant than for the SA-like variant, the same happens with the average values obtained with the static routing model simulation.

5. Conclusions and Further Work

In this work we began by reviewing a hierarchical bi-level multiobjective routing model in MPLS networks with alternative routing, with two classes of services (with different priorities in the optimization model) and different types of traffic flows in each class. A specialized heuristic strategy, HMOR-S2, for finding "good" compromise solutions to this very complex routing optimization problem, was also reviewed.

Sensitivity tests performed on HMOR-S2 showed that in particular cases there were "better" solutions to the routing problem that the basic heuristic was unable to find. This realization motivated the need to devise new variants that could possibly find solutions "better" than the ones obtained with the HMOR-S2 basic heuristic. Two different variants of this heuristic HMOR-S2 were put forward by introducing meta-heuristic techniques, namely SA and TS techniques.

These variants were applied to a test network used in a benchmarking case study [45] that uses a lexicographic optimization routing approach, including admission control for BE traffic, based on a deterministic traffic representation, with the expected revenues associated with QoS and BE traffic as objective functions. The analytical results ob-

tained with the variants were compared with the optimal values for the QoS service expected revenue in the benchmarking study and with the values obtained with the basic heuristic HMOR-S2. The results show that the introduction of meta-heuristic techniques, in particular SA and TS, in the specialized basic heuristic, does not necessarily lead to better results. However, the introduction of these techniques is advantageous in the search for improvements of the final solution obtained with the basic heuristic. In fact, a run of the basic heuristic HMOR-S2 followed by a run of either the variants tends to provide improved results for the routing problem, especially in the case of the TS variant.

A discrete-event simulation platform was used for a more exact evaluation of the results of the heuristic in a stochastic environment closer to real network working conditions. In most cases, the analytical results obtained with the HMOR-S2 are not inside the 95% confidence interval of the static routing model simulation results, although they are of similar magnitude, due to the inaccuracies intrinsic to the analytic/numerical resolution, namely those associated with the simplifications of the traffic model, and the associated error propagation.

Finally note that these variants have added a greater complexity to the basic heuristic. The computational burden of the resolution has also increased. These remain the major limitations of this type of routing method and restrain its potential practical application, at present, to networks with a limited number of nodes, such as the core and intermediate (metro-core) level networks of low dimension.

Further work on this model will focus on the search for possible simplifications and improvements in the heuristic resolution approaches. Also the extension of the model to broader routing principles such as probabilistic load sharing or traffic splitting might be studied and tested.

Acknowledgements

This work was partially supported by programme POSI of the EC programme cosponsored by FEDER and national funds.

References

[1] A. P. Wierzbicki, "Telecommunications, multiple criteria analysis and knowledge theory", *J. Telecommun. Inform. Technol.*, vol. 3, pp. 3–13, 2005.

[2] J. Clímaco and J. Craveirinha, "Multicriteria analysis in telecommunication network planning and design – problems and issues", in *Multiple Criteria Decision Analysis – State of the Art Surveys*, J. Figueira, S. Greco, and M. Ehrgott, Eds., Int. Ser. Oper. Res. & Manage. Sci., vol. 78. New York: Springer, 2005, pp. 899–951.

[3] J. C. N. Clímaco, J. M. F. Craveirinha, and M. M. B. Pascoal, "Multicriteria routing models in telecommunication networks – overview and a case study", in *Advances in Multiple Criteria Decision Making and Human Systems Management: Knowledge and Wisdom*, Y. Shi, D. L. Olson, and A. Stam, Eds. Amsterdam: IOS Press, 2007, pp. 17–46.

[4] J. Craveirinha, R. Girão-Silva, and J. Clímaco, "A meta-model for multiobjective routing in MPLS networks", *Cent. Eur. J. Oper. Res.*, vol. 16, no. 1, pp. 79–105, 2008.

[5] J. Craveirinha, R. Girão-Silva, J. Clímaco, and L. Martins, "A hierarchical multiobjective routing model for MPLS networks with two service classes – analysis and resolution approach", Res. Rep. 5/2007, INESC-Coimbra, Oct. 2007.

[6] D. Mitra, J. A. Morrison, and K. G. Ramakrishnan, "Optimization and design of network routing using refined asymptotic approximations", *Perform. Eval.*, vol. 36–37, pp. 267–288, 1999.

[7] L. Martins, J. Craveirinha, and J. Clímaco, "A new multiobjective dynamic routing method for multiservice networks: modelling and performance", *Comp. Manage. Sci.*, vol. 3, no. 3, pp. 225–244, 2006.

[8] R. Girão-Silva, J. Craveirinha, and J. Clímaco, "Hierarchical multiobjective routing in MPLS networks with two service classes – a heuristic solution" (accepted for publication in *Int. Trans. Oper. Res.*, 2009).

[9] S. Kirkpatrick, C. D. Gellat Jr, and M. P. Vecchi, "Optimization by simulated annealing", *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[10] F. Glover and M. Laguna, "Tabu search", in *Modern Heuristic Techniques for Combinatorial Problems*, Adv. Top. Comp. Sci. Oxford: Blackwell Sci. Publ., 1993, pp. 70–150.

[11] *Model-Based Decision Support Methodology with Environmental Applications*, A. P. Wierzbicki, M. Makowski, and J. Wessels, Eds. Dordrecht: Kluwer, 2000.

[12] F. Kelly, "Notes on effective bandwidths", in *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary, and I. Ziedins, Eds. Roy. Stat. Soc. Lect. Notes, vol. 4. Oxford: Oxford University Press, 1996, pp. 141–168.

[13] L. Martins, J. Craveirinha, and J. Clímaco, "A new multiobjective dynamic routing method for multiservice networks – modelling, resolution and performance", Res. Rep. 2/2005, INESC-Coimbra, Feb. 2005.

[14] H. M. Elsayed, M. S. Mahmoud, A. Y. Bilal, and J. Bernussou, "Adaptive alternate-routing in telephone networks: optimal and equilibrium solutions", *Inform. Decis. Technol.*, vol. 14, pp. 65–74, 1988.

[15] J. Craveirinha, L. Martins, T. Gomes, C. H. Antunes, and J. N. Clímaco, "A new multiple objective dynamic routing method using implied costs", *J. Telecommun. Inform. Technol.*, vol. 3, pp. 50–59, 2003.

[16] L. Martins, J. Craveirinha, J. N. Clímaco, and T. Gomes, "Implementation and performance of a new multiple objective dynamic routing method for multiexchange networks", *J. Telecommun. Inform. Technol.*, vol. 3, pp. 60–66, 2003.

[17] F. P. Kelly, "Routing in circuit-switched networks: Optimization, shadow prices and decentralization", *Adv. Appl. Probab.*, vol. 20, no. 1, pp. 112–144, 1988.

[18] A. Faragó, S. Blaabjerg, L. Ast, G. Gordos, and T. Henk, "A new degree of freedom in ATM network dimensioning: optimizing the logical configuration", *IEEE J. Selec. Areas Commun.*, vol. 13, no. 7, pp. 1199–1206, 1995.

[19] J. Craveirinha, L. Martins, and J. N. Clímaco, "Dealing with complexity in a multiobjective dynamic routing model for multiservice networks – a heuristic approach", in *Proc. 15th Mini-EURO Conf. MUDSM 2004*, Coimbra, Portugal, 2004.

[20] L. Martins, J. Craveirinha, J. Clímaco, and T. Gomes, "Modeling and performance analysis of a new multiple objective dynamic routing method for multiexchange networks", Res. Rep. ET-N8-5 – 11/2002, INESC-Coimbra, July 2002.

[21] R. Girão-Silva, J. Craveirinha, and J. Clímaco, "Hierarchical multiobjective routing in MPLS networks with two service classes – a heuristic resolution approach", Res. Rep. 8/2008, INESC-Coimbra, July 2008.

[22] K. A. Dowsland, "Simulated annealing", in *Modern Heuristic Techniques for Combinatorial Problems*, Adv. Top. Comp. Sci. Oxford: Blackwell Sci. Publ., 1993, pp. 20–69.

[23] E. I. Oyman and C. Ersoy, "Multicast routing using simulated annealing", in *Proc. COMCON'5*, Crete, Greece, 1995, pp. 419–424.

[24] E. I. Oyman and C. Ersoy, "Multipoint communication using simulated annealing", in *Proc. Birinci Bilgisayar Aglari Symp. BAS'96*, Istanbul, Turkey, 1996, pp. 136–143.

[25] Z. Kun, W. Heng, and L. Feng-Yu, "Distributed multicast routing for delay and delay variation-bounded Steiner tree using simulated annealing", *Comput. Commun.*, vol. 28, no. 11, pp. 1356–1370, 2005.

[26] S. Shimizu, T. Miyoshi, and Y. Tanaka, "Multicast network design by the use of heuristic algorithms", in *Proc. APSITT'99*, Ulaanbaatar, Mongolia, 1999.

[27] T. Miyoshi, S. Shimizu, and Y. Tanaka, "Fast topological design with simulated annealing for multicast networks", in *Proc. ISCC'02*, Taormina, Italy, 2002, pp. 959–966.

[28] M. Randall, G. McMahon, and S. Sugden, "A simulated annealing approach to communication network design", *J. Comb. Optim.*, vol. 6, no. 1, pp. 55–65, 2002.

[29] T. Thomadsen and J. Clausen, "Hierarchical network design using simulated annealing", Tech. Rep. IMM-2002-14, DTU, Sept. 2002.

[30] M. Rios, V. Marianov, and C. Abaroa, "Design of heterogeneous traffic networks using simulated annealing algorithms", in *Proc. ICOIN 2005*, C. Kim, Ed., *LNCS*, vol. 339. Heidelberg: Springer-Verlag, 2005, pp. 520–530.

[31] J. M. de Kock and A. E. Krzesinski, "Computing an optimal virtual path connection network by simulated annealing", in *Proc. SATNAC'98*, Cape Town, South Africa, 1998, pp. 611–617.

[32] Y. Cui, K. Xu, J. Wu, Z. Yu, and Y. Zhao, "Multi-constrained routing based on simulated annealing", in *Proc. IEEE ICC'03*, Anchorage, USA, 2003, vol. 3, pp. 1718–1722.

[33] B. Zhang, C. Huang, and M. Devetsikiotis, "Simulated annealing based bandwidth reservation for QoS routing", in *Proc. IEEE ICC 2006*, Istanbul, Turkey, 2006.

[34] X. Yao, "Call routing by simulated annealing", *Int. J. Electron.*, vol. 79, no. 4, pp. 379–387, 1995.

[35] R. Girão-Silva, J. Craveirinha, and J. Clímaco, "Hierarchical multiobjective routing in MPLS networks with two service classes – a meta-heuristic resolution approach", Res. Rep. 13/2008, INESC-Coimbra, Sept. 2008.

[36] F. Glover and M. Laguna, "Tabu search", http://www.dei.unipd.it/~fisch/ricop/tabu_search_glover_laguna.pdf

[37] A. Hertz, E. Taillard, and D. de Werra, "A tutorial on tabu search", http://www.dei.unipd.it/~fisch/ricop/tabu_search_tutorial.pdf

[38] J. Xu, S. Y. Chiu, and F. Glover, "Probabilistic tabu search for telecommunications network design", *Comb. Optim.*, vol. 1, no. 1, pp. 69–94, 1996.

[39] J. Xu, S. Y. Chiu, and F. Glover, "Tabu search for dynamic routing communications network design", *Telecommun. Syst.*, vol. 8, pp. 55–77, 1997.

[40] M. Laguna and F. Glover, "Bandwidth packing: a tabu search approach", *Manage. Sci.*, vol. 39, no. 4, pp. 492–500, 1993.

[41] M. Gendreau, J.-F. Larochelle, and B. Sansò, "A tabu search heuristic for the Steiner tree problem", *Networks*, vol. 34, no. 2, pp. 162–172, 1999.

[42] J. Shen, F. Xu, and P. Zheng, "A tabu search algorithm for the routing and capacity assignment problem in computer networks", *Comput. Oper. Res.*, vol. 32, no. 11, pp. 2785–2800, 2005.

[43] T. F. Noronha and C. C. Ribeiro, "Routing and wavelength assignment by partition colouring", *Eur. J. Oper. Res.*, vol. 171, no. 3, pp. 797–810, 2006.

[44] S. Routray, A. M. Sherry, and B. V. R. Reddy, "Bandwidth optimization through dynamic routing in ATM networks: genetic algorithm & tabu search approach", *T. Eng. Comput. Technol.*, vol. 12, pp. 171–175, Mar. 2006.

[45] D. Mitra and K. G. Ramakrishnan, "Techniques for traffic engineering of multiservice, multipriority networks", *Bell Labs Tech. J.*, vol. 6, no. 1, pp. 139–151, 2001.

[46] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, Ind. Eng. Manage. Sci. 2nd ed. New York: McGraw-Hill, 1991.

[47] J. Craveirinha, T. Gomes, S. Esteves, and L. Martins, "A method for calculating marginal variances in teletraffic networks with multiple overflows", in *Proc. First Euro-Japanese Worksh. Stoch. Risk Model. Finan. Insur. Product. Reliab.*, J. Janssen and S. Osaki, Eds., Brussels, Belgium, 1998, vol. II.

Rita Girão-Silva graduated in electrical engineering (telecommunications) from the University of Coimbra, Portugal, in 1999. She has recently submitted her Ph.D. thesis in electrical engineering (telecommunications and electronics) to the University of Coimbra and awaits the Ph.D. final defense. She is a Teaching Assistant at the Department of Electrical and Computer Engineering of the University of Coimbra, and a researcher at INESC-Coimbra. Her research areas include routing models in telecommunications networks and multiobjective optimization.

e-mail: rita@deec.uc.pt
Department of Electrical Engineering Science
and Computers
University of Coimbra
Pólo II, Pinhal de Marrocos
P-3030-290 Coimbra, Portugal

Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)
Rua Antero de Quental, 199
P-3000-033 Coimbra, Portugal

José Manuel Fernandes Craveirinha is full Professor in telecommunications at the Department of Electrical Engineering and Computers of the Faculty of Sciences and Technology of the University of Coimbra, Portugal, since 1997. He obtained the following degrees: undergraduate diploma in electrical engineering science (E.E.S.) – telecommunications and electronics at IST, Lisbon Technical University (1975); M.Sc. (1981) and Ph.D. in E.E.S. at the University of Essex (UK) (1984) and Doct. of Science ("Agregado") in E.E.S. telecommunications at the University of Coimbra (1996). Previous positions were: Associate Professor and Assistant Professor at FCTUC, Coimbra Univ., Telecommunication R&D Engineer (at CET-Portugal Telecom). He coordinated a research group in Teletraffic Engineering & Network Planning at INESC-Coimbra R&D Institute since 1986 and was Director of this institute in 1994–99. He is author and co-author of more than 100 scientific and technical

publications in teletraffic modeling, reliability analysis, planning and optimization of telecommunication networks. His main present interests are in reliability analysis models and algorithms and multicriteria routing models for optical and multiservice-IP/MPLS networks.
e-mail: jcrav@deec.uc.pt
Department of Electrical Engineering Science
and Computers
University of Coimbra
Pólo II, Pinhal de Marrocos
P-3030-290 Coimbra, Portugal

Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)
Rua Antero de Quental, 199
P-3000-033 Coimbra, Portugal

João Carlos Namorado Clímaco is full Professor at the Faculty of Economics of the University of Coimbra, Portugal, and President of the Scientific Committee of the INESC-Coimbra. He obtained the M.Sc. degree in control systems at the Imperial College of Science and Technology, University of London (1978); the "Diploma of Membership of the Imperial College of Science and Technology" (1978); the Ph.D. in optimization and systems theory, electrical engineering, University of Coimbra (1982); and the title of "Agregado" at the University of Coimbra (1989). He was, in the past, Vice-President of ALIO – Latin Ibero American OR Association, Vice-President of the Portuguese OR Society and Member of the International Executive Committee of the International Society on Multiple Criteria Decision Making. Actually he is Member of the IFIP WG 8.3 on Decision Support Systems. He belongs to the editorial board of the following scientific journals: "Journal of Group Decision and Negotiation" (JGDN), "International Transactions in Operational Research" (ITOR), "Investigação Operacional" (IO) – "Journal of the Portuguese OR Society" – and "ENGEVISTA" (a Brazilian journal). He is author and co-author of 95 papers in scientific journals and 30 papers in specialized books. His current major interests of research are: multiple criteria decision aiding, multiobjective combinatorial problems, and management and planning of telecommunication networks and energy systems.
e-mail: jclimaco@inescc.pt
Faculty of Economics
University of Coimbra
Av. Dias da Silva 165
P-3004-512 Coimbra, Portugal

Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)
Rua Antero de Quental, 199
P-3000-033 Coimbra, Portugal

Paper

Propagation Path Loss Modeling in Container Terminal Environment

Ryszard J. Katulski, Jacek Stefański, and Jarosław Sadowski

Abstract— This paper describes novel method of path loss modeling for radio communication channels in container port area. Multi-variate empirical model is presented, based on multidimensional regression analysis of real path loss measurements from container terminal environment. The measurement instruments used in propagation studies in port area are also described.

Keywords— path loss modeling, radio propagation.

1. Introduction

Container port area should be treated as a very difficult radio waves propagation environment, because lots of containers made of steel are causing very strong multipath effect and time-varying container arrangement in stacks of different height changes the path loss value in time. Path loss modeling for such area is still complex task and hasn't yet been considered in scientific research. But as the total amount of cargo carried yearly in containers by land and sea increases, the only effective way of controlling such huge number of containers is to build efficient electronic container supervision systems [1]. Nowadays almost all the major container ports have some kind of radio container monitoring, based on available radio communication standards (GSM/GPRS, UMTS, TETRA, WiFi, WiMAX, ZigBee, Bluetooth, many different RFID systems or other solutions in unlicensed frequency band) working in frequency range from about 0.4 GHz to 5 GHz. It should be noted that ITU-R did not present any special recommendation for propagation path loss prediction for radio link in container terminal environment. Differences in spatial arrangement and structure between container stacks and typical urban or industry area can cause relevant path loss prediction errors in case of use inadequate path loss model, so the special survey of propagation phenomenon in container terminal area becomes crucial.

This paper presents new analytical approach to path loss modeling in case of propagation in container port environment, based on empirical results from measurement campaign in Gdynia Container Terminal (Poland). Precise classification of propagation environment and selection of parameters which influence the propagation mechanism in essential way, allowed to define adequate multi-variate error function for multidimensional regression analysis. As a result of this research, new analytical relation between propagation path parameters and path loss in container terminal scenario was proposed. All measurements were carried with assumption, that propagation model has to be adequate for path loss prediction in case of radio communication between container monitoring unit and base station in container terminal area.

2. Measuring Equipment

Block diagram of primary equipment set used in propagation measurements in container terminal scenario is presented in Fig 1.

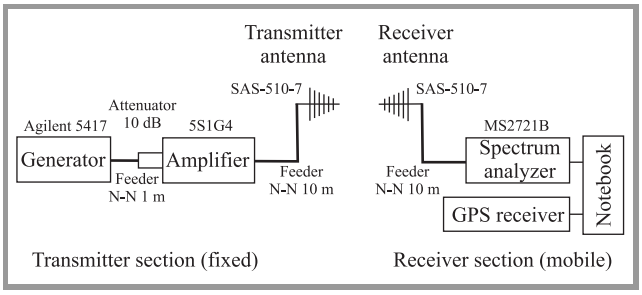


Fig. 1. Block diagram of primary measuring equipment set.

Propagation path loss measuring equipment concept was based on fixed reference signal transmitter and mobile receiver equipment placed in many different positions in the area of container terminal. Harmonic signal without modulation, with frequency in range 0.5 GHz to 4 GHz, was emitted by transmit antenna situated in various places in port. Power amplifier input was protected by precise 10 dB attenuator. The receiver section was made of handheld signal spectrum analyzer working as a sensitive received signal power meter, global positioning system (GPS) receiver and notebook with special software. All the receiver section components were battery powered. Log-periodic directional wideband antennas of the same type were used in both transmit and receiver side. These antennas were calibrated by producer and have precise parameters in whole frequency range of interest.

Firstly the measurement plan assumed four reference signal frequencies: 1, 2, 3 and 4 GHz, but during the measurement campaign additional frequency of 0.5 GHz was also put into investigation. Because the power amplifier used in transmitter section works properly only in frequency range 800 MHz to 4.2 GHz, schematic diagram of transmitter section in case of measurement at frequency 500 MHz was slightly modified: additional attenuator and power amplifier had to be removed and the output of signal generator was directly connected to antenna via 10 m long feeder.

3. Calibration Procedure

In order to precisely compute the propagation path loss from power level of signal detected by handheld spectrum analyzer, radio link power budget equation have to include parameters of all the components from Fig. 1. As the antenna's power gain at all the frequencies of interest is known (measured by manufacturer) and transmitter power level is being kept constant (output power was low enough to avoid any interference to existing radio communication systems), equivalent isotropic radiated power (EIRP) can be simply computed for every frequency.

To ensure that accuracy of measurements doesn't vary with frequency, the transmitter and receiver section was calibrated in the Gdańsk University of Technology laboratory. Firstly, the attenuation of transmitter section feeders at all the frequencies of interest was measured using vector network analyzer. The results are compared in Table 1.

Table 1
Attenuation of transmitter section feeders

Frequency [GHz]	1	2	3	4
Feeder loss between generator and additional attenuator [dB]	0.25	0.59	1.25	1.70
Feeder loss between amplifier and antenna [dB]	3.27	4.89	6.15	7.10

Although the power amplifier has smooth gain adjustment, authors decided to set the amplification to fixed value of 38 dB (amplifier setting, real amplification value was not measured) and determine the signal generator output power that is necessary to achieve required signal power at the input of transmit antenna. In laboratory conditions, spectrum analyzer from receiver section together with precise attenuator 20 dB was used instead of antenna as a power meter (Fig. 2).

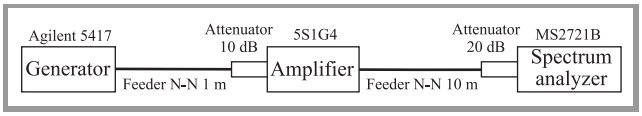


Fig. 2. Transmitter section calibration set schematic diagram.

As the receiver antenna gain in whole band of interest was precisely measured by producer, the only part of receiver section from primary block diagram (Fig. 1) with unknown parameters is the feeder between antenna and handheld spectrum analyzer.

Table 2
Attenuation of receiver section feeder

Frequency [GHz]	1	2	3	4
Feeder loss between antenna and spectrum analyzer [dB]	3.24	4.86	6.25	7.7

The MS2721B spectrum analyzer is able to measure and present received signal power level directly in dBm. Using the same device during calibration phase and in final measurement campaign should compensate eventual received signal power measurement errors. The receiver section feeder attenuation values are presented in Table 2.

Obviously, similar but not the same calibration procedure was repeated at frequency 0.5 GHz after measurement campaign to obtain the power level at the input of transmitter antenna and attenuation of feeders for this specified frequency.

Because the receiver antenna has directional spatial characteristic, path loss measurement procedure required pointing the antenna in direction of transmitter in case of line of sight (LOS) condition or in direction of maximum received signal power in case of non line of sight (NLOS) for every position of receiver section. To simplify the search of maximum signal direction, both transmit and receive antennas were fastened to movable masts with tripods, which allow to change azimuth of reception while height of antenna above terrain remained unchanged. Directional antennas were used for two reasons: firstly because authors would like to take advantage of antenna's power gain, and secondly – authors did not have omni-directional antennas working in frequency range up to 4 GHz with precisely known gain.

As the transmitter output power was kept low, the value of EIRP was far below 15 W limit. According to Polish law, electromagnetic radiation sources with EIRP less than 15 W are objects that do not affect environment or human, so nobody from the measurement team was exposed to harmful electromagnetic radiation.

To improve measurement speed and accuracy, data from spectrum analyzer (received signal power) and GPS receiver (geographic coordinates and time of each measurement) were collected by notebook. Special software running on computer with Linux operating system allowed to define the time between successive measurements, frequency and bandwidth of received signal, type of applied power detector, additional averaging of results, etc. It was also possible to record signal spectrum in each measurement point.

4. Path Loss Measurements in Container Terminal

With the help from administrative of Gdynia Container Terminal, complex survey of propagation aspect in container port was made in term from June to September 2007. Almost 5000 data sets were collected during measurements campaigns, which means about thousand measurement points for each analyzed frequency. The analyses were made in different weather conditions – sunny, cloudy and rainy days with temperature from 5°C to 20°C. In order to reduce path loss measurement errors caused by small scale fading, value of received signal power was calculated using several consecutive measurements and receiving equipment

was always moving. Measurements were carried out in accordance with ITU-R recommendation SM.1708 [2].
Exemplary results of propagation path loss measurements in area of container terminal are shown on map in Fig. 3, where black rectangles symbolize stacks of containers, dots symbolize location of successive measurement points.

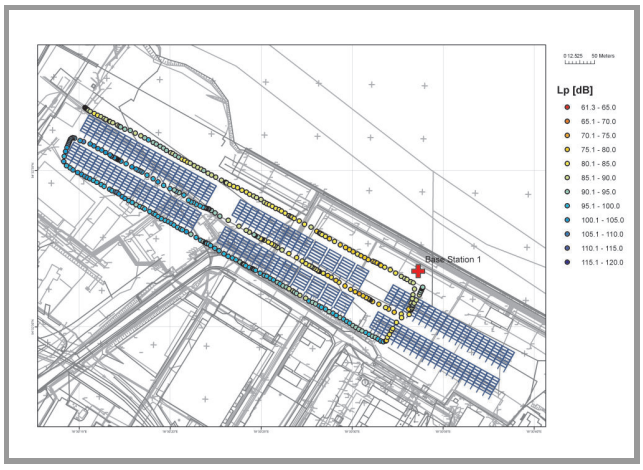


Fig. 3. Exemplary map of measurement points in container terminal.

Similar maps can be plotted for other propagation scenarios and different frequencies of interest.

5. A Novel Multi-Variate Empirical Path Loss Model

Upon the results of almost 5000 propagation path measurements in real container terminal environment, a novel analytical model was developed using multidimensional linear regression analysis with multiple independent variables. For the sake of this analysis a multi-variate error function was defined [3], [4]. The following parameters, which should affect the value of propagation path loss in port area, were chosen as independent variables in error function [5], [6]:

- frequency f;
- propagation path length d;
- path type qualification: line of sight or non line of sight condition;
- difference between transmitter antenna height above terrain level h_T and average height of container stack h_{av}, but two possible cases should be investigated separately: $h_T \geq h_{av}$ and $h_{av} > h_T$.

Because the container terminal, in which all the measurements were made, was permanently used for container transportation, safety restrictions forced authors to limit the height of receiver antenna h_R to fixed value equal 2 m. Due to fixed value of receiver antenna height, proposed propagation models do not include this height as a variable parameter.

As a result of defined error function analysis, regression coefficients for respective propagation cases were computed. Based on this, analytical formulas of propagation path loss in container terminal area can be presented.
Propagation path loss in dB in line of sight scenario:

- in case, when $h_T \geq h_{av}$ (LOS1):

$$L_{LOS1} = 55.2 + 20\lg f + 5.8\lg d - 22.1\lg(h_T - h_{av}), \tag{1}$$

- otherwise, when $h_{av} > h_T$ (LOS2):

$$L_{LOS2} = 41.9 + 20\lg f + 25.9\lg d + 4.2\lg(h_{av} - h_T). \tag{2}$$

Propagation path loss in non line of sight scenario:

- in case, when $h_T \geq h_{av}$ (NLOS1):

$$L_{NLOS1} = 32.6 + 20\lg f + 7.9\lg d + 0.8\lg(h_T - h_{av}), \tag{3}$$

- otherwise, when $h_{av} > h_T$ (NLOS2):

$$L_{NLOS2} = 38.6 + 20\lg f + 13\lg d + 5.9\lg(h_{av} - h_T). \tag{4}$$

The frequency f in Eqs. (1)–(4) should be in MHz, propagation distance d in km, height of transmit antenna and average height of container stock in m. Figures 4–7 presents propagation loss as a function of distance for exemplary frequency 2 GHz in all four scenarios.
Mean error (ME) and mean square error (MSE) are commonly being used to verify accuracy of path loss models. These errors are defined by expression (5) and (6), respectively:

$$ME = \frac{1}{N}\sum_{i=1}^{N}\left(L_{meas,i} - L_{reg,i}\right), \tag{5}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(L_{meas,i} - L_{reg,i}\right)^2}, \tag{6}$$

where $L_{meas,i}$ is the value of measured path loss in ith position of receiver equipment ($i = 1,\ldots,N$), $L_{reg,i}$ mean path loss value computed using Eqs. (2) to (5) for ith position, and N is the total number of considered results. Mean error value reflect the expected average difference between path loss values obtained using proposed model and real path loss measurement results, while mean square error is the ratio of dispersion of measured path loss values and describes how good the propagation model matches experimental data.
Mean errors and mean square errors for all the considered propagation path variants separately (different height of transmitter antenna, line of sight condition) and summary for all measurement results together, are presented in Table 3.
The propagation path loss calculated using proposed analytical model fits very well the results from measurement campaign for all propagation path variants, which is confirmed by very low values of mean errors and acceptably low values of mean square errors.

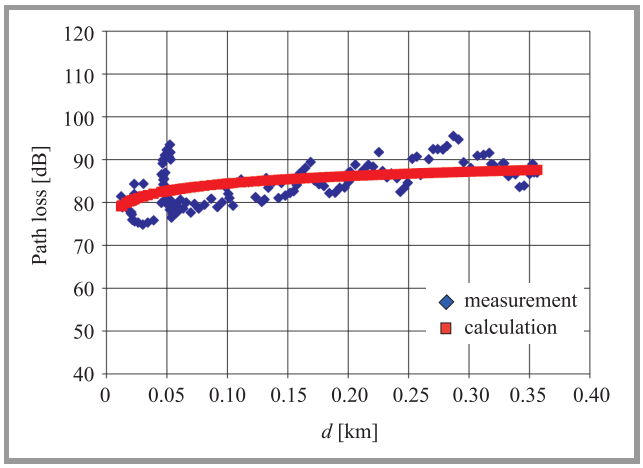


Fig. 4. Propagation path loss at frequency 2 GHz – scenario LOS1.

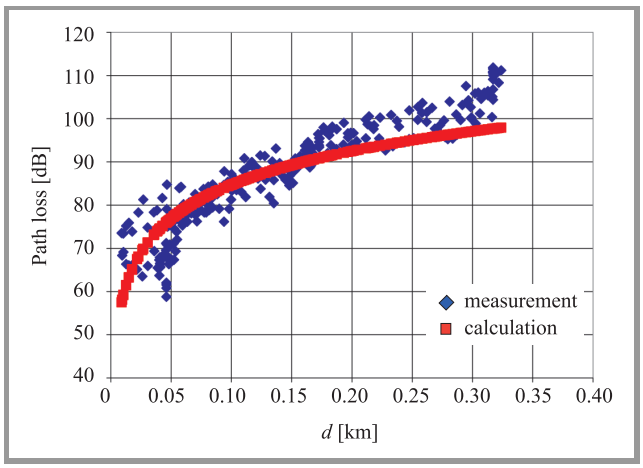


Fig. 5. Propagation path loss at frequency 2 GHz – scenario LOS2.

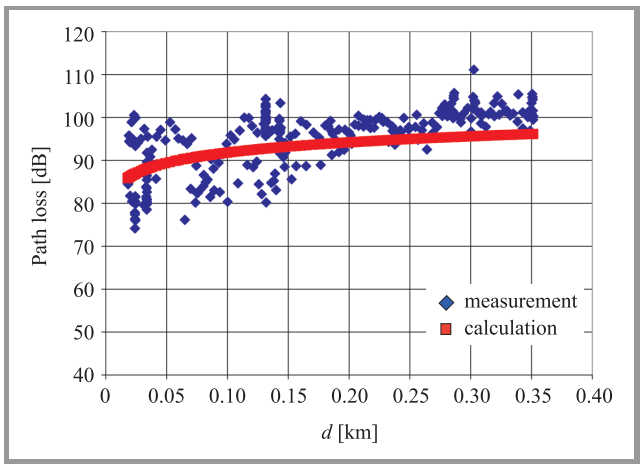


Fig. 6. Propagation path loss at frequency 2 GHz – scenario NLOS1.

Proposed model is valid for propagation path length up to 400 m, longer scenarios were not checked during measurement campaigns. Prediction of path loss at different frequencies between 0.5 GHz and 4 GHz should be enough accurate, because the difference in calculated path loss caused by rounding frequency of interest to nearest measured frequency is not greater than about half the value of mean square error.

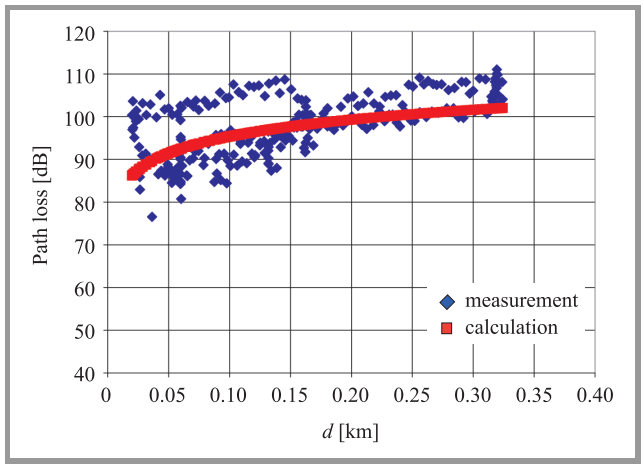


Fig. 7. Propagation path loss at frequency 2 GHz – scenario NLOS2.

Table 3
Mean errors and mean square errors for proposed container environment propagation model

LOS				NLOS				Summary	
LOS1		LOS2		NLOS1		NLOS2			
ME	MSE	ME	MSE	ME	MSE	ME	MSE	ME	MSE
0.00	8.51	0.01	6.02	0.00	6.73	0.00	6.28	0.00	6.82
ME = 0.01, MSE = 7.22				ME = 0.00, MSE = 6.49					

6. Conclusions

Radio propagation analysis in container terminal scenario, presented in this paper, was the first such measurement in Poland and unique in the worldwide area of radio communication research.

Upon the analysis of path loss measurement data, the novel container port area propagation model was proposed. This model has been verified in real propagation conditions in wide frequency range from 0.5 GHz to 4 GHz and can be used to predict propagation path loss in case of designing radio communication systems for container ports or even other container related propagation environments.

Acknowledgments

This research project is supported by the Polish Ministry of Science and Higher Education under grant no. R02 012 01. Authors would also like to thank the administration of the Gdynia Container Terminal for support during realization of this project.

References

[1] R. Katulski, R. Niski, J. Stefanski, and J. Zurek, "Concept of the container monitoring system in Polish harbours", in *Proc. IEEE Conf. Technol. Homel. Secur. Enhan. Transp. Secur. Effic.*, Boston, USA, 2006, pp. 145–148.

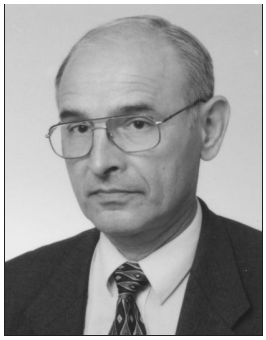
[2] "Field-strength measurements along a route with geographical coordinate registrations", ITU-R Rec. SM.1708, 2005.

[3] A. Kiedrowski, "Specyfika propagacji fal radiowych w systemie dostępowym w warunkach miejskich". Ph.D. thesis, Warsaw, National Institute of Telecommunications, 2004 (in Polish).

[4] R. Katulski and A. Kiedrowski, "Empirical formulas for determination of the propagation loss in urban radio access links", in *Proc. IEEE Veh. Technol. Conf.*, Dallas, USA, 2005, pp. 1742–1746.

[5] R. Katulski and A. Kiedrowski, "Modelowanie tłumienia propagacyjnego w systemach dostępowych, w warunkach zabudowy miejskiej", *Kwart. Elektr. Telekomun.*, vol 52, no. 2, pp. 193–210, 2006 (in Polish).

[6] R. Katulski and A. Kiedrowski, "Calculation of the propagation loss in urban radio-access systems", *IEEE Anten. Propag. Mag.*, vol. 50, no. 6, pp. 65–70, 2008.



Ryszard J. Katulski received his M.Sc., Ph.D. and D.Sc. degrees in radio communication in 1975, 1984 and 1999, respectively. He works for the Department of Radio Communications of the Gdańsk University of Technology, Poland, as a Professor in the field of wireless communications and electromagnetic compatibility, with special interest in mobile systems. He is an author of more than 200 papers and reports presented during international and domestic conferences. He is a member of EMC Section of the Electronics and Telecommunications Committee of the Polish Academy of Science and the Association of Polish Electrical Engineers.
e-mail: rjkat@eti.pg.gda.pl
Gdańsk University of Technology
Narutowicza st 11/12
80-233 Gdańsk, Poland

Jacek Stefański received the M.Sc degree in radio communication in 1993. Since that time he works in the Gdańsk University of Technology, Poland, firstly as an Assistant Professor, and after receiving the Ph.D. degree in 2000 as an Associate Professor in the Department of Radio Communications. Since 2005 he is also working as an Associate Professor in the National Institute of Telecommunications in Gdańsk. His scientific research concerns the theory and techniques of mobile communication. He is the author and co-author of over 130 papers and 15 technical reports.
e-mail: jstef@eti.pg.gda.pl
Gdańsk University of Technology
Narutowicza st 11/12
80-233 Gdańsk, Poland

Jarosław Sadowski received the M.Sc. degree in radio communication in 2002. After that he gained experience in radio communication equipment designing and installing, during his work in industry, firstly as a designer and later as a design and assembly team manager. Since 2007 he works as Assistant in the Department of Radio Communications of the Gdańsk University of Technology, Poland. Main scope of his work affect ultrawideband technology.
e-mail: jarsad@eti.pg.gda.pl
Gdańsk University of Technology
Narutowicza st 11/12
80-233 Gdańsk, Poland

Paper

Model for Balancing Aggregated Communication Bandwidth Resources

Piotr Pałka, Kamil Kołtyś, Eugeniusz Toczyłowski, and Izabela Żółtowska

Abstract—In this paper we present a multicommodity bandwidth exchange model for balancing aggregated communication bandwidth resources (BACBR) that allows us to aggregate similar offers. In this model offers submitted to sell (or buy) the same, similar, or equivalent network resources (or demands for end-to-end connections) are aggregated into single commodities. BACBR model is based on the balancing communication bandwidth trade (BCBT) model. It requires much less variables and constraints then original BCBT, however the outcomes need to be disaggregated. The general model for disaggregation is also given in the paper.

Keywords*— *aggregation, auctions, bandwidth market, market clearing, multicommodity trade.

1. Introduction

The multicommodity exchange models are promising tools that allows to answer emerging requirements for efficient, optimized trading mechanisms well suited on the competitive bandwidth markets. A complex and dynamic bandwidth trading environment is broadly believed to be developed [1]–[4], as new technological and conceptual opportunities are rapidly appearing. For new markets, requiring thousands of bids and offers to be auctioned, the today most popular communication bandwidth trading tools, such as bilateral agreements, or current simple auctions and exchanges (that aim mainly in facilitating buyer-seller contacts), are not sufficient.

For the purpose of modeling trade of bandwidth resources in the communication networks, we assume that the network consists of nodes connected by links. The inter-node link may represent a network resource, that can be an elementary commodity on the bandwidth market. However, network resources being traded can be more complex and can be composed of many parallel links, or end-to-end node connections represented by paths or subnetworks.

In this paper we present a multicommodity bandwidth exchange model for balancing aggregated communication bandwidth resources (BACBR), that considers aggregation of offers submitted to buy or sell the same, similar, or equivalent commodities, related to the network resources. The bandwidth trading is considered from the viewpoint of many network operators, service providers and other wholesale active market players, buying and selling bandwidth.

We believe that the current research proposals for auctioning bandwidth are still insufficient to address diverse market participants needs and requirements. One such a need is the end-to-end network paths trading under competition – when multiple parallel link resources can be offered for sale, or multiple end-to-end connections are bidding.

To cope with the problem of providing bidders with possibility of submitting offers for bundles of elementary commodities when auctioning bandwidth, researchers have proposed two approaches: simultaneous, single link auctions [5]–[8] and combinatorial auctions [9]. In the first approach special, iterative mechanisms are required to coordinate individual links-auctions. The second approach requires buyers to specify the particular links that constitute a desired path. Both approaches lead to welfare inefficiency, as was shown in [10].

The paper is organized as follows. Section 2 presents the proposed model. The mathematical statements both of balancing communication bandwidth trade (BCBT) and new BACBR models are given in Section 3. We also show, that the aggregated model BACBR has all positive features of the simple BCBT model, such as maximization of global economic surplus and possibility of placing buy offers not for bundled links, but for end-to-end connections. Moreover, the BACBR requires much less variables and constraints then original BCBT. What is also important, the market prices for aggregate commodities can be determined on competitive grounds. As the detailed realization of particular offers is not given in the solution of BACBR model so the disaggregation is needed. The process of disaggregation may also have some advantages as it can consider various individual constraints and requirements. The analysis of disaggregation techniques as well as general model for disaggregation is given in Section 4. In Section 5 we summarize our findings.

2. The Proposed Model

The auction BACBR model stated in this paper falls into a class of the multicommodity exchange models, that provide efficient resources allocation solving global economic welfare maximization problem [10]–[13]. Multicommodity means that market entities (further called bidders) can trade with bundles (packages) of different commodities. The BCBT model proposed in [10] allows bidders to place buy offers not for bundled links, but rather for end-to-end connections. Therefore buyer does not have to know which links to choose to best allocate the demanded capacity. It is the decision model that allocates the most efficient links to paths.

Basic BCBT model interprets individual bandwidth buy and sell offers as separate elementary commodities that correspond to traffic demand and network links, respectively. Then, in case of many market participants offering bandwidth on links connecting the same network nodes or demanding the same connections paths, the trade is perform upon a multigraph – see Figs. 1 and 2. It means that in the case of BCBT model, only one offer is submitted on the particular commodity.

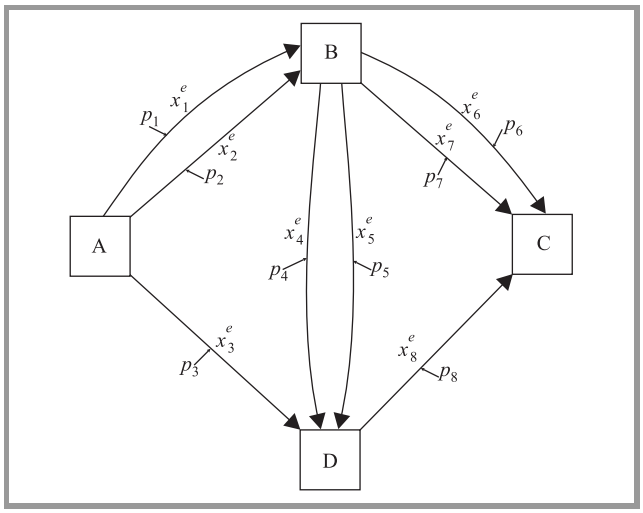


Fig. 1. Resource graph for network links modeled as the multigraph. One offer concerns one particular network link.

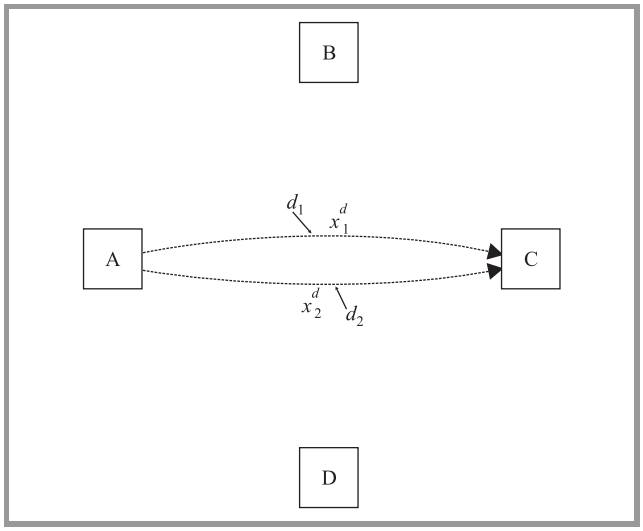


Fig. 2. Resource graph for network demands modeled as a multigraph. One offer concerns particular network demand.

In real world, with possibly thousands of bids being auctioned, such an approach may become inefficient. Thus an aggregation of the BCBT model is required to reduce the complexity.

The aggregated BACBR model considers aggregate commodities structure modeled as a simple graph – see Figs. 3 and 4. It means that in the BACBR model multiple offers

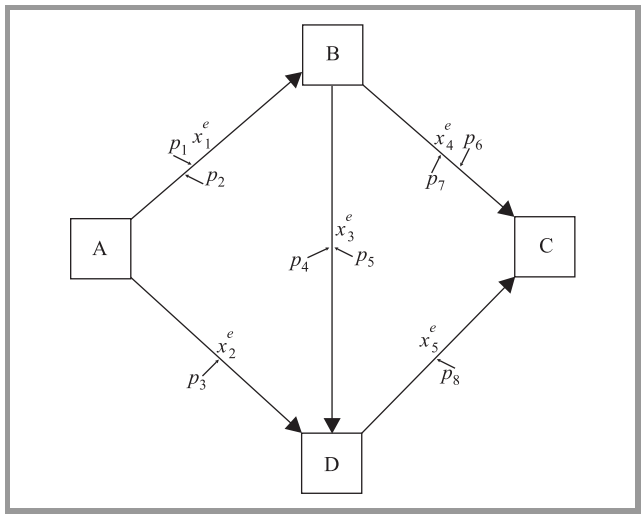


Fig. 3. Resource graph for network demands modeled as the simple graph. Multiple offers for the particular network resources exist.

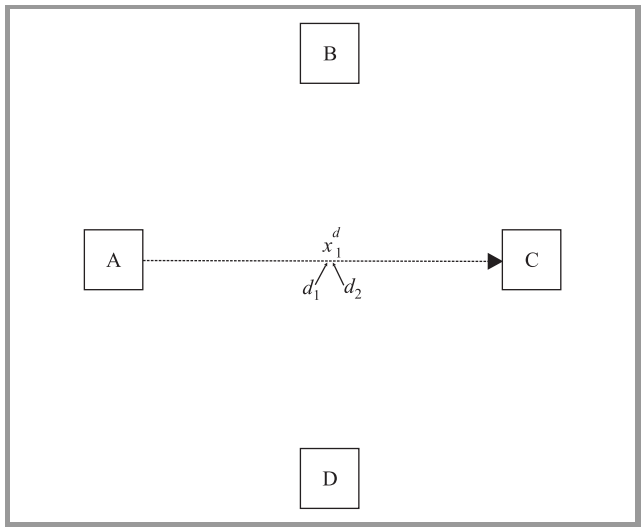


Fig. 4. Resource graph for network demands modeled as the strict graph. Multiple offers for particular network resources exists.

can be submitted on aggregate link or end-to-end connection, considered as an aggregate commodity.

3. Mathematical Model

We assume that the communication network consists of nodes connected by links. The inter-node link may represent a network resource (bandwidth), that can be an elementary commodity offered for sale on the bandwidth market. However, network resources being traded can be more complex and can be composed of many parallel links, or end-to-end node connections represented by paths or subnetworks.

Every buy offer concerns a point-to-point bandwidth connection between a pair of specified locations in a communication network. The locations form a set of network

nodes V. The connections (and links) are unidirectional, i.e., they have source and sink nodes. First we briefly report the conventional model BCBT to allow us for the extension discussion.

3.1. The BCBT Model

The objective of BCBT model [10], [12] is the maximization total economic welfare Eq. (1), which is the sum of total buyers and sellers surpluses. The constraints (2) and (3) set upper and lower bounds on particular network links (x_e) and particular end-to-end network demands (x_d). The non-negative variable x_{ed} Eq. (5) is interpreted as a bandwidth capacity allocated to network link e to serve end-to-end demand d. Also, the sum of capacities allocated to all network demands $\sum_{d\in D} x_{ed}$ served by particular network link e, should not exceed the realization x_e of the link Eq. (4). Finally, the sum of all capacities, provided with incidence matrix a_{ve}, allocated to all network links, serving particular network demand, should not exceed the realization of the end-to-end demand x_d Eq. (6):

$$\hat{Q} = \max\left(\sum_{d\in D} E_d x_d - \sum_{e\in E} S_e x_e\right), \quad (1)$$

$$0 \le x_d \le h_d, \quad \forall_{d\in D}, \quad (2)$$

$$0 \le x_e \le y_e, \quad \forall_{e\in E}, \quad (3)$$

$$\sum_{d\in D} x_{ed} \le x_e, \quad \forall_{e\in E}, \quad (4)$$

$$0 \le x_{ed}, \quad \forall_{e\in E, d\in D}, \quad (5)$$

$$\sum_{e\in E} a_{ve} x_{ed} = \begin{cases} x_d & v = s_d \\ 0 & v \ne s_d, t_d \\ -x_d & v = t_d \end{cases}, \quad \forall_{v\in V, d\in D}, \quad (6)$$

where:

indices:

- $d = 1,2,...,D$ buy offers – demands for bandwidth,
- $v = 1,2,...,V$ network nodes,
- $e = 1,2,...,E$ sell offers – network resources;

parameters:

- a_{ve} $= 1$ if link e originates in node v,
  $= -1$ if e terminates in node v,
  $= 0$ otherwise,
- s_d source node for demand d,
- t_d sink node for demand d,
- h_d required capacity of demand d,
- E_d offered unit price for demand d,
- y_e offered capacity of network link e,
- S_e offered unit price for network link e;

variables:

- x_{ed} bandwidth flow serving demand d allocated to network link e,
- x_d contracted bandwidth capacity for demand d,
- x_e contracted bandwidth capacity for network link e.

The x_e and x_d are, respectively, values of realized bandwidth on the link e and the demand d. They are also the accepted offers for link e and demand d – in the BCBT model sell offers correspond network links and buy offers correspond demand paths resulting in a multigraph. It means, that for a single commodity only one offer can be submitted. However, in case of competitive market with many participants demanding the same connections paths or offering bandwidth on links connecting the same network nodes, the size of the resource graph would be enormous.

3.2. The BACBR Model

Now we propose a nontrivial extension to BCBT model – the model for balancing aggregated communication bandwidth resources, where multiple offers for selling a single network aggregate resource (i.e., network link), or for buying the same connections, are handled in an aggregate manner.

Let us replace the x_e variable by the sum of all adequate bandwidth realizations p_l of sell offers concerning particular network link e:

$$x_e = \sum_{l\in S(e)} p_l, \quad \forall_{e\in E}. \quad (7)$$

Variable p_l is a realization of lth offer for selling link $(e : l \in S(e), S(e') \cap S(e'') = \emptyset, \forall_{e',e''\in E, e'\ne e''})$, $S(e) \subset S$ is a subset of sell offers S concerning particular link e. Thus, we obtain the aggregation of all sell offers submitted on specified link e ($\sum_{l\in S(e)} p_l$).

Analogously, let us replace the x_d variable by the sum of all adequate bandwidth realizations d_m of buy offers concerning particular end-to-end connection d:

$$x_d = \sum_{m\in B(d)} d_m, \quad \forall_{d\in D}. \quad (8)$$

Variable d_m is a realization of mth offer for buying demand $(d : m \in B(d), B(d') \cap B(d'') = \emptyset, \forall_{d',d''\in D, d'\ne d''})$, $B(d) \subset B$ is a subset of buy offers B concerning particular demand d. Thus, we obtain the aggregation of all buy offers submitted on specified connection d ($\sum_{m\in B(d)} d_m$).

Last, we need to change the notation of offer parameters: the offer price we denote as s_l for lth sell offer, and e_m for mth buy offer. The maximal volume of bandwidth, associated with the lth offer we denote as $p_l^{\max}$, analogously, the maximal volume of bandwidth, associated to mth offer we denote as $d_m^{\max}$.

Finally, we obtain the following mathematical model:

$$\hat{Q} = \max\left(\sum_{m\in B} e_m d_m - \sum_{l\in S} s_l p_l\right), \quad (9)$$

$$0 \le p_l \le p_l^{\max}, \quad \forall_{l\in S}, \quad (10)$$

$$0 \le d_m \le d_m^{\max}, \quad \forall_{m\in B} \quad (11)$$

$$\sum_{d\in D} x_{ed} \leq \sum_{l\in S(e)} p_l, \quad \forall_{e\in E}, \tag{12}$$

$$0 \leq x_{ed}, \quad \forall e\in E, \forall d\in D, \tag{13}$$

$$\sum_{e\in E} a_{ve}x_{ed} = \begin{cases} \sum_{m\in B(d)} d_m & v = s_d \\ 0 & v \neq s_d, t_d \\ -\sum_{m\in B(d)} d_m & v = t_d \end{cases}, \forall_{v\in V, d\in D}, \tag{14}$$

where:
indices:

$d = 1,2,...,D$	demands for bandwidth,
$v = 1,2,...,V$	network nodes,
$e = 1,2,...,E$	network links,
$l = 1,2,...,S$	offers for selling,
$S(e)$	offers for selling particular link e,
$m = 1,2,...,B$	offers for buying,
$B(d)$	offers for buying particular demand d;

parameters:

a_{ve}	$= 1$ if link e originates in node v,
	$= -1$ if e terminates in node v,
	$= 0$ otherwise,
s_d	source node for demand d,
t_d	sink node for demand d,
s_l	selling price for lth offer,
e_m	buying price for mth offer,
$p_l^{\max}$	maximal volume for lth offer,
$d_m^{\max}$	maximal volume for mth offer;

variables:

x_{ed}	bandwidth flow serving demand d allocated to the link e,
p_l	contracted bandwidth capacity for selling offer l,
d_m	contracted bandwidth capacity for buying offer m.

4. Aggregation and Disaggregation

Paper [5] considers the auction-based pricing of network bandwidth, where the utilities of particular participants are aggregated to obtain multicommodity flow problem with aggregated user. Authors propose disaggregation as a set of distributed auctions each for one commodity (i.e., network path). Paper [8] assumes that the exchange may concern aggregated resources, like bulk bandwidth for aggregate flows, and virtual paths, virtual private networks, or edge capacity.

Aggregating of participants' offers is useful from the market operator's point of view. When a growing number of offers is submitted on the same link (or the same demand), the liquidity and competitiveness on such market increases, as the concentration and market power decreases. Also, when there are multiple offers submitted on given bandwidth resource, it is much more easier to approach to the competitive price of such a resource.

As the detailed realizations of particular offers are not given in the solution of BACBR model, a disaggregation process is needed, which allows us to match the accepted individual buy and sell offers. From the business point of view, disaggregation of results of BACBR model assures that every buyer knows who will be responsible for its demand realization and every seller knows to whom the bandwidth is served. It may be specially important in the case of the market operator that is not concerned with the network operations or any access switches [2]. In the process of disaggregation it is possible to consider various individual constraints and requirements, not taken into account in the aggregated model.

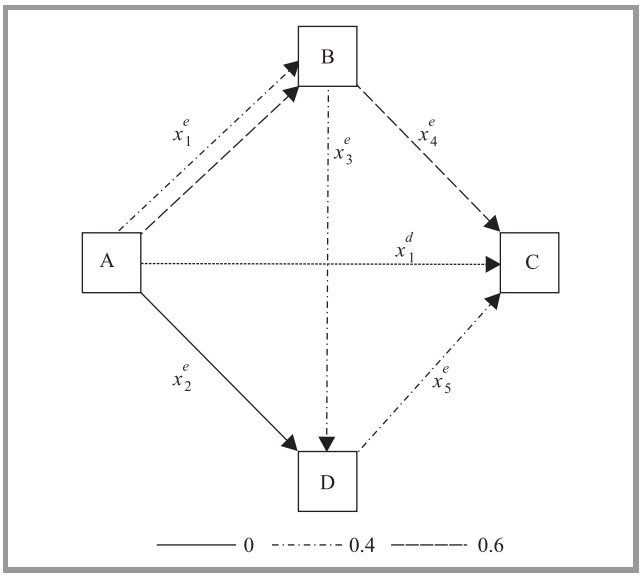


Fig. 5. Example solution for BACBR model. Realization of the path x_1^d by a bundle of links $x_1^e - x_3^e - x_5^e$, $x_1^e - x_4^e$.

Let us assume results of the BACBR model (Fig. 5). As we can see from Fig. 5, the demand x_1^d was realized by two link sequences: $x_1^e - x_3^e - x_5^e$ (A-B-D-C) with share equal to 0.4 and by the $x_1^e - x_4^e$ (A-B-C) with share 0.6 (see Fig. 5). Let us assume that bandwidth allocated on the path x_1^d is equal to 10 units of bandwidth (Table 1). The bandwidth allocated to link x_1^e is equal to 10, on link x_2^e is equal to zero, for links x_3^e and x_5^e allocated bandwidth is equal to 4, and finally for link x_4^e we have 6 units of bandwidth allocated.

Table 1
Example solution for BACBR model

x_{ed}	x_1^e	x_2^e	x_3^e	x_4^e	x_5^e
x_1^d	10	0	4	6	4

Up to now we know how particular network links serve particular demands. However, such aggregate allocation does not give us the answer to the question: how to assign particular realization of the accepted buy offers to particular accepted sell offers. In other words, results of the exemplary offer process do not give us the answer to

the question, how particular offers for buying demand x_1^d (i.e., offers d_1 and d_2 in Fig. 2) will be realized by particular sell offers ($p_1, \cdots, p_8$ in Fig. 1).

General model for disaggregation. After solving BACBR model, we obtain the following results: the volume realization of particular sell offers $p_l \forall_{l \in S}$, moreover we know how the sell offers are realized by the network links $e : l \in S(e)$. Respectively, we know the volume realization of particular buy offers $d_m \forall_{m \in B}$, moreover, we know how buy offers are realized by the appropriate demands $d : m \in B(d)$. Finally, we obtain the aggregated result, the realization of the demand by particular network links $x_{ed} \forall_{e \in E, d \in D}$.
Nevertheless, we need to disaggregate x_{ed} variable, and obtain specific realization of the particular accepted buy offers d_m (which is a portion of particular demand $d : m \in B(d)$ – the offer was submitted on this demand) by particular accepted sell offers p_l (which is a portion of the particular network resource $e : l \in S(e)$ – the offer was submitted on this link). Therefore, we need to determine variable $z_{e_l d_m}$ that satisfy the following equation:

$$\sum_{l \in S(e)} \sum_{m \in B(d)} z_{e_l d_m} = x_{ed}, \quad \forall_{e \in E, d \in D}. \tag{15}$$

To obtain correct disaggregation, we need to find appropriate values of $z_{e_l d_m}$.

The first stage of disaggregation. This stage assumes searching for complete and coherent flows for every accepted buy offer d_m. This problem can be decomposed for every aggregated demand $d \in D$ as the subset of particular buy offers d_m corresponds to only one demand d. Therefore, every accepted buy offer d_m should flow from its source node s_d to its sink node t_d. We obtain the following equation:

$$\sum_{e \in E} a_{ve} y_{ed_m} = \begin{cases} d_m & v = s_d \\ 0 & v \neq s_d, t_d \\ -d_m & v = t_d \end{cases}, \quad \forall v \in V, \forall m \in B(d). \tag{16}$$

To ensure that particular flows for demand will constitute an aggregated solution, the sum of all flows y_{ed_m} has to be equal x_{ed} for every link $e \in E$:

$$\sum_{m \in B(d)} y_{ed_m} = x_{ed}, \quad \forall_{e \in E}. \tag{17}$$

The variable y_{ed_m} is a result of partial disaggregation of parameter x_{ed}. In Table 2 and in Fig. 6 we can see exemplary result of the first stage of disaggregation. We can treat the y_{ed_m} variable as realization of mth buy offer (which belongs to the demand d).

Table 2
Results of first stage of disaggregation; buy offers are correctly disaggregated

y_{ed_m}		p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
		6	4	0	1	3	5	1	4
d	x_{ed}	x_1^e		x_2^e	x_3^e		x_4^e		x_5^e
$d_1 = 7$	x_1^d	7		0	2.8		4.2		2.8
$d_2 = 3$		3		0	1.2		1.8		1.2

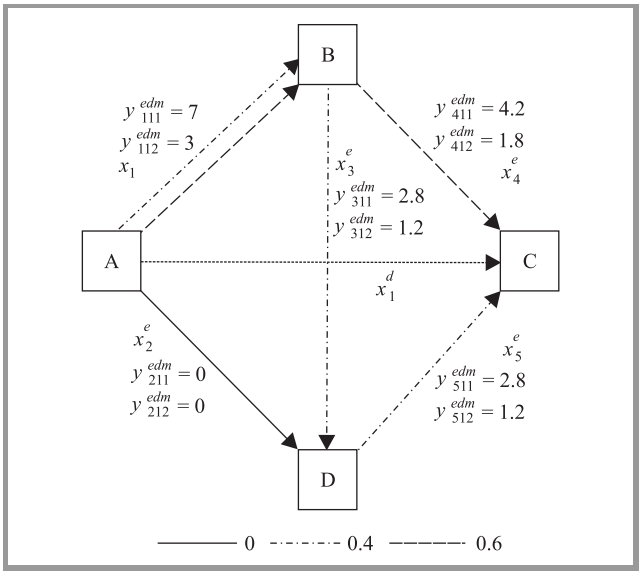


Fig. 6. Exemplary solution of the first stage of disaggregation model. We can see y_{ed_m} variables (y_{xxx}^{edm}) which are realization of mth buy offer (which belongs to the demand d).

The first stage of disaggregation is a general model with some degrees of freedom, so the number of feasible solutions can be often enormous.

The second stage of disaggregation. The solution of the first stage of disaggregation gives us values of variables y_{ed_m} for each $e \in E$. In the second stage we obtain the disaggregated variables $z_{e_l d_m}$, which correspond to the realization of particular accepted buying offer d_m by the particular accepted selling offer p_l. Note that the selling offer p_l belongs to the network link $e : l \in S(e)$, because it was submitted for that link.

Table 3
Results of the second stage of disaggregation

$z_{e_l d_m}$		p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
		6	4	0	1	3	5	1	4
d	x_{ed}	x_1^e		x_2^e	x_3^e		x_4^e		x_5^e
$d_1 = 7$	x_1^d	4.2	2.8	0	0.7	2.1	3.5	0.7	2.8
$d_2 = 3$		1.8	1.2	0	0.3	0.9	1.5	0.3	1.2

General model for the second stage of disaggregation is the following allocation problem (for separate $e \in E$):

$$y_{ed_m} = \sum_{l \in S(e)} z_{e_l d_m}, \quad \forall_{m \in B}, \tag{18}$$

$$\sum_{m \in B} z_{e_l d_m} = p_l, \quad \forall_{l \in S(e)}. \tag{19}$$

The first equation (18) is responsible for disaggregation of variables y_{ed_m} into variables $z_{e_l d_m}$. The second equa-

tion (19) is responsible for correct allocation of variables $z_{e_l d_m}$ to obtain values of accepted offers. In Table 3 and in Fig. 7 we can see exemplary results of the second stage of disaggregation. We can see, that after solving both stages, we obtain correct disaggregation, i.e., variables $z_{e_l d_m}$.
As in the previous stage, this disaggregation is also general, so there are possible multiple feasible solutions.

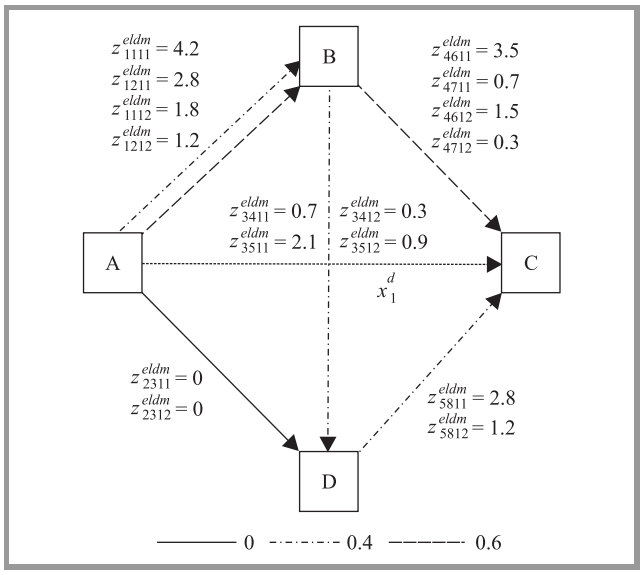


Fig. 7. Exemplary solution of the second stage of disaggregation model. We can see $z_{e_l d_m}$ variables (z^{eldm}_{xxxx})which are realization of mth buy offer (which belongs to the demand d) by the lth sell offer (which belongs to the link e).

Combined disaggregation. Two stages of disaggregation can be combined into one process. We can replace y_{ed_m} from Eq. (16) with Eq. (18), what results in:

$$\sum_{e\in E} a_{ve} \sum_{l\in S(e)} z_{e_l d_m} = \begin{cases} d_m & v = s_{d_m} \\ 0 & v \neq s_{d_m}, t_{d_m} \\ -d_m & v = t_{d_m} \end{cases}, \quad \forall v \in V, \forall m \in B. \tag{20}$$

Equations (20) plus (19) states general single-stage model for disaggregation. As there are many feasible solutions to the general model, many additional specific requirements can be incorporated into the process. Below we present only one simple example of disaggregation method.

Proportional disaggregation. The simplest disaggregation is the proportional method. It divides accepted buying offers proportionally, according to proportionality between selling offers:

$$z_{e_l d_m} = \frac{p_l d_m}{\sum_{l\in S(e)} p_l} x_{ed} = \frac{p_l d_m}{\sum_{m\in B(m)} d_m}. \tag{21}$$

As the result of such disaggregation we obtain the following results (see Table 3). We can observe that selling offer p_1, which is a portion of link x^e_1, serves the buying offer d_1, which is a portion of demand x^d_1, with 4.2 units of bandwidth.

5. Summary

The proposed BACBR model for balancing aggregated communication bandwidth resources assumes aggregation of particular offers and resources, which results in more concise aggregate optimization problem for clearing the multicommodity auction. The solution to the model determines aggregated results, so the need for disaggregation process appears. We have described two phases of the general disaggregation model and showed that it may be performed in one combined disaggregation process. A simple proportional disaggregation method was also proposed. Our future line of research includes exploiting some freedom in the disaggregation methods to take into account various individual participants constraints and requirements.

Acknowledgements

The authors acknowledge the Ministry of Science and Higher Education of Poland for partially supporting the research through project PBZ-MNiSW-02/II/2007.

References

[1] M. Bitsaki, G. Stamoulis, and C. Courcoubetis, "Auction-based bandwidth trading in a competitive hierarchical market", in *Worksh. Next Gener. Internet Netw.*, Rome, Italy, 2005, pp. 372–379.

[2] A. Iselt, A. Kirstadter, and R. Chahine, "Bandwidth trading – a business case for ASON?", in *Telecommun. Netw. Strat. Plann. Symp. NETWORKS 2004*, Vienna, Austria, 2004, pp. 63–68.

[3] R. Rabbat and T. Hamada, "Revisiting bandwidth-on-demand enablers and challengers of a bandwidth market", in *Netw. Oper. Manage. Symp. NOMS 2006*, Vancouver, Canada, 2006, pp. 1–12.

[4] R. G. Prinz, "Leasing cost optimizations for networks with on-demand leased lines", in *6th Int. Conf. Inform. Commun. Sig. Process.*, Singapore, 2007, pp. 1–5.

[5] S. Bessler and P. Reichl, *A Network Provisioning Scheme Based on Decentralized Bandwidth Auctions*. Operations Research/Computer Science Interfaces Series. New York: Springer, 2006.

[6] C. Courcoubetis, M. P. Dramitinos, and G. D. Stamoulis, "An auction mechanism for bandwidth allocation over paths", in *Int. Teletraf. Congr. ITC-17*, Salvador da Bahia, Brazil, 2001, pp. 1163–1174.

[7] M. Dramitinos, G. D. Stamoulis, and C. Courcoubetis, "An auction mechanism for allocating the bandwidth of networks to their users", *Comput. Netw.*, vol. 51, pp. 4979–4996, 2007.

[8] A. Lazar and N. Semret, "Design and analysis of the progressive second price auction for network bandwidth sharing", *Telecommun. Syst.* (Special Issue on Network Economics), 1999.

[9] Ch. Kaskiris, R. Jain, R. Rajagopa, and P. Varaiya, "Combinatorial auction bandwidth trading: an experimental study", in *Developments on Experimental Economics. Lecture Notes in Economics and Mathematical Systems*, vol. 590. Berlin: Springer, 2007, pp. 181–186.

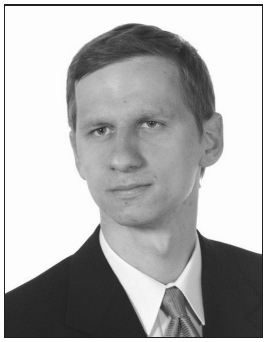
[10] W. Stańczuk, J. Lubacz, and E. Toczyłowski, "Trading links and paths on a communication bandwidth markets", *J. Univer. Comput. Sci.*, vol. 14, no. 5, pp. 642–652, 2008.

[11] P. Kacprzak, M. Kaleta, K. Kołtyś, P. Pałka, E. Toczyłowski, and I. Żółtowska, "Multicommodity mechanisms in bandwidth trading", in *Worksh. Socio-Econom. Asp. Fut. Gener. Internet*, Karlskrona, Sweden, 2008.

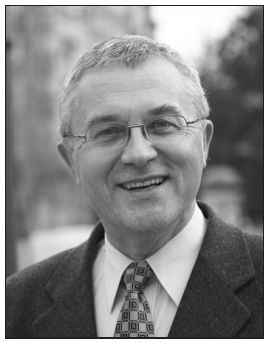
[12] P. Kacprzak, M. Kaleta, P. Pałka, K. Smolira, E. Toczyłowski, and T. Traczyk, "Application of open multi-commodity market data model on the communication bandwidth market", *J. Telecommun. Inform. Technol.*, no. 4, pp. 45–50, 2007.

[13] K. Kołtyś, P. Pałka, E. Toczyłowski, and I. Żółtowska, "Multicommodity auction model for indivisible network resource allocation", in *7th Int. Conf. Decis. Supp. Telecommun. Inform. Soc. DSTIS 2008*, Warsaw, Poland, 2008 or *J. Telecommun. Inform. Technol.*, no. 4, pp. 60–66, 2008.

Piotr Pałka received the M.Sc. degree in 2005 from the Warsaw University of Technology, Poland. Currently he prepares his Ph.D. thesis in computer science at the Institute of Control and Computation Engineering at the Warsaw University of Technology. His research interest is incentive compatibility on the infrastructure markets. His current research is focused on application of multicommodity turnover models.
e-mail: P.Palka@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Kamil Kołtyś received the M.Sc. degree in computer science in 2007 from the Warsaw University of Technology, Poland. Currently he prepares his Ph.D. thesis in computer science at the Institute of Control and Computation Engineering at the Warsaw University of Technology. His research interests include decision support, optimization and bandwidth trading. His current research is focused on application of multicommodity turnover models to network resource allocation.
e-mail: K.J.Koltys@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Eugeniusz Toczyłowski is Professor, the Head of Operations Research and Management Systems Division, at the Institute of Control and Computation Engineering at the Warsaw University of Technology, Poland. He received the M.Sc. degree in 1973, Ph.D. in 1976, D.Sc. in 1989, and the title of Full Professor in 2004. His main research interests are centered around the operations research models and methods, including structural approaches to large scale and discrete optimization, auction theory and competitive market design under constraints, multicommodity trading models, and design of management information systems.
e-mail: E.Toczylowski@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Izabela Żółtowska received the M.Sc. degree in 2000 and the Ph.D. degree in 2006 from the Warsaw University of Technology, Poland. She is an Assistant Professor at the Institute of Control and Computation Engineering at the Warsaw University of Technology. Her research focuses on the optimization models applied on the restructured competitive markets.
e-mail: I.Zoltowska@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

Incorporating Customer Preference Information into the Forecasting of Service Sales

Piotr Rzepakowski

Abstract— Customers change their preferences while getting more familiar with services or being motivated to change their buying habits. Different sources of motivation induce customers to change their behavior: an advertisement, a leader in a reference group, satisfaction from services usage and other experiences, but usually those reasons are unknown. Nevertheless, people vary in susceptibility to suggestions and innovations, and also in preference structure change dynamics. Historical information about the preference structure gives additional information about uncertainty in forecasting activity. In this work the conjoint analysis method was used to find customer preference structure and to improve a prediction accuracy of telecommunication services usage. The results have shown that prediction accuracy increases about by one percent point, what results in a 20 percent increase after using proposed algorithm modification.

Keywords— conjoint analysis, consumer behavior, decision analysis, forecasting, marketing tools, multiple criteria analysis, preference measurement.

1. Introduction

The goal is to forecast services usage without complete knowledge and deep understanding of the domain, including lack of knowledge about predictor variables and intervention effects. Some of intervention effects [1] like customer relationship management activities are usually known but that information can be difficult to obtain. On the other hand, other factors, such as: an advertisement, an influence of a leader in a reference group, satisfaction from services usage, and other experiences can change a customer behavior, but usually this information is unavailable or the influence is unidentified. Taking into account this lack of konowledge, we make an assumption that an analyst has only substantial knowledge about business relationships and constraints which affect the customer activity. His knowledge must be good enough to identify which attributes describing users behavior differentiate them.

Usually, forecasting of time series, when only historical time series are known, are solved by univariate time series models which describe the behavior of a variable in terms of its own past values. Mostly, the exponential smoothing models (ESM) with or without seasonal effects are used [2]. In this work we consider user preference information to improve the exponential smoothing forecasting algorithm. Moreover, we make an assumption that data which were used to create time series are those which can be used for forecasting and for forecasting improvement.

In the summary of the progress made over the past quarter century with respect to methods reducing a forecast error [3] we can find seven well-established approaches which had been shown to improve prediction accuracy. The four of them: combing forecasts, Delphi, causal models, and trend-damping help with time series data. Additionally, other methods such as: segmentation, rule-based forecasting, damped seasonality, decomposing by causal forces and a damped trend with analogous data, were mentioned to be promising for those data. The author indicates also relatively untested methods: prediction markets, a conjoint analysis, diffusion models, and game theory. One of the conclusions from the summary is that, in general, the methods that have ignored theory, prior evidence, and domain knowledge have had a poor record in forecasting. That is why the general structure of the data should be analyzed.

Let us consider two promising methods: segmentation and decomposition by causal forces. The segmentation method is presented as an advantageous one because forecasting errors in different segments may offset one another. The author stresses also problems that can occur, if segments are based on small samples and noisy data, segment forecasts might contain very large errors. However, three reported comparative studies on segmentation that had been conducted since 1975 brought good results. The causal forces method also seems to be worth considering in the analysis of complex series. Complex series are defined as those in which causal forces derive series in opposite directions. If components of a complex series can be forecast more accurately than global series, it helps to decompose the problem by causal forces.

We have combined those two methods with the conjoint analysis to improve ESM models. It is known that forecasting in subgroups shouldn't bring worse prediction accuracy as long as values come from a stationary stochastic process [2]. Furthermore, if time series is known to follow a univariate autoregressive integrated moving-average (ARIMA) model, a forecast made using disaggregated data is, in terms of a mean square error (MSE), at least as good as using aggregated data. However, analyzed stochastic processes are not stationary and the disaggregation can deteriorate accuracy. On the other hand, a good subgroup selection can also improve forecasting exactness [4].

As a consequence, the main idea is to perform forecasting in subgroups defined dynamically by the customers' preference information gained from the conjoint analysis.

The proposed method has been verified on artificialy generated telecommunication services usage data. The best conjoint analysis model was chosen from models definied to identify telecommunication customers' preferences, and run on behavioral data [5]. The following values are forecasted: the number and duration of voice calls, the number of short message service (SMS) usages, the number of multimedia messaging service (MMS) usages, and the number and amount of general packet radio service (GPRS) usages. All the above mentioned values must be predicted within dimensions defined further in table in Subsection 4.2. The 18 months' history of the original telecommunication behavioral data – call data records (CDR) – are aggregated monthly by attributes defined in Table 1.

Table 1
Attributes of call data

Attribute	Levels
Service	Voice
	SMS
	MMS
	GPRS
Location	Home
	Roaming
Net	To on-net
	To off-net (mobile operators)
	To fixed operators
	To international operators
Tariff	Tariff [1–120]
Day type	Working days
	Weekend or holiday
Duration class	0 seconds
	15 seconds
	60 seconds
	240 seconds
Volume	Real values
Count	Integer values

In Section 2 the exponential smoothing models are introduced. Next, in Section 3 the preference identification method is described. In Section 4, a forecasting improvement is proposed. Results are presented in Section 5 and in Section 6 conclusions are drawn, and a plan for future work is proposed.

2. Exponential Smoothing Models

An exponential smoothing is a pure time series technique. This means that the technique is suitable when data have only been collected for series that are going to be forecasted. The exponential smoothing can therefore be applied when there are not enough variables measured to achieve good causal time series models, or when the quality of data is such that causal time series models give poor forecasts. In comparison, more general multivariate ARIMA models allow to predict values of a dependent time series with a linear combination of its own past values, past errors (also called shocks or innovations), and current and past values of other time series. Exponential smoothing takes the approach that recent observations should have relatively more weight in forecasting than distance observations. "Smoothing" implies predicting an observation by a weighted combination of previous values and "exponential" smoothing means that weights decrease exponentially as observations get older. In exponential smoothing only the slowly changing level is being modeled, nevertheless, it can be extended to different combinations of trend and seasonality:

– simple,

– double (Brown),

– linear (Holt) trend,

– damped-trend linear,

– no seasonality,

– additive seasonality,

– multiplicative seasonality.

Additionally, transformed versions of these models can be defined:

– logarithmic,

– square root,

– logistic,

– Box-Cox.

Given a time series $Y_t : 1 \leq t \leq n$, the underlying model assumed by the smoothing models has the following (additive seasonal) form:

$$Y_t = \mu_t + \beta_t t + s_p(t) + \varepsilon_t, \tag{1}$$

where:
μ_t – represents the time-varying mean term,
β_t – represents the time-varying slope,
$s_p(t)$ – represents the time-varying seasonal contribution for one of the p seasons,
ε_t – are disturbances.
Different smoothing models are presented in Table 2.

Table 2
Exponential smoothing models

Smoothing model	Equation
Simple	$Y_t = \mu_t + \varepsilon_t$
Double (Brown)	$Y_t = \mu_t + \beta_t t + \varepsilon_t$
Linear (Holt)	$Y_t = \mu_t + \beta_t t + \varepsilon_t$
Damped-trend linear	$Y_t = \mu_t + \beta_t t + \varepsilon_t$
Seasonal	$Y_t = \mu_t + s_p(t) + \varepsilon_t$
Winters – additive	$Y_t = \mu_t + \beta_t t + s_p(t) + \varepsilon_t$
Winters – multiplicative	$Y_t = (\mu_t + \beta_t t)s_p(t) + \varepsilon_t$

2.1. Smoothing State and Smoothing Equations

The smoothing process starts with an initial estimate of the smoothing state, which is subsequently updated for each observation using the smoothing equations. Depending on the smoothing model, the smoothing state at time t will consist of the following:
L_t – smoothed level that estimates μ_t,
T_t – smoothed trend that estimates β_t,
$S_{t-j}, j = 0, ..., p-1$, are seasonal factors that estimate $s_p(t)$.
The smoothing equations determine how the smoothing state changes as time progresses. Knowledge of the smoothing state at time $t-1$ and that of the time series value at time t uniquely determine the smoothing state at time t. The smoothing weights determine the contribution of the previous smoothing state to the current smoothing state. The smoothing equations for each smoothing model are listed in Table 3.

Table 3
Equations for the smoothing models

Smoothing model	The error-correction form, The k-step prediction equation
Simple	$L_t = L_{t-1} + \alpha\varepsilon_t$
	$\hat{Y}_t(k) = L_t$
Double (Brown)	$L_t = L_{t-1} + T_{t-1} + \alpha\varepsilon_t$
	$T_t = T_{t-1} + \alpha^2\varepsilon_t$
	$\hat{Y}_t(k) = L_t + ((k-1) + 1/\alpha)T_t$
Linear (Holt)	$L_t = L_{t-1} + T_{t-1} + \alpha\varepsilon_t$
	$T_t = T_{t-1} + \alpha\gamma\varepsilon_t$
	$\hat{Y}_t(k) = L_t + kT_t$
Damped-trend linear	$L_t = L_{t-1} + \phi T_{t-1} + \alpha\varepsilon_t$
	$T_t = \phi T_{t-1} + \alpha\gamma\varepsilon_t$
	$\hat{Y}_t(k) = L_t + \sum_{i=1}^{k} \phi^i T_t$
Seasonal	$L_t = L_{t-1} + \alpha\varepsilon_t$
	$S_t = S_{t-p} + \delta(1-\alpha)\varepsilon_t$
	$\hat{Y}_t(k) = L_t + S_{t-p+k}$
Winters – additive	$L_t = L_{t-1} + T_{t-1} + \alpha\varepsilon_t$
	$T_t = T_{t-1} + \alpha\gamma\varepsilon_t$
	$S_t = S_{t-p} + \delta(1-\alpha)\varepsilon_t$
	$\hat{Y}_t(k) = L_t + kT_t + S_{t-p+k}$
Winters – multiplicative	$L_t = L_{t-1} + T_{t-1} + \alpha\varepsilon_t / S_{t-p}$
	$T_t = T_{t-1} + \alpha\gamma\varepsilon_t / S_{t-p}$
	$S_t = S_{t-p} + \delta(1-\alpha)\varepsilon_t / L_t$
	$\hat{Y}_t(k) = (L_t + kT_t)S_{t-p+k}$

In order to use the multiplicative version of Winters method, the time series and all predictions must be strictly positive. Additionally, coefficient α, δ, γ must fulfill stability conditions [6].
Almost all exponential smoothing models have ARIMA equivalents presented in Table 4. ARIMA is more general than ESM and allows to predict values of a dependent time series with a linear combination of its own past values, past errors (also called shocks or innovations), and current and past values of other time series (predictor time series).

Table 4
ARIMA equivalent models

Smoothing model	ARIMA equivalent
Simple	ARIMA $(0,1,1)$
Double (Brown)	ARIMA $(0,2,2)$
Linear (Holt)	ARIMA $(0,2,2)$
Damped-trend linear	ARIMA $(1,1,2)$
Seasonal	ARIMA $(0,1,p+1)(0,1,0)_p$
Winters – additive	ARIMA $(0,1,p+1)(0,1,0)_p$
Winters – multiplicative	No equivalent

2.2. Prediction Errors

Predictions are made based on the last known smoothing state. Predictions made at time t for k steps ahead are denoted $\hat{Y}_t(k)$ and the associated prediction errors are denoted $\varepsilon(k) = Y_{t+k} - \hat{Y}_t(k)$.
The one-step-ahead predictions refer to predictions made at time $t-1$ for one time unit into the future, that is $\hat{Y}_{t-1}(1)$, and the one-step-ahead prediction errors are more simply denoted $\varepsilon_t = \varepsilon_{t-1}(1) = Y_t - \hat{Y}_{t-1}(1)$. The one-step-ahead prediction errors are also the model residuals, and the statistic related to the one-step-ahead prediction errors is the objective function used in smoothing weight optimization.

Table 5
The variance of the prediction errors

Smoothing model	$\varepsilon_t(k)-$ variance
Simple	$\mathrm{var}(\varepsilon_t)[1+\sum_{j=1}^{k-1}\alpha^2]$
Double (Brown)	$\mathrm{var}(\varepsilon_t)[1+\sum_{j=1}^{k-1}(2\alpha+(j-1)\alpha^2)^2]$
Linear (Holt)	$\mathrm{var}(\varepsilon_t)[1+\sum_{j=1}^{k-1}(\alpha+j\alpha\gamma)^2]$
Damped-trend linear	$\mathrm{var}(\varepsilon_t)\left[1+\sum_{j=1}^{k-1}(\alpha+\frac{\alpha\gamma\phi(\phi^j-1)}{(\phi-1)})^2\right]$
Seasonal	$\mathrm{var}(\varepsilon_t)\left[1+\sum_{j=1}^{k-1}\psi_j{}^2\right]$
Winters – additive	$\mathrm{var}(\varepsilon_t)\left[1+\sum_{j=1}^{k-1}\psi_j{}^2\right]$
Winters – multiplicative	$\mathrm{var}(\varepsilon_t)\left[1+\sum_{i=0}^{\infty}\sum_{j=1}^{p-1}(\frac{\psi_{j+ip}S_{t+k}}{S_{t+k-j}})^2\right]$

The variance of the prediction errors counted as presented in Table 5 is used to calculate the confidence limits.

3. Conjoint Analysis for Preference Identification

For preference identification, wchich are going to be used for spliting customers into homogenous segments, we used the conjoint analysis method running on behavioral data [5].

The conjoint analysis process consists of:

- selection of utility factors,
- conjoint measure definition,
- conjoint model definition,
- questionnaire preparation,
- questionnaire data acquisition,
- statistical analysis,
- data interpretation.

For utility factors we get some attributes from the behavioral data. The questionnaire preparation step is not required because the historical data are analyzed. Hence, the questionnaire data acquisition step changes to the behavioral data preparation one.

3.1. Selection of Utility Factors

Attributes differentiating the cost of services mostly were chosen to be utility factors. Among them there are: service, location, network, day types, and duration class attributes with categories presented in Table 6. Original CDR were transformed to determine chosen attributes. Next, the data were aggregated and statistics of call frequencies for each aggregation were calculated.

Table 6
Utility factors

Attribute	Levels
Service	Voice
	SMS
	MMS
	GPRS
Location	Home
	Roaming
Net	To on-net
	To off-net (mobile operators)
	To fixed operators
	To international operators
Day type	Working days
	Weekend or holiday
Duration class	0 seconds
	15 seconds
	60 seconds
	240 seconds

3.2. The Conjoint Measure Definition

The dependency between utility factors is defined by the conjoint measure. It consists of intercept coefficient μ and part-worth utilities associated with attributes. If some attributes are correlated then the interaction between those attributes are added to the conjoint measure. Interactions between pairs usually suffice but sometimes interactions of higher orders, for example, between three variables are used. For presented telecommunication task, the conjoint measure is defined by Eq. (2). In that example part worth utilities are presented by α vectors of utilities for attribute values, β vectors of utilities for all combinations of values associated with two attributes and γ vector of utilities for a combination of values taken from service, net, and day type attributes. For the presented telecommunication task, we used a measure consisting of linear terms and correlation between all pairs of attributes extended by interactions between three attributes. Finally, the conjoint measure consists of factors presented in Table 7 and is defined as follows:

$$\begin{aligned} y = \mu & \\ & + \alpha_{service} + \alpha_{location} + \alpha_{net} + \alpha_{day\ type} + \alpha_{duration\ class} \\ & + \beta_{service*location} + \beta_{service*net} + \beta_{service*duration\ class} \\ & + \beta_{service*day\ type} + \beta_{location*net} + \beta_{location*duration\ class} \\ & + \beta_{location*day\ type} + \beta_{net*day\ type} + \beta_{net*duration\ class} \\ & + \gamma_{service*net*day\ type} \\ & + \varepsilon. \end{aligned} \tag{2}$$

Table 7
Conjoint measure factors

Attribute	Levels
Service	4
Location	2
Net	4
Day type	2
Volume	4
Service*location	8
Service*net	16
Service*day type	8
Service*duration class	7
Location*net	8
Location*day type	4
Location*duration class	8
Net*day type	8
Net*duration class	16
Service*net*day type	32
Total	131

3.3. Conjoint Model Definition

The conjoint model is a statistical model which represents dependencies between utility of a profile and its attributes and is defined by Eq. (3). Now α coefficient represent utilities associated with all conjoint factors α, β and γ defined earlier. Because all of attributes of conjoint measure are categorical, dummy variables x created to represent no met-

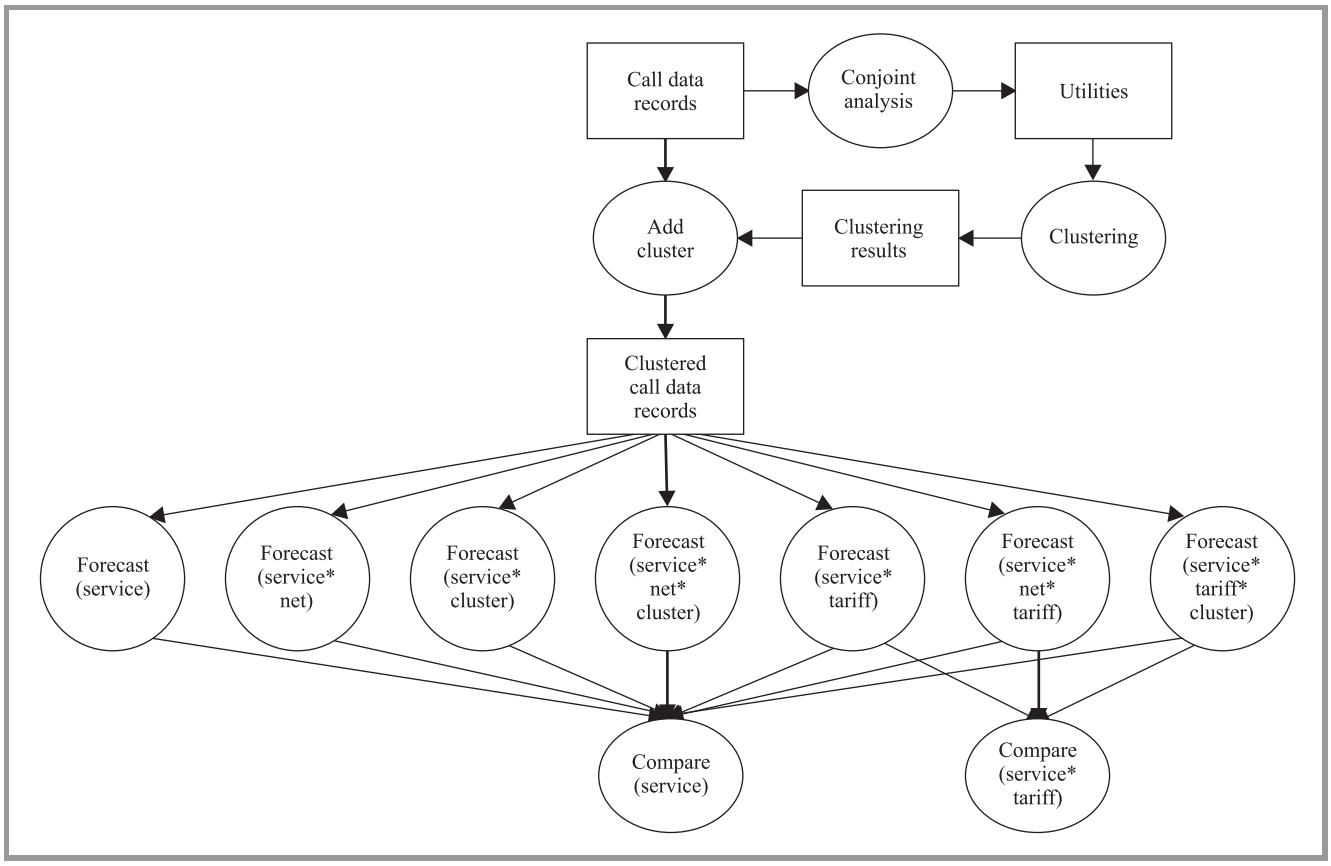


Fig. 1. Forecasing procedure.

ric information. One attribute with k levels was replaced by $k-1$ binary attributes

$$y = \alpha^T x + \varepsilon. \tag{3}$$

After adding dummy variables, regression techniques can be used for part worth utilities identification. Dependent variable y in the regression model represents the utility of a profile. In the analyzed problem it was calculated as the number of events, which means that it has binomial distribution. That problem cannot be solved simply by linear regression as regression techniques require normal distribution of dependent variable. However, binomial distribution can be simply transformed to the normal one by logarithmic function. In consequence, general linear model (GLM) was defined as

$$\ln(y) = \alpha^T x + \varepsilon. \tag{4}$$

4. Forecasting Improvement

Forecasting improvement is done by data disaggregation and the criteria of splitting the data are the main point of this improvement. In fact, the data are split using information about customer preferences. This proposition is supported by hypothesis which states, that customers who have similar preferences behave similarly and the variance of a service usage in a group is lower than in the whole population. In the presented method, preferences come from a behavioral data and can be treated as aggregated representation of the way in which customers use services. As a consequence of this idea, analyses are carried out as follows:

- at first, customer segmentation is done on preferences to a service usage;
- next, forecasts are made in segments;
- finally, a forecast in the whole population is calculated as a sum of forecasts in subgroups.

In conduct analysis, forecasts using different disaggregation methods are compared on two levels: on the service aggregation level and the combinations of service and tariff aggregations. The process of forecasting using various disaggregation methods is presented in Fig. 1. At first, the CDR data are used to find coustomers part-worth utilities. After that, customers are clustered into homogenous groups using calculated utilities. Information about a customer group is added to each record in the CDR. Then a customer segment identifier can be used in data aggregation to make forecasting in subgroups.

4.1. Customer Segmentation on Preferences to Service Usage

Consumer preferences were determined by running a conjoint analysis procedure on behavioral data as it has been shown in Section 3. Those preferences were computed on 12 months' data. Next, clustering was done to split consumers into homogenous groups.

There are two types of clustering: partition clustering and hierarchical clustering. Partition clustering attempts to directly decompose data set into a set of disjoint clusters. Hierarchical clustering, on the other hand, proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. For a huge amount of data hierarchical clustering is not practically applicable, thus we used partition clustering implemented in statistic analytical software (SAS) as a FASTCLUS procedure. In the used partition clustering, the number of clusters has to be given as an input to the procedure. The procedure was run many times to make: 200, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000 groups clustering.

4.2. Forecasting

To check how preference clustering influences the forecasting accuracy, we made comparisons of forecasting made in aggregations presented in Table 8.

Table 8
Dimension intersections

Intersection	Number of forecasts
Service	4
Service*net	16
Service*tariff	800
Service*net*tariff	3200
Service*cluster	4*clusters
Service*net*cluster	16*clusters
Service*tariff*cluster	800*clusters
Service*net*tariff*cluster	3200*clusters

The high-performance forecasting (HPF) procedure from SAS was used for forecasting. This procedure provides an automatic way to generate forecasts for each time series.

The best model is automatically choosen from the exponential smoothing models presented in Section 2. And the mean absolute percent error (MAPE) good-of-fit statistic is used to measure how models fit data:

$$MAPE = \frac{100}{\mathrm{n}} \sum_{t=1}^{n} \left| \frac{\mathrm{y_t} - \hat{\mathrm{y}}_\mathrm{t}}{\mathrm{y_t}} \right|. \tag{5}$$

The summation ignores observations, where $y_t = 0$.

5. Analytical Results

Analitycal results are summarized in two subsections. The first one concerns the conjoint analysis and the second one the forecasting process.

5.1. The Conjoint Analysis

The conjoint analysis was performed on 12 months' data. Statistics R^2 presented in Table 9 show that the model is well fitted to the data. The average value of R^2 is 95% and the standard deviation is very low.

Table 9
Analysis of variance for the coujoint model

Statistic	Avg	Std
R^2	0.95	0.11
$adj - R^2$	0.82	0.29
p-value	0.05	0.15

Table 10
Relative importance statistics in population [%]

Attribute/statistic	Avg	Std
Service	12.0	10.2
Location	2.2	5.8
Net	10.9	9.2
Day type	7.2	9.8
Duration class	13.3	17.6
Service*location	2.2	5.7
Service*net	10.1	9.3
Service*day type	5.2	5.4
Service*duration class	3.9	6.2
Location*net	1.6	4.5
Location*day type	0.9	2.9
Location*duration class	1.3	4.0
Net*day type	5.3	5.1
Net*duration class	13.3	8.8
Service*net*day type	5.8	7.4

Comparing standard deviations to average values of importances illustrated in Table 10, we find that customers have different menners and different features of services are important for them. These statistics show, that spliting customers into more homogenous grups is worth considering, what is also shown in Subsection 5.2.

5.2. Forecasting Comparison

Prediction accuracy in clusters has been compared to forecasting made in data disaggregated by attributes available a priori (Figs. 2–12): the tariff plan and the net including intersections defined in Table 8. An optimal number of clusters were found from figures presenting prediction accuracy of statistics at the total level drawn in different number of clusters. In three out of five time series, clustering brougth good results and only prediction of GPRS

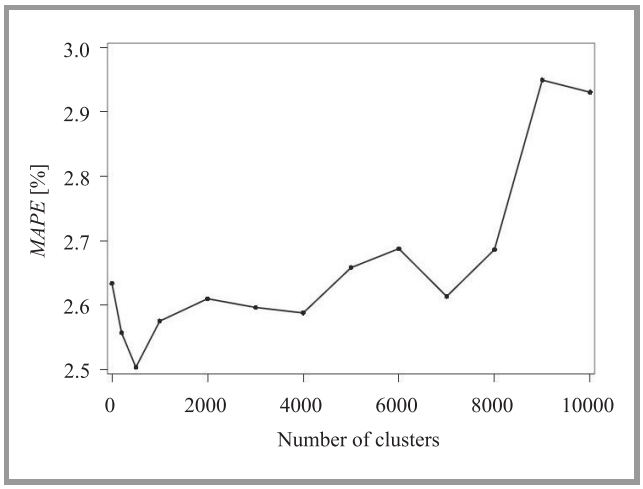


Fig. 2. Prediction accuracy of the total duration of voice events for different number of clusters verified on 15 month data.

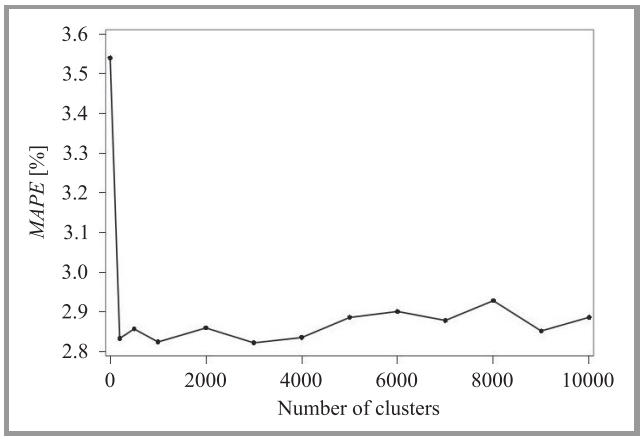


Fig. 3. Prediction accuracy of the total number of voice events for different number of clusters verified on 15 month data.

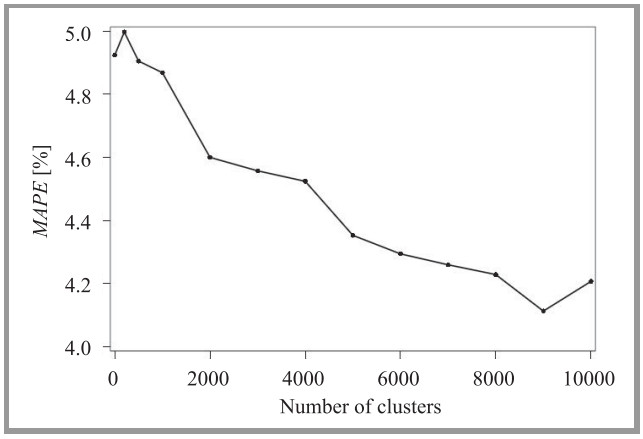


Fig. 4. Prediction accuracy of the total number of SMS events for different number of clusters verified on 15 month data.

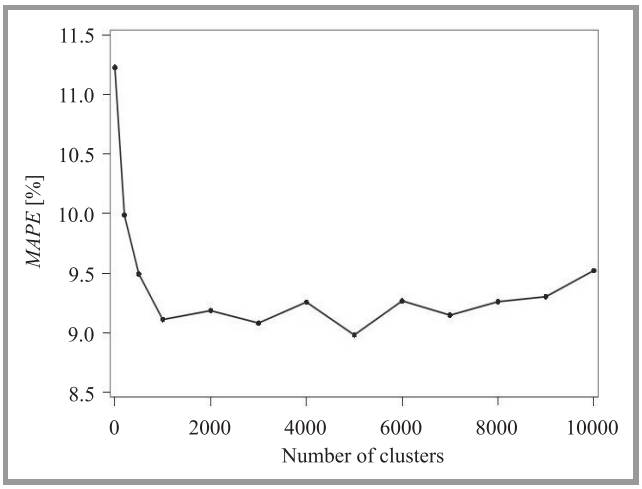


Fig. 5. Prediction accuracy of the total number of MMS events for different number of clusters verified on 15 month data.

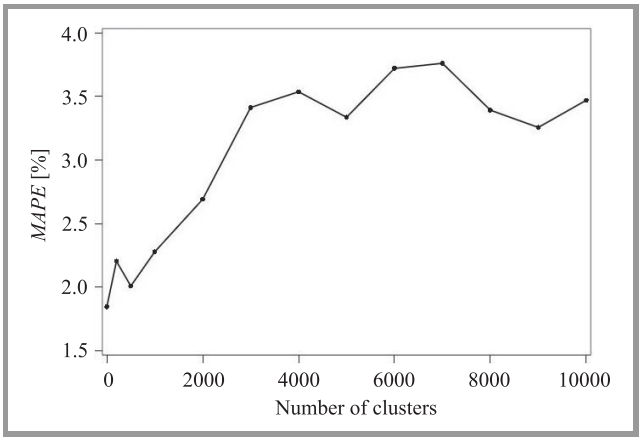


Fig. 6. Prediction accuracy of the total number of GPRS events for different number of clusters verified on 15 month data.

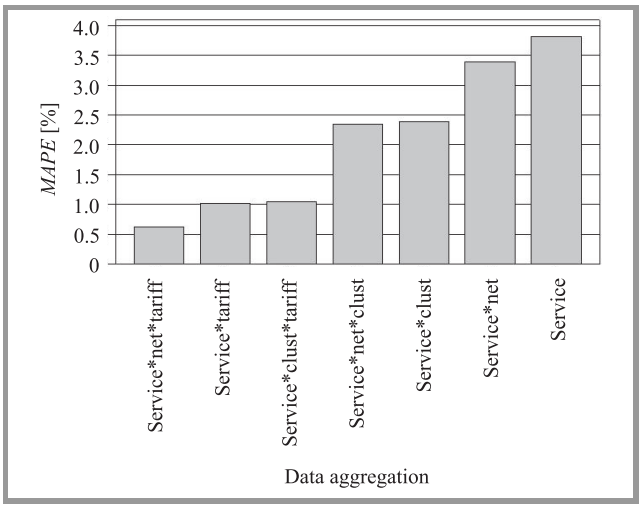


Fig. 7. Prediction accuracy of the total number of voice events (500 clusters).

usages (Fig. 6) and duration of voice calls (Fig. 2) in clusters brought worse results. Probably this is caused by the conjoint analysis model not properly suited to GPRS data. Prediction of the total number of voice calls is better without clustering (Figs. 7–10), however, when predictions of the same value are compared at tariff aggregations, results

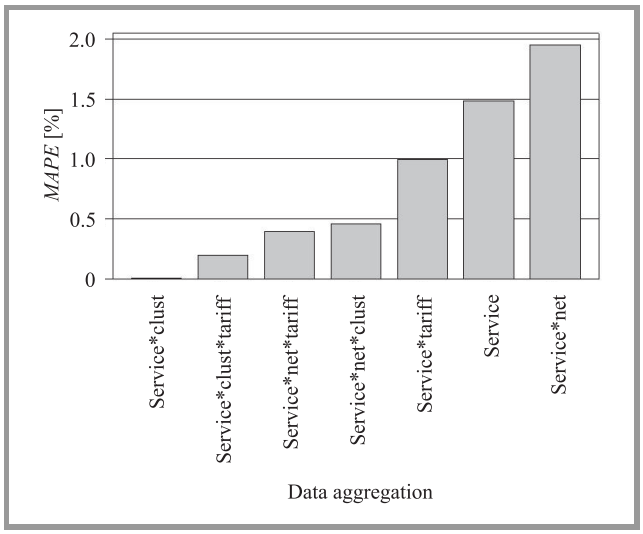


Fig. 8. Prediction accuracy of the total duration of voice events (500 clusters).

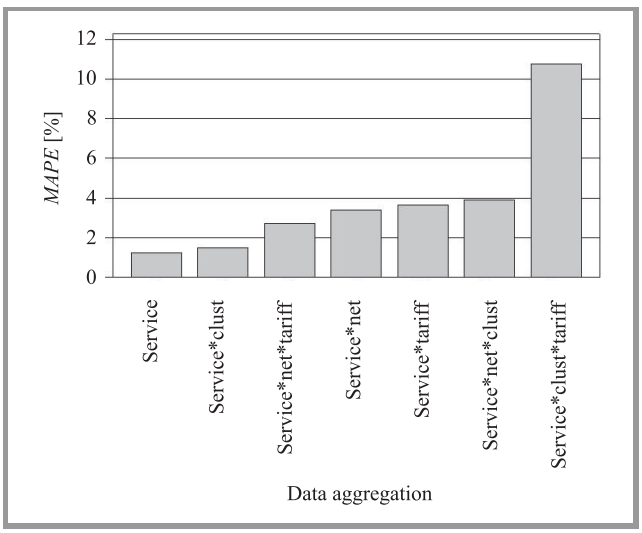


Fig. 10. Prediction accuracy of the total number of the MMS events (1000 clusters).

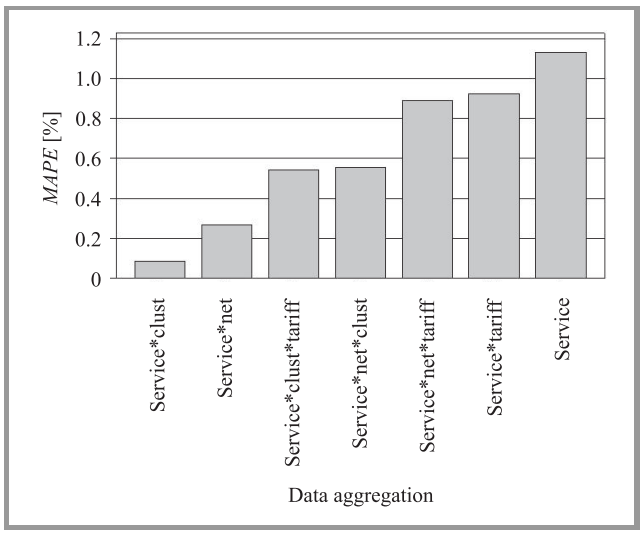


Fig. 9. Prediction accuracy of the total number of SMS events (10000 clusters).

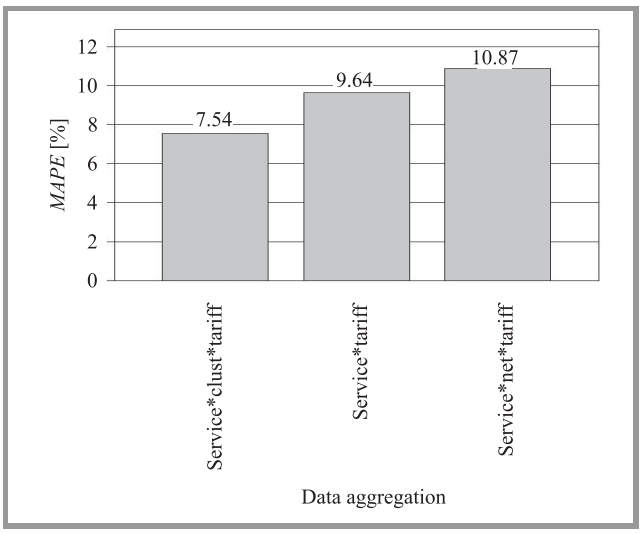


Fig. 11. Prediction accuracy of the total duration of voice events calculated at the tariff level (500 clusters).

are much better what is shown in Fig. 12 and accordingly in Fig. 11 for voice duration. An optimal number of clusters for prediction statistics at the total level as well as a prediction accuracy increase are summarized in Tables 11 and 12.

From Fig. 13, which presents what kind of forecasting models were used, we can find that only logarithmic transformation is applied to telecommunication data and the number of transformations decreases while the number of clusters

Table 11
Optimal number of clusters

Service	Optimal number of clusters	Accuracy increase [p.p.]
Voice – duration	500	1.5
Voice – count	500	0.7
SMS	10000	0.8
MMS	1000	2.0
GPRS	0	0

Table 12
Total service usage prediction

Service	Best data aggregation	Accuracy change after clustering [p.p.]
Voice – duration	Service*cluster	0.4
Voice – count	Service*net*tariff	−0.4
SMS	Service*cluster	0.2
MMS	Service*cluster	1.2
GPRS	Service*net	−1.6

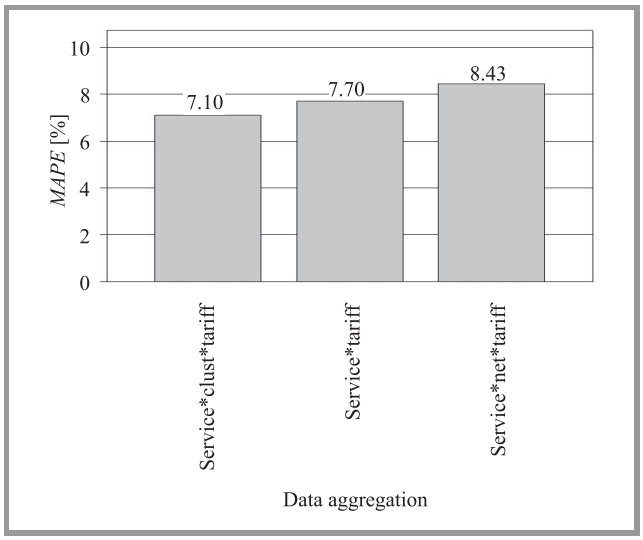


Fig. 12. Prediction accuracy of the total number of voice events calculated at the tariff level (500 clusters).

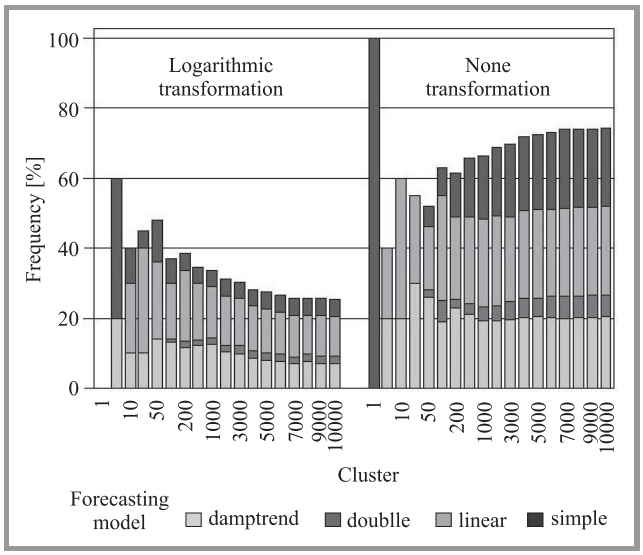


Fig. 13. Statistics of forecasting models applied to the number of voice events prediction.

increases. We can also observe that the number of models called simple, increases as the number of clusters is going up. It shows that forecasting in disaggregated data results is simple and usually more accurate models.

6. Conclusions and Future Research

Analytical results have shown that clustering with the optimal number of clusters, can increase model prediction accuracy. However, good results can be achieved only when the preference model used to identify customers' preferences describes dependencies in data appropriately. The used conjoint measure is not sufficiently suited to data and does not describe GPRS usage properly. The week conjoint measure causes a lack of the prediction accuracy increase in the GPRS time series. On the other hand, poor prediction made after clustering at the top level does not have to cause poor prediction at the lower level. This feature was shown on the number of voice call prediction example, where predictions at the tariff level were much better then at the service level.

In future work more sophisticated forecasting model should be considered. It would be worth to knowing if the multivariate ARIMA models get better results then ESM with proposed improvement. Probably customer preferences could be incorporated into ARIMA models as intervention effects and would also give positive results, what is going to be verified in future work.

References

[1] A. Vag, "Simulating changing consumer preferences: a dynamic conjoint model", *J. Busin. Res.*, vol. 60, no. 8, pp. 904–911, 2007.

[2] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting", *Int. J. Forecast.*, vol. 22, no. 3, pp. 443–473, 2006.

[3] J. S. Armstrong, "Findings from evidence-based forecasting: methods for reducing forecast error", *Int. J. Forecast.*, vol. 22, pp. 583–598, 2006.

[4] M. J. del Moral and M. J. Valderrama, "A principal component approach to dynamic regression models", *Int. J. Forecast.*, vol. 13, no. 2, pp. 237–244, 1997.

[5] P. Rzepakowski, "Supporting telecommunication product sales by conjoint analysis", *J. Telecommun. Inform. Technol.*, no. 3, pp. 28–34, 2008.

[6] SAS Institute Inc., Cary, NC: SAS Institute Inc. SAS/ETS® 9.2: User's Guide, 2008.

Piotr Rzepakowski received the M.Sc. degree in computer science from the Warsaw University of Technology, Poland, in 2003. Currently he is a Ph.D. candidate in computer science at the Warsaw University of Technology (Institute of Control and Computation Engineering). He is employed by National Institute of Telecommunications in Warsaw. Has taken part in projects related to data warehousing and analysis for a telecommunication operator. His research focuses on modeling, decision support, data mining, and customer preferences identication issues.

e-mail: Piotr.Rzepakowski@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

e-mail: P.Rzepakowski@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Paper

Multiobjective Approach to Localization in Wireless Sensor Networks

Michał Marks and Ewa Niewiadomska-Szynkiewicz

Abstract— Wireless sensor network localization is a complex problem that can be solved using different types of methods and algorithms. Nowadays, it is a popular research topic. What becomes obvious is that there are several criteria which are essential when we consider wireless sensor networks. Our objective is to determine accurate estimates of nodes location under the constraints for hardware cost, energy consumption and computation capabilities. In this paper the application of stochastic optimization for performing localization of nodes is discussed. We describe two phase scheme that uses a combination of the trilateration method, along with the simulated annealing optimization algorithm. We investigate two variants of our technique, i.e., centralized and distributed. The attention is paid to the convergence of our algorithm for different network topologies and trade-off between its efficiency and localization accuracy.

Keywords— ad hoc network, localization, simulated annealing, stochastic optimization, wireless sensor network.

1. Introduction to Localization Techniques

Recent advances in wireless communications and electronics have enabled the development of low-cost, low-power and multi-functional sensors that are small in size and communicate in short distances. Cheap, smart sensors, networked through wireless links are deployed in various environments and are used in large number of practical applications, such as environmental information (light, pollution, temperature, etc.), traffic or health monitoring, intrusion detection, etc., [1], [2]. Typical sensor network consists of a large number of nodes – densely deployed sensor devices.

A sensor node by itself is strongly constrained by a low battery power, limited signal processing, limited computation and communication capabilities, and a small amount of memory; hence it can sense only a limited portion of the environment. However, when a group of sensor nodes collaborate with each other, they can accomplish a much bigger task efficiently. In order to do that nodes networked through wireless must gather local data and communicate with other nodes. The information sent by a given sensor is relevant only if we know what location it refers to. Location estimation allows applying the geographic-aware routing, multicasting and energy conservation algorithms. It makes self-organization and localization capabilities one of the most important requirement in sensor networks.

The simplest way to determine a node location is to equip this node with a global positioning system (GPS) or install it at a point with known coordinates. Because of the cost, size of sensors and constraints on energy consumption most sensors usually do not know their locations, only a few nodes, called anchors are equipped with GPS adapters. Location of other nodes, called non-anchors, are unknown. In such model the techniques that estimate the locations of non-anchors based on information about positions of anchors are utilized.

In this paper we define the mathematical model of the distance-based localization, and propose a two phase localization algorithm that uses a combination of the trilateration method, along with the stochastic optimization. We consider two possible implementations: centralized and distributed ones. The efficiency of proposed method strongly depends on the values of control parameters specific to the optimization algorithm. We report the results of numerical tests performed for various values of these parameters. We discuss the results obtained both for centralized and distributed scheme in terms of accuracy and energy efficiency. Finally, we model the localization task as a multiobjective optimization problem, maximizing the localization accuracy while minimizing the localization time.

2. Localization Problem Formulation

Let us formulate the mathematical model of the localization problem for distance-based approaches. There is a network of N nodes (sensors) in $\Re^k$ with bidirectional communication constraints as the edges. Positions of M nodes (anchors) are known. The Euclidean physical distance d_{ij} between the ith and jth nodes can be measured if $(i,j) \in N_i$, where $N_i = \{(i,j) : ||x_i - x_j|| = d_{ij} \leq r\}$ denotes a set of neighbors of node i, $x_i \in \Re^k$ and $x_j \in \Re^k$ true locations of nodes i and j, r is a fixed parameter called transmission range (radio range). Assuming that we have the measurements of distances between all pairs of nodes we can formulate the model of the localization problem that minimizes the sum of squares of errors in sensor positions for fitting the distance measurements:

$$\min_{\hat{x}} \left\{ J(\hat{x}) = \sum_{i=M+1}^{N} \sum_{j \in N_i} (\hat{d}_{ij} - \tilde{d}_{ij})^2 \right\}, \tag{1}$$

where

$$\hat{d}_{ij} = ||\hat{x}_i - \hat{x}_j||, \quad \hat{x}_i \in \Re^k, \quad \hat{x}_j \in \Re^k. \tag{2}$$

The $\hat{d}_{ij}$ denotes an estimated distance between nodes i and j, $\hat{x}_i$ an estimated position of node i and $\hat{x}_j$ an estimated position of a neighbor of node i, $\tilde{d}_{ij}$ a measured distance between nodes i and j.

3. Properties of Localization Techniques

Let us now turn to focus on the properties of localization procedures. Even if we restrict the localization task to distance-based localization with anchors, there is still a number of facets that should be taken into account in design process.

3.1. Centralized versus Distributed Computation

First of all it is necessary to determine if any required computations should be performed locally, by the participants, on the basis of some locally available measurements or all measurements should be reported to a central station that computes positions of nodes in the network and distributes them back to the participants? There are two main issues that should be considered: scaling and efficiency.

Centralized algorithms are designed to run on a central machine with plenty of computational power. Each sensor node gathers the measurements of distances between its and all the neighbors and passes them to the central station where the positions of nodes are calculated. The computed positions are transmitted back into the network. Centralized algorithms overcome the problem of nodes computational limitations by accepting the communication cost of moving data back to the central station. This trade-off becomes less effective as the network grows larger, because it unduly stresses nodes near the base station. Furthermore, the data transmission to the central station involves time delays, so the centralized techniques can not be acceptable in many applications (e.g., mobile nodes).

In contrast, distributed algorithms are designed to run in the network where computation takes place at every node. Each node is responsible for determining its position using information about neighbors. It offers a significant reduction in computation requirements because the number of neighbors is usually not very big (between ten and twenty), so the number of connections is usually a few orders of magnitude less. The use of a distributed computation model is also tolerant to node failures, and distributes the communication cost evenly across the sensor nodes. On the other hand, distributed algorithms implementation is often connected with the loss of information and because of that the results which can be obtained are usually less accurate.

3.2. Speed versus Accuracy

The most important figure of merit for a localization system is the accuracy of its results. Of course the obtained accuracy depends on the selected method, range estimation error, the number of anchors, etc. In case of many methods, especially based on optimization techniques, the accuracy is also dependent on computation time. The open question is when the computation should be stopped and how to decrease the calculation effort?

3.3. Complexity of the Algorithm versus Energy Conservation

In our analysis we consider localization algorithms based on the stochastic optimization. It is obvious they are more complicated than one-hop localization techniques or simple multi-hop localization techniques based only on connectivity described in [3]. Intuitively, the more complex localization algorithm is the better accuracy can be obtained. It is true if we consider only the localization accuracy. However, we have to realize that more complex algorithm is connected with higher energy consumption for data processing and data transmission.

4. Criteria for Distance-Based Localization

Multiple criteria can be formulated for distance-based localization. In our analysis we decided to stress the importance of four criteria which are essential for wireless sensor nodes. The majority of them are connected with economical or technical constraints such as hardware cost, low battery power and limited computation capabilities.

4.1. Localization Accuracy

To evaluate the performance of tested algorithms we used the mean error between the estimated and the true location of the non-anchor nodes in the network, defined as follows:

$$LE = \frac{1}{N-M} \cdot \frac{\sum_{i=M+1}^{N} (||\hat{x}_i - x_i||)^2}{r^2} \cdot 100\%, \qquad (3)$$

where x_i denotes the true position of the sensor node i in the network, $\hat{x}_i$ estimated location of the sensor node i and r the radio transmission range. The location error LE is expressed as a percentage error. It is normalized with respect to the radio range to allow comparison of results obtained for different size and range networks.

4.2. Hardware Cost

Each sensor node is equipped with radio. It is necessary to communicate with other nodes. For example CC2420 radio module, which is very popular, allows the programmer to measure the received signal strength (RSS) that can be used to calculate inter-nodes distances $\tilde{d}_{ij}$ used in the performance function defined in Eq. (1). But many authors says, that this measure is inaccurate [4]. We can obtain more accurate results if we decide to use additional hardware, for example, a sensor board equipped with light, temperature, acoustic signals sensors. Acoustic signals in conjunction with the standard radio module allows to use time difference of arrival (TDoA) technique, which is assumed to be more accurate than RSS. However, additional hardware can significantly increase the sensor node cost (typical sensor board costs approximately the same as a simple node).

4.3. Energy Consumption

In our analysis we consider only the energy consumed at sensor nodes, and we do not take into account the energy consumption for the base station, which is assumed not to be energy constrained. At each sensor node energy is consumed for data processing and data transmission. Energy consumed for data processing depends on the quantity of processed data and the complexity of the performed operations.

4.4. Localization Time

The same as energy also localization time is related to data processing and data transmission. The communication time depends on a network size, efficiency of multi-hop transmission, complexity of the localization technique, and computational power. It is not the aim of our work to improve communication algorithms, but we would like to show how localization algorithms can be improved in order to achieve satisfying accuracy in a short time.

5. The TSA Scheme Description

5.1. Centralized TSA Method

In [5] we proposed the localization technique that uses a combination of the geometry of triangles (trilateration), along with the stochastic optimization. This algorithm operates in two phases.

In the first phase the initial localization is provided. Trilateration uses the known locations of a few anchor nodes, and the measured distance between a given non-anchor and each anchor node. To accurately and uniquely determine the relative location of a non-anchor on a 2D plane using trilateration alone, generally at least three neighbors with known positions are needed. Hence, all nodes are divided into two groups: group A containing nodes with known location (in the beginning only the anchor nodes) and group B of nodes with unknown location. In each step of the algorithm node i, where $i = M+1,\ldots,N$ from the group B is chosen. Next, three nodes from the group A that are within node i radio range are randomly selected. If such nodes exist the location of node i is calculated based on inter-nodes distances between three nodes selected from the group A and the measured distances between node i and these three nodes. The localized node i is moved to the group A. Otherwise, another node from the group B is selected and the operation is repeated. The first phase stops when there are no more nodes that can be localized based on the available information about all nodes location. It switches to the second phase.

Due to the distance measurement uncertainty the coordinates calculated in the first phase are estimated with non-zero errors. Hence, the solution of the first phase is modified by applying stochastic optimization methods. Two techniques, i.e., simulated annealing and genetic algorithm were considered. The numerical results obtained for simulated annealing (SA) were much more promising (see [5], [6]) w.r.t. calculated location accuracy and speed of convergence. So, we decided to focus on this approach. We called our method TSA (trilateration and simulated annealing). The structure of the SA algorithm used in the second phase of TSA is presented in Algorithm 1.

Algorithm 1: Simulated annealing algorithm used in TSA

```
T = T_0, T_0 – initial temperature
Δd = Δd_0, Δd_0 – initial move distance
while T > t_f do
  for i = 1 to P·(N − M) do
      select a node to perturb
      generate a random direction and move a node
      at distance Δd
      evaluate the change in the cost function, ΔJ
      if (ΔJ ≤ 0) then
          //downhill move ⇒ accept it
          accept this perturbation and update
          the solution
      else
          //uphill move ⇒ accept with probability
          pick a random probability rp = uniform(0,1)
          if (rp ≤ exp(−ΔJ/T)) then
              accept this perturbation and update
              the solution
          else
              reject this perturbation and keep the old
              solution
          end if
      end if
  end for
  change the temperature: T_new = α·T, T = T_new
  change the distance Δd_new = β·Δd, Δd = Δd_new
end while
```

From the numerical experiments it was observed that the increased value of the location error is usually driven by incorrect location estimates calculated for a few nodes. The additional functionality (correction) was introduced to the second phase to remove incorrect solutions involved by the distances measurement errors. The detailed description of the correction algorithm can be found in [5].

5.2. Distributed TSA Method

From the numerical experiments performed for the centralized TSA method it was observed that centralized TSA provides quite accurate location estimates even in the case of unevenly distributed nodes with known positions.

However, in this approach we have to gather the measurements of distances between all pairs of network nodes

in a single computer to solve the optimization problem Eq. (1). The data transmission to the central station involves time delays and it can not be used in some application, e.g., mobile networks. In contrast to the centralized method we proposed a fully distributed method where computations take place at every node. In this implementation each node is responsible for determining its position using local information about its neighbors.

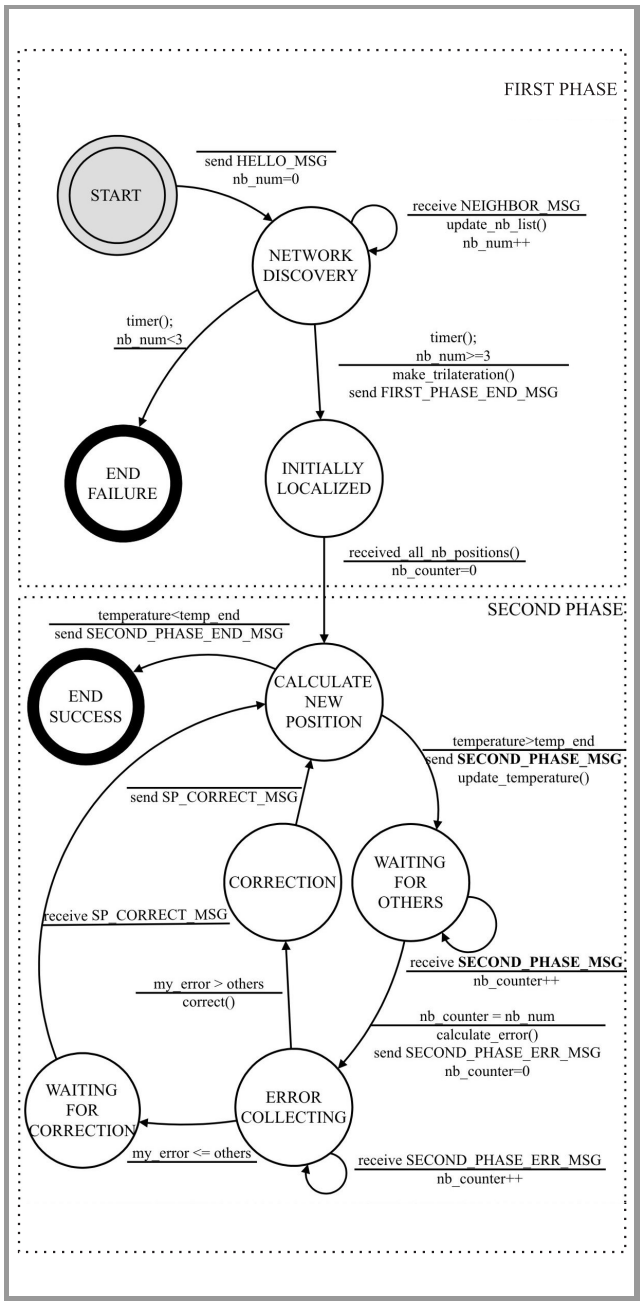


Fig. 1. The state diagram for distributed TSA method.

The state diagram for distributed TSA algorithm is presented in Fig. 1. The estimated position of each node is calculated in parallel. Every P iterations of SA algorithm the neighboring nodes exchange the messages with the current results of calculations.

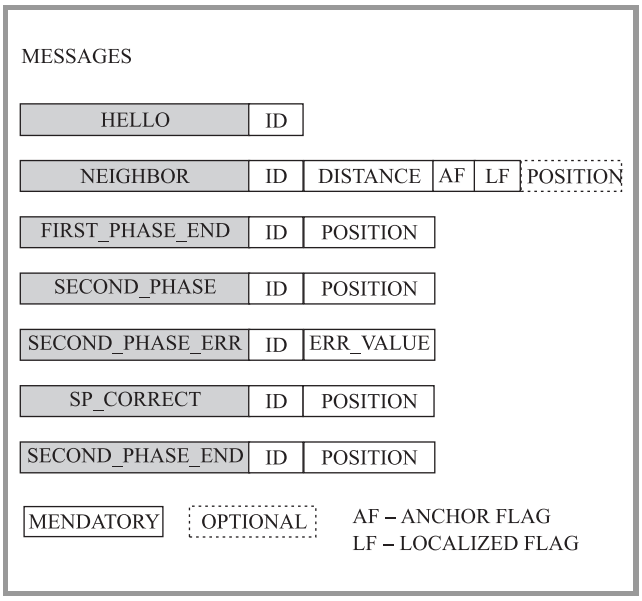


Fig. 2. The scheme of exchanged messages.

The messages structure is illustrated in Fig. 2. Next, the nodes update their location estimates. Many transmissions are needed to obtain a reasonable solution.

6. TSA Scheme Evaluation

We performed many numerical tests to cover a wide range of network system configurations including size of the network (200 – 10000 nodes) and anchor nodes deployment. Especially the anchor nodes deployment seems to be important to evaluate the proposed approaches to sensor network localization. Therefore we prepared a few test problems. Figure 3 depicts four network topologies: a, b with evenly distributed anchor nodes (a – random distribution, b – anchor nodes placed near the edges of a sensor field) and c, d with anchor nodes deployed only in a part of the region to be covered by sensors.

To solve the localization problem Eq. (1) we needed the values of the measured distances between pairs of nodes. In real applications the measured distance $\tilde{d}_{ij}$ between two neighbor nodes is produced by measurement methods described in literature [7], [8]. These methods involve measurement uncertainty; each distance value $\tilde{d}_{ij}$ represents the true physical distance d_{ij} corrupted with a noise describing the uncertainty of the distance measurement. For the purpose of numerical experiments we supposed that this disturbance is described by introducing Gaussian noise with a mean of 0 and a standard deviation of 1 added to the true physical distance d_{ij}:

$$\tilde{d}_{ij} = d_{ij}\,(1.0 + randn() \cdot nf), \tag{4}$$

where nf denotes a noise factor.

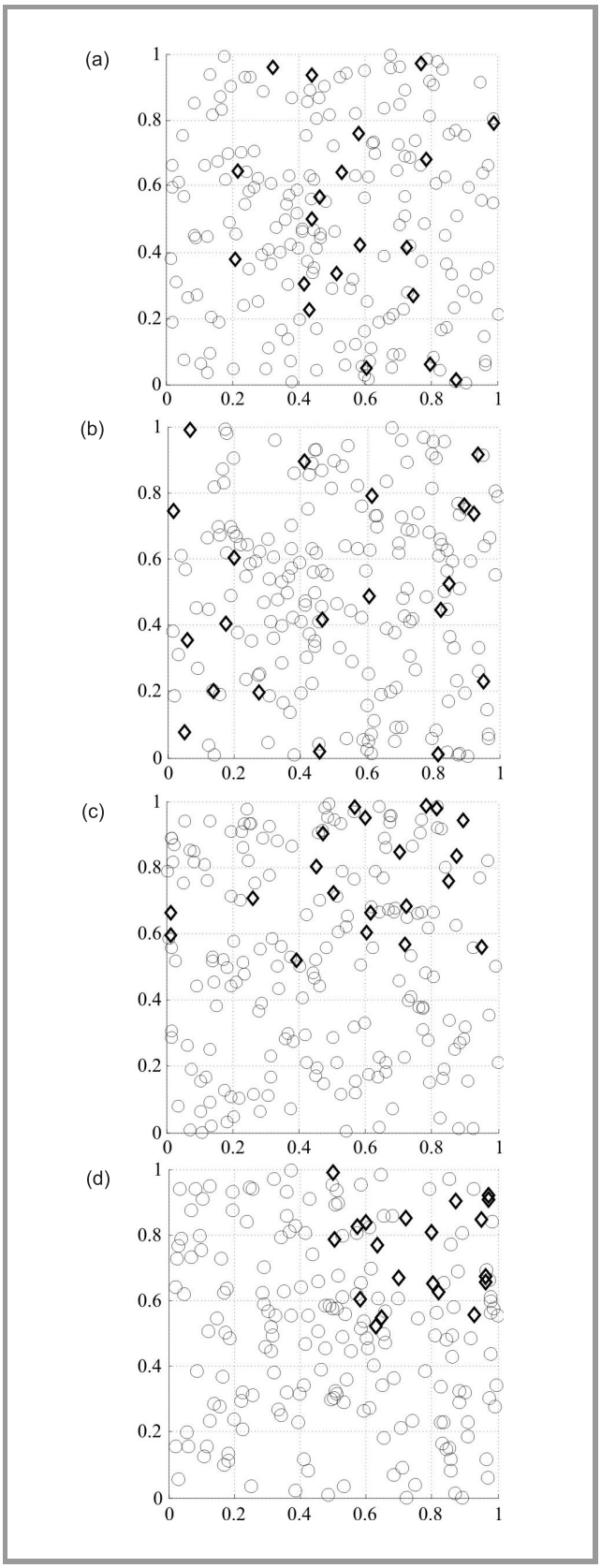


Fig. 3. Test problems four network topologies: a, b with evenly distributed anchor nodes (a – random distribution, b – anchor nodes placed near the edges of a sensor field) and c, d with anchor nodes deployed only in a part of the region to be covered by sensors.

6.1. Centralized versus Distributed TSA Methods

Figure 4 presents the solution quality difference between centralized and distributed algorithms for two test networks (b) and (c) depicted in Fig. 3. The obtained results confirm that from the perspective of location estimation accuracy, centralized algorithm provides more accurate location estimates than distributed one. As a final result we can say that for evenly distributed anchors we obtain quite accurate solution using both methods, otherwise the results of location estimation are much worse in case of distributed version of our scheme.

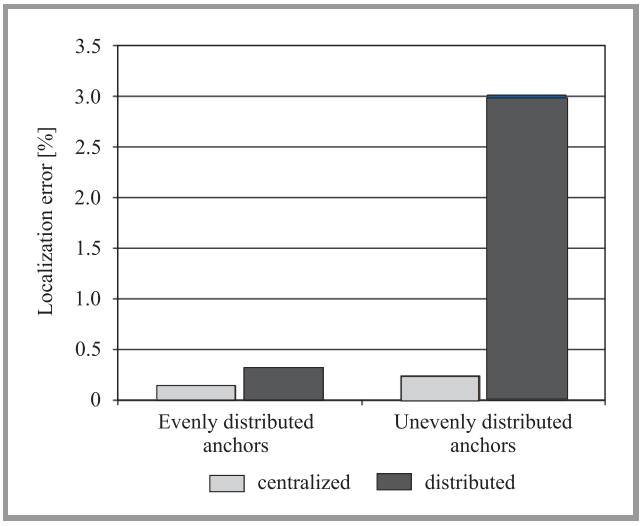


Fig. 4. Localization error for centralized and distributed scheme; test problems b and c.

Distributed version of localization algorithm has many advantages that were discussed in Subsection 3.1. However, distributed algorithm performance is often connected with the loss of information, which was confirmed in simulations (see Fig. 4). There are two reasons of that: loss of information due to parallel computation and loss of information due to the incomplete network map.

6.2. Complexity of the Algorithm versus Energy Use

Let us now turn to the structure of our algorithm. It operates in two phases. In the first phase the auxiliary solution (initial localization) is provided. The solution of the first phase is modified by applying stochastic optimization method in the second phase. Two aspects are worth considering here. First of all what does it mean that auxiliary solution is provided, and how far this solution can be improved in the second phase? The second question is, how the stochastic optimization implies the energy consumption?

The results obtained for centralized algorithm after the first phase and the second phase (final result) are collected in Table 1. The simulations were performed for four test networks depicted in Fig. 3.

From this table we can see that stochastic optimization greatly improves the solution quality. It is obvious that the TSA algorithm needs many iterations to achieve a stable solution. The cost of each iteration, in energy terms, is different for centralized and distributed TSA scheme. Centralized algorithm in large networks requires each sensors measurements to be sent over multiple hops to a central processor, while distributed algorithm requires only local information exchange between neighboring nodes but many such local exchanges may be required, depending on the number of iterations needed to arrive at a stable solution.

Table 1
Localization accuracy for different tasks

Test problem	Localization error (LE) [%]	
	I phase	II phase
a	6.3544	0.1447
b	8.8331	0.1414
c	25.0953	0.1961
d	57.0212	0.3248

In case of centralized implementation energy consumption for localization is asymmetric, because the multi-hop transmission stresses nodes near the central station more than any others. Fortunately this is not a problem because localization task generates only one packet per node which must be transmitted to the base station. In most cases this packet can be transmitted without fragmentation, because of the small amount of data.

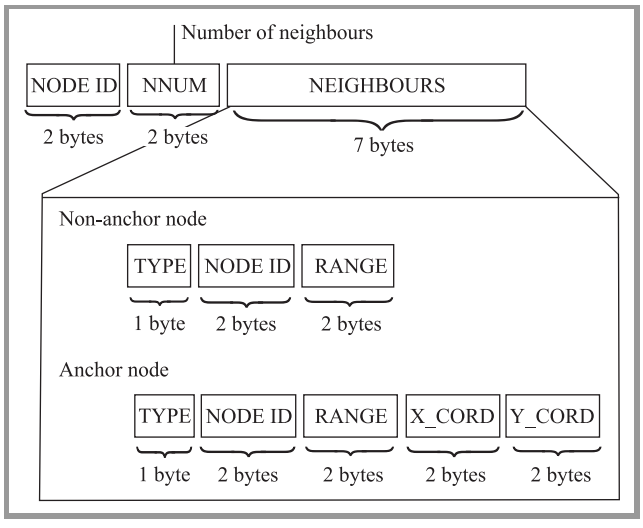


Fig. 5. Localization packet.

Figure 5 presents the localization packet structure. From this figure we can see that even for a node with 10 neighbors the packet size doesn't exceed the fragmentation boundary (approximately 100 bytes) – more detailed information can be found in [9].

Energy consumption becomes a bigger problem for distributed algorithms which require many local information exchanges between neighboring nodes. In the second phase many iterations is needed and each iteration is connected with "SECOND_PHASE_MSG" sending. The problem is depicted by the loop in Fig. 1. The critical message is marked in bold.

6.3. TSA Parameters Tuning

Robustness for anchor nodes deployment. In the paper [6] we have reported the comparison of the results obtained for the TSA method and some other methods. Our scheme seems to be very promising. However, its efficiency and robustness strongly depend on control parameters $\alpha, \beta, \Delta d_0, t_f$ specific to the simulated annealing algorithm used in the second phase of TSA, and depicted in Algorithm 1. All these parameters influence the speed of convergence and accuracy of the solution. To obtain the general purpose algorithm the values of them should be tuned for various network topologies.

We performed the experiments for four test problems presented in Fig. 3. Our aim was to calculate the values of SA parameters: $\alpha, \beta, \Delta d_0, t_f$, depicted in Algorithm 1, which minimize the localization error Eq. (3) for all considered tasks. We solved a decision problem defined as an optimization problem with four criteria (localization errors for tasks a, b, c and d), where all criteria are minimized:

$$\min_{\mathbf{z}}(LE_a(\mathbf{z}), LE_b(\mathbf{z}), LE_c(\mathbf{z}), LE_d(\mathbf{z})), \tag{5}$$

where $\mathbf{z} = [\alpha, \beta, \Delta d_0, t_f]$ denotes a vector of decision variables to be selected within the feasible set, which consists of 48000 elements. Model Eq. (5) specifies that we are interested in minimization of all objective functions and allows us to eliminate insufficient solutions leading to a dominated outcome vectors. After the elimination the Pareto frontier consists of 196 undominated solutions.

In order to select the preferred solution we used a quasi-satisfying approach to multiple criteria optimization – the reference point method [10]–[12]. The model of preferences was created by introducing the reference levels.

Table 2
Aspiration and reservation levels

Reference vector	LE_a	LE_b	LE_c	LE_d
Aspiration levels $\mathbf{r}^a$	0.10	0.10	0.25	0.50
Reservation levels $\mathbf{r}^r$	1.00	1.00	2.00	4.00

We considered two reference vectors: vector of aspiration levels $\mathbf{r}^a$ and vector of reservation levels $\mathbf{r}^r$, which specified acceptable and required values for the localization error (see Table 2).

Depending on the specified reference levels, the partial achievement function s_i can be built and interpreted as a measure of the decision maker satisfaction with the current value of outcome the ith criterion. It is a strictly increasing function of outcome LE_i with value $s_i = 1$ if $LE_i = r_i^a$, and $s_i = 0$ for $LE_i = r_i^r$. We used the piece-wise linear partial achievement function with strong dissatisfaction connected with outcomes worse than the reservation level and s_i value slightly greater than 1 for outcomes better than the aspiration level.

Having all the outcomes transformed into a uniform scale of individual achievements they can be aggregated to form a scalarizing achievement function Eq. (6). Maximization of the scalarizing achievement function generates an efficient solution to the multiple criteria problem:

$$\max_{\mathbf{z}} \left[\min_{i=a,b,c,d} s_i(LE_i) + \varepsilon \cdot \sum_{i=a,b,c,d} s_i(LE_i) \right]. \qquad (6)$$

The solution obtained by solving the problem Eq. (6) was equal:

$$\mathbf{z} = \begin{bmatrix} \alpha \\ \beta \\ \Delta d_0 \\ t_f \end{bmatrix} = \begin{bmatrix} 0.94 \\ 0.98 \\ 0.26 \\ 10^{-13} \end{bmatrix}.$$

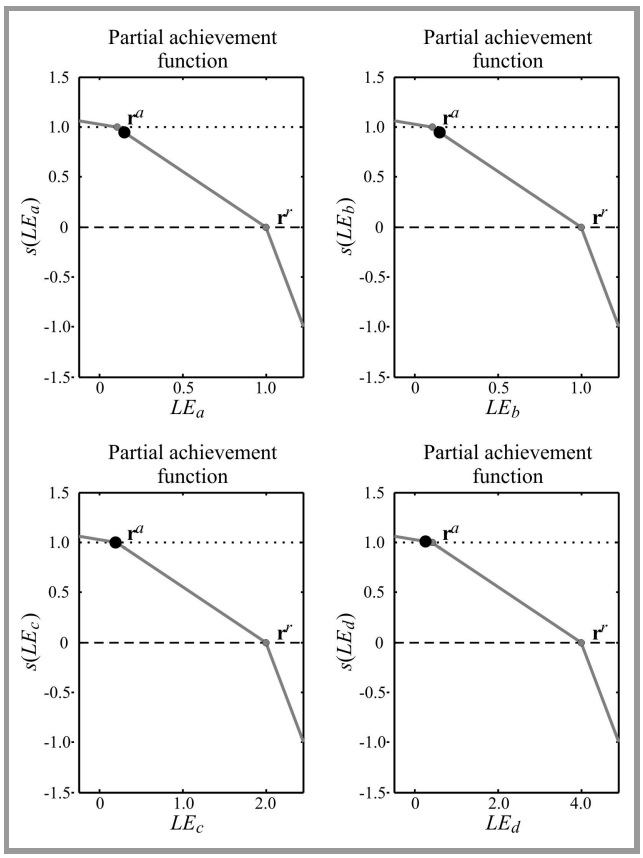


Fig. 6. Partial achievement functions.

The corresponding objective and partial achievements values are collected in Table 3.

Table 3
Values in criterion space for selected solution

Task	Localization error LE	Partial achievement value $s(LE)$
a	0.1447	0.9503
b	0.1414	0.9540
c	0.1961	1.0077
d	0.3248	1.0125

Partial achievements functions are also depicted in Fig. 6. The solution is marked with the dot for each partial achievement function.

A trade-off between efficiency and accuracy. Time consumed on localization in case of centralized algorithm increases proportionally to the network dimension, as it can be seen in Table 4.

Table 4
Localization error and computation times for different network sizes

Number of nodes	Localization error LE [%]	Computation time [s]
200	0.11	1.4
500	0.15	7.6
1000	0.29	29.4

The trade-off between efficiency and accuracy is expected. To decrease the calculation effort the optimal value of another SA control parameter (P) have to be estimated. In the SA implementation used in the second phase of the TSA scheme at each value of the coordinating parameter T (temperature), $P(N-M)$ non-anchor nodes are randomly selected for modification (where N denotes the number of sensors in the network, M the number of anchors, and P a reasonably large number to make the system into thermal equilibrium). The parameter P plays the important role – it influences the estimated location accuracy and calculation time.

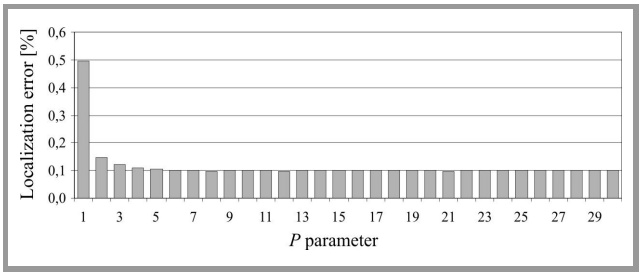


Fig. 7. Localization error for various values of P.

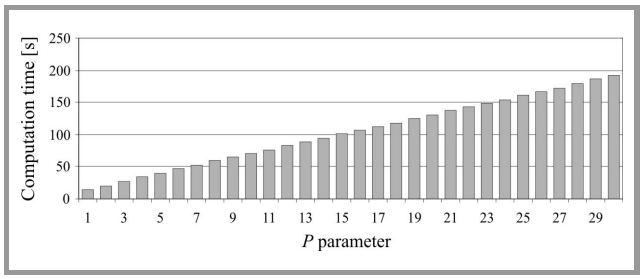


Fig. 8. Computation times for various values of P.

Figures 7 and 8 present the results of numerical tests performed for the network with 2000 nodes and various values of P.

To calculate the optimal value of the parameter P for a given network we can solve the two-criterion optimization problem:

$$\min_{P}(\Delta t, LE), \tag{7}$$

where Δt denotes a calculation time, LE a localization error defined in Eq. (3).

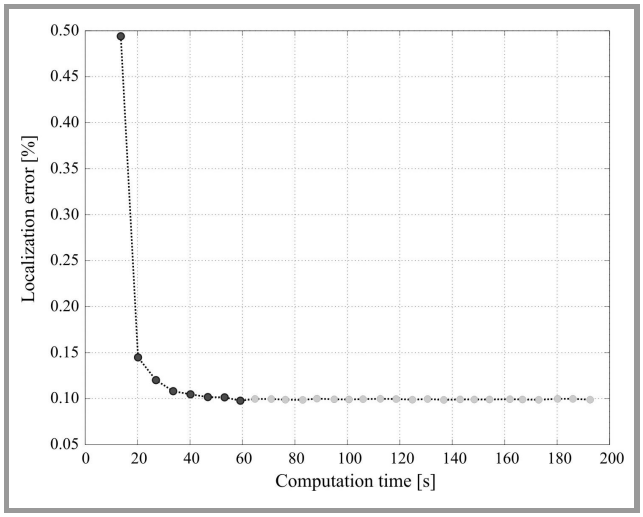


Fig. 9. The solution of the problem (7) for the network with 2000 nodes.

Figure 9 illustrates the Pareto frontier for the network of 2000 nodes. In order to select the preferred solution we also used the reference point method. As an aspiration and reservation level we assumed that computation time can not exceed ten seconds, and localization error must be less then 1% (see Table 5).

Table 5
Aspiration and reservation levels

Reference vector	Calculation time [s]	Localization error [%]
$\mathbf{r}^a$	10	0.10
$\mathbf{r}^r$	60	1.00

In Table 6 values of partial achievement functions for both criteria and the scalarizing achievement function for all undominated solutions are presented. We can see that the best value of achievement is for the solution calculated for $P = 2$.

Table 6
Undominated solutions and corresponding achievement function

P	Computation time [s]	Localization error LE	PAF*		SAF**
			$s(t)$	$s(LE)$	
1	13.6	0.4940	0.9280	0.5622	0.5624
2	**20.2**	**0.1448**	**0.7960**	**0.9503**	**0.7962**
3	27.0	0.1201	0.6600	0.9777	0.6602
4	33.6	0.1081	0.5280	0.9910	0.5282
5	40.2	0.1047	0.3960	0.9948	0.3961
6	46.8	0.1017	0.2640	0.9981	0.2641
7	53.2	0.1014	0.1360	0.9985	0.1361
8	59.2	0.0977	0.0160	1.0006	0.0161

* PAF – partial achievement function,
** SAF – scalarizing achievement function.

The optimal values of parameter P corresponding to the solutions of the task Eq. (7) for different networks are illustrated in Table 7. Because TSA should be the general purpose localization scheme that can be used to different dimension problems we solved the more general problem for five networks with various dimensions (automatically the number of criteria was ten). The preferred solution was obtained for $P = 4$. The results of calculations performed for network with 200 to 10000 nodes and $P = 4$ are presented in Table 8.

Table 7
Optimal values of parameter P for different size of network

Number of nodes	200	500	1000	2000	4000
Calculated P	4	4	4	2	2

Table 8
Localization errors and computation times for different sizes of network

Number of nodes	LE [%]	t [s]
200	0.1275	0.4
500	0.4124	2.2
1000	0.1387	8.0
2000	0.1081	33.6
4000	0.1086	125.8
5000	0.1581	189.8
10000	0.1193	790.4

7. Summary and Conclusions

In this paper we outline the main properties and criteria that should be considered while estimating the location of nodes with unknown positions in the sensor network. We stressed the importance of such criteria like localization accuracy, hardware cost, energy consumption and calculation capabilities. The main objective was to develop the efficient and robust localization algorithm. We presented and evaluated the hybrid scheme that combines simple geometry of triangles and stochastic optimization technique. The big effort was on tuning the parameters of the optimization algorithm. Finally, we demonstrated that our method provides quite accurate location estimates in the sensible computing time even in the case of unevenly distributed nodes with known positions.

Acknowledgement

This work was supported by Ministry of Science and Higher Education grant 4/TINFO/2008.

References

[1] P. H. Bauer, "New challenges in dynamical systems: the networked case", *Int. J. Appl. Math. Comput. Sci.*, vol. 18, no. 3, pp. 271–278, 2008.

[2] G. Mao, B. Fidan, and B. D. O. Anderson, *Sensor Network and Configuration: Fundamentals, Techniques, Platforms and Experiments*. Berlin: Springer, 2006, pp. 281–316.

[3] Y. Shang, W. Ruml, Y. Zhang, and M. Fromherz, "Localization from connectivity in sensor networks", *IEEE Trans. Paral. Distrib. Syst.*, vol. 15, no. 11, pp. 961–974, 2004.

[4] N. Patwari, A. O. Hero III, M. Perkins, N. S. Correal, and R. J. O'Dea, "Relative location estimation in wireless sensor networks", *IEEE Trans. Sig. Proces.*, vol. 51, no. 8, pp. 2137–2148, 2003.

[5] M. Marks and E. Niewiadomska-Szynkiewicz, "Genetic algorithm and simulated annealing approach to sensor network localization", in *Proc. KAEiOG'07 Conf.*, Będlewo, Poland, 2007, pp. 193–202.

[6] M. Marks and E. Niewiadomska-Szynkiewicz, "Two-phase stochastic optimization to sensor network localization", in *SENSORCOMM 2007 Proc. Int. Conf.*, Valencia, Spain, 2007, pp. 134–139.

[7] G. Mao, B. Fidan, and B. D. O. Anderson, "Wireless sensor network localization techniques", *Comp. Netw. Int. J. Comput. Telecommun. Netw.*, vol. 51, no. 10, pp. 2529–2553, 2007.

[8] *Handbook of Sensor Networks: Algorithms and Architectures*, I. Stojmenović, Ed. Wiley Series on Parallel and Distributed Computing. New York: Wiley, 2005.

[9] J. W. Hui and D. E. Culler, "Extending ip to low-power, wireless personal area networks", *IEEE Internet Comp.*, vol. 12, no. 4, pp. 37–45, 2008.

[10] W. Ogryczak, "WOWA enhancement of the preference modeling in the reference point method", in *Proc. MDAI Conf.*, Barcelona, Spain, 2008, pp. 38–49.

[11] W. Ogryczak, A. Wierzbicki, and M. Milewski, "Fair and efficient network dimensioning with the reference point methodology", *J. Telecommun. Inform. Technol.*, no. 4, pp. 21–30, 2006.

[12] A. P. Wierzbicki, "A mathematical basis for satisficing decision making", *Math. Model.*, no. 3, pp. 391–405, 1982.

Michał Marks received his M.Sc. in computer science from the Warsaw University of Technology, Poland, in 2007. Currently he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2007 he works at the Research and Academic Computer Network (NASK). His research area focuses on global optimization, multiple criteria optimization, decision support and machine learning.

e-mail: mmarks@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

e-mail: M.Marks@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Ewa Niewiadomska-Szynkiewicz received her Ph.D. in 1996, D.Sc. in 2006 in control and computation engineering from the Warsaw University Technology, Poland. Since 1987 she works at the Warsaw University of Technology, as the Head of Complex Systems Group, and since 2000 at the Research and Academic Computer Network (NASK), as the Head of Network Modeling and Simulation Group. Her research interests focus on complex systems modeling and control, computer simulation, global optimization, parallel calculations and computer networks.

e-mail: ens@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

e-mail: ewan@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

Comparative Study of Wireless Sensor Networks Energy-Efficient Topologies and Power Save Protocols

Ewa Niewiadomska-Szynkiewicz, Piotr Kwaśniewski, and Izabela Windyga

Abstract— Ad hoc networks are the ultimate technology in wireless communication that allow network nodes to communicate without the need for a fixed infrastructure. The paper addresses issues associated with control of data transmission in wireless sensor networks (WSN) – a popular type of ad hoc networks with stationary nodes. Since the WSN nodes are typically battery equipped, the primary design goal is to optimize the amount of energy used for transmission. The energy conservation techniques and algorithms for computing the optimal transmitting ranges in order to generate a network with desired properties while reducing sensors energy consumption are discussed and compared through simulations. We describe a new clustering based approach that utilizes the periodical coordination to reduce the overall energy usage by the network.

Keywords— ad hoc network, energy conservation protocols, topology control, wireless sensor network.

1. Introduction to Ad Hoc and Wireless Sensor Networks

An ad hoc network is a wireless decentralized structure network comprised of nodes, which autonomously set up a network. No external network infrastructure is necessary to transmit data – there is no central administration. Freely located network nodes participate in transmission. Network nodes can travel in space as time passes, while direct communication between each pair of nodes is usually not possible. Generally, ad hoc network can consist of different types of multi functional computation devices.

Wireless sensor network (WSN) is most often set up in an ad hoc mode by means of small-size identical devices grouped into network nodes distributed densely over a significant area. These devices, each equipped with central processing unit (CPU), battery, sensor and radio transceiver networked through wireless links provide unparalleled possibilities for collection and transmission of data and can be used for monitoring and controlling environment, cities, homes, etc. In most cases WSNs are stationary or quasi-stationary, while node mobility can be ignored. There is no prearrangement assumption about specific role each node should perform. Each node makes its decision independently, based on the situation in the deployment region, and its knowledge about the network. In the case of networks comprising several hundreds or thousands of nodes, it is necessary to choose an architecture and technology which will enable relatively cheap production of individual devices. For this reason, WSNs need some special treatment as they have unavoidable limitations, for example, limited amount of power at their disposal. Each battery powered device, participating in WSN needs to manage its power in order to perform its duties as long, and as effective as possible. Wireless sensors are thus characterized by low processing speed, limited memory and communication range.

Wireless sensor networks [1]–[3] can be used in different environments and situations and perform tasks of different kinds. Their application will condition the network topology and the choice of technology for its production. The network protocols used in the case of networks whose operating range covers a single building will differ from those operating within large space areas. The construction of a network capable of performing its task requires obtaining information on the devices (nodes) it comprises. The crucial data is the following: geographical location of network nodes, admissible power of radio transmitter and options for control of signal power, estimated number of network nodes, number of nodes that can be lost before the network is declared non-operational, assumed network functionality (maximization of nodes operational time, maximization of throughput, etc.).

In our paper we discuss the approaches to design the optimal w.r.t. minimal energy consumption WSN topologies. The short description of communication methods, energy conservation techniques (power save protocols) and algorithms for computing the optimal transmitting ranges in order to generate a network with desired properties while reducing sensors energy consumption (topology control protocols) is provided. Power save protocols attempt to save nodes energy by putting its radio transceiver in the sleep state. Topology control protocols are responsible for providing the routing protocols with the list of the nodes' neighbors, and making decisions about the ranges of transmission power utilized in each transmission. We analyze the properties of two location based distributed topology control protocols, and report the results of simulation experiments covering a wide range of network system configurations. Finally, we discuss the idea of our novel location based power save scheme utilizing hierarchical structure with periodic coordination of network nodes activity.

2. Communication Methods

Communication protocols used in modern wireless networks like IEEE 802.11 or Bluetooth (IEEE 802.15.1) enable ad hoc mode operation. However, for the protocols to operate in this mode in practice, several basic issues must be solved [2]–[4]. The most important ones are:

- **Limited resources**. Nodes comprised by the network are often small battery-fed devices, which means their power source is limited. The network's throughput is also limited.
- **Poor quality of connection**. The quality of wireless transmission depends on numerous external factors, like weather conditions or landform features. Part of those factors change with time.
- **Small communication range**.

Small communication range in WSN networks results in communication limitations. Each node communicates only with the nodes present in its closest vicinity – the neighbors. For this reason, the natural communication method in wireless sensor networks is the multihop routing. When using multihop routing, it is assumed that the receiving node is located outside the transmitter's range. Contrary to single-hop networks, the transmitter must transmit data to the receiver by means of intermediate nodes. This is a certain limitation that hinders the implementation of routing algorithms but enables the construction of network of greater capacity. Multihop network enables simultaneous transmission via many independent routes. Independence of routes reduces the interference between individual nodes, which additionally enhances the wireless transmission speed in comparison to single-hop networks, where devices share a common space.
Individual WSN network node can collect data recorded by sensors but do not have enough power to process it. Moreover, analyzes require collection of information from many points. Therefore, efficient inter-node communication is necessary in order to transfer data to the base station.

3. Topology Control

Transmission of data package between two network nodes x_i and x_j requires power proportional to d_{ij}^2, where d_{ij} denotes the Euclidean distance between sender and receiver. Lets assume that instead of performing direct transmission, a relay node x_k is used. In such case two transmissions need to be performed: from a source node x_i to a relay node x_k (distance d_{ik}) and from the node x_k to the destination node x_j (distance d_{kj}). Lets consider a triangle $x_i x_k x_j$, also let α be an angle at vertex x_k. By elementary geometry we have:

$$d_{ij}^2 = {d_{ik}}^2 + {d_{kj}}^2 - 2 d_{ik} d_{kj} \cos\alpha, \tag{1}$$

when $\cos\alpha \leq 0$, total amount of energy spent to transmit a data package is smaller when a relay node is used.

Generally, short transmissions in the network are desired. They involve smaller power consumption and cause less interference in a network, simultaneously effected, transmissions, thus increasing the network throughput. In general, the goal of topology control (TC) [3] is to identify the situation when the using of the relay node is more energy-efficient than direct transmission and create the network topology accordingly. Topology control assumes that the nodes have impact on the power used to transmit a message. The basic task of TC algorithm consists in attributing the level of power used to send messages to every node in order to minimize the amount of power received from the power source, while at the same time maintaining the coherence of the network.

3.1. Topology Control Protocols

Topology control protocols are responsible for providing the routing protocols with the list of nodes' neighbors, and making decisions about the ranges of transmission power utilized in each transmission. The open systems interconnection (OSI) network model assumes that routing task is dealt with the network layer. On the other hand all functions and procedures required to send data through the network are stored in the OSI data link layer. Therefore the topology control layer is placed partially in the OSI network layer and the OSI data link layer, as presented in Fig. 1.

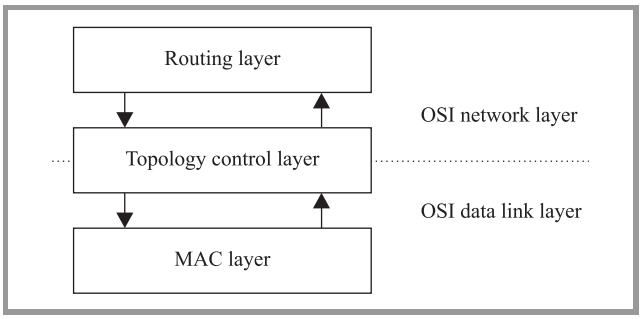


Fig. 1. Placement of topology control layer in the OSI stack.

Topology control protocols may utilize various information about a network, nodes localization and resources [3]–[5]. We can divide these protocols into several groups.

- **Homogeneous topology control protocols** assume that each node uses the same value of transmission power, which reduces the problem to simpler task of finding the minimal level of transmit power such that certain network property is achieved.
- **Location based topology control protocols** utilize the information about geographical location of nodes in the deployment area.
- **Neighbor based topology control protocols** assume that no information about location of nodes is available but each node can determine set of its neighbors and build an order on this set. Order may be based on round trip time, link quality or signal strength.

3.2. Location Based Protocols

We implemented and tested two location based protocols: R&M developed by Rodoplu and Meng, described in [6] and LMST (local minimum spanning tree) proposed by Li, Wang and Song in [7]. The short description of these techniques is provided.

The R&M and LMST protocols. Let N be a set of n wireless nodes deployed in the certain region and forming WSN. Assuming that R_i denotes the maximal transmission range assigned to ith node we can generate the communication graph $G = (N, E)$ induced by R on a given WSN. The E denotes a set of directed edges, and the directed edge $[x_i, x_j]$ exists if x_i and x_j are neighbors, i.e., $d_{ij} \leq R_i$, where d_{ij} denotes the Euclidean distance between sender and receiver. The communication graph G obtained when all the nodes transmit at maximum power is called *maxpower graph*.
Let us consider the situation when all nodes transmit the collected data to one (or more) master node(s) x_m – a base station(s). We can formulate the minimum energy all-to-one communication problem of calculating the optimal reverse spanning tree T of maxpower graph G rooted at x_m:

$$\min_{T} \sum_{x_i \in N, i \neq m} C(x_i, Pred_T(x_i)), \tag{2}$$

where $Pred_T(x_i)$ denotes the predecessor of ith node in the spanning tree T and $C(\cdot)$ the energy cost of transmission from x_i to its predecessor.
The R&M protocol calculates the most energy-efficient path from any node to the master node. It is composed of two phases.

- **Phase 1**. The goal is to compute the enclosure graph of all nodes in WSN. Each node sends a broadcast message, at maximum power, containing its ID and location information. As such message is received by x_i from any neighbor node, x_i identifies the set of nodes locations for which communicating through relay node is more energy efficient than direct communication (the relay region of x_i). Next, x_i checks if the newly found node is in the relay region of any previously found neighbors. A node is marked *dead* if it lays in the relay region of any neighbor of x_i, and *alive* otherwise. After receiving broadcast messages from all neighbors, the set of nodes marked with *alive* identifier creates the enclosure graph of x_i.

- **Phase 2**. In the second phase the optimal, i.e., minimum-energy reverse spanning tree rooted at the master node is computed. The Bellman-Ford algorithm [8] for shortest path calculation is used on the enclosure graph that was determined in the phase 1. Each node computes the minimal cost, i.e., minimal energy to reach the master node given the cost of its neighbors, and broadcasts the message with this value at its maximum power. The operation is repeated every time a message with a new cost is received. After all nodes determine the minimum energy neighbor link, the optimal topology is computed.

The second considered protocol LMST can be used to WSN with nodes equipped with transceivers with the same maximum power. LMST operates in three phases.

- **Phase 1**. Each node sends a broadcast message, at maximum transmit power, containing its ID and location information to its one hop neighbor in the maxpower graph.

- **Phase 2**. The topology is generated. Each node determines a set of its neighbors, calculates Euclidean distance to every neighbor, and finally creates a minimum spanning tree based on its neighbors and computed distances (edge weights in the MST). Final network topology is derived from local MST created by all nodes. Neighbor set of each node consists of nodes, which are its direct neighbors in its local MST. Unfortunately, created topology may contain unidirectional links. Two approaches are proposed to solve this problem: it is assumed that all of them are bidirectional links or all unidirectional links are removed.

- **Phase 3**. Transmission power required to reach every neighbor in a given topology is calculated based on the broadcast messages transmitted in the first phase. Based on the measurements of power of the broadcast messages and knowledge about power level used when transmitting the message, it is possible to compute power level needed to reach the target neighbor.

Simulation results. The performance of R&M and LMST in terms of energy conservation was investigated through simulation. We carried out a set of experiments for various wireless sensor network topologies. It was assumed that all data collected in sensors were transmitted to one base station. We compared the results obtained using both algorithms with those when energy consumption was not considered while routing calculation. The key metric for evaluating the listed methods was the energy consumption used for data transmission. All experiments were conducted using the popular software for network simulation – ns-2 [9]. We implemented R&M and LMST protocols based on modules provided in ns-2 library of classes. The sensor networks with 50 – 300 nodes simulating the commercially available MICA2 sensors [10] with randomly generated positions in a square regions 400×400 to 3000×3000 were considered in our experiments. The technical parameters of sensors were taken from [11], i.e., the radio power consumption for transmission was from 8.6 mA (RF transmission power -20 dBm) to 25.4 mA (RF transmission power 5 dBm), the initial energy resource of each node was assumed to be equal to 21 kJ.

The objective of the first series of simulations was to compare the topologies calculated using described algorithms. The results are presented in Figs. 2 – 4. The base station is marked with the bold dot in presented figures. Figures 3 and 4 show the topologies formed using the LMST and R&M protocols. The obtained results can be compared with the topology generated without utilizing any TC algorithm (Fig. 2).

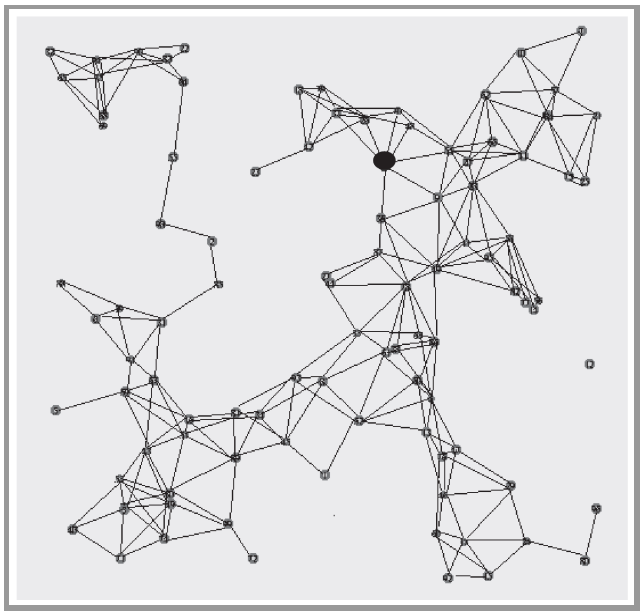


Fig. 2. Topology calculated without TC protocols.

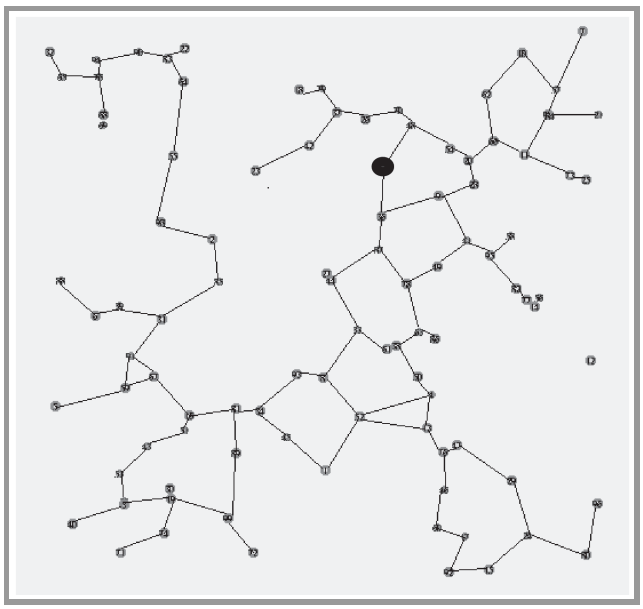


Fig. 3. Topology calculated using LMST method.

The second case study was related to simulation of data transmission in WSNs. Different sizes of networks were examined. In this experiment it was assumed that each node in WSN generates a single message that has to be delivered

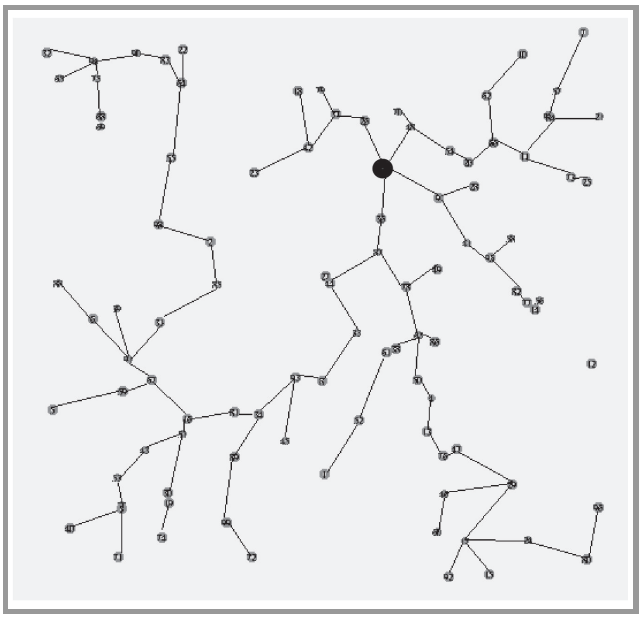


Fig. 4. Topology calculated using R&M method.

to the base station. In addition all nodes could play the role of relay nodes. The shortest path from each node to the destination was calculated taking into account topologies generated using R&M and two versions of LMST: LMST0 (topology can contain unidirectional links), LMST1 (topology contains only bidirectional links). The total energy consumed by all nodes for data transmission was divided by the number of nodes.

Figure 5 depicts the results of calculations, i.e., the average energy used by one node in WSN for data transmission.

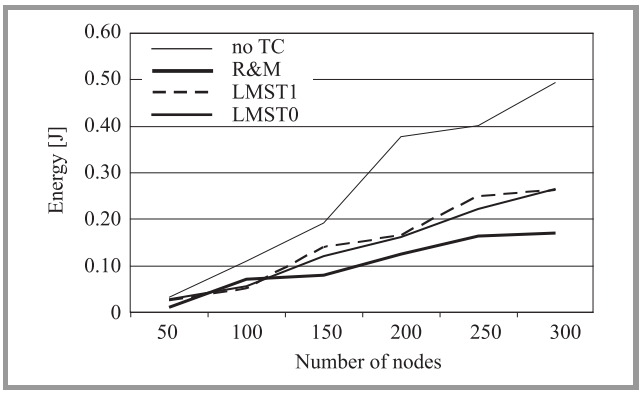


Fig. 5. Average energy consumption by one node for single transmission to the base station; different TC methods and network size.

Figures 6 and 7 show the average amount of energy used by one node for data transmission in case of different TC protocols, number of relay nodes transmitting to the base station and distance to the base station. WSN with 150 nodes was considered. It can be observed that in case of R&M and LMST protocols the energy usage for transmission in the whole network decreases while increasing the number

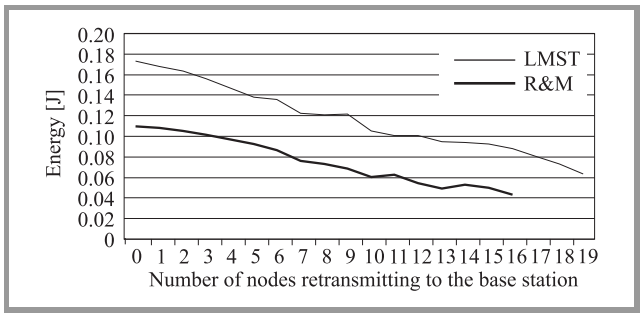


Fig. 6. Average energy usage for transmission w.r.t. the number of relay nodes; different TC methods.

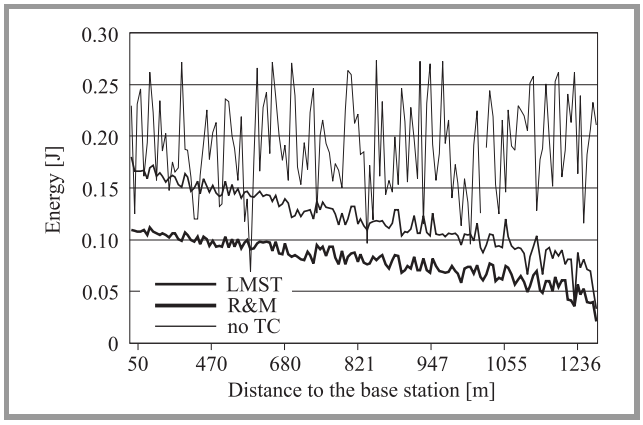


Fig. 7. Average energy usage for transmission w.r.t. the distance to the base station; different TC methods.

of relay nodes transmitting to the base station. It is obvious that the energy used for data transmission by nodes located far from the base station is smaller than those used by nodes closed to the master node, which have to retransmit a lot of messages (Fig. 7).

Table 1 contains the average number of messages generated by one node in WSN that can be transmitted to the base station up to its batteries are dead. The results obtained for different networks and topologies are compared.

Table 1
Average number of messages transmitted by one node to the base station

TC methods	Network size			
	150	200	250	300
Without TC	109 950	55 633	52 380	42 543
R&M	261 241	167 177	127 328	123 549
LMST0	173 893	130 485	94 130	78 850
LMST1	150 233	126 389	84 181	80 001

Discussion. The R&M and LMST protocols can be successfully used to calculate optimal topology in many WSN application scenarios. Both methods have to spent some energy to build the topology, which is concerned with beacon messages broadcasting in the first phase of their operation. However, the energy consumption for topology generation is small, i.e., LMST – 0.0011 J and R&M – 0.052 J for WSN of 50 nodes and energy resource of each node equal 21 kJ. Both protocols generate energy-efficient topologies (see Fig. 5). The energy consumption for data transmission in case of small size of the network (less than 150 nodes) is similar, while using topologies formed by R&M and LMST. In case of large size networks the R&M protocol seems to be much more efficient.

In summary, both techniques generate different topologies and have some advantages and drawbacks. In case of R&M we obtain more energy-efficient topologies but two potential drawbacks of the algorithm can be observed. The computation performed in the second phase of R&M requires the exchange of global information, which induces message overhead, and the explicit radio propagation model is used to compute the optimal topology. Hence, the calculated topology strongly depends on the accuracy of the channel model. Data transmission while applying the LMST protocol is more energy-intensive, but created topology is more robust and preserves connectivity in the worst case. In addition, it can be computed in a fully distributed fashion.

4. Energy Conservation

4.1. Power Consumption

The handling of the wireless transceiver contributes significantly to the node's overall energy consumption. Depending on the state of the transceiver, different levels of power consumption are being observed. Table 2 summarizes the sample power consumption of some 802.11 wireless interfaces.

In order to extend the working time of individual devices, it is frequent practice that some node elements are deactivated, including the radio transceiver. They remain inactive for most time and are activated only to transmit or receive messages from other nodes. Radio transceiver in WSN network node can operate in one out of four modes, which differ in the consumption of power necessary for proper operation: transmission – signal is transmitted to other nodes (greatest power consumption), receiving – message from other node is received (medium power consumption), stand-by (idle) – transceiver inactive, turned on and ready to change to data transmission or receiving (low power consumption), sleep – radio transceiver off.

Table 2
Aspiration and reservation levels

Interface	Power consumption [W]			
	transmit	receive	idle	sleep
Aironet PC4800	1.4–1.9	1.3–1.4	1.34	0.075
Lucent Bronze	1.3	0.97	0.84	0.066
Lucent Silver	1.3	0.90	0.74	0.048
Cabletron Romabout	1.4	1.0	0.83	0.13
Lucent WaveLAN	3.10	1.52	1.5	–

4.2. Power Save Protocols

The power-saving protocols used in sensor networks impose reduced consumption by putting the radio transceiver into the sleep mode. The use of such protocols involves the limitation of accessible band, and can also interrupt the data transfer in the network. Adequate choice of radio transceiver's switch-off time introduces further difficulty in the implementation of network protocols. The literature (e.g., [3]) present algorithms designed to limit the power consumption while simultaneously minimizing the negative impact on the network throughput and on the efficiency of data transmission routing. Different types of protocols are used depending on the application of the network. Two categories can be distinguished.

- **Synchronous power save protocols**, where it is assumed that nodes periodically wake up to exchange data packets. The sleep cycles of all nodes are globally synchronized. The main issue is to adjust length of sleep and wake phases that will minimize energy consumption and impact on a given network's throughput.
- **Topology based power save protocols**, where a subset of nodes which topologically covers whole network is selected. Nodes belonging to this set are not allowed to operate in the sleep mode. Other nodes are required to be periodically awake in order to receive incoming traffic.

Power save protocols should be capable of buffering traffic destined to the sleeping nodes and forwarding data in partial network defined by the covering set. The covering set membership needs to be rotated between all nodes in the network in order to maximize the life time of the network.

It was observed that grouping sensor nodes into clusters can reduce the overall energy usage in a network. Clustering based algorithms seems to be the most efficient routing protocols for wireless sensor network. Abbasi and Younis in the paper [12] present a taxonomy and general classification of clustering schemes. The survey of energy-efficient clustering based protocols can be found in [13]–[16].

We developed a new clustering based power save protocol that utilizes the periodical coordination mechanism to reduce the energy consumption of a network. The proposed algorithm is an extension of the popular geographic adaptive fidelity (GAF) protocol.

The GAF protocol. The GAF protocol described in [17] is a power save protocol that utilizes the information about the geographical location of the nodes. It relies heavily on the concept of *node equivalence*. The nodes A and B are *equivalent* with regard of data transmission between nodes C and D if and only if it is possible to use either node A or node B as a relay for the transmission between nodes C and D. The *node equivalence* is a feature that is not easily discovered. It is easy to notice, that nodes A and B, *equivalent* with regard of data transmission between nodes C and D do not have to be *equivalent* with regard of transmission between nodes D and E.

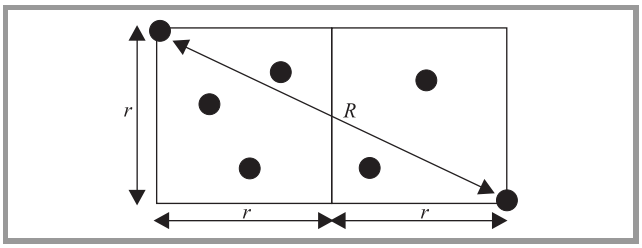


Fig. 8. Network grid construction for GAF protocol.

In order to solve this problem, the GAF protocol partitions the network using a geographic grid. The grid size r is defined such that each node in one grid square is in transmission range of all nodes within adjacent grid squares. The sample construction of such a grid is depicted in Fig. 8. With elementary geometry we have grid size of $R/\sqrt{5}$, where R denotes the maximal transmission range assigned to each node. The construction of such a grid allows the GAF protocol to preserve the original network connectivity.

The sole concept of the GAF protocol is to maintain only one node with its radio transceiver turned on per grid square. Such a node is called an *active* node and is responsible for relaying all the network traffic on behalf of its grid square. When there are more nodes in a grid square, the function of an *active* node is rotated between all the nodes in a grid square. The full graph of state transitions in the GAF algorithm is depicted in Fig. 9.

Each node starts operation in the *discovery* mode, meaning the node has its radio transceiver turned on and is pending to switch to *active* state. The node spends a fixed amount of time T_D in discovery state, when the time has passed, the node switches to the *active* state. After spending a fixed amount of time T_A in *active* state, the node switches back to the discovery state. Whenever a node changes state to *discovery* or *active*, it sends a broadcast message containing node ID, grid ID and the value of a *ranking function*. If a node in *discovery* or *active* state receives a message from a node in the same grid and a higher value of the ranking function, it is allowed to change its state to *sleep* and turn its radio transceiver off for T_S. The ranking function and timers T_D, T_A, T_S can be used to tune the algorithm. Usually the ranking function selects nodes with "longest ex-

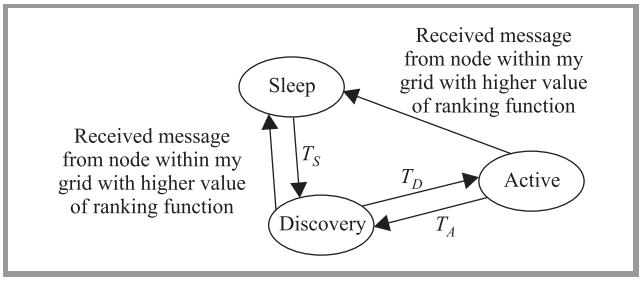


Fig. 9. State transitions in GAF protocol.

pected life time" as the active nodes. The GAF protocol can be easily adapted to mobility scenarios, in such a case the ranking function utilizes information about the time, when a node will leave the grid square.

The coordination-based power save protocol (CPSP). The typical wireless sensor network consists of large quantity of sensor nodes and a base station – a dedicated node which serves as a destination for messages generated by the sensor nodes. The base station is responsible for relaying information gathered by the network to the network operator. It can be assumed that the base station has significantly more resources than the sensor nodes and is directly connected to the power grid. The wireless sensor network is utilized to deliver messages generated by the sensor nodes to the base station. From the operator's point of view there is no difference between not having any nodes in the network and the nodes not being able to deliver their messages to the base station.

We propose to utilize the dedicated network node (or nodes) as a network coordinator (or coordinators) in order to ensure that the base station is able to receive messages from the network nodes for as long period of time as possible. The base station is a natural candidate to play a role of the coordinator. Our protocol assumes that the network is partitioned by a geographical grid in the same manner as in the GAF protocol. In addition we assume that not every network grid needs to maintain an active node. The cells that do not need to establish an active node are determined by the coordinator. The grids that must maintain an active node operate similarly to the grids in the GAF protocol. In remaining grids all nodes are put to sleep state until the next topology update.

The coordinator views the network grids as a graph. The nodes periodically send to the coordinator information about amount of power available to them, which enables the coordinator to assign weights to the edges in the graph. Periodically, the coordinator calculates minimum spanning tree on the graph with itself as the root of the tree. The leaves of the tree are network grids that do not need to maintain an active node. The structure of spanning tree was chosen in order to preserve the original network connectivity. The calculated network topology is sent to all network nodes using a dedicated broadcast algorithm.

The CPSP broadcast algorithm. The CPSP broadcast algorithm relies heavily on the structure of the network and the information it is supposed to deliver to all network nodes. In order to perform the broadcast transmission, extended GAF discovery messages are utilized. Each discovery message contains the sequence number of latest transmitted network map. Since each network grid is able to receive discovery messages originating from neighboring grids, it is able to determine whether it is necessary to broadcast the latest received packet. If the grid determines that the neighboring grid has newer information, it sends a discovery message for neighboring grids to hear it. The size of broadcasted messages is kept as small as possible, information which cells should maintain an active node is sent as a bitmap – one bit represents one network grid.

Simulation results. The coordinated power save protocol was implemented in the environment of the ns-2 network simulator [9]. The proposed protocol was compared with the plain GAF protocol and a network with no power save capabilities at all. Figure 10 shows the performance of examined algorithms on a network with 60 stationary nodes distributed uniformly over a 800 x 800 meter region. Figure 11 presents the performance of the proposed broadcast algorithm against the plain GAF protocol.

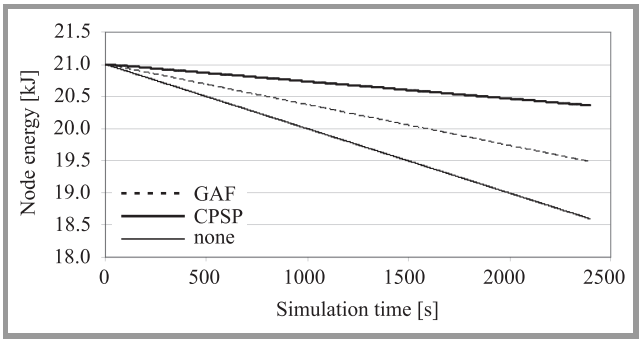


Fig. 10. Average energy consumption, various power save methods.

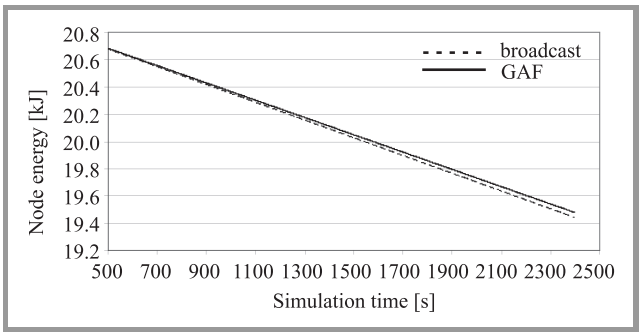


Fig. 11. Average energy consumption; CPSP broadcast and GAF comparison.

The initial energy resource of each node was assumed to be 21 kJ. Additionally it was assumed that the nodes utilize standard 802.11 radio transceiver. The traffic scheme utilized during simulation assumed random nodes sending messages to the base station at random moments of time. The messages sent to the base station were batches of 512 byte packets. The map of the network and the traffic scheme were generated using standard utilities shipped with the ns-2 network simulator.

The metric for evaluating the GAF and CPSP methods was the average amount of energy left in the node during the time of simulation. Although the main objective of CPSP algorithm is to optimize the lifetime of the network and the utilized metric does not directly show the performance of protocols in that area, it was chosen in order to be able to compare the proposed CPSP protocol with other power save solutions.

Discussion. The proposed coordinated power save protocol in its current state allows greater average energy savings than plain GAF protocol. The amount of energy saved is greater than in the GAF protocol due to larger number of sleeping nodes. The use of CPSP protocol introduces a slight overhead caused by the necessity of transmitting messages containing current statuses of nodes to the coordinator and broadcasting coordinator decisions to all nodes in the network. The proposed mechanism can be easily adapted to introduce a coordinator in a wireless sensor networks for other purposes than power saving.

5. Summary and Conclusions

The paper provides the short overview of the energy conservation techniques and algorithms for calculating energy-efficient topologies for WSNs. The efficiency of four location based approaches, i.e., two schemes for topology control and two power save algorithms are discussed based on the results of simulation experiments. The energy efficient method of introducing a coordinator to a WSN is presented. We show that our algorithm outperforms the results obtained for popular clustering based power save protocol GAF.

In general, the simulation results presented in the paper show that topology control and power save protocols effect the scheduling transmissions in a wireless sensor network, and confirm that all discussed approaches to reduce the energy consumption improve the performance of this type of network.

Acknowledgement

This work was supported by Warsaw University of Technology Research Program grant 2008.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirici, "A survay on sensor networks", *Commun. ACM*, pp. 102–114, Aug. 2002.

[2] A. Hac, *Wireless Sensor Network Design*. New York: Wiley, 2003.

[3] P. Santi, *Topology Control in Wireless Ad Hoc and Sensor Networks*. Chichester: Wiley, 2005.

[4] P. Kwaśniewski and E. Niewiadomska-Szynkiewicz, "Optimization and control problems in wireless ad hoc networks", in *Evolutionary Computation and Global Optimization*. Warsaw: WUT Publ. House, 2007, no. 160, pp. 175–184.

[5] J. Branch, G. Chen, and B. Szymański, "ESCORT: Energy-efficient sensor network communal routing topology using signal quality metrics", Lecture Notes in Computer Science, vol. 3420, Berlin/Heidelberg: Springer, 2005, pp. 438–448.

[6] V. Rodoplu and T. Meng, "Minimum energy mobile wireless networks", *IEEE J. Selec. Areas Mob. Comp.*, vol. 4, no. 3, pp. 310–317, 1999.

[7] N. Li, J. Hou, and L. Sha, "Design and analysis of an MST-based topology control algorithm", in *Proc. IEEE Infocom'03 Conf.*, San Francisco, USA, 2003, pp. 1702–1712.

[8] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs: Pentice-Hall, 1992.

[9] Ns2. The network simulator, http://www.isi.edu/nsnam/ns/

[10] MICA2, Crossbow Technology Inc., http://www.xbow.com/Products/productdetails.aspx?sid=174

[11] MPR/MIB user's manual, Crossbow Technology Inc., 2007, http://www.xbow.com/support/support_pdf_files/mpr-mib_series_users_manual.pdf

[12] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks", *Comput. Commun. Arch.*, vol. 30, no. 14–15, pp. 2826–2841, 2007.

[13] D. J. Dechene, A. E. Jardali, M. Luccini, and A. Sauer, "Wireless sensor networks – a survey of clustering algorithms for wireless sensor networks", Project Report, Department of Electrical and Computer Engineering, The University of Western Ontario, Canada, Dec. 2006.

[14] N. Israr and I. Awan, "Energy efficient intra cluster head communication protocol", in *Proc. 6th Ann. Postgrad. Symp. Converg. Telecommun. Netw. Broadcast.*, Liverpool, UK, 2006.

[15] N. Israr and I. Awan, "Coverage based inter cluster communication for load balancing in heterogeneous wireless sensor networks", *J. Telecommun. Syst.*, vol. 38, no. 3–4, pp. 121–132, 2008.

[16] O. Younis and S. Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering aproach for ad hoc sensor networks", *EEE Trans. Mob. Comp.*, vol. 3, no. 4, pp. 366–379, 2004.

[17] Y. Xu, J. Heidemann, and D. Estrin, "Geography-informed energy conservation for ad hoc routing", in *Proc. IEEE Ann. Int. Conf. Mob. Comp. Netw.*, Rome, Italy, 2001.

Piotr Kwaśniewski received his M.Sc. in computer science from the Warsaw University of Technology, Poland, in 2005. Currently he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2007 he works at the Polish Airports State Enterprise. His research area focuses on wireless sensor networks, mobile networks, topology control, energy efficient protocols.
e-mail: P.Kwasniewski@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Izabela Windyga received her B.Sc. in computer science from the Warsaw University of Technology, Poland, in 2008. Currently she is a M.Sc. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Her research area focuses on wireless sensor networks and topology control.
e-mail: I.Windyga@stud.elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Ewa Niewiadomska-Szynkiewicz – for biography, see this issue, p. 67.

Paper

Parallel and Distributed Simulation of Ad Hoc Networks

Andrzej Sikora and Ewa Niewiadomska-Szynkiewicz

Abstract— Modeling and simulation are traditional methods used to evaluate wireless network design. This paper addresses issues associated with the application of parallel discrete event simulation to mobile ad hoc networks design and analysis. The basic characteristics and major issues pertaining to ad hoc networks modeling and simulation are introduced. The focus is on wireless transmission and mobility models. Particular attention is paid to the MobASim system, a Java-based software environment for parallel and distributed simulation of mobile ad hoc networks. We describe the design, performance and possible applications of presented simulation software.

Keywords— ad hoc network, distributed simulation, mobile network, software systems.

1. Introduction

Ad hoc networks are the ultimate technology in wireless communication that allow network nodes located within its transmission range to communicate directly to each other without the need for an established infrastructure such as base station, and centralized administration. For communicating with nodes located beyond the transmission range, the node needs to use intermediate nodes to relay messages hop by hop, thus, in general, routes between mobile nodes may include multiple hops.

A mobile ad hoc network (MANET) [1] is formed through the cooperation of an arbitrary set of independent nodes – mobile, wireless devices. The nodes are free to move randomly and organize themselves. The network's wireless topology may change rapidly and unpredictably. There is no prearrangement assumption about specific role each node should perform. Each node makes its decision independently, based on the situation in the deployment region and its knowledge about the network. Mobile ad hoc networks may operate in a standalone fashion, or may be connected to the Internet. The above description outlines the features of a typical MANET application scenario:

- **Wireless network**: nodes communicate wirelessly and share the same media (e.g., radio).
- **Heterogenous network**: a typical MANET is composed of heterogenous devices.
- **Infrastructureless network**: nodes operate in peer-to-peer mode, act as autonomous routers, and generate independent data; a network does not depend on any fixed infrastructure. MANETs are easy to deployment.
- **Dispersed network and multihop routing**: nodes composing the network are geographically dispersed, thus, multihop communication is necessary – each node may act as a router.
- **Time varying topology**: the topology is dynamic in nature due to the constant movement of the participating nodes.

By exploiting ad hoc wireless technology, various portable devices and fixed equipment can be connected together, forming a sort of ubiquitous network. MANETs enable devices to create and join networks on the fly – any time and anywhere for a given application. Potential applications of wireless ad hoc networks are numerous. Among them, we can cite following: delivery of location-aware information, traffic or health monitoring, intrusion detection, ubiquitous Internet access, etc.

Ad hoc architecture has many benefits, however its flexibility come at a price. A number of complexities and design constraints are concerned with the features of wireless communication (limited transmission range, limited link bandwidth and quality of transmission, constrained resources), mobility and multihop nature of the network [2]–[6].

Currently research effort is directed toward these specifics and constraints in mobile ad hoc networks. Although mathematical modeling and analysis allow to solve many problems and bring some insights into the design of MANETs, the complexity and scale of modern ad hoc networks limit the applicability of purely analytic approaches. Thus, computer simulation can significantly help to obtain crucial performance characteristics.

Computer simulation has been widely recognized as an important tool for researchers and engineers that allow to design and analyze the behavior and performance of cable and wireless networks, and verify new ideas (new protocols, mechanisms, network services, etc.) [7]–[10]. The main difficulty in large scale networks simulation is the enormous computation power, i.e., speed and memory requirements needed to execute all events involved by internodes communication and nodes' mobility. As a consequence, the developments of methods to speed up calculations has recently received a great deal of interest. Parallel and distributed discrete event simulation has already proved to be very useful when performing the analysis of different network systems [3], [8], [9], [11]. It allows to reduce the computation time of the simulation program, and to better reflect the structure of the simulated physical system. Parallel execution of computations can improve the scalability of the network simulator both in term of network size and

execution speed, enabling large scale networks and more network traffic to be simulated in real time.
In this paper, we discuss some guidelines related to wireless, mobile, and ad hoc networks modeling and simulation. We model MANET application using discrete event systems methodology (DEVS) and address the challenges to design high-performance simulation of MANETs' systems. Finally, we describe organization, implementation, usage, and practical application of our ad hoc networks simulator called MobASim.

2. Mobile Ad Hoc Network Modeling

In the performance evaluation of an ad hoc network application, simulations should be done under a variety of modeling parameters and conditions, in order to capture effects of the simulated real life system. In MANETs a correct model design should evaluate a priori any possible relationship among simulated area, network topology, mobility levels, wireless transmission, power consumption, etc.

2.1. Mobility Models

Modeling of movement of network nodes plays the crucial role in almost every simulation experiments of MANETs. The dynamic topologies due to nodes' mobility introduces adaptive behavior of users, control mechanisms, and communication protocols. The mobility models should resemble the real life movements, and at the same time be simple enough for simulation. In general, two types of mobility models have been adopted in the simulation of MANETs [3], [6], [12].

- **Motion traces** that provide accurate information about mobility patters and behavior of the nodes in the considered environment (e.g., streets, highways). Traces define the positions of nodes in time, so they require long files depending on the time granularity of samples. It is good description of steady-state mobility if the motion samples are collected for considered time intervals.
- **Synthetic models** are analytical random-motion models that describe mobility without using real traces. We can distinguish several less and more realistic synthetic models. The random mobility model is a discrete implementation of a Brownian-like motion. In the random waypoint model each node chooses uniformly at random a destination point and velocity, and moves toward it along a straight line. The random direction model is similar to the previous one, but in this model each node chooses uniformly at random a direction.

The *map-based mobility models* are used for applications in which nodes are constrained to move within defined paths. Most of presented models describe an obstacle-free movement.

2.2. Wireless Transmission Modeling

A simulation of wireless communication, including propagation, mobility, and interference is very difficult and computationally expensive task. The main problem in wireless communication modeling is estimation of the size of the transmission area of a transmitter. This area can be defined as the area where the transmitted signal between any two nodes u and v propagates and can be correctly detected and decoded. We can define the signal degradation $PL(d)$ with a distance d:

$$PL(d) = \frac{P_t}{P_r}, \tag{1}$$

where d denotes the distance between nodes u and v, P_t power used by u to transmit the signal and P_r power of the signal received by v. $PL(d)$ is called "path loss" with a distance d.
A path loss modeling is difficult but very important task. If we know the model of $PL(d)$ we can predict the occurrence of a radio channel between any two nodes in the network. Over time, many less and more detailed propagation models have been introduced [3], [6], [13], [14]. In practice, three techniques for path loss estimation are extensively used: long-distance path loss models, log-normal shadowing, and fading models. The long-distance models predict variations of the signal intensity over large distances. They have been developed as a combination of analytical and empirical methods. In these models the average large-scale path loss is expressed as a function of a distance d raised to a certain exponent n ("distance-power gradient"), which indicates the rate at which the path loss increases with a distance:

$$PL(d) = PL(d_0)\left(\frac{d}{d_0}\right)^n, \tag{2}$$

$$PL(d)[\mathrm{dB}] = PL(d_0)[\mathrm{dB}] + 10n\log\left(\frac{d}{d_0}\right), \tag{3}$$

where d_0 denotes a close-in reference distance determined from measurements close to transmitter, d a distance between transmitter and receiver.
The log-normal shadowing model considers the fact that the transmission area of a transmitter may be different at two different locations, which leads to measure signals that are different than the average value calculated by Eq. (3). In this model path loss at distance d is modeled as random variable with log-normal distribution:

$$PL(d)[\mathrm{dB}] = PL(d_0)[\mathrm{dB}] + 10n\log\left(\frac{d}{d_0}\right) + X_\sigma, \tag{4}$$

where X_σ is a zero-mean Gaussian distributed random variable with standard deviation σ (all in dB).
The fading models predict variations of the signal intensity over very short distance.

3. MobASim – Software System for Ad Hoc Networks Simulation

The MobASim system provides a framework for mobile ad hoc networks simulation performed on parallel computers or computer clusters. It can help testing of various technologies designed for ad hoc networks application scenarios. The considered network to be simulated is described by different parameters defined by the user, thus we can perform the experiments for various topologies, wireless devices, mobility models, routing protocols, localization capabilities, etc. In this section we present the design and implementation of MobASim and comparison of our project to the other existing tools for ad hoc networks simulation.

3.1. Related Works and Comparison

Today, many software tools for wireless networks simulation are proposed. Some popular network simulators like OPNET [15], ns-2 [16], OMNeT++ [17] or GloMoSim [18] can simulate ad hoc networks. The others are dedicated to MANETs [19] or wireless sensor networks [10] simulation. The simulators provide the facility to simulate protocols in different layers, nodes mobility, energy consumption and various ad hoc networks application scenarios. Different tools are optimized for different purposes.

However, most of available simulators require costly shared-memory supercomputers to run even medium size network simulation. We are involved to large scale network systems simulation and their practical applications, and our goal was to develop scalable simulator operating in real time. Hence, to provide high performance and scalability we utilized the paradigm of federating disparate simulators [20] and asynchronous distributed simulation technology [21]–[23]. This is the main difference between our software and the other tools. The other reason for developing a new simulator was the complicated architecture of available systems and limitations in results visualization and user-system interaction. In case of OPNET, OMNeT++ or ns-2 systems a user must read a large number of manuals to learn how to use the tool. The source coding is specialized and it is not easy to implement a given example and add modules developed by the user.

Moreover, many systems do not support both the user interactions during the experiments and animation of network topology changes. Users set configuration parameters before starting the simulation, and they can see computation results after the experiment is terminated. In addition, most existing ad hoc networks simulators focus on the MAC protocols implementation with the lack of the radio management and mobility modeling. Usually only simplified wireless transmission models and obstacles free simple mobility models are provided (ns-2, OMNeT++).

The MobASim is a general purpose federated simulator which elements can be easily reused in many computations. The process of implementing a given application for MobASim is quite straightforward and convenient especially thanks to GUI (graphical user interface) and dedicated language ASimML – the XML (extensible markup language) schema specification. MobASim supplies the library of classes to implement the user's modules, which are specific to a given application. Hence, the current version of our simulator provides different models of radio management, mobility models handling obstacles, user-friendly interface and tools for results visualization and animation. The open design of MobASim architecture, easy usage, and its extensibility to include external modules, was chosen in the hope that the system will be a useful platform for research and education in ad hoc networks modeling and testing. The software will be free available for researchers and students.

3.2. MobASim Overview

The discrete event systems methodology is applied to model mobile ad hoc network operation, i.e., the process being modeled is understood to advance through events [21], [23]. The major concept is defined as follows:

- **System**: a collection of entities that interact together over time to accomplish a set of goals or objectives.
- **Model**: a representation of the system in terms of its entities and their events, attributes and objectives.
- **Entity**: component of the system that requires the explicit representation in the model.
- **System state**: a collection of variables which values define the sate of the system at a given point of time.
- **Event**: an instantaneous occurrence in time that alters the state of the system.

In our application the *system* denotes the wireless, mobile, and ad hoc network, *entities* are components of this network responsible for different functionalities. We distinguish three types of such components:

- **Node**: a mobile device that performs the assigned task. It can change dynamically its position in the deployment region and can interact with other nodes in the system.
- **Communication manager**: an object that models the wireless communication between all nodes.
- **Mobility manager**: an object responsible for tracking the nodes on the map and collision avoidance.

The MANET simulator developed in the MobASim system consists of logical processes (LPs) implementing the operation performed by three listed types of entities: *nodes*, *communication managers* and *mobility managers*. Hence,

LPs are divided into three groups of computing processes, adequately responsible for:

- N – tasks to be performed by mobile nodes,
- CM – internode wireless communication and the network communication topology updating,
- MM – mobile nodes movement and providing the access to information about the terrain (deployment region) and other nodes location in the network.

Each process from the group MM can implement one of three mobility models. It is possible to combine various models in one simulator, i.e., the model of mobility can switch w.r.t. the current state of the node. The processes from the group CM implement one of two wireless communication models.

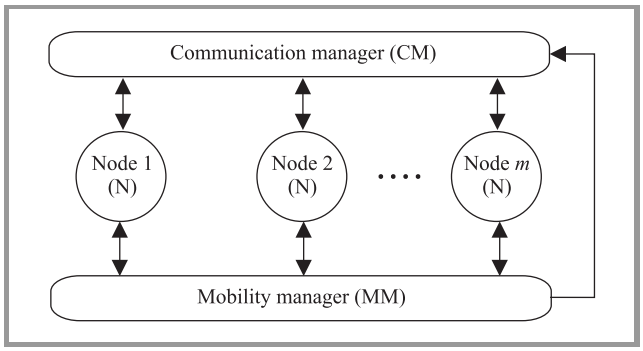


Fig. 1. The architecture of MobASim application.

The structure of a given application (MANET simulator) implemented in MobASim is presented in Fig. 1. We can see that every simulator of MANET is composed of one logical process from the group CM, one process from the group MM, and several processes from the group N. The number of N processes is equal to the number of wireless devices in the simulated network.

3.3. Mobility Models in MobASim

The popular commercial and publicly released software tools for networks simulation support mobility models based on motion traces, see OPNET [15], ns-2 [16], GloMoSim [18]. These models describe an obstacle-free movement. The user has to provide accurate information about mobility patterns. Our MobASim simulator provides three types of mobility models. In all cases the obstacles are accounted. The obstacles are generated by the user or are localized based on a real map. They are accounted for also when simulating the radio signal propagation. It is assumed that wireless signal is obstructed by the obstacles.
The state of each mobile node is described by four state variables:

– location within the deployment region,
– orientation (an angle between X axis and the direction of node movement),
– speed of movement,
– energy stored in the node.

It is assumed that generated movement paths are dynamically changed taking into account the state of the nodes and surroundings (obstacles and neighbouring nodes). All data concerned with the deployment region and all nodes in the network are stored in the data base served by the logical process MM (mobility manager). The DEVS methodology is used to implement mobility models. The following types of events are defined:

- MStart – start the movement,
- MC – continue the movement,
- MStop – stop the movement,
- LT – track the changes in the node location,
- DU – update the MobASim data base (changes in nodes' location and their surrounding),
- CA – alert the node to a collision (the movement directions of at least two nodes are crossed),
- CO – a collision between two nodes was occurred.

All presented events are served by logical processes from two groups: N (node) serves, respectively, MStart, MC, MStop, LT, DU, CA, CO events, and MM (mobile manager) serves DU, CA and CO events. Hence, the motion trajectory is generated dynamically and results the following events execution: MStart, several MC and CA, and finally MStop. The number of MC events depends on the distance to the destination point, node velocity, obstacles occurrence and sampling intervals. In the case of collision the CO event is executed.
Three types of mobility models are implemented. Types MM1 and MM2 are modified versions of the random waypoint model (RWP). Each node chooses at random a destination point and velocity, and next moves toward the direction. In addition, in our model the user can define the specific destination instead of random generation. The main difference to RWP is that in case of MM1 and MM2 the obstacles are considered. Hence, the shortest (if possible) path is calculated from the current position of the node to the destination to avoid the collision, while in case of the RWP model the node moves toward the destination along a straight line. The differences between models MM1 and MM2 are such that in case of MM1 we assume the access to the information about the whole deployment region, i.e., we know the locations of all obstacles and current positions of all neighbor nodes in the network. In case of MM2 we assume the restricted access to the data about the environment. Only the knowledge about the obstacles and other nodes located in the surrounding of a given node is available. The path has to be dynamically changed after possibility of collision identifying. It is not guaranteed that the shortest path will be realized in case of this model.

Both in models MM1 and MM2 nodes are free to move within the deployment region. In case of model MM3 (map-based mobility) nodes are constrained to move within specified paths. All these paths are stored in the MobASim data base.
Each MobASim application can implement all described models: MM1, MM2 and MM3. The mobility model can dynamically change w.r.t. the current state of the node.

3.4. Wireless Transmission Modeling in MobASim

Most of the available software platforms for mobile ad hoc network simulation implement only large-distance wireless transmission model Eq. (3) in its simplest version. MobASim simulator implements two of the transmission models described in the Subsection 2.2: long-distance Eq. (3) and shadowing Eq. (4).
The medium access control (MAC) layer is of fundamental importance in wireless ad hoc networks. MAC protocols are responsible for controlling the access to wireless channel. MobASim provides the implementation of MAC protocols from three categories based on the method that they handle the hidden and the exposed terminal problems: class 1 – protocol assuming random access to the wireless channel (the hidden and exposed node problem is unsolved), class 2 – the protocol solves the hidden node problem but leaves the exposed node problem unsolved, class 3 – the protocol solves both the hidden node and the exposed node problems, but requires the deployment of an additional signaling channel. The MobASim user can choose the protocol suitable to designed application. The currently available version of MobASim implements the simplified models of the physical layer and the interference management. We assume that the accurate model of MAC layer can be adopted from the other open source simulators, if necessary.

4. MobASim System Design and Implementation

The MobASim system is completely based on Java. At the heart of its technology is the asynchronous simulation Java (ASimJava) library – collection of Java-based procedures that can be used to develop general purpose discrete-event parallel and distributed simulators designed as federations of disparate simulators, utilizing runtime infrastructure (RTI) to interconnect them. Each simulator is described in terms of logical processes that communicate with each other through message-passing. LPs simulate the real life physical processes. The federation paradigm described in [20] allows to perform parallel or distributed calculations, i.e., each simulator can be executed in a separate processor or machine. The goal is to speed up calculations and perform real time simulation. The synchronous and asynchronous variants of simulation are provided [21], [23]. ASimJava technology was described in details in [11], its application to computer networks simulation in [8].

Composition and implementation of MobASim. The MobASim software provides tools to build simulators utilizing ASimJava library and runtime infrastructure, thus we can develop our application as a federation of simulators implementing the subnetworks that compose the considered MANET or a federation of simulators of independent, geographically dispersed MANETs or WSNs (wireless sensor networks) that cooperate from time to time (see Fig. 2). When consider the simulation of mobile networks we have to generate a map of deployment area. The MobASim user can define the simple objects in the domain as polygons. For more detailed description of a terrain to be considered the MobASim simulator provides the interface to the GeoTools toolkit. The GeoTools [24] is an open source Java coded library containing standard methods for the manipulation of geospatial data. All geographical information are stored in the MobASim database.

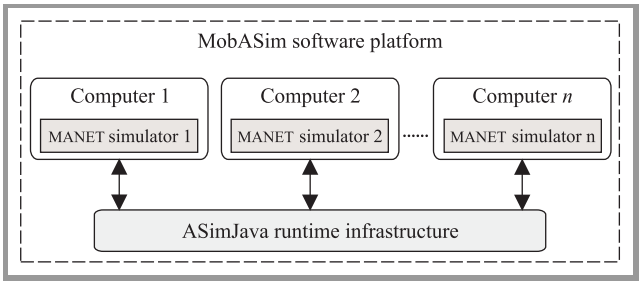


Fig. 2. A federation of simulators in MobASim.

In summary, the MobASim simulator is composed of: a runtime platform of ASimJava, a set of libraries of functions for parallel discrete event systems calculation provided in ASimJava, a set of libraries of functions for mobile and wireless applications, and a set of tools mainly to support the interaction with the user and visualisation tool for the runtime monitoring (see Figs. 3 and 4).
The user GUI is organized in a set of nested windows. The setting windows are used to facilitate the configuration phase. The network is constructed graphically. The user can enter parameters concerned with the whole network (number of nodes, wireless transmission model), network nodes (radio communication range, minimal and maximal speed, mobility model, routing protocol, MAC protocol, energy reserve, etc.), and deployment area (type of geographical data). The dedicated setting windows are used to insert the parameters specific to chosen mobility and wireless communication models provided in the system. Finally, the user is asked to configure the experiment (simulation time, number of processes, number of machines, etc.). After completing the initial settings, MobASim starts the simulation experiment. The results of simulation – time varying topology (animation of nodes) and adequate statistics are displayed. The configuration of the system to be simulated can be loaded and saved into the disc file in the XML format.

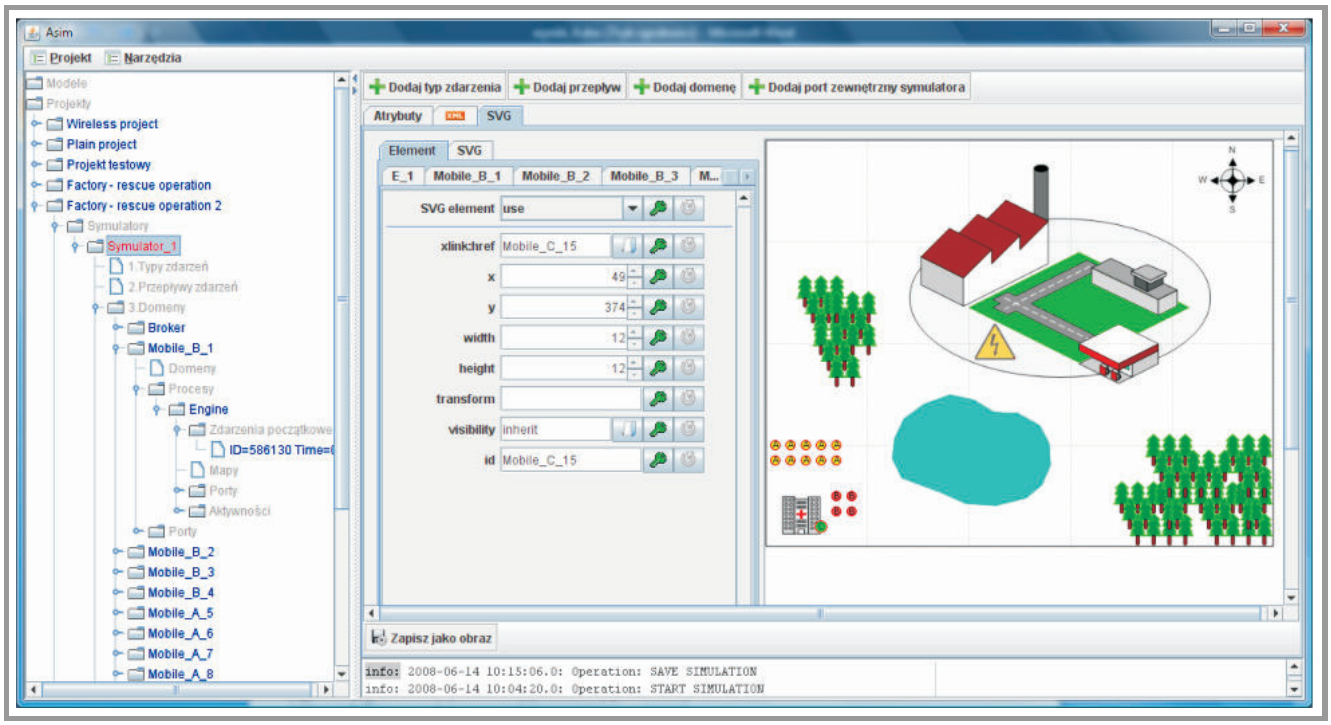


Fig. 3. MobASim graphical user interface.

5. Simulation Experiments

In order to evaluate the efficiency of MobASim, and indicate the usefulness of such software systems to support the decision making process in case of real-life problems the simulations of several ad hoc network topologies were performed. In this paper the application of our tool to support the design of MANET for the rescue action is discussed. Let us consider the following situation. The explosion in the factory devastated its surrounding. Most of the communication infrastructure, i.e., base stations for cable networks, wired phone lines, etc., was destroyed. Thus, two of priorities in the disaster management are to organize on-line monitoring of the situation on the disaster scene and to organize a relief effort for explosion victims by dispatching several rescue teams to the disaster area. The efforts of the rescue teams should be coordinated. It can be achieved only if rescuers are able to communicate, both within their team and the members of the other teams. To carry out these goals it is necessary to reinstall the communication infrastructure as quickly as possible. It can be done by deploying temporary communication equipment (vehicles equipped with transceivers), and creating an ad hoc network. By using multihop wireless communication and mobile nodes acting as communication relay stations, even relatively distant rescuer will be able to communicate. The communication will be possible without the need for rebuilding the fixed communication infrastructure.

The ad hoc network designed for reestablishing the communication for the discussed example consists of three types of nodes (see Fig. 4):

- A: the mobile node – wireless router (e.g., vehicle equipped with a transceiver) that provides the communication between nodes B and C;
- B: the mobile node – the rescue unit (e.g., the rescuer equipped with a transceiver) working in the disaster area;
- C: the base station – the rescue center that coordinates the rescue action, controls the nodes A and B and collects the data (monitoring of the situation) transmitted by nodes B.

The ad hoc network of ten wireless routers (nodes A), four rescue units (nodes B) and one rescue center (node C) was used for the rescue action. For the purpose of simulations we assumed following values of parameters used in wireless communication model: distance power gradient $n = 2$ in Eq. (2) and standard deviation $\sigma = 6$ dB in Eq. (4). The map of the deployment area was generated based on MobASim GUI and saved to the MobASim database.

Several simulation experiments were performed. The objective was to design the mobile ad hoc network that provide the continuous communication with all rescuers during the rescue action.

During the simulations the bandwidth of all links and current traffic are calculated, and the critical paths are pointed. The animation of time varying network topology – all nodes moving from the initial position to the destination, avoiding the obstacles – are displayed in MobASim main window. The user can keep track how the communication network created by a set of nodes A adopts to the new positions

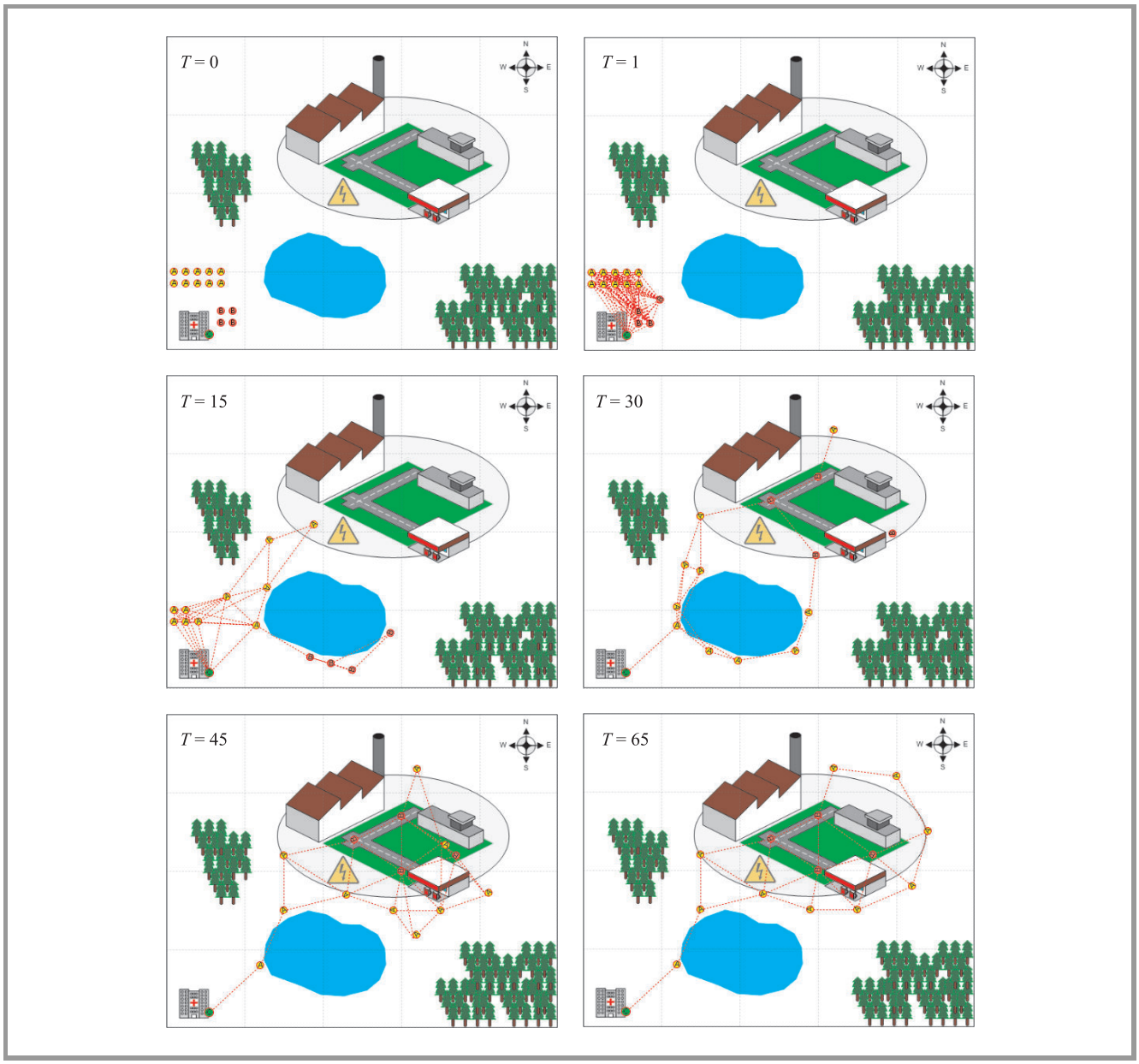


Fig. 4. MobASim simulator: application of ad hoc network.

of rescue teams (nodes B). The current network connectivity is marked by lines connecting the nodes. They appear when the communication between two nodes is possible (the distance is less than the radio range and the node is not a hidden one).

In our experiments we considered various range assignments of radio transmitters in nodes A and B. We tested the efficiency of the mobility manager algorithm implemented in the rescue center (node C), which goal is to calculate the paths that substitute the critical ones. It allows to create the robust and optimal network topology for the current time period.

The results are presented in a figure and two tables. Figure 4 shows the dynamically changing network topology during the entire network operational lifetime. The snapshots of initial, temporary and final topologies calculated for 0, 1, 15, 30, 45 and 65 steps of simulated time units are presented.

The initial and final positions of all nodes in the network are collected in Table 1.

Next, we compared the results obtained for MAC protocols from three categories: class 1, 2 and 3. MAC protocols restrict the number of simultaneous signal transmissions per unit of area and consequently restrict the number of interfering nodes. The number of wireless connections and interfering nodes in case of various classes of MAC protocol for several time steps (every 5 units of simulated time) are presented in Table 2. It can be seen from the table that, as expected, the number of interfering nodes is highest in MAC class 1 and lowest in MAC class 2. For the cal-

culation of the interference power in ad hoc and sensor networks, the density and the distribution of the interfering nodes must be known. If the density of nodes increases

Table 1
Simulation results

Node name	Time [s]	Initial position (x, y) [m]	Destination position (x, y) [m]
B1	0	(85,350)	(240,160)
B2	2	(85,365)	(300,130)
B3	4	(70,350)	(300,200)
B4	6	(70,365)	(370,180)
A1	7	(70,300)	(320,70)
A2	9	(70,315)	(150,180)
A3	11	(55,300)	(230,230)
A4	13	(55,315)	(120,320)
A5	15	(40,300)	(150,250)
A6	17	(40,315)	(290,250)
A7	19	(25,300)	(350,250)
A8	21	(25,315)	(420,220)
A9	23	(10,300)	(440,150)
A10	25	(10,315)	(400,80)
C1	–	(55,380)	(55,380)

Table 2
Results for various categories of MAC protocol

Time step	Wireless connections	Interfering nodes (MAC classes)		
		class 1	class 2	class 3
1	105	1	0	1
5	99	2	0	1
10	75	3	0	1
15	45	3	0	3
20	30	6	1	3
25	28	6	1	3
30	24	6	1	4
35	25	5	1	4
40	28	5	0	4
45	30	5	0	3
50	32	6	0	2
55	29	6	0	3
60	26	6	0	3
65	25	6	0	3

the number of nodes falling within the prohibited transmission areas increases. The density of interfering nodes is not expected to increase linearly with the increase in the density of nodes.

From the simulation results we see that by using multihop wireless communication and mobile nodes, the communication between the rescue center and rescue teams is possible without the need for reestablishing the fixed communication infrastructure.

6. Summary and Conclusions

The evolution of wireless, mobile ad hoc networks and improved designs will strongly depend on the ability to predict their performance using analytical and simulation methods. In this paper we described the software platform MobASim for mobile ad hoc networks simulation. MobASim was designed to be powerful, effective, scalable, flexible, and easy to use ad hoc network simulator. It can support researches and engineers during the design and implementation of MANETs applications and verification of new MANET's technologies. The tool is especially useful in large scale applications in which the speed of simulation is of essence, such as real time ad hoc networks simulation. MobASim is a general purpose federated simulator, which elements can be easily reused in many computations. The federated approach to parallel and distributed simulation of networks, provided functionality, easy usage and its extensibility to include other open source modules or modules developed by the user, which are specific to a given application, make different our tool from the popular software systems for simulation.

Acknowledgment

This work was supported by TINFO Project 2008.

References

[1] MANET Group, http://www.ietf.org/html.charters/manet-charter.html

[2] G. Aggelou, *Mobile Ad Hoc Networs. From Wireless LANs to 4G Networks*. New York: McGraw-Hill, 2005.

[3] S. Basagni, M. Conti, S. Giordano, and I. Stojmenovic, *Mobile Ad Hoc Networking.* New York: Wiley, 2004.

[4] S. Basagni and A. Capone, "Recent research directions in wireless ad hoc networking", *Ad Hoc Netw.*, vol. 5, iss. 8, pp. 1205–1348, 2007.

[5] A. Hac, *Wireless Sensor Network Design*. New York: Wiley, 2003.

[6] P. Santi, *Topology Control in Wireless Ad Hoc and Sensor Networks*. New York: Wiley, 2005.

[7] M. Małowidzki, "Network simulators: a developer's perspective", in *Proc. Int. Symp. Perform. Eval. Comput. Telecommun. Syst. SPECTS'04*, San Jose, USA, 2004.

[8] A. Sikora and E. Niewiadomska-Szynkiewicz, "FR/ASimJava simulator: a federated approach to parallel and distributed network simulation in practice", *J. Telecommun. Inform. Technol.*, no. 4, pp. 53–59, 2006.

[9] B. K. Szymanski, A. Saifee, A. Sastry, Y. Liu, and K. Mandnani, "Genesis: a system for large-scale parallel network simulation", in *Proc. Paral. Distrib. Simul. PADS 2002*, Washington, USA, 2002.

[10] B. K. Szymanski and G. G. Chen, "Sensor network component based simulator", in *Handbook of Dynamic System Modeling*, P. Fishwick, Ed. Boca Raton: Taylor and Francis Publ., 2007, pp. 35-1–35-16.

[11] E. Niewiadomska-Szynkiewicz and A. Sikora, "ASim/Java: A Java-based library for distributed simulation", *J. Telecommun. Inform. Technol.*, no 3, pp. 12–17, 2004.

[12] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research", *Wirel. Commun. Mob. Comp. (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications*, vol. 2, no. 5, pp. 483–502, 2002.

[13] B. A. Forouzan, *Data Communications and Networking*. New York: McGraw-Hill, 2004.

[14] *Mobile Radio Communications*, R. Steele, Ed. New York: IEEE Press, 1994.

[15] OPNET Modeler, http://www.opnet.com/products/modeler/home.html

[16] Network Simulator ns-2, http://www.isi.edu/nsnam/ns/

[17] OMNeT++, http://www.omnetpp.org/

[18] GloMoSim, http://pcl.cs.ucla.edu/projects/glomosim/

[19] T. Facchinetti, G. Buttazzo, and L. Almeida, "A flexible visual simulator for wireless ad-hoc networks of mobile nodes", in *Proc. IEEE Conf. Emerg. Technol. Fact. Autom.*, Catania, Italy, 2005, vol. 1, pp. 50–54.

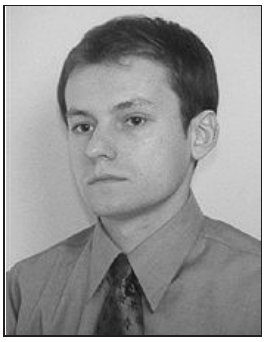
[20] S. L. Ferenci, K. S. Perumalla, and R. M. Fujimoto, "An approach for federating parallel simulators", in *Proc. 14th Worksh. Paral. Distrib. Simul. PADS 2000*, Bologna, Italy, 2000.

[21] S. Ghosh and T. S. Lee, *Modeling and Asynchronous Distributed Simulation*. New York: IEEE Press, 2000.

[22] D. M. Nicol and R. Fujimoto, "Parallel simulation today", *Ann. Oper. Res.*, vol. 53, pp. 249–285, 1994.

[23] B. P. Zeigler, H. Praehofer, and T. G. Kim, *Theory of Modeling and Simulation*. London: Academic Press, 2000.

[24] GeoTools The Open Source Java GIS Toolkit, http://geotools.codehaus.org/



Andrzej Sikora received his M.Sc. in computer science from the Warsaw University of Technology, Poland, in 2003. Currently he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2005 he works at the Research and Academic Computer Network (NASK). His research area focuses on parallel and distributed simulation, computer networks, ad hoc networks and database systems.
e-mail: A.Sikora@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

e-mail: A.Sikora@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Ewa Niewiadomska-Szynkiewicz – for biography, see this issue, p. 67.

Paper

The Non-Didactic Aspects of e-Learning Quality

Ewa Stemposz, Andrzej Jodłowski, and Alina Stasiecka

Abstract— The paper presents a research on the quality of e-learning from the non-didactic point of view. It illustrates a discussion about measures developed on the basis of statistical analysis of data gathered from e-learners who evaluated the quality of e-learning applications and systems. The main contribution of the paper is the proposal for the quality metrics with the features concerning e-learning platforms in the technological and human aspects.

Keywords— e-learning, metrics, quality.

1. Introduction – the Quality of e-Learning

E-learning is currently a very dynamically developing form of distance learning, carried out with the use of up-to-date communication and information technologies. One of its learning forms is learning through Internet/Intranet that utilizes the access of teachers and students to a global/local computer network.

That kind of education can gain advantage over traditional teaching methods mainly on the grounds of freedom of access to information (knowledge) – unlimited time and unlimited place of learning and also for the reason that e-learning enables learners to assimilate new information at a pace and in the way adjusted to one's needs and abilities. Despite unquestionable merits of e-learning, there appear many problems related to its propagation.

1. Technological possibilities of educational environment – lack of Internet connection and/or insufficient technical parameters of those connections.

2. Resources – the HTML (hypertext markup language) file format is the basic content format of distance training and courses that are available through the Internet and Intranets. E-trainings seldom take other forms, e.g., a teleconference or a videoconference. E-resources are usually custom-made, so they do not support any common e-learning standard. For example, online course materials used in higher education are created in colleges and at universities by the teaching staff responsible for the course. Therefore, schools do not in fact order materials from other producers.

3. Direct participants of that process, i.e., teachers and learners – research shows resistance to the introduction of new technologies. It also confirms that there is a strong need for interaction among course participants, which is often missing in that form of learning.

We think problems that are related to e-resources, as well as those related to e-teachers/e-learners require direct attention. Those issues involve to make an attempt to solve them through ensuring adequate quality level of e-learning processes. In our opinion, quality is undeniably one of the vital issues concerning education process by e-learning techniques.

There exist many different definitions of learning quality that are dependent on needs and expectations of participants of that process. However, it is difficult to call those definitions as precise. For example, the definition of quality in ISO 9000[1] standard is as follows:

"A quality is a characteristic that a product or service must have. For example, products must be reliable, useable, and repairable; similarly, service should be courteous, efficient, and effective. These are some of the characteristics that a good quality product/service must have. In short, a quality is a desirable characteristic. However, not all qualities are equal. Some are more important than others. The most important qualities are the ones that customers want. So providing quality products and services is all about meeting customer requirements. It's all about meeting the needs and expectations of customers. So a quality product or service is one that meets the needs and expectations of customers."

There arises a fundamental problem from such a general definition. How to identify the minimal possible set of the most important quality criteria which could encompass the needs and expectations of all interested parties? Which way to discipline the e-learning processes so as not to limit creativity, flexibility, and abilities of e-learning participants?

Basing on the division of those problems into three groups, we propose to consider the quality of e-learning education in three general aspects.

1. **In technological aspect**, related to computing environment, where education processes and e-learning platforms are embedded, concerning, i.e.:

   – the expectations regarding the scope of design, implementation and development quality for e-learning systems, including the development of associated standards also;

   – the activities encompassing adaptation and integration of computer technologies with existing e-learning systems and associated standards;

[1]ISO 9000: 2005, "Quality management systems, fundamentals and vocabulary".

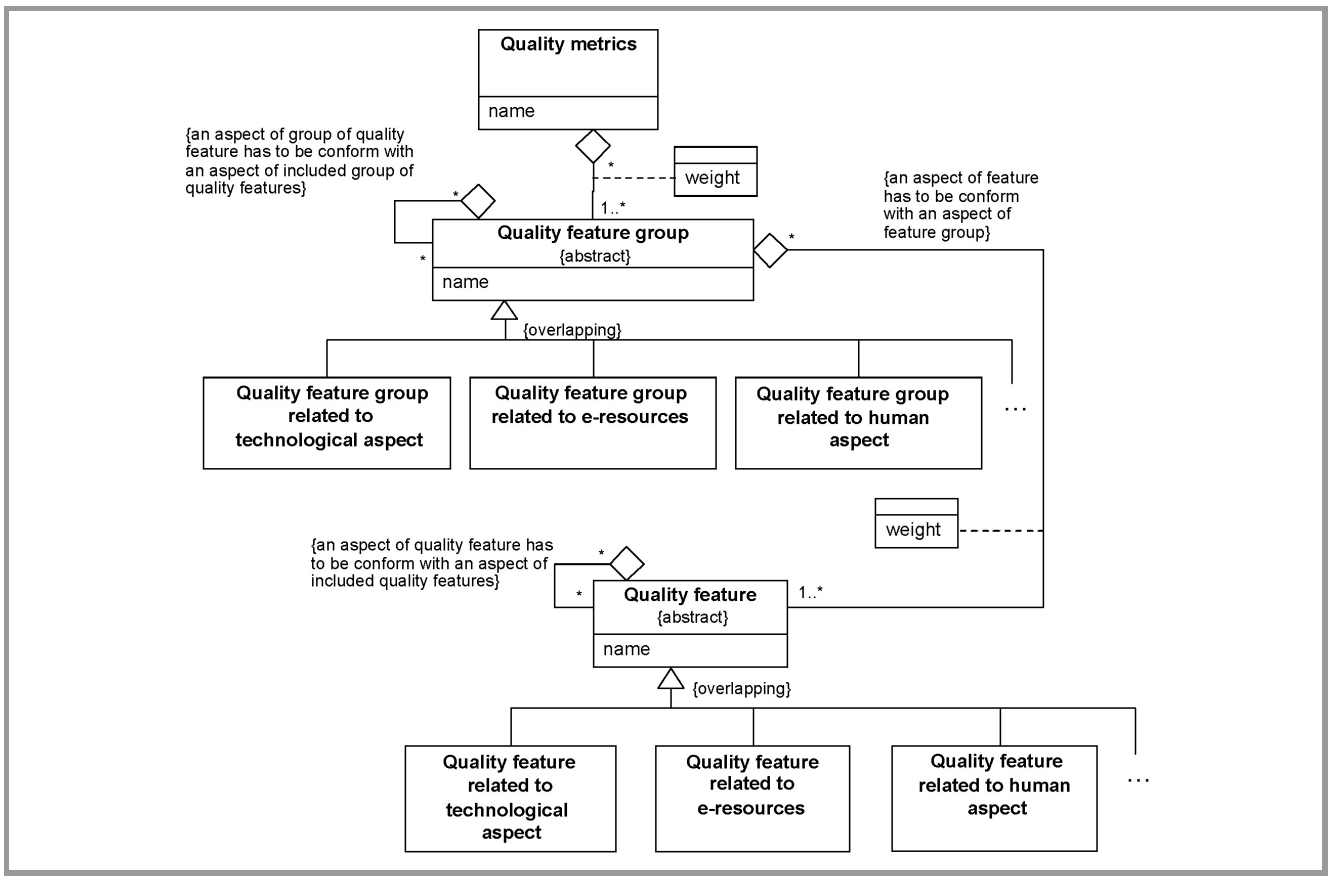


Fig. 1. The idea of quality metrics.

– user expectations related to working platforms (e-learning systems) including, e.g., support for personalization and customization, ensuring proper security level, data protection, complying with the needs of learners related to unlimited access to materials, ensuring data recovery after failure, support for interoperability with other platforms, ease of use, work speed.

2. **In e-resources aspect** related to requirement descriptions and estimation of quality of e-learning materials, both in didactic aspect, e.g., conformance to teaching model(s) [1], [2], as well as in non-didactic aspect considering, e.g., amount and quality of multimedia used, or quality of process of e-resource development.

3. **In human aspect** – we classify e-learning process participants into two general groups as follows:

   – direct participants: suppliers and designers of e-learning systems, teachers, methodology specialists, trainers, students;

   – and indirect participants: authorities, accreditation, standardization, law establishing, and law regulating institutions, etc.

Realization processes ensuring the quality of e-learning should involve all participants. Quality is influenced both by qualifications of a team designing a course and teachers who realize it. One cannot also forget the degree of involvement of teachers and students in the learning processes.

Various e-learning quality elements can be shown using a graphical diagram (in UML (unified modeling language) notation), see Fig. 1. *Quality feature* denotes an element which influences the quality of e-learning. To ensure the clarity of the diagram, most of class attributes are omitted except for *name* attributes within classes: *Quality metrics, Quality features group, Quality feature*.

2. The Quality of e-Learning in Technological and Human Aspects

Our prior research was focused on the quality of e-learning from a didactic point of view [3]–[6]. The next stage of our considerations included the analysis of e-resources quality in the non-didactic aspect and the research on the quality of platforms (applications and e-learning systems).

In this paper, we make an attempt to identify measures of quality features from a technological point of view and from a human point of view. Both analyses were per-

formed on the basis of studies of the quality of e-learning applications. The first step was to create a questionnaire concerning technological and human aspects of e-learning. The questionnaire ought to have provided data with reference to the quality of existing e-learning applications and with reference to expectations of potential users to such applications [7]. The questionnaire consisted of 26 questions concerning issues about graphical interfaces and e-learning. The respondents were mostly students of computer engineering (32 persons).

The questions are classified into three groups.

A. The questions concerning respondents; they focused attention on the effectiveness of e-learning.

B. The questions characterizing features that are desired for platforms and e-resources; they could be used to build a quality metrics.

C. The questions concerning processes of interface design for e-learning platforms.

Further research used the data gathered from the questionnaire that were related to desirable features of platforms and e-resources (the B group) only. The questions included in the A group and C group were passed over.

The questions from the B group concerned both the actual state (they should show what platforms and e-resources were used) – the B1 subgroup, as well as user expectations (what platforms and e-resources should look like) – the B2 subgroup. Further works were based on those question belonging to the B, and B2 subgroups.

In order to perform statistical analysis of data gathered from the questionnaire, we constructed a set of measures that characterized e-learning platforms. Successive measures corresponded with features characterized by questions from questionnaire, where features were denoted by labels: "name-and-number-of-group.question-number-within-group", e.g., b1.1, b2.4 – see Table 1.

The data gathered from the questionnaire concerning the set of features from Table 1 were subjected to the statistical analysis using the gradational data analysis of the GradeStat program [8].

We considered two groups of features:

– features related to the technological aspect;

– features concerning the e-resource aspect.

Because the questionnaire, in fact, omits the human aspect (only one feature) – we did not examine separately the group of features related to that aspect. The analysis was performed with regard to the classification of features into the B1 ("present state") and B2 ("expectations") subgroups, where the B2 group included features both from technological and e-resources aspects (because the latter comprised one feature only).

For those groups mentioned above we computed overrepresentation maps with the use of the GradeStat program. Further analysis led us to specify sets of characteristics which differentiated and undifferentiated the elements of the population.

2.1. Analysis of Overrepresentation Maps – Features Related to the Technological Aspect

Figure 2 presents the overrepresentation map for features related to questions from the B1 group.

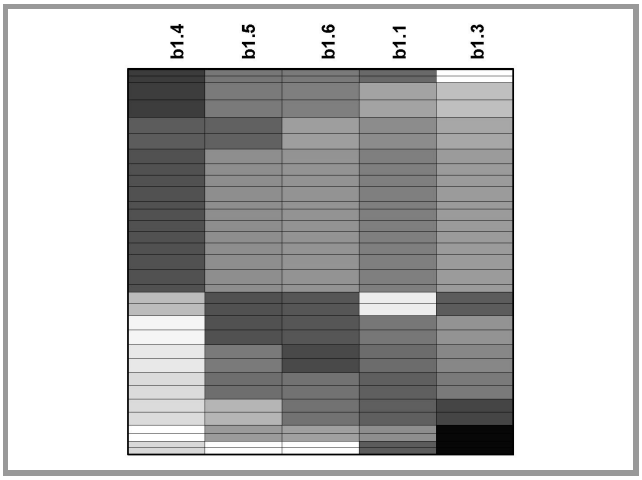


Fig. 2. The overrepresentation map for features from the B1 group – technological aspect.

On the basis of the overrepresentation maps, the cluster analysis was performed. Figure 3 illustrates the dependence of Rho* values[2] on a cluster count that was evaluated for columns. Basing on that diagram it was assumed that the cluster count for columns should be equal 3.

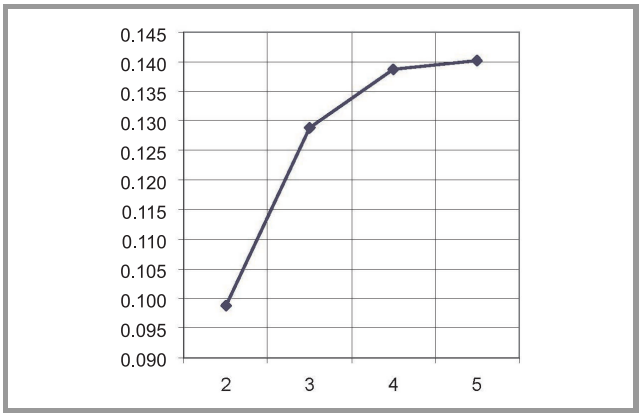


Fig. 3. The dependence of Rho* on the cluster count for columns – technological aspect, the B1 group.

The overrepresentation map containing 3 feature clusters is presented in Fig. 4.

[2]Rho* – spearman's rank correlation coefficient.

Table 1

Set of quality measures

Questionnaire question	Feature characterizing a platform/an e-resource	Domain
TECHNOLOGICAL ASPECT		
How do you estimate interfaces of platforms that you used to work?	**b1.1 platform interface**	{**1 =** slow, illogical design, uncomfortable, not easy to use **2 =** not very well designed, allowing the use of platform **3 =** I have not learnt that way **4 =** well designed, required some improvements for the quality work **5 =** very well designed}
How fast do pages, graphics, audio, video, and other materials load? How do you estimate the general work speed?	**b1.3 work speed**	{**1 =** definitely slow **2 =** rather to slow **3 =** don't know **4 =** sufficient fast **5 =** definitely fast}
How do you estimate an audiovisual attractiveness of e-learning applications that you used?	**b1.4 audiovisual attractiveness**	{**1 =** definitely low **2 =** rather low **3 =** no opinion **4 =** rather high **5 =** definitely high}
Have applications well-designed navigation (with a readable menu, site map, etc.) and well-organized courses (with clear structure; how lessons and are materials subdivided into chapters, exercises, etc.)?	**b1.5 navigation**	{**1 =** poor design **2 =** not very well designed **3 =** don't know **4 =** mostly well designed **5 =** definitely well designed}
For application you used, was interface consistent in such aspects as navigation, background colors, font colors, or within header, content, text, link, material and label elements?	**b1.6 interface consistency**	{**1 =** inconsistent **2 =** partially consistent **3 =** don't know **4 =** mostly consistent **5 =** fully consistent}
In your opinion, what features have the biggest influence on the reliability of an Internet application?	Influence on application reliability **b2.1a objectivity and extensiveness of content** **b2.1b reputation of author(s)** **b2.1c professional graphic design** **b2.1d links to other sites** **b2.1e lack of advertising banners** **b2.1f visit count**	{**1 =** inessential **2 =** little importance **3 =** important **4 =** vital}
What features best characterize the usability of Internet application? (according to ISO 9241, the usability is defined as a measure of performance, efficiency and user satisfaction, i.e., in shorthand as a measure of service ergonomics)	Importance of features characterizing a platform usability **b2.2a good navigation design** **b2.2b content essentiality** **b2.2c work performance** **b2.2d platform-independent layout** **b2.2e professional graphic design** **b2.2f simplicity of use**	{**1 =** inessential **2 =** little importance **3 =** important **4 =** vital}

Continuation of Table 1

Questionnaire question	Feature characterizing a platform/an e-resource	Domain
In your opinion, how important is graphical user interface (GUI) for everyday work when using the same application?	**b2.4 significance of graphical interface**	{**1 =** lack of influence, a form of use does not matter **2 =** little importance **3 =** no opinion **4 =** important **5 =** crucial}
Choose maximum 5 features that are the most important, in your opinion, for the user interface. If there is any not quite clear description, trust your intuition and your first impressions.	**b2.5a clarity/simplicity/ cleanliness** **b2.5b forseeability/acquitance/ compatibility with other systems** **b2.5c easy-to-use/comfortableness** **b2.5d configurability/flexibility** **b2.5e visual attractiveness of graphical design** **b2.5f consistency** **b2.5g communication directness/ awareness and control** **b2.5h performance/speed** **b2.5i error tolerance/reversibility**	{**1 =** inessential **2 =** important}
Would you like that an e-learning platform could be able to facilitate relationships among learners and teachers in a similar way as on community portals, e.g., grono.net, nasza-klasa, facebook?	**b2.7 possibility to build community relationships**	{**1 =** no **2 =** no, no opinion **3 =** yes}
ASPECT RELATED TO E-RESOURCE		
How do you estimate the quality and the design of e-learning resources?	**b1.2 e-resource**	{**1 =** mediocre **2 =** sufficient **3 =** don't know **4 =** well **5 =** very well}
In your opinion, what features have the biggest influence on reliability of an Internet application?	**b2.1a objectivity and extensiveness of content essentiality**	{**1 =** inessential **2 =** little importance **3 =** important **4 =** vital}
What features best characterize the usability of Internet application? (according to ISO 9241, the usability is defined as a measure of performance, efficiency and user satisfaction, i.e., in shorthand as a measure of service ergonomics)	**b2.2b content essentiality**	{**1 =** inessential **2 =** little importance **3 =** important **4 =** vital}
In your opinion, what kind of elements should usually supplement textual content of courses?	**b2.3a graphics** **b2.3b audio** **b2.3c video** **b2.3d animation** **b2.3e interludes/interactive games** **b2.3f only text**	{**1 =** never **2 =** rarely **3 =** often **4 =** always}
HUMAN ASPECT		
In your opinion, what features have the biggest influence on reliability of an Internet application?	**b2.1b reputation of author(s)**	{**1 =** inessential **2 =** little importance **3 =** important **4 =** vital}

Analyzing the overrepresentation maps shown in Fig. 4 we chose the most external columns corresponding to the most differentiated features: b1.4 and b1.3. On the basis of

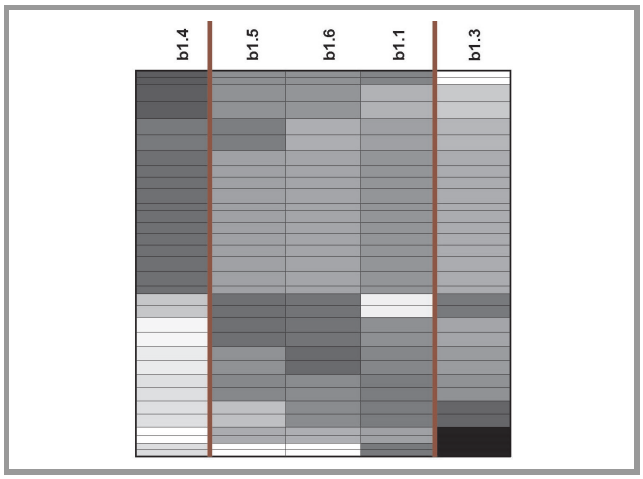


Fig. 4. The overrepresentation map with marked 3 clusters – technological aspect, the B1 group.

those features, one can find that persons who estimated high the attractiveness of platforms regarding a visual aspect (overrepresentation of the b1.4 feature) at the same time estimated low the loading speed (underrepresentation of the b1.3 feature).

As non-differentiated features we chose columns in the middle of the overrepresentation map (b1.1, b1.5, and b1.6). On those grounds one can find that the majority of respondents estimated as important (non-differentiated) the following features: the quality of the interface of the e-learning platform (b1.1), the well-designed navigation of an e-learning application (b1.5), and the interface consistency (b1.6). It is interesting that the attractiveness of e-learning applications in the audiovisual aspect and with respect of the working speed (i.e., loading speed of pages, graphics, audio, video, etc.) were definitely important for the minority of respondents (b1.3, b1.4).

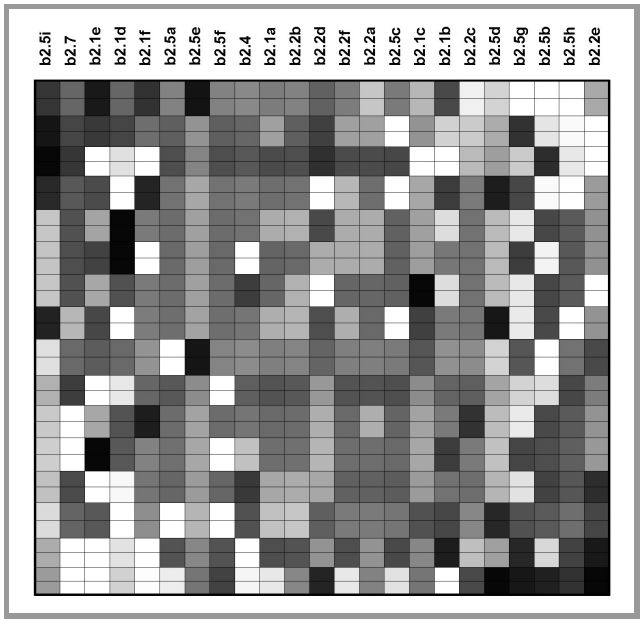


Fig. 5. The overrepresentation map for the B2 feature group.

Next, the B2 feature group was analyzed analogically. The overrepresentation map for them is presented in Fig. 5.

As previously, we performed the cluster analysis in order to find two subset of features: non-differentiating and differentiating for the features of the B2 group. Figure 6 illustrates the dependency of Rho* values (for the columns).

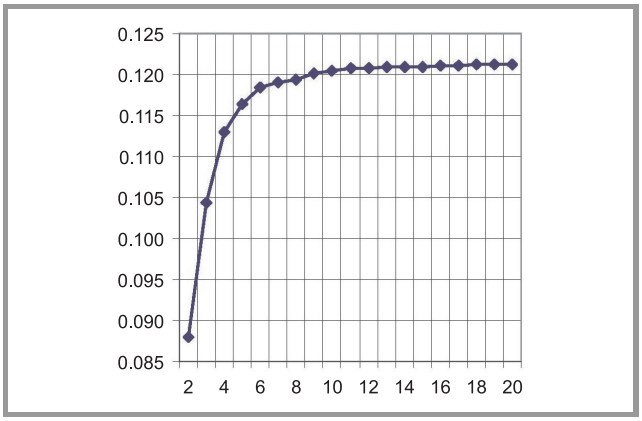


Fig. 6. The Rho* for the different values of the number of clusters – technological aspect, the B2 group.

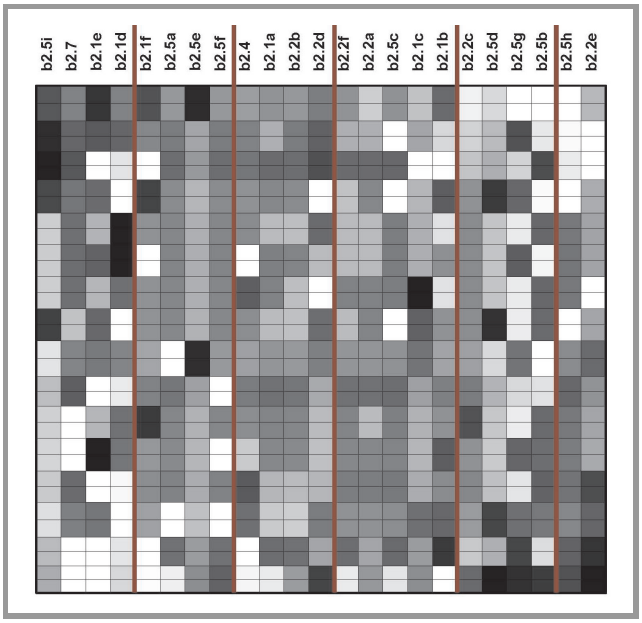


Fig. 7. The overrepresentation map with the chosen numer of clusters – technological aspect, the B2 group.

On the basis of that diagram, 6 clusters for the columns were chosen. The overrepresentation map with the determined number of clusters is shown in Fig. 7.

Analyzing the map presented in Fig. 7 we can distinguish two separate groups of features:

– differentiating: two left-most and two right-most clusters;

– non-differentiating: two clusters in the middle of the map.

On the basis of the differentiating features, we can notice that for a small group of respondents the following features are important:

- within the group of features that are most important for good interface: b2.5a – clarity/simplicity, cleanliness, b2.5e – visual attractiveness of graphical design, b2.5f – consistency, and b2.5i – error tolerance/reversibility;
- within the group concerning the reliability of an internet application: b2.1e – lack of advertising banners, b2.1d – links to other sites, and b2.1f – visit count;
- also b2.7 – a possibility to build community relationships.

On the other hand, the respondents don't pay attention to the following differentiating features:

- b2.2c – work performance, b2.2e – professional graphic design;
- features characterizing the interface: b2.5b – forseeability/familiarity/compatibility with other systems, b2.5d – configurability/flexibility, b2.5g – communication directness/awareness and control, and b2.5h – performance/speed.

To estimate the quality of e-learning platforms from the technological point of view, the non-differentiating features should be taken into consideration:

- the group of features with the greatest importance for the application reliability, i.e., b2.1a – objectivity and extensiveness of content essentiality, b2.1b – reputation of author(s), b2.1c – professional graphic design;
- the group of features characterizing the internet applications with the best usability, i.e., b2.2a – good navigation design, b2.2b – content essentiality, b2.2f – simplicity of use;
- the group of features, the most important for good interface, i.e., b2.5c – easy-to-use/comfort/ convenience, and b2.4 – significance of graphical interface.

2.2. Analysis of Overrepresentation Maps – the e-Resource Aspect

In Fig. 8, we present the results of the analysis performed using the GradeStat overrepresentation map for the features concerning the e-resource aspect (the B2 group).

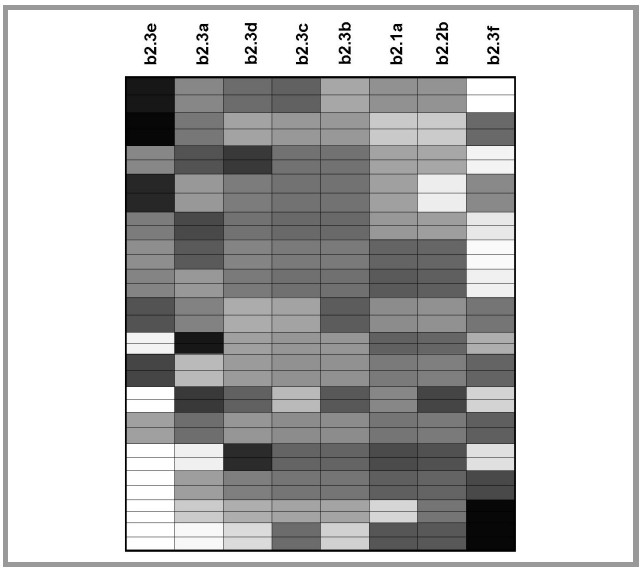


Fig. 8. The overrepresentation map for the B2 features group (e-resource aspect).

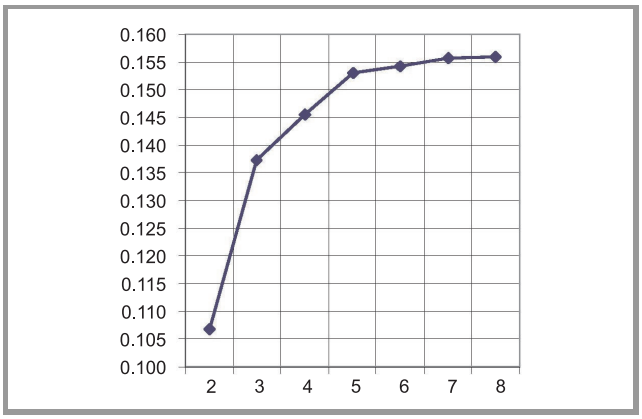


Fig. 9. The Rho* for the different values of the number of clusters – e-resource aspect, the B2 group.

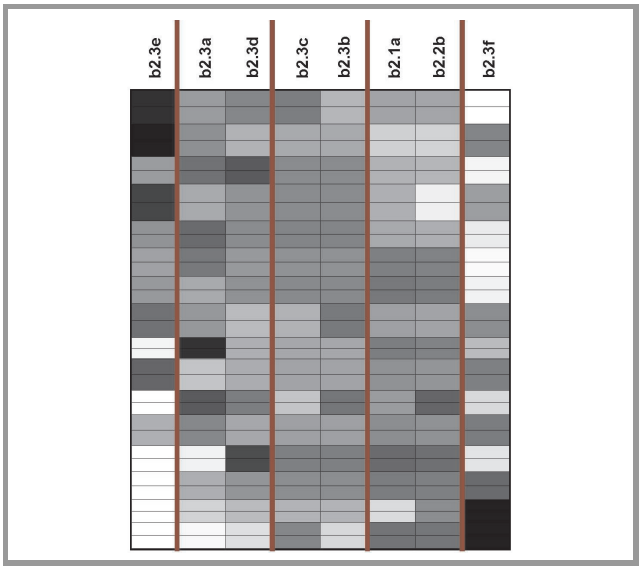


Fig. 10. The overrepresentation map with the chosen number of clusters – e-resource aspect, the B2 group.

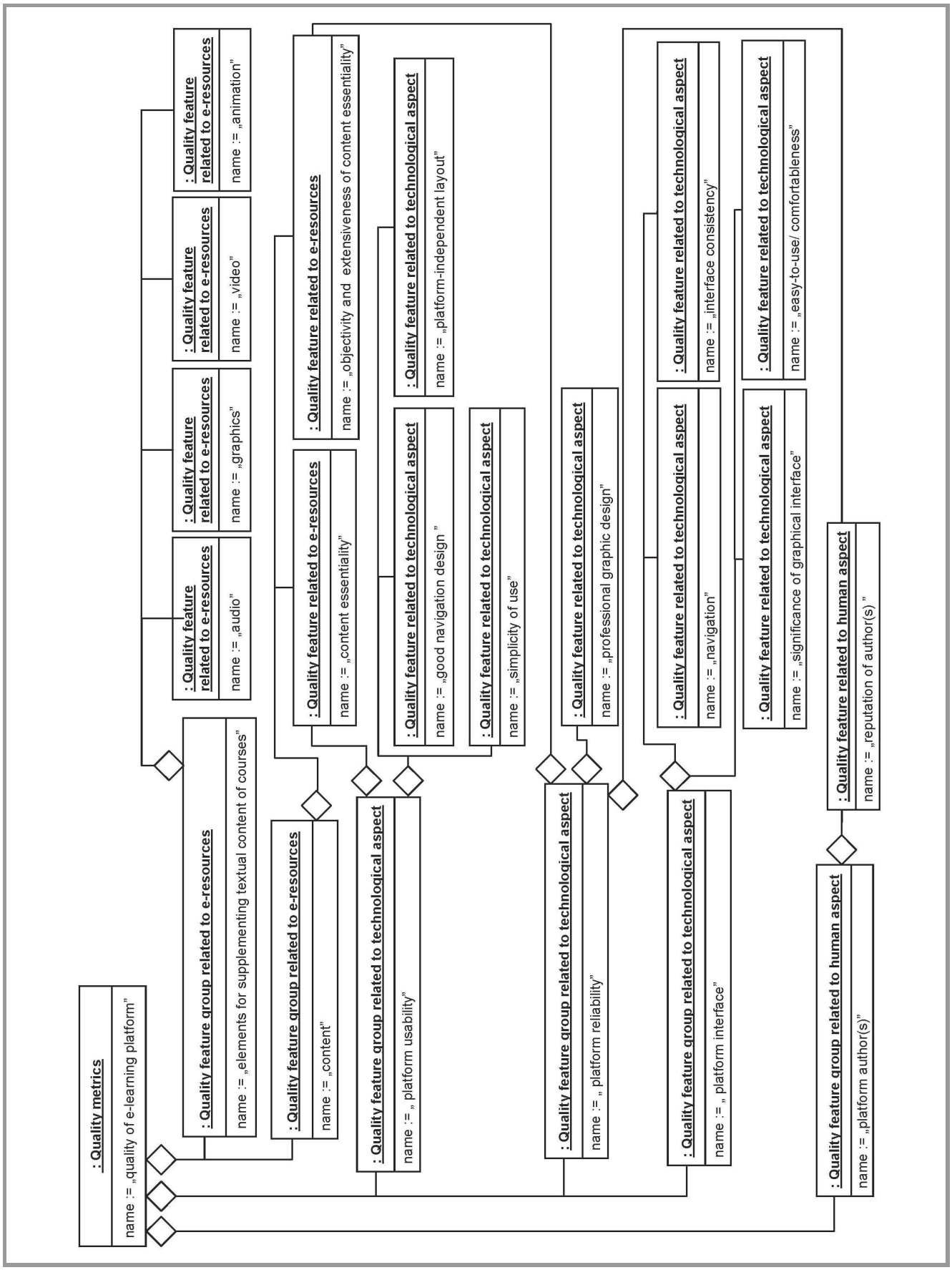


Fig. 11. A proposal for the quality metrics for the e-learning platforms.

The cluster analysis helps us to determine two groups of features: the features which do not differentiate the examined population (the middle columns of the map) and those which differentiate the population (the left-most and the right-most columns). After the analysis of Rho* variations (see Fig. 9) we chose 5 clusters for the columns within the overrepresentation map (see Fig. 10).

Finally, we can specify two following groups of features:

– differentiating features: b2.3e – interludes/interactive games, and b2.3f – only text;

– non-differentiating features: b2.3a – graphics, b2.3b – audio, b2.3c – video, b2.3d – animation, and also b2.1a – objectivity and extensiveness of content essentiality, b2.1b – reputation of author(s).

2.3. The Quality Metrics for the e-Learning Platforms

As a result of our studies, we propose the quality metrics for the e-learning platforms (Fig. 11), conformant to the idea of quality metrics (Fig. 1).

In Fig. 11 the attribute *weight* was omitted. At the moment the weights are equal 1 for all quality features. Of course, in the future we should find weights for the particular features, testing the metrics on the existing e-learning platforms.

3. Conclusions and Further Research

In our previous publications concerning the quality of e-learning we focused on the research on e-learning resources. This paper discusses two other quality aspects, i.e., the technological aspect and the, so-called, human aspect, which in our opinion, are vital to the quality of e-learning.

After the analysis of the data gathered from the questionnaire with regard to non-didactic features, for both aspects we specified the most important features, those having the biggest influence on the quality of e-learning. That constitutes the quality metrics in the non-didactic aspect.

To the most important features related to the technological aspect were ranked, i.e., the quality of the interface of the e-learning platform, in particularly, the well-designed navigation and the interface consistency. Regarding the human aspect, the following features are identified as distinctive, i.e., content essentiality, the reputation of author(s), multimedia form/s of e-materials, or clarity, simplicity, and attractiveness of graphical interface.

Further work will be necessary to establish the weights of measures and to the augmented quality metrics for e-resources and e-learning platforms with regard to non-didactic features.

References

[1] E. Stemposz and A. Stasiecka, "Determining a set of measures for quality estimation of e-resources conformant to the model/models defined in traditional education", in *Proc. 6th WSEAS Int. Conf. E-ACTIVITIES'07*, Puerto de la Cruz, Spain, 2007.

[2] B. Joyce, E. Calhoun, and D. Hopkins, *Przykłady modeli uczenia się i nauczania* (Models of learning – tools for teaching). Warsaw: WSiP, 1999 (transl. – in Polish).

[3] A. Stasiecka, E. Stemposz, and W. Dąbrowski, "Didactic aspects influence on quality of e-learning resources", in *Proc. WSEAS Int. Conf. Circ. Syst. Commun. Comput.*, Athens, Greece, 2005 (*WSEAS Trans. Inform. Sci. Appl.*, iss. 7, vol. 2, pp. 1002–1008, 2005).

[4] A. Stasiecka, J. Plodzien, and E. Stemposz, "Measures for estimating the quality of e-learning materials in the didactic aspect", in *Proc. Conf. Web Inform. Syst. Technol. WEBIST 2006*, Lisbon, Portugal, 2006, pp. 204–212.

[5] E. Stemposz, A. Stasiecka, and A. Jodłowski, "The proposal of metadata for defining the quality of e-learning", in *Proc. Conf. ENMA 2007 Eng. Math.*, Bilbao, Spain, 2007.

[6] A. Stasiecka, E. Stemposz, and A. Jodłowski, "E-resources versus traditional teaching models", *J. Telecommun. Inform. Technol.*, no. 3, pp. 74–81, 2008.

[7] M. Maliszewski, "Metodyka projektowania interfejsu użytkownika w e-learningowych aplikacjach WWW". M.Sc. thesis. Warsaw, Polish-Japanese Institute of Information Technology, Dec. 2008 (in Polish).

[8] "Program for grade data analysis", GradeStat, 2007, http://gradestat.ipipan.waw.pl/

Ewa Stemposz was born in Wasilków/Białystok, Poland. She graduated master degree at the Faculty of Electronics, Technical University of Warsaw. She has been employed at the Institute of Computer Science of the Polish Academy of Sciences, Warsaw (from 1988) and the Polish-Japanese Institute of Information Technology, Warsaw (from 1994), where she has been engaged in computer graphics, computer vision, data basis, software engineering, object analysis, project management, and e-learning.

e-mail: ewag@pjwstk.edu.pl
Polish-Japanese Institute of Information Technology
Koszykowa st 86
02-208 Warsaw, Poland

e-mail: ewag@ipipan.waw.pl
Institute of Computer Science
Polish Academy of Sciences
J. K. Ordona st 21
01-237 Warsaw, Poland

Andrzej Jodłowski received his M.Sc. degree from the Warsaw University of Technology, Poland, in 1995, in the area of applied computer sciences. He received his Ph.D. from the Institute of Computer Science of the Polish Academy of Sciences, Warsaw, in 2003, in the area of object-oriented database technologies. Currently he is a Lecturer at the Warsaw University of Life Sciences. His research interests concern object-oriented databases, software engineering, and e-learning.

e-mail: Andrzej_Jodlowski@sggw.pl
Warsaw University of Life Sciences
Nowoursynowska st 166
02-787 Warsaw, Poland

Alina Stasiecka was born in Warsaw, Poland. She finished the Electronics and Information Technology Faculty of the Warsaw University of Technology. She was working in the Institute of Computer Science of the Polish Academy of Sciences, Warsaw (years 1986–2001). Since 2001 she has been employed at the Polish-Japanese Institute of Information Technology, Warsaw, where she has been engaged in software engineering and e-learning.

e-mail: alas@pjwstk.edu.pl
Polish-Japanese Institute of Information Technology
Koszykowa st 86
02-008 Warsaw, Poland

Paper

Heuristic Analysis of Transport System Efficiency Based on Movement of Mobile Network Users

Grzegorz Sabak

Abstract— The paper describes results of introductory research focused on possibility to use location data available in a mobile network for the analysis of transport system status and efficiency. The details of a system capable of detecting abnormal traffic situation (accidents, heavy congestion) are described. This system (called VASTAR) uses a neural network to learn and store certain characteristic of the analyzed part of a road system. Based on a measured divergence from normal characteristic, a notification about non-typical situation is triggered. The results of a computational experiment using real-world location data and simulation of abnormal situation are provided. The proposed system can be a relatively low cost way to improve competitiveness of a mobile network operator by allowing him to offer new type of informational service. It could also aid municipal authorities by providing support for decisions regarding road traffic control and management and be used by emergency services as a monitoring an alarming tool for detecting abnormal road traffic situations when other means of observation are unavailable.

Keywords— congestion detection, decision support systems, neural networks, transport system.

1. Introduction

It is a common observation frequently found in newspapers that transport service is vital for both operations of the countries' economies and quality of life of the ordinary people. We all frequently experience either a pleasure of traveling along scenic highway or strong inconvenience of being stuck in a traffic jam or in a crowded bus traveling slowly through streets of a city centre.
Other domain of technology that became extremely important nowadays is mobile communications. Having mobile phone and being almost instantly accessible is assumed for all people living active professional or social life.
These two domains are similar in many ways. Firstly, their purpose is a change of place of availability of either goods (transport) or information (communication) with high efficiency. This typically means low cost and high speed. Secondly, problems encountered in research in these two areas are to a certain extent similar (e.g., routing, scheduling, congestion). Lastly, we all have an indispensable need of using them regularly, and we are actually doing that almost all the time.
These observations lead to the concept of analysis of transportation system state based on the behavior of users of mobile communication network. Since some of the required information is already available at relatively low cost in a mobile network, it may be a valid opportunity for business and government entities. This research is aimed to exploit this opportunity using information technology tools and techniques.

2. Location Services

Typically, among other value added services, the mobile network operator provides a set of services based on the customer's location. Customer's location, means position of a mobile station (MS) consisting of a user equipment and a SIM card. User equipment is usually one of the following: a mobile phone, a PC data card, or a dedicated GSM module installed in a vehicle. Business case of location services (LCS) assumes that information or service which would be strongly related to the current situation of the customer (his location in this case) is so valuable for him, that he is willing to pay a premium.
In general, there are four types of location services [1]:

- Commercial LCS – all services that are available to customers (e.g., SMS, MMS or WAP services offering content dependent on the location).
- Internal LCS – in which location information is used for internal purposes of a mobile network operator.
- Emergency LCS – enforced by law, provide location information in case of emergency situations.
- Lawful intercept LCS – also typically enforced by law, supports various legally sanctioned services.

Regardless the type, implementation of all LCS require presence of a gateway mobile location centre (GMLC) in the network. The GMLC plays key role as it actually enables different applications to access customer location information. Different interfaces can be used by LCS client to use location service, however mobile location protocol (MLP) is most widely adopted, being now considered sufficiently standarized.
In GSM networks, different positioning methods are available to choose from. Most of them are based on the features of a radio access network such as cell identification (cell ID), radio signal strength, and various time variables. They are not described in this paper, but detailed specifications are available, e.g., in [1], [2].

The cell ID method is especially important and shall be explained in more detail. This positioning method is most widely used and does not require significant investment in operator's access network or his core network. Cell ID parameter is used frequently in a mobile network. In this method, all cells in the network are mapped to geographical coordinates which point to a location considered to be the most likely position of the mobile station active in the cell.

3. Problem Formulation

3.1. Introductory Definitions

Let *an event* will be a pair $e = \langle t, P \rangle$, where t is a timestamp and P is a list of parameters (attributes) defining event type.
The sequence of events ordered according to ascending timestamp will be called *an event stream* and denoted as

$$\mathscr{E} = \{e_1, e_2 \ldots, e_k\}.$$

Given two event streams $\mathscr{E}_1$ and $\mathscr{E}_2$ an operation of *stream junction* $\oplus$ can be defined. This operation creates a new event stream that contains all events from both streams and preserves the order of timestamps.
Let an *event stream filter* will be a function $f : e \to \{0,1\}$. Each event mapped to value 1 will be considered as passed through the filter. All other events will be considered as blocked by the filter. Although filter is a function defined for sequence elements, for convenience we will denote it also as $f(\mathscr{E})$, which indicates that all sequence elements are processed by filter.
An operation $F(\mathscr{E}, f)$ removing all events for which $f(e) = 0$ from an event stream $\mathscr{E}$ will be called a *filtering operation*. The result of applying a filtering operation to a stream of events is also a stream of events.
In order to be able to model input data and define a problem let us define the following:

- $x, y \in R$: geographical coordinates; no particular coordinate system is assumed;
- AREA $\subset R^2$: area on which the problem is defined;
- MS: a set of all mobile stations in a mobile network;
- $p = \langle x, y \rangle \in$ AREA: location of a mobile station;
- $\hat{p} = \langle \hat{x}, \hat{y}, r \rangle$: mobile station location estimate defined as a circle (with a centre in $\langle \hat{x}, \hat{y} \rangle$ and radius r) in which the station is located.

Let *a location event* be defined as an event $l = \langle t, P \rangle$ with $P = \langle m, \hat{p} \rangle$, where $m \in$ MS is a mobile station which was located and $\hat{p} \in$ AREA is the location estimate of the mobile station.
A sequence of location events

$$\mathscr{L} = \{l_1, l_2 \ldots, l_k\}$$

will be called *a location event stream*.

3.2. Definition of Problems

Based on the definitions above, two problems can be defined. These problems will become a base for future considerations.

Trip time prediction. Given AREA, MS, and $\mathscr{L}$ at the moment t^* estimate time $\hat{t}$ that is required for mobile station $m \in$ MS to move from location p_1 to p_2 ($p_1, p_2 \in$ AREA) assuming that mobile station is minimizing trip time.

Congestion detection. Given AREA, MS, and $\mathscr{L}$ find the place and time when high congestion happened.

3.3. Further Definitions

For further considerations let *a move event* be an event $v = \langle t, P \rangle$, where $P = \langle m, t_b, \hat{p}_b, t_e, \hat{p}_e \rangle$, where: $m \in$ MS is a mobile station, t_b is a starting moment of the move, $\hat{p}_b = \langle \hat{x}_b, \hat{y}_b, r_b \rangle$ is a location estimate at the start of the move, t_e is an ending moment of the move, $\hat{p}_e = \langle \hat{x}_e, \hat{y}_e, r_e \rangle$ is a location at end of the move.
A sequence of movement events

$$\mathscr{V} = \{v_1, v_2 \ldots, v_k\}$$

will be called *a move event stream*.
For each move event the following variables can be calculated:

- movement duration $t_{dur} = t_e - t_b$,
- distance $d = \sqrt{(\hat{x}_e - \hat{x}_b)^2 + (\hat{y}_e - \hat{y}_b)^2}$.

Creation of the move event stream. The move event stream is created from a location event stream according the following procedure. For every $l_i = \langle t_i, m_i, \hat{p}_i \rangle \in \mathscr{L}$.

1. Let j be the highest number, where $m_j = m_i$ and $j < i$. The $\hat{p}_j$ is the last known location estimate of mobile station m_i. The t_j is the timestamp associated with the location event.
2. If j cannot be determined (which means that $\hat{p}_i$ is the first location event of mobile station m_i) continue with processing of $\mathscr{L}$.
3. If j can be determined create a movement event $v_k = \langle t_k, m_k, t_{b_k}, \hat{p}_{bk}, t_{e_k}, \hat{p}_{ek} \rangle$, where $t_k = t_i$, $m_k = m_i$, $t_{b_k} = t_j$, $\hat{p}_{bk} = \hat{p}_j$, $t_{e_k} = t_i$, and $\hat{p}_{e_k} = \hat{p}_i$.

This procedure can be effectively implemented using, for example, a sorted list of last known location estimates.

4. The VASTAR Project

Analysis of the mobile network subscribers movement provides unique, nearly real-time information about the status of the transport system they use. To achieve this, a number

of problems must be dealt with and solutions to them must be found.
Potential attractive applications of a system performing such analysis include.

- Provision of city or highway traffic data supporting urban traffic management and control (UTMC) process with no additional costs of traffic measurement units. Obviously, limited accuracy of such data would probably mean that traditional methods would have to be used in key areas.
- Support tool for monitoring city and highway traffic able to detect abnormal situations (e.g., congestions, accidents). This would be especially useful in areas without video surveillance.
- Business opportunity for mobile network operators to attract their subscribers to the website presenting up to date traffic information.
- Decision support system for local authorities optimizing investment in transport infrastructure.

This list defines objectives for a design of traffic decision supporting system. In Section 3, model of available location data was defined. In this section an architecture of such system is proposed. Let this information system be called VASTAR (name can be resolved to value added services for streets and roads) which for the sake of convenience can be shortened to VA∗. The VA∗'s high level logical architecture presented further does not limit its functionality to solving only one kind of problems but rather provides a framework which can support different types of analysis.

Elements constituting VA∗ are presented in Fig. 1. The functions of different modules are described in the following paragraphs, and an application of this architecture to the task of detecting abnormal road conditions is described later in Section 6.

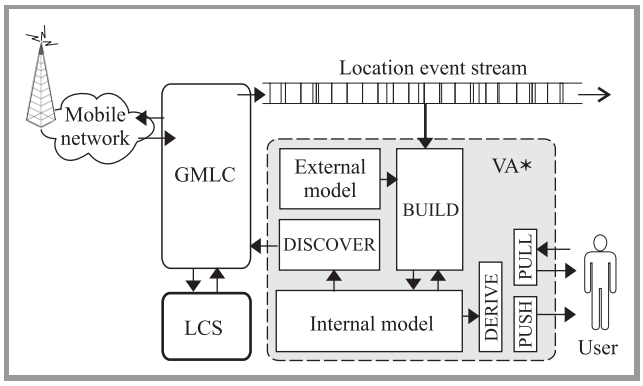


Fig. 1. Architecture of the VA∗ system.

External model. The purpose of the external model is to provide the VA∗ with information regarding transport system which is important to the application and known a priori to the user. This could include, for example, model of the transport infrastructure (roads, their types, nodes, etc.) or other key assumptions to the problem being solved.

Internal model. Represents VA∗'s knowledge about the analyzed transport system, including information about objects and their attributes.

BUILD module. This module is responsible for the analysis of incoming data stream, processing it, and acquiring information needed to perform tasks requested by users. This information is stored in the system by modifications of parameters of the internal model.

DERIVE module. The module performs analysis of information stored in the internal model in order to provide responses to the queries submitted by users via PULL/PUSH modules.

PUSH/PULL modules. These modules are responsible for processing queries from the users (PULL) and for notification about the events in the transport system (e.g., expected congestion) based on the requirements provided by the users (PUSH).

DISCOVER module. This functional element of the VA∗ shall have an ability to request a location of the specific mobile station. Its goal is to improve information quality of the location event stream from the point of view of the system's application.

5. Analysis of Real-World Data

In order to assess informational value of available data a basic statistical analysis was performed. A sample of the GMLC logs was used as an input data. The GMLC operates in the network of mobile network operator in Poland which uses cell ID as a location method. The consequence of limiting input data only to location streams processed by GMLC is that only a small fraction of all mobile stations are "observed". However, by doing this we avoid costs of collecting location data from network elements, which can be quite problematic. The sample used contained GMLC log entries from eight consecutive days from 25 Sept. to 2 Oct. 2008.
The key findings of analysis are as follows.

- The observed mobility of mobile stations, understood as a number of locations visited by a MS is different for working days (higher) and weekends (lower).
- The mobility of mobile stations is not equal throughout the day, but significantly lower between 10PM and 5AM.
- In a real world, location errors are encountered. For example, mapping of the cell ID to geographical coordinates can be incorrect.

- Some locations are visited more frequently then others (majority of them are actually near important transport routes, e.g., Warsaw-Katowice express road).
- It is visible that there is a process running in the network that is monitoring location of a set of mobile stations. This can be accounted to one of location services offered by the operator.
- Only a part of collected data is related to real movements of mobile stations. In many cases mobile stations are not changing their locations or a *virtual* movement between two or three locations is observed.

Virtual movement of mobile stations. In the land mobile networks it is very common that multiple cells overlap. This is a result of the requirement that holes in the coverage are eliminated. A situation when two cells cover the place where mobile station is located is shown in Fig. 2.

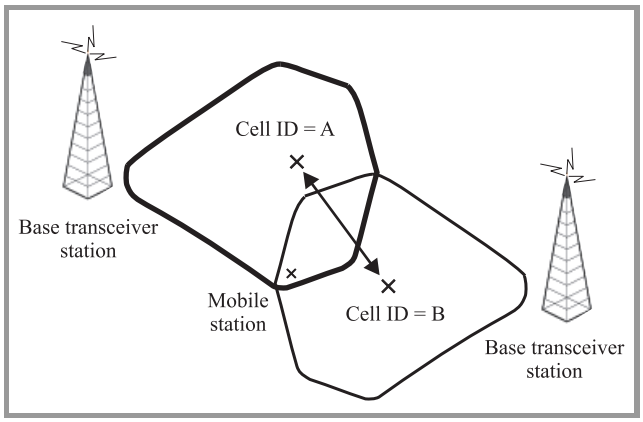


Fig. 2. Two overlapping sector cells.

In this case, despite the fact that real location of the mobile station remains the same, changing conditions (physical environment factors and mobile network status) cause mobile station to be reported as present in either cell A or cell B. The MS appears to move back and forth between centers of the cells A and B. Similar situation can be encountered when three or more cells cover location of the mobile station.

Summary. Analysis of available data shows that information about users of road system which can be used to evaluate status of this system is available in the GMLC logs. However, there are difficulties which need to be addressed.

- Lack of information about accuracy of the location data. When location data comes from the GMLC, radius of the circle representing position estimate is not available. Typically for cell ID method accuracy is in the 50 m – 1 km and 1 km – 35 km ranges in the urban and rural areas, respectively [3].
- Lack of information which MS are actual road system users. Location events found in GMLC logs do not necessarily refer to the positions of the mobile stations attached to the vehicles or people using transportation system. A way to distinguish between MS of his type and all other MS has to be proposed or a method that can effectively deal with this kind of "information noise" has to be used. "Information noise" is a data present in the event stream, but not useful for the system.
- The virtual movement of mobile stations. This introduces additional "information noise" and causes situation when static mobile station is seen by the VA$*$ as changing its position.
- High volumes of data to be processed. Extensive use of the location services in the mobile network may require processing millions of records, which may be a problem when a real-time operation of the system is required.
- Incorrect data due to errors in the positioning process. As in all real world systems, especially such complex and depending on the physics of nature as mobile network, incorrect position reports cannot be excluded. In some cases an error of location estimate can reach several hundred of kilometers.

The nature of these problems indicate that some kind of heurisitics method should be used to deal with them.

6. Accident Detection

Based on the high level architecture described in Section 4, a monitoring system capable of detecting abnormal traffic situations can be built.
Accidents or heavy congestions are generally situations when velocity of vehicles drops down dramatically. This means that in a given period of time the distances by which vehicles move are significantly smaller. It can be assumed that for every move event the distance d between estimated starting and finishing points depends on the following variables:

– starting point coordinates (x, y);

– direction of the move a;

– time of the day t_{day};

– duration of the move t_{dur}.

Let this dependency be called *a move length characteristics* and be denoted as $d(x, y, a, t_{day}, t_{dur})$. This characteristic is directly related to vehicles' average speed (the lower speed the lower value of characteristic). An assumption is made, that abnormal traffic conditions can be detected by monitoring changes of this characteristic.

In the accident detection application the VA∗ determines the move length characteristic of a road system defined by the given external model. Later on, during operation it monitors how much this characteristics for the current moment differs form the known one. A difference exceeding a pre-defined threshold may indicate abnormal conditions, e.g., accident or high congestion.

6.1. Internal and External Models of a Road System

In order to verify the proposed approach a computational experiment was conducted. In this experiment the analysis was limited to only a single road segment. In this case the road is represented as one line section. For the sake of further possible generalization of knowledge, the move events starting and finishing points are transformed to the new coordinate system. This coordinate system is defined by setting coordinates of the road segment in a way that it becomes a line section with ends at points $(0,0)$ and $(0,1)$. The effects of such transformation (for the road segment and for one of the move events) are shown in Fig. 3.

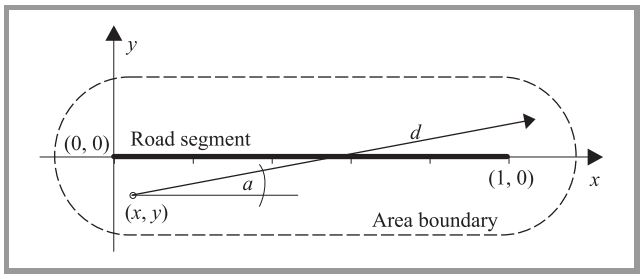


Fig. 3. Transformation of road segment.

In this application the VA∗ calculates move distance characteristics within an DERIVE module and stores it in the internal model. This characteristics is obtained from movement events created from location event stream available from the GMLC. In the internal model a neural network [4]–[6] (dual layer perceptron) is used to approximate this characteristic. The architecture of such network is shown in Fig. 4.

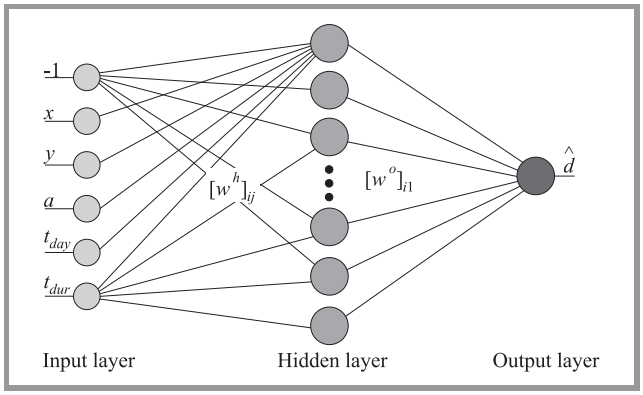


Fig. 4. Architecture of neural network.

The neural network used had six inputs nodes with the following input variables:

- constant value -1 (which makes activation thresholds in neurons unnecessary);
- x coordinate of the starting point;
- y coordinate of the starting point;
- a: angle between the move vector and $0x$ axis,
- t_{day}: moment of the day the move began;
- t_{dur}: duration of the move.

All nodes in a hidden layer use a sigmoid activation function. The output value of the ith neuron in a hidden layer of the perceptron is calculated according to the following formula:

$$o_i^h = g(w_{i1}^h \cdot -1 + w_{i2}^h \cdot x + w_{i3}^h \cdot y + w_{i4}^h \cdot a + w_{i5}^h \cdot t_{day} + w_{i6}^h \cdot t_{dur}),$$

where w_{ij}^h is a weight associated with jth input of a ith neuron, and $g()$ is a sigmoid activation function (logistic curve) defined by equation:

$$g(x) = \frac{1}{1+\exp^{-2\beta x}}.$$

Parameter β is called a slope factor. Plots of the sigmoid function for different values of β are shown in Fig. 5.

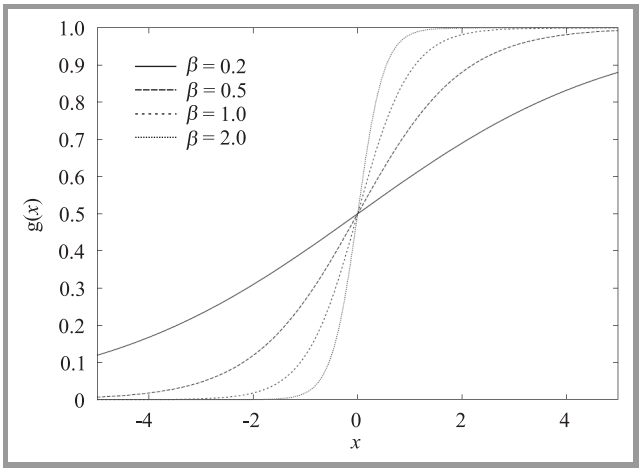


Fig. 5. Activation function.

There is only one node in an output layer. The output value of this node is calculated acording to the following equation:

$$\hat{d} = \sum_{i=1}^{n\_hidden} w_i^o \cdot o_i^h,$$

where *n_hidden* is a total number of neurons in the hidden layer, and w_i^o is a weight associated with ith input of the output neuron. For the output neuron activation function is not used. Because of that, the perceptron is able to learn function values not limited to any specific range, e.g., $(0,1)$ in case of a sigmoid function.
The output value of the network is the estimated value of a move length characteristics, denoted as $\hat{d}$.

6.2. Learning Process

From the VA∗'s architecture point of view, the development of the internal model is done by modification of the weights of a neural network. An important, but quite straightforward in implementation, function of the BUILD module is filtering out all move events that does not start within defined distance from the road segment (5 km in this case). During the learning, weights were adjusted according to an *error backpropagation* [4] algorithm with the step α.

When learning is completed the VA∗ knows, with some level of error, how mobile stations behave in the analyzed area. As new move events come in the stream a moving window average of error is calculated. When this value exceeds an average value of the error by a certain factor the notification is triggered that the traffic conditions on this part of the road differ from typical ones. The above describes principles of the DERIVE module operation.

6.3. Model of an Abnormal Situation

It was not possible to gather information regarding occurrence of real-world accidents which happened during period when data was collected, so an option to simulate such situation by modification of the movement stream was chosen.

Let t^A_{start} and t^A_{stop} be the start and end time of an abnormal situation period. Let $\gamma \in \langle 0,1 \rangle$ be a congestion intensity. For every move event $v_k = \langle t_k, m_k, \hat{p}_{b_k}, \hat{p}_{e_k} \rangle$.

- $t_k < t^A_{start}$ – move event is not changed, it is assumend that accident did not influence the move.
- $t^A_{start} \leq t_k \leq t^A_{stop}$, the move can be impacted by the accident. With the probability:

$$p_{block} = \gamma \frac{t_k - t^A_{start}}{t^A_{stop} - t^A_{start}}$$

  $\hat{p}_{e_k}$ is set to $\hat{p}_{b_k}$.
- $t^A_{stop} < t_k$ – move event remains unchanged.

The probability of impacting the move event is greater at the beginning of a congestion period and it decreases as congestion is coming to its end. For a given t_k, p_{block} is proportional to γ.

This model assumes that during the abnormal period mobile station movements are impacted in a following way: mobile station move is blocked with a probability depending on the relative time that elapsed from the beginning of the congestion (Fig. 6).

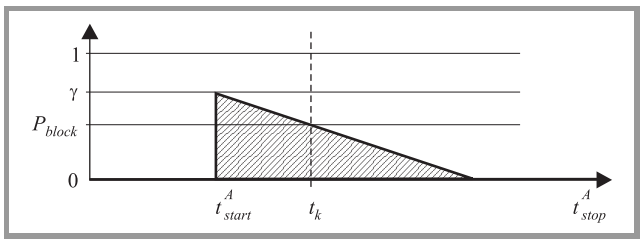


Fig. 6. Probability of blocking a move.

6.4. Computational Results

In this experiment the analysis of location events which took place during workdays between 25 Sept. and 2 Oct. 2008 was performed. The area covered was limited to the part of the Warsaw-Katowice express road between Tomaszów Mazowiecki and Rawa Mazowiecka. The summary of the main configuration parameters is presented in Table 1.

Table 1
Experiment configuration and parameters

Parameter	Value
Tomaszów node latitude	51°32'58.10" N
Tomaszów node longitude	19°59'16.54" E
Rawa node latitude	51°46'4.80" N
Rawa node longitude	20°15'24.22" E
Days used for learning	25 – 26 Sept., 29 Sept. – 2 Oct.
Max. dist. from road model	5 km
Simulated congestion period	25 Sept. 3PM – 4PM
Learning step α	0.2
Slope factor β	2.0
No. of neurons in hidden layer	20

The number of neurons in the hidden layer and other parameters of neural network learning process were determined by a series of experiments. The learning process in case of different parameter values is presented in Figs. 7, 8, and 9 (the results were averaged over 10 repetition of the learning process).

The conclusion of these experiments is that 20 neurons in the hidden layer, $\alpha = 0.2$, and $\beta = 2.0$ result in a good learning speed and accuracy.

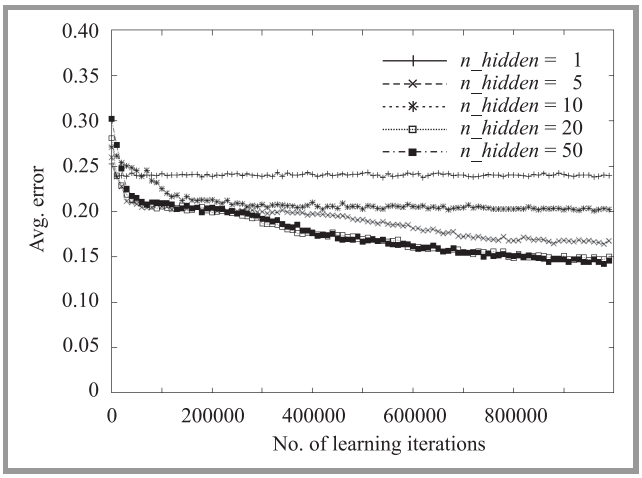


Fig. 7. Average error for different number of hidden neurons.

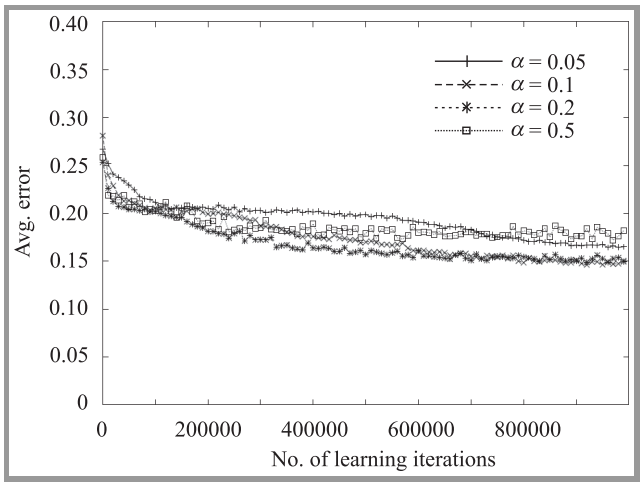


Fig. 8. Average error for different α.

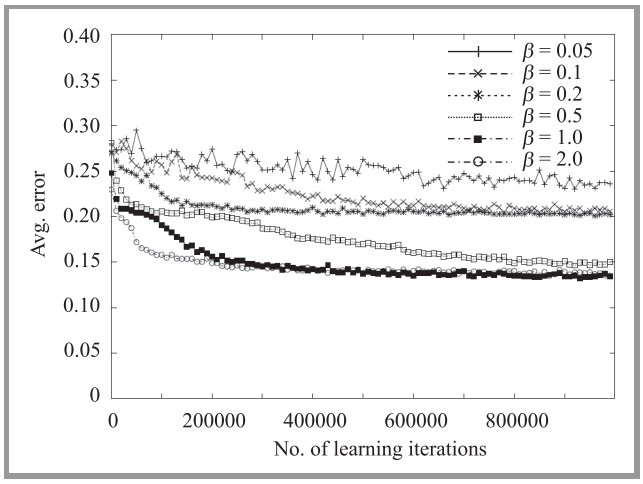


Fig. 9. Average error for different β.

In order to verify whether congestion can be detected by monitoring divergence between observed and estimated values of the move length characteristics, the following calculations were performed. For $\gamma \in \{0.25, 0.5, 0.75 1.0\}$ and different durations of congestion period (*dur*), a ratio of the average $\hat{d}$ estimation errors with congestion simulation and without simulation was calculated. This calculation was repeated 10 times for different, randomly chosen values of t^A_{start}. The average values are shown in Table 2.

For $\gamma = 1.0$ the error calculated for the move stream with congestion simulation is 22–30% higher in comparison to the average error in the case without congestion. As values of γ get smaller, the error ratio gets closer to 1 which means that this two cases became difficult (or even impossible) to be distinguish.

The conclusion of this computational experiment is that for high congestions it is possible to detect such situations by monitoring level of the average error of the move length characteristic estimation.

Table 2
Error ratio values

$dur[h]$	$\gamma = 1$	$\gamma = 0.75$	$\gamma = 0.5$	$\gamma = 0.25$
1	1.292630	1.250090	1.177957	1.128258
2	1.227444	1.151583	1.088391	1.026890
3	1.247943	1.188703	1.104728	1.059480
4	1.303445	1.252754	1.165497	1.097391
5	1.224714	1.156673	1.113232	1.019861

7. Conclusions and Further Work

In this paper a concept to use mobile stations location estimates to monitor status of a transport system was proposed and explained. A high level architecture of a decision support system (called VA∗) was proposed, its main functional modules outlined and their purpose described. This architecture represents an approach to analysis of input data (available as a sequence of location events) which can be used in different applications, depending on models and methods of analysis and reasoning used.

The analysis of a sample of a real-world data logged in a mobile network showed that such data can be used for this purpose. However, lack of location accuracy and substantial "information noise" present in the input data limit the number of methods which can be used.

The computational experiment, although limited to a small geographical area and using a simulation of the accident situation, showed that this approach can be effectively used for congestion detection. In this experiment a neural network was used as a function approximation tool. However, not unlike other heuristic methods some effort to tune parameters (in order) to achieve the best results is required.

Further research should include not only a neural network parameter tuning but also:

- verification whether other function approximation tools can give better results;
- defining reasoning about congestion detection from move length characteristics as a statistical hypothesis and verification of such hypothesis through statistical testing;
- what other move event stream characteristic can be proposed and how they change in case of accidents or congestion situations;
- analyzis how effective the usage of different strategies for DISCOVER module can be and how it can improve overall VA∗'s performance;
- verification whether data mining methods can be effectively used to deal with data available;
- how VA∗ can support objectives listed in Section 4 other than congestion detection.

A particular effort is planned to be made to determine what are the conditions influencing quality (e.g., speed, accuracy) of the results of VA∗'s operation.

References

[1] 3G TS 23.171 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Functional stage 2 description of location services in UMTS, Release 1999.

[2] W. Michalski, "Przegląd metod określania lokalizacji abonentów w ruchomych publicznych sieciach komórkowych GSM/UMTS z uwzględnieniem dokładności dostarczanej informacji, technicznych możliwości wdrożenia oraz czynników ekonomicznych i prawnych. Etap 1: Charakterystyka metod służących do określania lokalizacji abonentów w sieciach GSM i UMTS". Warsaw, Instytut Łączności, 2007 (in Polish).

[3] CGALIES Final Report, Report on implementation issues related to access to location information by emergency services (E112) in the European Union, 2002.

[4] W. S. McCulloh and W. Pitts, "A logical calculus of ideas immanent in nervous activity", *Bull. Math. Biophys.*, no. 5, pp. 115–133, 1943.

[5] F. Rosenblatt, *Principles of Neurodynamics*. Washington: Spartan, 1962.

[6] J. Hertz, A. Krogh, and R. G. Palmer, *Wstęp do teorii obliczeń neuronowych*. Warsaw, WNT, 1995 (transl. in Polish).

Grzegorz Sabak received the M.Sc. degree in computer science in 1998 from the Warsaw University of Technology, Poland. Currently he is working toward a Ph.D. degree. His research interests include application of heuristic methods to real-world problems, machine learning, and artificial intelligence. In his professional work he is responsible for design and management of value added services for one of the mobile network operators in Poland.

e-mail: grzegorz.sabak@home.pl
Faculty of Cybernetics
Military University of Technology
Gen. S. Kaliskiego st 2
00-908 Warsaw, Poland

Paper

Analytical Modeling of the WCDMA Interface with Packet Scheduling

Maciej Stasiak, Piotr Zwierzykowski, and Janusz Wiewióra

Abstract— The article presents the application of a new analytical model of the full-availability group carrying a mixture of different multi-rate traffic classes with compression property for modeling the WCDMA radio interface with packet scheduling. The proposed model can be directly used for modeling of the WCDMA interface in the UMTS network servicing different traffic classes. The described model can be applied for a validation of the efficiency of the WCDMA interface measured by the blocking probability and the average carried traffic for particular traffic classes.

Keywords— analytical model, radio interface, UMTS.

1. Introduction

The increase in popularity of data transfer services in mobile networks of the second and the third generations has been followed by an increase in the interest in methods for dimensioning and optimization of networks servicing multi-rate traffic. In traffic theory, the issues on the problem are increasingly becoming part of the mainstream analysis [1]–[11]. This situation is primarily caused by the special conditions in the construction of these networks, and by the construction of the infrastructure of the access radio network in particular – as its development or extension needs a precise definition and assessment of clients' needs and is relatively time-consuming. Cellular network operators define, on the basis of service level agreement (SLA), a set of the key performance indicator (KPI) parameters that serve as determinants in the process of network dimensioning and optimization [12]. Dimensioning can be presented as an unending and on-going process of analyzing and designing of the network. To make this work effective it is thus necessary to work out algorithms that would, in a reliable way, model the parameters of a designed network [13].

One of the mechanisms that should be analyzed in view of performance (expectations) are radio access algorithms. This article discusses and analyzes packet scheduling that is used for transmission of background and interactive traffic (the guaranteed minimum bandwidth is not a requirement), but also for the streaming class, which requires the minimum bandwidth, not being at the same time very sensitive to delays. The conversational traffic class is carried without scheduling on dedicated channels [14].

The paper has been divided into five sections. Section 2 recalls basic model of a full-availability group (FAG) with multi-rate traffic which is used in the model presented in Section 3. Section 3 describes an analytical model of the full-availability group with traffic compression. Section 4 shows application of the model in the universal mobile telecommunication system (UMTS) network for modeling of the wideband code division multiple access (WCDMA) interface with packet scheduling. This section also includes the results obtained in the study of the system. The final Section 5 sums up the discussion.

2. Model of the FAG

Let us assume that the total capacity of the full-availability group with multi-rate traffic is equal to V basic bandwidth units (BBUs). The group is offered M independent classes of Poisson traffic streams having the intensities: $\lambda_1, \lambda_2, ..., \lambda_M$. The class i call requires t_i BBUs to set up a connection. The holding time for calls of particular classes has an exponential distribution with the parameters: $\mu_1, \mu_2, ..., \mu_M$. Thus, the mean traffic offered to the system by the class i traffic stream is equal to:

$$A_i = \lambda_i / \mu_i. \tag{1}$$

The demanded resources in the group for servicing particular classes can be treated as a call demanding an integer number of (BBUs) [15]. The value of BBU, i.e., R_{BBU}, is calculated as the greatest common divisor of all resources demanded by traffic classes offered to the system:

$$R_{BBU} = \mathrm{GCD}(R_i, ..., R_M), \tag{2}$$

where R_i is the amount of resources demanded by class i call in kbit/s.

The multi-dimensional Markov process in FAG can be approximated by the one-dimensional Markov chain which can be described by Kaufman-Roberts recursion [16], [17]:

$$n\,[P_n]_V = \sum_{i=1}^{M} A_i t_i\,[P_{n-t_i}]_V\,, \tag{3}$$

where $[P_n]_V$ is the probability of state n BBUs being busy, and t_i is the number of BBUs required by a class i call:

$$t_i = \lfloor R_i / R_{BBU} \rfloor. \tag{4}$$

On the basis of formula (3) the blocking probability E_i for class i stream can be expressed as follows:

$$E_i = \sum_{n=V-t_i+1}^{V} [P_n]_V, \tag{5}$$

where V is the total capacity of the group and is expressed in BBUs ($V = \lfloor V_{phy} / R_{BBU} \rfloor$, where V_{phy} is the physical capacity of group in kbit/s).

The diagram in Fig. 1 corresponds to formula (3) for the system with two call streams ($M = 2$, $t_1 = 1$, $t_2 = 2$). The $y_i(n)$ symbol denotes the so-called *reverse transition rate* of a class i service stream outgoing from state n. This parameter can be calculated on the basis of the local equations of equilibrium in the Markov chain [16], [18]:

$$y_i(n) = \begin{cases} A_i [P_{n-t_i}]_V / [P_n]_V & \text{for } n \le V, \\ 0 & \text{for } n > V. \end{cases} \tag{6}$$

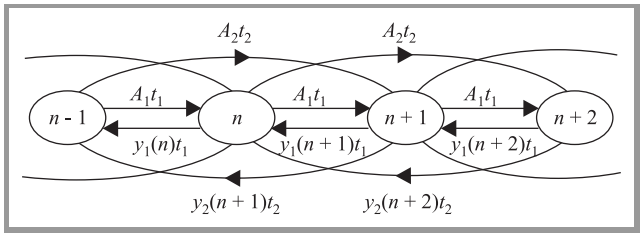


Fig. 1. Fragment of a diagram of the one-dimensional Markov chain in a multi-rate system ($M = 2$, $t_1 = 1$, $t_2 = 2$).

The reverse transition rate determines the average number of class i calls serviced in the state n.

3. The FAG with Compression

The following section recall the basics assumptions of the model of a full-availability group with traffic compression which was firstly described in [11].

Let as assume now that a full-availability group services a mixture of different multi-rate traffic streams with the compression property. This means that in the traffic mixture there are such calls in which a change in demands (requirements) follows evenly as the result of the overload of the system.

In this group it is assumed that the system services simultaneously a mixture of different multi-rate traffic classes, while these classes are divided into two sets: classes whose calls can change requirements (demands) while being serviced and classes that do not change their demands in their service time.

In the considered model the following notation is used:

- $\mathbb{M}_k$ denotes a set of classes with the possibility of compression, while $M_k = |\mathbb{M}_k|$ is the number of compressed traffic classes.
- $\mathbb{M}_{nk}$ is a set of classes without compression, and $M_{nk} = |\mathbb{M}_{nk}|$ denotes the number of classes without compression.

It was assumed in the model that all classes undergoing compression were compressed in the same degree. The measure of a possible change of requirements is *maximum compression coefficient* that determines the ratio of the maximum demands to minimum demands for a given traffic classes. The coefficient $K_{\max}$ can be determined on the basis of the dependence:

$$\forall_{j \in \mathbb{M}_k} \; K_{\max} = \frac{t_{j,\max}}{t_{j,\min}}, \tag{7}$$

where $t_{j,\max}$ and $t_{j,\min}$ denote, respectively, the maximum and the minimum number of basic bandwidth units demanded by a call of class j.

We assume that the system will be treated as a full-availability group with multi-rate traffic. The occupancy distribution in such a system can be expressed by the recursive Kaufman-Roberts formula (3), under the assumption that the amount of required resources by calls of the classes with compression property is minimum. In the case of a system carrying a mixture of traffic streams that undergo and do not undergo compression, the occupancy distribution (3) will be more conveniently expressed by dividing the two types of traffic[1]:

$$n [P_n]_V = \sum_{i=1}^{M_{nk}} A_i t_i [P_{n-t_i}]_V + \sum_{j=1}^{M_k} A_j t_{j,\min} \left[P_{n-t_{j,\min}} \right]_V, \tag{8}$$

where $t_{j,\min}$ is the minimum number of demanded BBUs in a given occupation state of the system by a call of class j that belongs to the set $\mathbb{M}_k$.

The blocking (loss) coefficient in the full-availability group will be determined by the dependence (6) that, in the considered case, will take on the following form:

$$E_i = B_i = \begin{cases} \sum\limits_{n=V-t_i+1}^{V} [P_n]_V & \text{for} \quad i \in \mathbb{M}_{nk}, \\ \sum\limits_{n=V-t_{i,\min}+1}^{V} [P_n]_V & \text{for} \quad i \in \mathbb{M}_k. \end{cases} \tag{9}$$

In equations (8) and (9), the model is characterized by the parameter $t_{i,\min}$ which is the minimum number of BBUs demanded by a call of class i under the conditions of maximum compression. Such an approach is indispensable to determine the blocking probabilities in the system with compression, since blocking states will occur in the conditions of maximum compression. The maximum compression determines such occupancy states of the system in which further decrease in the demands of calls of class i is not possible.

In order to determine a possibility of the compression of the system it is necessary to evaluate the number and the kind of calls serviced in a given occupancy state of the system. For this purpose we can use formula (5) that makes it possible to determine the average number of calls of class i serviced in the occupancy state n BBUs. This dependence,

[1]Further on in the paper, the terms "a set of classes with the possibility of compression" and "class with the possibility of compression", will be simplified to a "a set of classes with compression" and, respectively, a "class with compression".

under the assumption of the maximum compression, can be written in the following way:

$$y_i(n) = \begin{cases} \frac{A_i\left[P_{n-t_i}\right]_V}{[P_n]_V} & \text{for} \quad i \in \mathbb{M}_{nk}, \\ \frac{A_i\left[P_{n-t_{i,\min}}\right]_V}{[P_n]_V} & \text{for} \quad i \in \mathbb{M}_k. \end{cases} \tag{10}$$

On the basis of formula (10), knowing the demands of individual calls, we can thus determine the total average carried traffic in state n, under the assumption of the maximum compression:

$$\begin{aligned} Y_{\max}(n) &= Y^{nk}(n) + Y^k_{\max}(n) \\ &= \sum_{i=1}^{M_{nk}} y_i(n)t_i + \sum_{j=1}^{M_k} y_j(n)t_{j,\min}, \end{aligned} \tag{11}$$

where $Y^k_{\max}(n)$ is the average number of busy BBUs in state n occupied by calls that undergo compression, whereas $Y^{nk}(n)$ is the average number of busy BBUs in state n occupied by calls without compression.

Let us assume that the value of the parameter $Y^{nk}(n)$ refers to non-compressed traffic and is independent of the compression of the remaining calls. The real values of carried traffic, corresponding to state n (determined in the condition of maximum compression), will depend on the number of free BBUs in the system. We assume that the real system operates in such a way as to guarantee the maximum use of the resources, i.e., a call of a compressed class always tends to occupy free resources and decreases its maximum demands in a least possible way. Thus, the real traffic value $Y(n)$, carried in a system in a given state corresponding to state n (determined in maximum compression) can be expressed in the following way[2]:

$$Y(n) = Y^{nk}(n) + Y^k(n) = \sum_{i=1}^{M_{nk}} y_i(n)t_i + \sum_{j=1}^{M_k} y_j(n)t_j(n). \tag{12}$$

The parameter $t_j(n)$ in formula (12) determines the real value of a demand of class j in state n:

$$\forall_{j\in\mathbb{M}_k} \; t_{j,\min} < t_j(n) \le t_{j,\max}. \tag{13}$$

The measure of the degree of compression in state n is the compression coefficient $\xi_k(n)$, which can be expressed in the following way:

$$t_j(n) = t_{j,\min}\xi_k(n). \tag{14}$$

Taking into consideration (14), the average number of busy BBUs occupied by calls with compression can be written thus:

$$Y^k(n) = \sum_{j=1}^{M_k} y_j(n)t_j(n) = \xi_k(n)\sum_{j=1}^{M_k} y_j(n)t_{j,\min}. \tag{15}$$

[2]Further on in the description, to simplify the description, we will use the term "in state n" instead of the description "a given state n in maximum compression".

We assume in the considered model that the system operates in such a way that guarantees the maximum use of available resources and this means that calls that undergo compression will always tend to occupy free resources, decreasing their demands in the least possible way. The other parameter of the considered system, beside the blocking (loss) probability, is the average number of busy BBUs in the system occupied by calls with compression (formula (15)). In order to determine this parameter, the knowledge of the compression coefficient $\xi_k(n)$ is indispensable. This coefficient can also be defined as the ratio of potentially available resources for the service of calls with compression to the resources occupied by these calls in the state of maximum compression. Thus, we can write (Fig. 2):

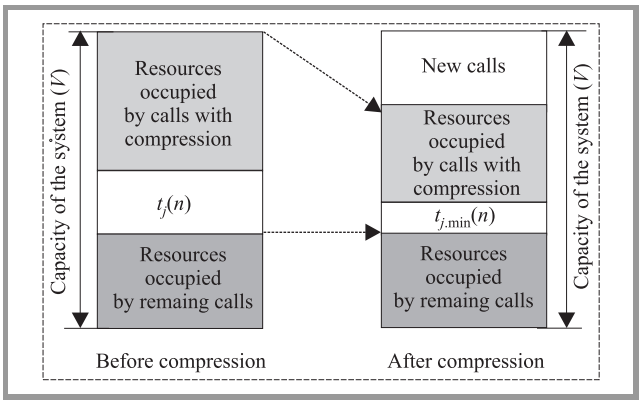


Fig. 2. Exemplary system with compression.

$$\xi_k(n) = \frac{V - Y^{nk}(n)}{Y^k_{\max}(n)} = \frac{V - Y^{nk}(n)}{n - Y^{nk}(n)}. \tag{16}$$

The numerator in formula (16) expresses the total amount of resources of the system which can be occupied by calls of the class with compression, whereas the denominator can be interpreted as the amount of resources which can be occupied by the calls of the class with compression, under the assumption that the system (FAG) is in the state n BBUs being busy. A constraint to the value of the coefficient (16) is the maximum compression coefficient determined on the basis of the dependence (7). This constraint can be taken into account by defining formally the compression coefficient in the following way:

$$\xi_k(n) = \begin{cases} K_{\max} & \text{for} \quad \xi_k(n) \ge K_{\max}, \\ \xi_k(n) & \text{for} \quad 1 \le \xi_k(n) < K_{\max}. \end{cases} \tag{17}$$

The compression coefficient determined by formula (17) is not dependent on the traffic class. This results from the adopted assumption in the model of the same degree of compression for all traffic classes that undergo the mechanism of compression.

Knowing the value of the compression coefficient in every state n, we can determine the average resources occupied by calls of class j with compression:

$$Y^k_j = \sum_{n=0}^{V} y_j(n)\left[\xi_k(n)t_{j,\min}\right][P_n]_V. \tag{18}$$

On the basis of the average resources occupied by calls of class j, we can determine the average resources occupied by calls of all traffic classes with compression:

$$Y^k = \sum_{j=0}^{M_k} Y_j^k. \tag{19}$$

Let us note that the value Y^k in formula (19) is the average carried traffic in the system by calls which undergo compression.

4. Application of the Model

4.1. UMTS Architecture

Let us consider the structure of the UMTS network illustrated in Fig. 3. The presented network consists of three functional blocks designated, respectively: UE (user equipment), UTRAN (UMTS terrestrial radio access network) and CN (core network). The following notation has been adopted in Fig. 3: RNC is the radio network controller, WCDMA is a radio interface and Iub is the interface connecting node B and RNC. In the dimensioning process

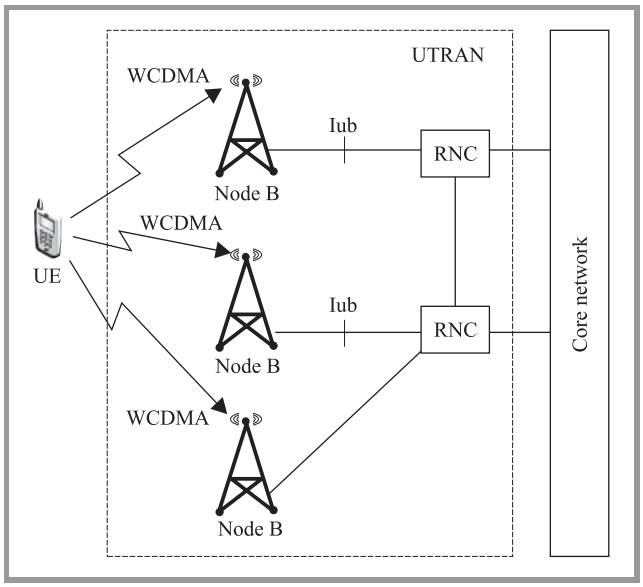


Fig. 3. Elements of the UMTS network structure.

for the UMTS network, an appropriate dimensioning of the connections in the access part (UTRAN), i.e., the radio interface between the user and node B, and the Iub connections between node B and the radio network controller, has a particular significance. The issues pertaining to Iub interface dimensioning are already discussed in the subject literature, for example in the earlier work of the present authors, e.g., [19], also models dedicated for radio interface dimensioning are widely discussed in the subject literature, for example in earlier works of the present authors, e.g., [11], [20]–[23], whereas those dealing with dimensioning of the WCDMA interface with the packet scheduling functionality have not been hitherto addressed in a satisfactory way.

4.2. Packet Scheduling

Packet scheduling is an important mechanism that should be included in the analysis of the efficiency of the WCDMA radio interface in the UMTS networks. In the relevant literature we can consider user-specific and cell-specific packet scheduling algorithms [14].

In user-specific packet scheduling, scheduler controls the use of transport channels and their bit rate depending on the volume of traffic, informing of a demand for packet bearers with appropriate bit rates.

Cell-specific scheduler is responsible for appropriate distribution of the capacity of the base station between users of non-real time services (i.e., background, interactive and streaming). Bit rates assigned to users are controlled every 100 ms – 1 s and if the load approaches the target load level, the scheduler can reduce the load by decreasing bit rates of the packet bearer. The change in the capacity for scheduled connections in relation to the resources assigned for non-scheduled connections is presented in Fig. 4.

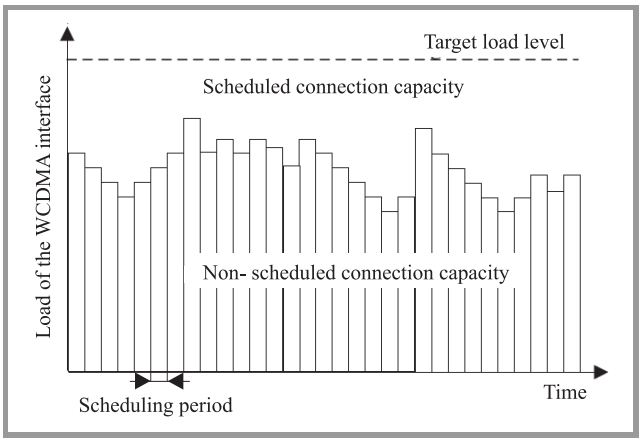


Fig. 4. Illustration of cell-specific packet scheduler.

An example of the operation of the algorithm is presented in Fig. 5, where the calls of non-real time connections (based on the user-specific scheduler) are admitted until target load level is reached and then, in the case of a con-

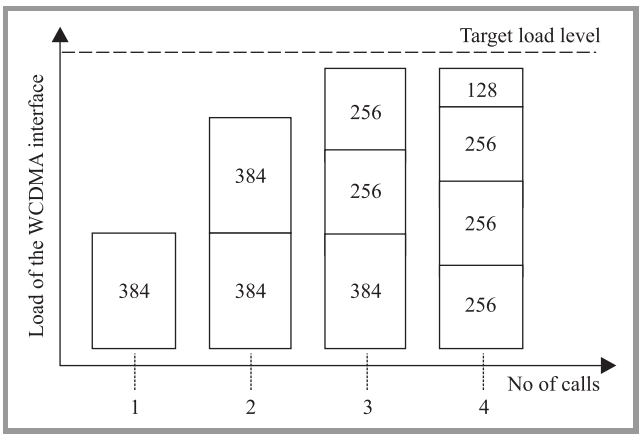


Fig. 5. An example of operation of the packet scheduler.

tinuing arrival process of new calls, connections are compressed.

Figure 5 shows relation between the load of the interface and the number of serviced calls. In Fig. 5 the first arriving call required 384 kbit/s and it was admitted for service, the second arriving call which required 384 kbit/s was also admitted for service. When the third call arrived it also required 384 kbit/s and was not admitted, but the compression mechanism of one of already admitted calls to 256 kbit/s was applied and the call is assigned the resources of 256 kbit/s. The last forth arriving call required 384 kbit/s and it was not admitted, but the reconfiguration of the resources ensues (as it is presented in the figure).

With compression mechanisms, one of the ways of the bit rate analysis is to base the evaluation on the average bit rate. The WCDMA interface with packet scheduling can be treated as the full-availability group with multi-rate traffic and compression property (Section 3).

4.3. Calculation Algorithm

On the basis of the considerations presented in Sections 2 and 3, the algorithm of blocking probability E_i and average occupied traffic Y_i^k calculations for the WCDMA interface may be written in the form of Algorithm 1.

Algorithm 1: Algorithm of blocking probabilities calculation in the downlink direction

1. Calculation of offered traffic load A_i of class i Eq. (1).
2. Designation of the value of t_{BBU} as the greatest common divisor Eq. (2).
3. Determination of the value of t_i as the integer number of demanded resources by class i calls Eq. (4).
4. Calculation of state probabilities $[P_n]_V$ in FAG Eq. (8).
5. Designation of the blocking probability of class i Eq. (9).
6. Determination of the reverse transition rate for class i Eq. (10).
7. Calculation of the average compression coefficient Eq. (17).
8. Determination of the average traffic of class i carried by WCDMA Eq. (18).

4.4. Numerical Study

The proposed analytical model of the WCDMA interface is an approximate one. Thus, the results of the analytical calculations of the WCDMA interface were compared with the results of the simulation experiments. The study was carried out for users demanding a set of following services in the downlink direction:

- Class 1: speech – $t_{1,\min} = 12$ kbit/s $= 12$ BBUs.
- Class 2: video – $t_{2,\min} = 64$ kbit/s $= 64$ BBUs.
- Cass 3: data 384/384 – $t_{3,\min}$ = 128 kbit/s = 128 BBUs (non-real time service).

In the presented study, it was assumed that:

- The hard capacity of the WCDMA interface in the downlink direction [13], [23].
- R_{BBU} was equal to 1 kbit/s.
- The coefficient $K_{\max}$ was equal to 3.
- The capacity of the WCDMA interface was limited to 80% of the physical capacity: $V_{DL} =$ 1600 kbit/s $=$ 1600 BBUs.
- The services were offered in the following proportions:

$$A_1 t_1 : A_2 t_2 : A_3 t_3 = 15 : 5 : 40.$$

It was assumed that the main part of traffic is generated by data service followed by speech service, while the smallest part of traffic comes from video service.

Figure 6 shows the dependency of the blocking probability in relation to traffic offered per BBU in the WCDMA interface. The presented results were obtained for the minimum value of required (demanded) resources for traffic classes with the compression property.

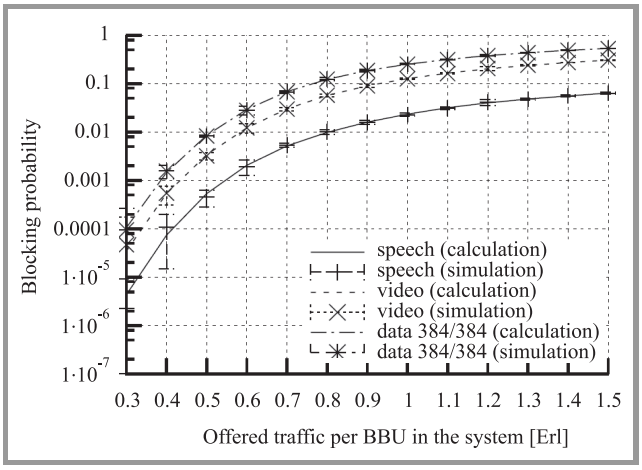


Fig. 6. Blocking probabilities for all traffic classes carried by the WCDMA interface.

Figures 8 and 7 present the influence of traffic offered per BBU in the WCDMA interface on the average carried traffic by WCDMA (Fig. 7), and on the value of the comp-

ression coefficient (Fig. 8). It can be noticed that the exponential dependence characterizes the plots corresponding to the traffic class with compression in both figures. The linear relation between compression coefficient and the average carried traffic (see Eq. (18)) explains the similar character of the curves in the both figures. The results con-

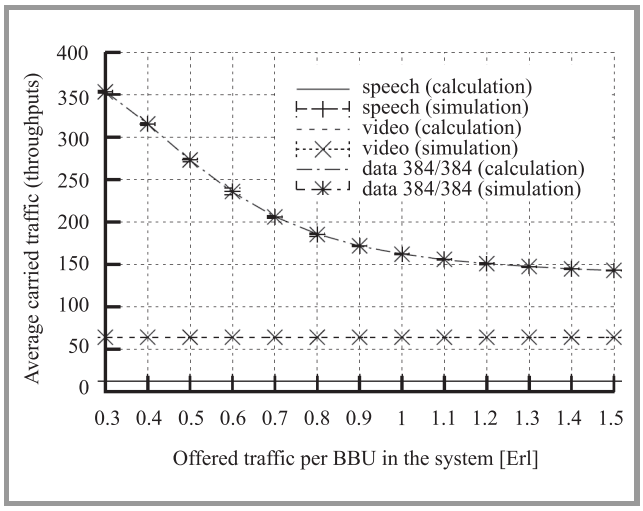


Fig. 7. Average carried traffic for particular classes serviced by the WCDMA interface.

firm strong dependence between the average carried traffic (throughput) and the load of the system – the more overloaded system the lower value of throughput. The character of the curves results from the decrease of the amount of resources required by a call of class with compression: the more overloaded system the smaller demands of the calls with compression.

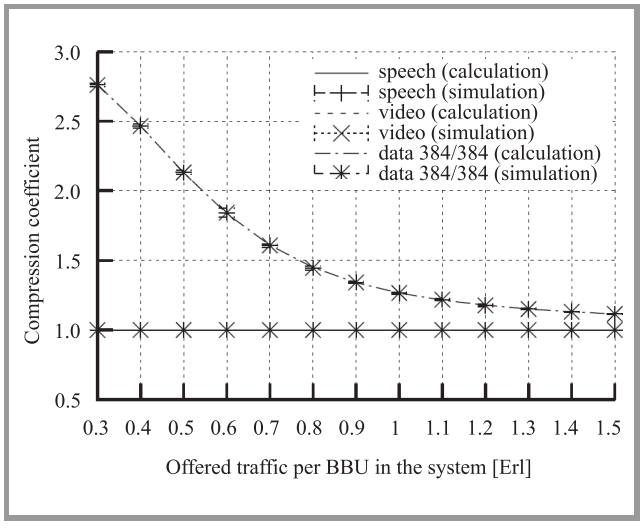


Fig. 8. Compression coefficient in relation to traffic offered to the WCDMA interface.

The results of the simulations are shown in the charts in the form of marks with 95% confidence intervals calculated after the *t*-student distribution. 95% confidence intervals of the simulation are almost included within the marks plotted in the figures.

5. Conclusions

This paper proposes a new analytical model with compression that finds its application in modeling the WCDMA interface with packet scheduler, in the UMTS network, carrying a mixture of different multi-rate traffic classes.

The presented analytical method allows to determine the blocking probability for all traffic classes serviced by the WCDMA interface. It should be noted that in the model we assume the "worst case" approach in the WCDMA modeling and dimensioning that makes our calculations independent of the way of operation of the scheduler [24], which underlines the universal character of the method.

It is worth emphasizing that the described analytical model could be used for a determination of the average carried traffic for particular traffic classes serviced by the WCDMA interface.

The KPI, being an indispensable element of SLA, can be defined differently depending on the kind of the receiver of information. Thus, KPI will be defined differently for engineering staff and differently for non-technical staff often involved in decision making concerning expenditures that are to ensure appropriate quality of services. While such parameter as the blocking probability is well understood by engineers, clients and non-technical staff may have some problems with the interpretation of the indicator and this group of users will rather prefer the average value as being more intuitive.

The average value of carried traffic is also very characteristic for some services such as data (e.g., file transfer protocol – FTP). With regards to the above factors, a necessity appears of a skilful use of the average value of carried traffic as the initial value in the process of designing and dimensioning of the UMTS networks without violating the basic merits of the adopted model that are necessary for a system to operance successfully. Thus, this parameter is an important factor in 3G network capacity calculations, i.e., in dimensioning and optimization of WCDMA and Iub interfaces.

References

[1] D. Staehle and A. Mäder, "An analytic approximation of the uplink capacity in a UMTS network with heterogeneous traffic", in *18th Int. Teletraf. Congr. ITC 18*, Berlin, Germany, 2003, pp. 81–91.

[2] V. B. Iversen and E. Epifania, "Teletraffic engineering of multi-band W-CDMA systems", in *Network Control and Engineering for Qos, Security and Mobility II*. Norwell: Kluwer, 2003, pp. 90–103.

[3] V. B. Iversen, V. Benetis, and H. N. Trung, "Evaluation of multi-service CDMA networks with soft blocking", in *ITC 16th Spec. Sem. Perfor. Eval. Mob. Wirel. Syst.*, Antwerp, Belgium, 2004, pp. 212–216.

[4] D. Staehle, "Analytic methods for UMTS radio network planning". Ph.D. thesis, Bayerische Julius-Maximilians-Universitat Wurzburg, 2004.

[5] K. Subramaniam and A. A. Nilsson, "Tier-based analytical model for adaptive call admission control scheme in a UMTS-WCDMA system", in *Proc. Veh. Technol. Conf.*, Stockolm, Sweden, 2005, vol. 4, pp. 2181–2185.

[6] K. Subramaniam and A. A. Nilsson, "An analytical model for adaptive call admission control scheme in a heterogeneous UMTS-WCDMA system", in *Proc. Int. Conf. Commun.*, Seoul, Korea, 2005, vol. 5, pp. 3334–3338.

[7] I. Koo and K. Kim, *CDMA Systems Capacity Engineering. Mobile Communications*. Boston, London: Artech House, 2005.

[8] I. Koo and K. Kim, "Erlang capacity of multi-service multi-access systems with a limited number of channel elements according to separate and common operations", *IEICE Trans. Commun.*, vol. E89-B, no. 11, pp. 3065–3074, 2006.

[9] G. A. Kallos, V. G. Vassilakis, and M. D. Logothetis, "Call blocking probabilities in a W-CDMA cell with fixed number of channels and finite number of traffic sources", in *Proc. 6th Int. Conf. Commun. Syst., Netw. Dig. Sig. Proces.*, Graz, Austria, 2008 pp. 200–203.

[10] V. G. Vassilakis and M. D. Logothetis, "The wireless Engset multirate loss model for the handofftraffic analysis in W-CDMA networks", in *Proc. 19th Int. Symp. Pers., Indoor Mob. Radio Commun.*, Cannes, France, 2008, pp. 1–6.

[11] M. Stasiak, P. Zwierzykowski, J. Wiewióra, and D. Parniewicz, "An approximate model of the WCDMA interface servicing a mixture of multi-rate traffic streams with priorities", in *Computer Performance Engineering*, Lecture Notes in Computer Science, vol. 5261. Berlin: Springer, 2008, pp. 168–180.

[12] J. Laiho, A. Wacker, and T. Novosad, *Radio Network Planning and Optimization for UMTS*. 2 ed. Chichester: Wiley, 2006.

[13] M. Stasiak, P. Zwierzykowski, and M. Głąbowski, *Modelowanie i wymiarowanie ruchomych sieci bezprzewodowych*. Warsaw: Wydawnictwa Komunikacji i Łączności, 2009 (in Polish).

[14] H. Holma and A. Toskala, *WCDMA for UMTS. Radio Access for Third Generation Mobile Communications*. Chichester: Wiley, 2000.

[15] *Broadband Network Teletraffic, Final Report of Action COST 242*, J. W. Roberts, V. Mocci, and I. Virtamo, Eds. Commission of the European Communities. Berlin: Springer, 1996.

[16] J. S. Kaufman, "Blocking in a shared resource environment", *IEEE Trans. Commun.*, vol. 29, no. 10, pp. 1474–1481, 1981.

[17] J. W. Roberts, "A service system with heterogeneous user requirements – application to multi-service telecommunications systems", in *Proceedings of Performance of Data Communications Systems and Their Applications*, G. Pujolle, Ed. Amsterdam: North Holland, 1981, pp. 423–431.

[18] M. Głąbowski, "Modelling of state-dependent multi-rate systems carrying BPP traffic", *Ann. Telecommun.*, vol. 63, no. 7–8, pp. 393–407, 2008.

[19] M. Stasiak, J. Wiewióra, and P. Zwierzykowski, "Analytical modelling of the Iub interface in the UMTS network", in *Proc. 6th Int. Conf. Commun. Syst., Netw. Dig. Sig. Proces.*, Graz, Austria, 2008.

[20] M. Stasiak, A. Wiśniewski, and P. Zwierzykowski, "Blocking probability calculation in the uplink direction for cellular systems with WCDMA radio interface", in *Proc. 3rd Polish-German Teletraf. Symp.*, P. Buchholtz, R. Lehnert, and M. Pióro, Eds., Dresden, Germany, 2004, pp. 65–74.

[21] M. Głąbowski, M. Stasiak, A. Wiśniewski, and P. Zwierzykowski, "Uplink blocking probability calculation for cellular systems with WCDMA radio interface, finite source population and differently loaded neighbouring cells", in *Proc. Asia-Pacific Conf. Commun.*, Perth, Australia, 2005.

[22] M. Głąbowski, M. Stasiak, A. Wiśniewski, and P. Zwierzykowski, "Uplink blocking probability calculation for cellular systems with WCDMA radio interface and finite source population", in *Performance Modelling and Analysis of Heterogeneous Networks*. Wharton: River Publ., 2009, pp. 301–318.

[23] M. Stasiak, A. Wiśniewski, P. Zwierzykowski, and M. Głąbowski, "Blocking probability calculation for cellular systems with WCDMA radio interface servicing PCT1 and PCT2 multirate traffic", *IEICE Trans. Commun.*, vol. E92-B, no. 4, pp. 1156–1165, 2009.

[24] H. Holma and A. Toskala, *HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications*. Chichester: Wiley, 2006.

Maciej Stasiak is a Professor at the Poznań University of Technology, Poland. He received the M.Sc. and Ph.D. degrees in electrical engineering from the Institute of Communications Engineering, Moscow, Russia, in 1979 and 1984, respectively. In 1996 he received D.Sc. degree from the Poznań University of Technology in electrical engineering. In 2006 he was nominated as Full Professor. Between 1983–1992 he worked in Polish industry as a designer of electronic and microprocessor systems. In 1992, he joined the Institute of Electronics and Telecommunications, Poznań University of Technology, where he is currently a Head of the Chair of Communications and Computer Networks at the Faculty of Electronics and Telecommunications at the Poznań University of Technology. He is the author or co-author of more than 200 scientific papers and three books. He is engaged in research and teaching in the area of performance analysis and modeling of multiservice networks and switching systems.
e-mail: stasiak@et.put.poznan.pl
Chair of Communications and Computer Networks
Poznań University of Technology
Polanka st 3/228
60-965 Poznań, Poland

Piotr Zwierzykowski received the M.Sc. and Ph.D. degrees in telecommunication from the Poznań University of Technology, Poland, in 1995 and 2002, respectively. Since 1995 he has been working in the Faculty of Electronics and Telecommunications, Poznań University of Technology. He is currently an Assistant Professor in the Chair of Communications and Computer Networks. He is the author or co-author over 120 papers. He is engaged in research and teaching in the area of performance analysis and modeling of multiservice switching systems.
e-mail: pzwierz@et.put.poznan.pl
Chair of Communications and Computer Networks
Poznań University of Technology
Polanka st 3/231
60-965 Poznań, Poland

Janusz Wiewióra received the M.Sc. degree in telecommunication from Warsaw University of Technology, Poland, in 2000. Between 1998–2000 he worked as an expert responsible for international frequency coordination for digital radio (DAB), medium wave band radio and long wave band radio in National Radiocommunication Agency. He is an epxert leading the BSS Optimization Section at the Polska Telefonia Cyfrowa sp. z o.o. (PTC) responsible for dimensioning and optimization the dual band 900/1800 GSM network and UMTS network. During his work at PTC he has kept in touch with Poznań University of Technology, where he was engaged in research in the area of performance analysis and modeling of multi-rate mobile networks. Until now he has published over 10 papers.

e-mail: jwiewiora@era.pl
BSS Optimization Section
Polska Telefonia Cyfrowa Sp. z o.o.
Regional Office Warsaw
Annopol st 3
03-236 Warsaw, Poland

Paper

Perspective for Using the Optical Frequency Standards in Realization of the Second

Karol Radecki

Abstract—The second is currently defined by the microwave transition in cesium atoms. Optical clocks offer the prospects of stabilities and reproducibilities that exceed those of cesium. This paper reviews the progress in frequency standards based on optical transitions, recommended by International Committee for Weights and Measures, as a secondary representation of the second. The operation of these standards is briefly described and factors affecting stability and accuracy of these and some new optical clocks are discussed.

Keywords—atomic clocks, atomic time scale, optical frequency standards.

1. Introduction

The best realization of the SI second today is served by cesium fountain primary frequency standards. The frequency accuracy of atomic time scale TAI realized by these standards is less than 10^{-15} [1]–[3].

Commercial cesium clocks installed in time laboratories realize the second with accuracy and long term stability of 10^{-14}. In laboratories, the active hydrogen masers are also installed, with short term instability better then 10^{-15}. Commercial cesium and hydrogen masers standards contribute to the reliability and frequency stability of the atomic time scale, but they do not contribute to the realization of the second.

Over the past decade metrologists at various time and frequency standards laboratories have investigate the so-called forbidden optical transitions in cold trapped atoms and single ions. As clock transitions they have two major advantages: their frequencies are five orders of magnitude higher than the cesium frequency and natural linewidths are in the region of 1 Hz. This leads to high quality factors of these lines. However, the observed linewidths are larger, in the range up to few hundred Hz. Because the instability of optical clock is inversely proportional to the quality factor of the observed spectral line, it could be possible to achieve the short term stability of a few orders of magnitude better, assuming the number of atoms and transition interrogation time is the same.

Optical frequencies can be measured precisely by the femtosecond comb [4] and compared to cesium frequency with high accuracy.

Optical clocks offer the prospects of stabilities and reproducibilities that exceeds those of cesium. Today some optical clocks, based on $^{88}Sr^{+}$, $^{199}Hg^{+}$, $^{171}Yb^{+}$, may be used as a secondary representation of the second [5], [6]. Recently, two optical clocks, based on $^{27}Al^{+}$ ions and neutral ^{87}Sr atoms, demonstrated systematic uncertainties which significantly exceed the current best evaluations of cesium primary standards. The progress in optical clocks is so rapid that in the near future the redefinition of the second will be most probably required.

2. Requirements for Optical Clock Transition

The main requirements for optical clock transition are a narrow natural line (linewidth less than 1 Hz) and the ability of their observation with the highest possible resolution. Transition frequency should also be unaffected by external electric and magnetic fields.

The clock transitions observed in number of laboratories worldwide are the weak, forbidden optical transitions in a single cold ion or cold atoms cloud.

In 2006 the International Committee for Weights and Measures (CIPM) recommended four optical transitions, which may be used as secondary representation of the second (Table 1) [5].

Table 1
Recommended optical clock transitions (2006)

Atom/ ion	Transition	Frequency of transition/ uncertainty
^{87}Sr	$5s^2\,{}^1S_0 - 5s5p\,{}^3P_0$	429 228 004 229 877 Hz/ $1.5 \cdot 10^{-14}$
$^{88}Sr^{+}$	$5s\,{}^2S_{1/2} - 4d\,{}^2D_{5/2}$	444 779 044 095 484 Hz/ $7 \cdot 10^{-15}$
$^{171}Yb^{+}$	$6s\,{}^2S_{1/2}(F=0) -$ $5d\,{}^2D_{3/2}(F=2)$	688 358 979 309 308 Hz/ $9 \cdot 10^{-15}$
$^{191}Hg^{+}$	$5d^{10}6s\,{}^2S_{1/2}(F=0) -$ $5d^9 6s^2\,{}^2D_{5/2}(F=2)$	1 064 721 609 899 145 Hz/ $3 \cdot 10^{-15}$

The CIPM has established the Working Group to review and discuss the uncertainty budget for possible optical candidates. It is required, that the selected frequency must have evaluated and documented uncertainty to the same level as it is required for primary standards contributing to international atomic time. In addition it is required, that

this uncertainty should be not worse than 10 times value that is for the best primary frequency standard.
Table 1 gives the values of recommended by CPIM 2006 unperturbed ground-state hyperfine the frequency transitions and estimated relative standard uncertainties. At present, due to progress in the optical clocks and the measurements systems, these parameters are evaluated more accurately.
The instability of the frequency standard that is operated in the interrogation cycles of duration T can be written as [7], [8]:

$$\sigma_y(\tau) \approx \frac{\mathrm{C}}{SNR \cdot Q}\sqrt{\frac{T}{N\tau}}, \tag{1}$$

where: C is the constant that depends on the interrogation scheme, Q is the resonance quality factor $Q = f_0/\Delta f$, Δf is the linewidth of resonance line centered at frequency f_0, SNR is signal to noise ratio ($SNR \approx 1$ if limited by quantum projection noise), T is the interrogation time (it should not be significantly larger than $1/\Delta f$ because of the stability degradation), N is the total number of atoms/ions.
If we assume quantum limited operation of the $^{199}Hg^+$ clock, $\Delta f = 10$ Hz, $N = 1$ ion and Rabi excitation pulse of $T = 100$ ms, then the expected instability is $\sigma_y(\tau) \approx 3 \cdot 10^{-15}\tau^{-1/2}$. Similarly for the ^{87}Sr optical lattice clock and $N = 10^4$ atoms, the instability is about $\sigma_y(\tau) \approx 7 \cdot 10^{-17}\tau^{-1/2}$. For comparison the instability of ^{133}Cs fountain clock with $\Delta f = 1$ Hz, $T = 1$ s and $N = 10^6$ atoms is expected to be $\sigma_y(\tau) \approx 5 \cdot 10^{-14}\tau^{-1/2}$.

3. Look into Possible Optical Time and Frequency Standards

The main requirement for optical frequency standard is the need for highly stable laser which is disciplined by the clock transition in the trapped cold ion or neutral atoms. This is the so-called forbidden transition with natural linewidth of 1 Hz or less. The wideband femtosecond comb [4] is applied for precise comparison of optical frequency of the resonance line with cesium microwave frequency.
Ion or atomic trap works in cycles. The measurement cycle comprises laser cooling, state preparation, excitation of the clock transition and detection. Clock transitions are excited in a weak external magnetic field using Rabi or Ramsey pulse technique.

3.1. Optical Clocks Based on Single Ions

Recommended by CIPM 2006 clock transitions are: the $5s\ ^2S_{1/2} \leftrightarrow 4d\ ^2D_{5/2}$ in $^{88}Sr^+$, the $5d^{10}6s\ ^2S_{1/2}(F = 0, m_F = 0) \leftrightarrow 5d^9 6s^2\ ^2D_{5/2}(F = 2, m_F = 0)$ in $^{199}Hg^+$ and the $6s\ ^2S_{1/2}(F = 0, m_F = 0) \leftrightarrow 5d\ ^2D_{3/2}(F = 2, m_F = 0)$ in $^{171}Yb^+$. There are quadrupole transitions with natural linewidths of 0.4 Hz, 1.1 Hz and 3.1 Hz, respectively.
Ions are confined in RF traps and laser cooled to the so-called Lamb-Dicke limit. This greatly reduces the Doppler broadening and frequency shift associated with ions motion relative to the excitation clock radiation.
Partial energy levels schemes of $^{199}Hg^+$ and $^{171}Yb^+$, are very similar (Figs. 1 and 2). Clock transitions are excited in a small external magnetic field ($\sim 1\ \mu$T) between $m_F = 0$ sublevels with no first order Zeeman shift.

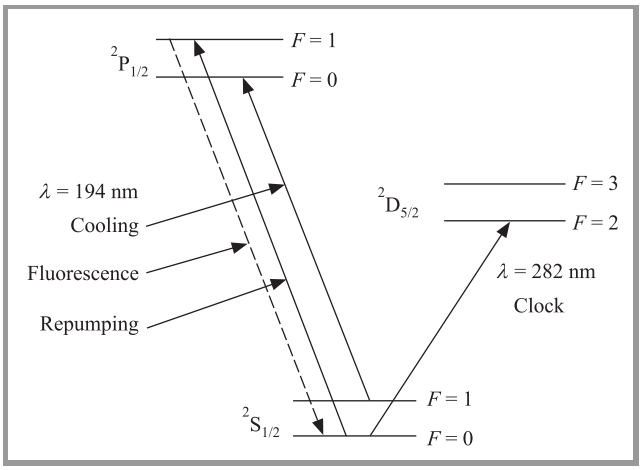


Fig. 1. Partial energy levels scheme of $^{199}Hg^+$.

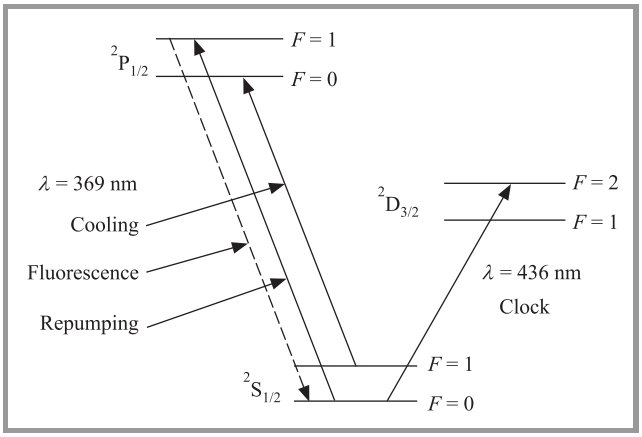


Fig. 2. Partial energy levels scheme of $^{171}Yb^+$.

A single trapped $^{199}Hg^+$ ion is laser cooling using $\lambda = 194$ nm. After cooling, ion is prepared in $^2S_{1/2}(F = 0, m_F = 0)$. To probe the transition, the $\lambda = 282$ nm laser radiation is used. The clock transition is observed on $\lambda = 194$ nm using quantum jumps technique. The line shape is measured from the statistics of many quantum jumps during discrete laser frequency sweeping across the clock transition. The measured linewidth of 6.5 Hz has been demonstrated with Rabi excitation pulse 120 ms long [9], [10].
Frequency instability of $5 \cdot 10^{-15}$ at $\tau = 1$ s was measured. In comparison with the $^{27}Al^+$ standard, the systematic fractional uncertainty of Hg^+ standard was estimated to be less than $3 \cdot 10^{-17}$ [11]. The systematic uncertainty through comparison with Cs NIST-F1 frequency standard is estimated at $7 \cdot 10^{-16}$ [12]. Optical cryogenic clock based

on $^{199}Hg^+$ are now investigated at National Institute of Standard and Technology (NIST), USA.

In the $^{171}Yb^+$ clock, the $\lambda = 369$ nm laser radiation with repumper sideband is used for cooling ion (Fig. 2). After cooling, the ion is prepared in $^2S_{1/2}(F = 0, m_F = 0)$. To probe the transition, the $\lambda = 436$ nm laser radiation is used. The clock transition is observed on $\lambda = 369$ nm using quantum jumps technique.

The measured linewidth of 30 Hz has been demonstrated with Rabi excitation pulse 30 ms long. The systematic uncertainty through comparison with PTB Cs standards is $1.5 \cdot 10^{-15}$ [13]. Optical clocks based on $^{171}Yb^+$ are investigated at Physikalisch Technische Bundesanstalt, Germany (PTB) and British National Physics Laboratory (NPL).

The partial energy levels scheme for the $^{88}Sr^+$ ion is shown in Fig. 3. The ion is laser cooled using radiations at both $\lambda = 422$ nm and $\lambda = 1092$ nm.

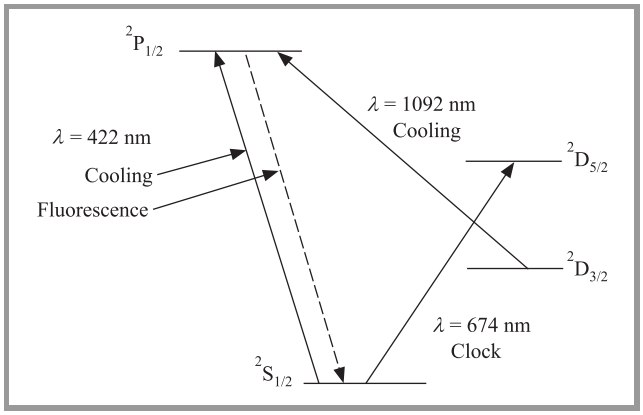


Fig. 3. Partial energy levels scheme of $^{88}Sr^+$.

In contrast to $^{199}Hg^+$ and $^{171}Yb^+$ ions the $^{88}Sr^+$ ion has the linear Zeeman sensitivity to magnetic external field. This field split the clock transition into five pairs of Zeeman components, symmetrically located about the line centre. By probing one pair of components, cancellation of linear shift is achieved. The clock operates by stabilizing the interrogation laser to the mean transition frequency of the pair $m_J = \pm 1/2$, $\Delta m_J = 0$ clock transitions.

The resonance linewidth is 9 Hz with Rabi excitation pulse 100 ms long [14]. The fractional systematic uncertainty through comparison with NPL primary Cs standard is estimated at $3 \cdot 10^{-15}$ [15].

Optical standards with $^{88}Sr^+$ are investigated at the NPL and National Research Council (NRC), Canada.

Recently, the optical clock transition $^1S_0 - {}^3P_0$ (with natural linewidth of 8 mHz) has been observed in $^{27}Al^+$ ion, which cannot be directly laser cooled ($\lambda = 167$ nm). The group at NIST [16] solved that problem by using sympathetic laser cooling of $^{27}Al^+$ through the $^9Be^+$ ion medium. Both ions are coupled together in the ion trap (by Coulomb interaction) and can be cooled using $\lambda = 313$ nm radiation in $^9Be^+$. The $^{27}Al^+$ ion is probed at $\lambda = 267.4$ nm clock transition.

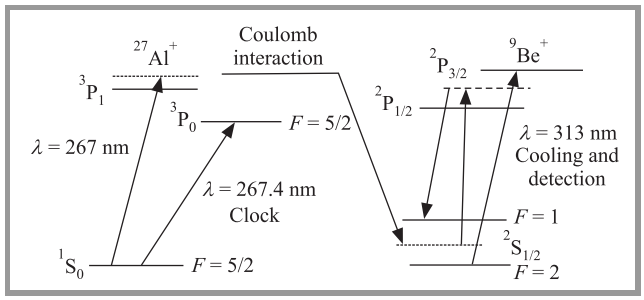


Fig. 4. Transfer of the $^{27}Al^+$ clock state to detectable states in $^9Be^+$ [16].

Clock transition information is sent to $^9Be^+$ using quantum logic technique (Fig. 4). Pulse sequence maps the $^{27}Al^+$ clock state $^1S_0$ to detectable states ($^2S_{1/2}\, F = 1$) in logic $^9Be^+$ ion through the ions motional state, using $^1S_0 - {}^3P_1$ vibrational excitation $\lambda = 267$ nm and Raman transition $\lambda = 313$ nm. Fluorescence photons on $\lambda = 313$ nm are counted if $^{27}Al^+$ ion is in $^3P_0$ state. The clock operates by stabilizing the interrogation laser to the mean transition frequency of the pair $m_F = \pm 5/2$, $\Delta m_F = 0$ clock transitions. The systematic uncertainty through the comparison with $^{199}Hg^+$ frequency standard is estimated at $2.3 \cdot 10^{-17}$ [11].

3.2. Optical Clocks Based on Neutral Atoms

Recommended clock transition $5s^2\ {}^1S_0 - 5s5p\ {}^3P_0$ ($\lambda = 698$ nm) in neutral ^{87}Sr atoms has the natural linewidth of 1 mHz. Partial energy levels scheme for ^{87}Sr atom are shown in Fig. 5.

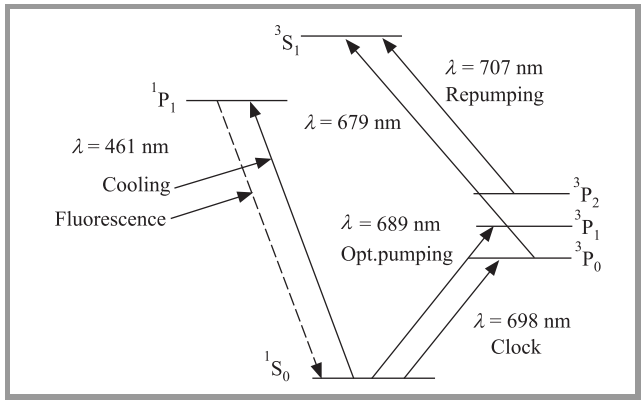


Fig. 5. Partial energy levels scheme of ^{87}Sr.

Neutral atoms are trapped and cooled in magneto optical trap (MOT) operated on $\lambda = 461$ nm transition. Two repumping lasers ($\lambda = 707$ nm and $\lambda = 679$ nm) are used to prevent atom loss into the $^3P_2$ state. In a second stage of MOT, the atoms are cooled on $\lambda = 689$ nm transition to a final temperature of 2.5 μK. After cooling the atoms are loaded into optical lattice trap. Optical lattice greatly reduce the motional effects of atoms and allow for extension interrogation times of probing laser [17].

The clock operates at two transitions $^1S_0(F = 9/2,\ m_F = \pm 9/2) \leftrightarrow {}^3P_0(F = 9/2,\ m_F = \pm 9/2)$ excited on $\lambda = 698$ nm and observed by measuring fluorescence on

$\lambda = 461$ nm. The clock centre frequency is found by taking the average frequency of both transition peaks. The resonance linewidth of 10 Hz with Rabi excitation pulse 80 ms long was observed [18].

The fractional systematic uncertainty of ^{87}Sr clock through comparison with ^{40}Ca NIST clock and NIST H-maser was evaluated at $1.5 \cdot 10^{-16}$ [19], [20].

Optical standards with ^{87}Sr are investigated at National Institute of Standard and Technology (NIST) USA, Laboratoire National de Métrologie et d'Essais (LNE-SYRTE), Physikalisch Technische Bundesanstalt Germany (PTB), National Metrology Institute of Japan (NMIJ) and University of Tokyo.

Optical clocks based on neutral ^{40}Ca, ^{199}Hg and ^{171}Yb atoms are also developed [21]–[24]. In contrast to ^{87}Sr and ^{171}Yb neutral mercury has low sensitivity to black body radiation and has the potential to achieve uncertainty at 10^{-18} level [25].

3.3. Stability and Accuracy

Recently evaluated (2007/2008) systematic uncertainties and short term stabilities ($\tau = 100$ s) for the optical clocks recommended by CIPM are summarized in Table 2. In the single ion frequency standards a significant uncertainty can arise from uncancelled electric quadrupole shift and quadratic Zeeman effect.

Table 2
Systematic uncertainties and stabilities
for various optical clocks

Optical clocks	^{87}Sr/ ^{40}Ca	^{88}Sr$^+$/ ^{133}Cs	^{171}Yb$^+$/ ^{171}Yb$^+$	^{191}Hg$^+$/ ^{27}Al$^+$	^{27}Al$^+$/ ^{191}Hg$^+$
$\sigma_y(\tau)$ 100 s	$6 \cdot 10^{-15}$	$3 \cdot 10^{-15}$	10^{-15}	$4 \cdot 10^{-16}$	$4 \cdot 10^{-16}$
u_B	$1.5 \cdot 10^{-16}$	$3 \cdot 10^{-15}$	$1.5 \cdot 10^{-15}$	$1.9 \cdot 10^{-17}$	$2.3 \cdot 10^{-17}$

The ^{199}Hg$^+$ and ^{27}Al$^+$ clock frequencies were measured relatively each other, and to the NIST-F1 cesium fountain [11], [16]. In both ion standards inaccuracies at $2 - 3 \cdot 10^{-17}$ were evaluated. The dominant uncertainties in the ^{199}Hg$^+$ standard are due to the AC quadratic Zeeman effect and the magnetic field orientation, but in the ^{27}Al$^+$ the dominant components are due to the micromotion and secular 2nd order Doppler shifts. The black-body radiation shift for the ^{199}Hg$^+$ standard is negligible because the ion trap is operated at liquid helium temperature (4.2K). However, the black-body radiation shift for the ^{27}Al$^+$ standard is unusually small at the normal operating temperature ($\sim 10^{-17}$ at 300K).

Similarly low level uncertainty at 10^{-16} level was evaluated for the ^{87}Sr clock compared with ^{40}Ca optical clock [18], [20]. The dominant systematic uncertainty arose from lattice laser field, the room temperature black body radiation and interatomic collisions.

The ^{88}Sr$^+$ and ^{171}Yb$^+$ optical clocks have been evaluated in comparison with Cs primary atomic clock. Experiments which allow for the tests of frequency stability and evaluation of systematic frequency shifts by comparing two identical clocks are currently underway.

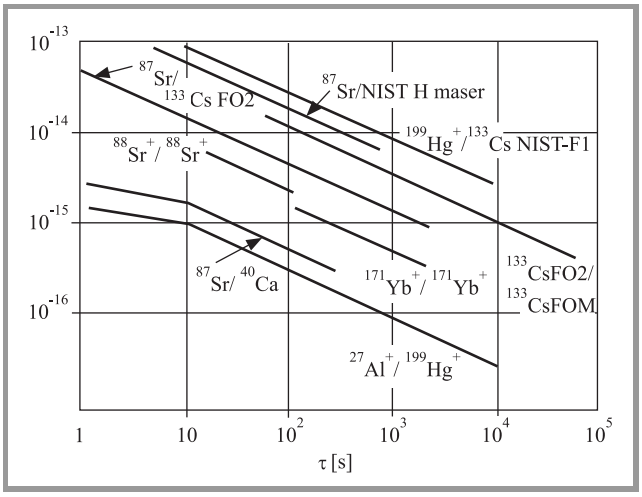


Fig. 6. Short term stability diagrams for optical clocks.

Figure 6 shows short term stability graphs for cesium fountains (FOM/FO2) and presently investigated optical clocks as a function of averaging time. The combined short term instability between FO2 and FOM (LNE-SYRTE) is $8.4 \cdot 10^{-14} \tau^{-1/2}$. Recently measured fractional frequency instability of the ^{27}Al$^+$/^{199}Hg$^+$ optical frequency comparison is $\sigma_y(\tau) \approx 4 \cdot 10^{-15} \tau^{-1/2}$ for measurement duration $\tau > 10$ s. Under assumption that both clocks contribute the same uncorrelated noise to the statistical measurement uncertainty, the short term stability of $2.8 \cdot 10^{-15} \tau^{-1/2}$ for each clock is derived [11].

4. Summary

Narrow optical transitions observed in many atoms and ions are now promising candidates for next generation of high performance frequency standards. Recent advances in optical frequency measurements technique allow to achieve very high accuracy of remote optical clocks comparison over kilometer distances. Through this comparison, the uncertainty of optical clocks placed in different laboratories can be evaluated at the 10^{-16} or at better level.

Optical clocks based on recommended by CIPM 2006 transitions are still in progress. To date optical standards based on ^{199}Hg$^+$ ion, neutral ^{87}Sr atoms and new one based on ^{27}Al$^+$ ion, have demonstrated systematic uncertainties which significantly exceed (10 times) the current best evaluations of cesium primary standards.

Presently it is not clear what kind of clocks will be the best: single trapped ion or neutral atoms lattice clock [26]. Lattice clocks combine the advantages of trapped single ions and the large number of neutral atoms: long storage times and the good signal-to-noise ratio. These clocks require

precise and long term compensation of the large frequency shift associated with the lattice laser field. Promising candidate for reaching the ultimate performance of lattice clock is neutral mercury because of a low sensitivity to blackbody radiation (20 times smaller than Sr).

It seems that the optical clocks with instabilities and inaccuracies at 10^{-18} level are expected in the time and frequency laboratories over the next several years. The progress in optical clocks is so rapid that in the near future the redefinition of the second will be most probably required.

References

[1] E. F. Arias and G. Petit, "Estimation of the duration of the scale unit of TAI with primary frequency standards", in *Proc. Freq. Contr. Symp.*, Vancouver, Canada, 2005, pp. 244–246.

[2] E. F. Arias, "The metrology of time", *Phil. Trans. R. Soc. A*, vol. 363, no. 1834, pp. 2289–2305, 2005.

[3] R. Wynands and S. Weyers, "Atomic fountain clocks", *Metrologia*, vol. 42, no. 3, pp. S64–S79, 2005.

[4] T. Udem and F. Riehle, "Frequency combs applications and optical frequency standards", *Riv. Nuovo Cim.*, vol. 30, no. 12, pp. 564–602, 2007.

[5] "Recommendation CCTF 2 concerning secondary representation of the second", in *Rep. 17th Meet. CCTF*, Sevres, France, 2006, p. 40.

[6] F. Riehle, "On secondary representation of the second", in *XXIXth URSI Gener. Assem. Conf.*, Chicago, USA, 2008, p. 21 (abstracts A0 1.2).

[7] S. G. Porsev *et al.*, "Determination of Sr properties for a high accuracy optical clock", *Phys. Rev. A*, vol. 78, no. 3, pp. 032508-1–032508-9, 2008.

[8] L. Holberg *et al.*, "Optical frequency standards and measurements", *IEEE J. Quant. Electron.*, vol. 37, no. 12, pp. 1502–1513, 2001.

[9] U. Tanaka *et al.*, "Optical frequency standard based on $^{199}Hg^{+}$ ion", *IEEE Trans. Instrum.*, vol. 52, no. 2, pp. 245–249, 2003.

[10] W. H. Oskay *et al.*, "Single-atom optical clock with high accuracy", *Phys. Rev. Lett.*, vol. 97, no. 2, pp. 020801-1–020801-4, 2006.

[11] T. Rosenband *et al.*, "Frequency ratio of Al^{+} and Hg^{+} single-ion optical clocks; metrology at the 17th decimal place", *Science*, vol. 319, no. 5871, pp. 1808–1812, 2008.

[12] J. E. Stalnaker *et al.*, "Optical to microwave frequency comparison with fractional uncertainty of 10^{-15}", *Appl. Phys. B*, vol. 89, no. 2–3, pp. 167–176, 2007.

[13] C. Tamm *et al.*, "$^{171}Yb^{+}$ single ion optical frequency standard at 688 THz", *IEEE Trans. Instrum.*, vol. 56, no. 2, pp. 601–604, 2008.

[14] G. P. Barwood *et al.*, "Observation of a sub-10-Hz linewidth $^{88}Sr^{+}$ $^{2}S_{1/2} - ^{2}D_{5/2}$ clock transition at 674 nm", *IEEE Trans. Instrum.*, vol. 56, no. 2, pp. 226–229, 2007.

[15] H. S. Margolis *et al.*, "Hertz-level measurement of the optical clock frequency in a single $^{88}Sr^{+}$ ion", *Science*, vol. 306, no. 5700, pp. 1355–1358, 2004.

[16] T. Rosenband *et al.*, "Observation of the $^{1}S_0 \rightarrow ^{3}P_0$ clock transition in $^{27}Al^{+}$", *Phys. Rev. Lett.*, vol. 98, no. 22, pp. 220801-1–220801-4, 2007.

[17] M. Boyd, "High precison spectroscopy of strontium in a optical lattice: towards a new standard for frequency and time". Ph.D. thesis, University of Colorado, Department of Physics, 2007.

[18] G. K. Campbell *et al.*, "The absolute frequency of the ^{87}Sr optical clock transition", *Metrologia*, vol. 45, no. 5, pp. 539–548, 2008.

[19] A. D. Ludlow *et al.*, "Sr lattice clock at 10^{-16} fractional uncertainty by remote optical evaluation with a Ca clock", *Science*, vol. 319, no. 5871, pp. 1805–1808, 2008.

[20] A. D. Ludlow, "The strontium optical lattice clock: optical spectroscopy with sub-Hertz accuracy". Ph.D. thesis, University of Colorado, Department of Physics, 2008.

[21] M. Petersen *et al.*, "Doppler-free spectroscopy of the $S_0 - ^{3}P_0$ optical clock transition in laser-cooled fermionic isotopes of neutral mercury", *Phys. Rev. Lett.*, vol. 101, no. 18, pp. 183004-1–183004-4, 2008.

[22] G. Wilpers *et al.*, "Absolute frequency measurement of the neutral ^{40}Ca optical frequency standard at 657 nm based on microlelvin atoms", *Metrologia*, vol. 44, no. 2, pp. 146–151, 2007.

[23] Z. W. Barber *et al.*, "Optical lattice induced light shifts in an Yb atomic clock", *Phys. Rev. Lett.*, vol. 100, no. 10, pp. 103002-1–103002-4, 2008.

[24] H. Katori *et al.*, "Optical lattice clocks with non-interacting bosons and fermions", *Rev. Laser Eng.*, vol. 36, no. APLS, pp. 1004–1007, 2008.

[25] H. Hachisu *et al.*, "Trapping of neutral mercury atoms and prospects for optical lattice clocks", *Phys. Rev. Lett.*, vol. 100, no. 5, pp. 053001-1–053001-4, 2008.

[26] F. Riehle, "Optical atomic clocks: challenges and possible solutions", in *Worksh. Quant. Phys. Atoms and Phot. IFRAF*, Les Houches, France, 2009.

Karol Radecki received the M.Sc. and Ph.D. degrees in electronic engineering from the Warsaw University of Technology, Poland, in 1970 and 1977, respectively. Since 1970 he has been with the Warsaw University of Technology, where he currently is an Assistant Professor at the Institute of Radioelectronics. His field of interest covers atomic frequency standards and electronic orientation systems for blind people. He is the URSI Commission A (Electromagnetic Metrology) National Chairman.

e-mail: K.Radecki@ire.pw.edu.pl
Institute of Radioelectronics
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

PHY Abstraction Methods for OFDM and NOFDM Systems

Adrian Kliks, Andreas Zalonis, Ioannis Dagres, Andreas Polydoros, and Hanna Bogucka

Abstract— In the paper various PHY abstraction methods for both orthogonal and non-orthogonal systems are presented, which allow to predict the coded block error rate (BLER) across the subcarriers transmitting this FEC-coded block for any given channel realization. First the efficiency of the selected methods is investigated and proved by the means of computer simulations carried out in orthogonal muticarrier scenario. Presented results are followed by the generalization and theoretical extension of these methods for non-orthogonal systems.

Keywords— orthogonal and non-orthogonal multicarrier systems, PHY abstraction methods.

1. Introduction

In the past where multi-modal operation was not an option, the role of performance evaluation (analytically or by simulation) was to simply check whether a given signal design met the pre-specified performance requirements. The average performance of a system was quantified by using the topology and the channel macro-characteristics in order to compute a geometric (or average) signal-to-interference plus noise ratio (SINR) distribution across the cell. If there were degrees of freedom either for transmitter-based signal design or for receiver-based algorithmic choice, then the role of performance evaluation was to pick the right set of parameter values so as to optimize a performance metric. In that sense, performance evaluation started becoming an integral part of the system design process itself, and the motivation thus arose to have simple analytic forms for these performance results which would make them amenable to easy parametric optimization.

Once the design aspect advances to become multi-modal and multi-parametric at both sides of the transmission link (e.g., current orthogonal frequency division multiplexing (OFDM) based systems: 3rd generation partnership project-long term evolution (3GPP-LTE), worldwide interoperability for microwave access (WiMAX)), the task of link-performance evaluation becomes not only germane to the design procedure itself, but the effective and efficient representation of this parameterized performance in ways that are compact (parsimonious) yet accurate comprises a main challenge of the optimization task.

Compact-description models are also of great interest in the context of evaluation methodologies (EVM's) which are currently being developed for various systems in the respective standardization bodies (e.g., IEEE 802.16m Task Group [1]). The goal of this type of physical-layer (PHY) abstraction is to determine the performance of a given link and thus avoid the need for extensive simulation. This "simulation-shortcut" accelerates the corresponding system-level simulations where a large number of physical-layer-related links need to be taken into account. The abstraction should be accurate, computationally simple, relatively independent of channel models, and extensible to interference models and multi-antenna processing.

A very novel and challenging task is to define the proper (PHY) abstraction methods for the non-orthogonal mutlicarrier (NOMC) systems, which are gaining the interest in the area of considered future wireless communication techniques. In the case of non-orthogonal frequency division multiplexing (NOFDM) signals, the impulses used at the transmitter overlap each other both in time and in frequency domain, thus they are not orthogonal. The shape and the signaling time of the applied impulses can be chosen without any restrictions besides ensuring that the pulses used at the receiver are biorthogonal to pulses used on the transmitter side. The (NOFDM) systems are the part of the larger set of generalized mutlicarrier (GMC) systems, where all of the transmit parameters can be in general chosen without any specified restriction. Thus a GMC signal set includes the orthogonal multicarrier signals as well.

The remainder of the paper is organized as follows: first the idea of (PHY) abstraction methodology is described and some possible abstraction methods are presented. These are followed by some results obtained for (OFDM) scenario. Finally, the main features of (NOFDM) systems are presented and the proposals for modification of some abstraction methods for (NOFDM) case are described. The whole paper is summarized in the last section.

2. PHY Abstraction Methodology

Physical-layer abstraction methodology for predicting instantaneous link performance for OFDM systems has been an active area of research and has received considerable attention in the literature [2]–[11]. The content in this section is based on the evaluation methodology document [1] of the ongoing work in IEEE 802.16m Task.

In a coded OFDM system, the coded block is transmitted over many subcarriers usually over a frequency selective channel, resulting in unequal channel gains for the subcarriers, and thus non-uniform and time-varying post-processing SINR values just prior to decoding. The task of the PHY abstraction methodology is to predict the coded block error rate (BLER) across the OFDM subcarriers transmitting this forward error correction (FEC) coded block for any given

channel realization (not averaged over the channel statistics). To do that, the vector of post-processing SINR values at the input to the FEC decoder are also considered as the input to the PHY abstraction methodology. As the link-level BLER curves are always generated based on a frequency-flat channel for various SINR's, an effective SINR (ESINR) is required to map the system-level SINR vector on these link-level BLER curves to determine the resulting BLER. This mapping is termed effective SINR mapping (ESM). The PHY abstraction is thus tantamount to compressing the vector of received SINR values to a single ESINR value, which can then be further mapped to a BLER number.

Several ESM approaches for predicting the instantaneous link performance have been proposed in the literature, including: mean instantaneous capacity [2]–[4], exponential-effective SINR mapping (EESM) [5]–[8] and mutual information effective SINR mapping (MIESM) [9], [10]. Each of these approaches uses a different function to map the vector of SINR values to a single number. In general, any ESM PHY abstraction method can be described via the following equation:

$$SINR_{eff} = \Phi^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\Phi\left(SINR_n\right)\right), \tag{1}$$

where: $SINR_{eff}$ is the effective SINR, $SINR_n$ is the SINR in the nth subcarrier, N is the number of symbols in a coded block, or the number of subcarriers used in an OFDM system, and Φ is the invertible function that defines the specific ESM.

Another important abstraction step is the per-tone SINR computation. All PHY abstraction metrics are computed as a function of post-processing per-tone SINR values across the coded block at the input to the decoder. The post-processing per-tone SINR is therefore dependent on the transmitter/receiver multiple input multiple output-space time coding (MIMO-STC) structure used to modulate/demodulate the symbols.

2.1. PHY Abstraction Methods for OFDM Systems

2.1.1. Mutual Information Based Effective SINR Mapping – Received Bit Mutual Information Rate (RBIR)

The computation of the mutual information per coded bit can be derived from the received symbol-level mutual information; this approach is termed received bit mutual information rate (RBIR). For a soft-input/soft-output (SISO)/soft-input/soft-output (SIMO) system the symbol mutual information (SI) is given by [1]

$$SI\left(SINR_n, m(n)\right) = \log_2 M - \frac{1}{M}\sum_{m=1}^{M} E_U\left\{\log_2\left(1 + \sum_{k=1,k\neq m}^{M} \mathrm{e}^{-\frac{|X_k - X_m + U|^2 - |U|^2}{1/SINR_n}}\right)\right\}, \tag{2}$$

where: U is zero mean complex Gaussian with variance $\frac{1}{2SINR_n}$ per component, $SINR_n$ is the post-equalizer (SINR) at the nth symbol or subcarrier and $m(n)$ is the number of bits at the nth symbol (or subcarrier).

Assuming N subcarriers are used to transmit a coded block, the normalized mutual information per received bit (RBIR) is given by

$$RBIR = \frac{\sum_{n=1}^{N} SI\left(SINR_n, m(n)\right)}{\sum_{n=1}^{N} m(n)}. \tag{3}$$

2.1.2. Mutual Information Based Effective SINR Mapping – Mean Mutual Information per Bit (MMIB)

The mutual information can be defined on the bit channel, which is referred as the mutual information per coded bit. The bit channel is obtained by defining the mutual information between bit input into the quadrature amplitude modulation (QAM) mapping and log-likelihood ratio (LLR) output at the receiver. The concept of "bit channel" encompasses SIMO/MIMO channels and receivers. The main difference between the bit and symbol level mutual information (MI) definitions is that the bit LLR reflects the demodulation process to compute LLR, which was not reflected in the symbol-level. The MMIB can be expressed as [1]

$$MI = \frac{1}{mN}\sum_{n=1}^{N}\sum_{i=1}^{m} I_{m,b_i(n)}\left(SINR_n\right) = \frac{1}{N}\sum_{n=1}^{N} I_m\left(SINR_n\right). \tag{4}$$

The mean mutual information is dependent on the SINR on each modulation symbol (index n) and the code bit index i (or ith bit channel), and varies with the constellation order m. In order to construct a numerical approximation can be used, for details refer to [1].

2.1.3. Exponential-Effective SINR Mapping (EESM)

The EESM abstraction method is given by [1]

$$SINR_{eff} = -\beta \ln\left(\frac{1}{N}\sum_{n=1}^{N} \mathrm{e}^{-\frac{SINR_n}{\beta}}\right), \tag{5}$$

where: β is a value for optimization/adjustment that depends on the modulation and coding scheme (MCS) and the encoding block length.

2.2. Evaluation of Selected ESM Methods

In this section a preliminary assessment of the abovementioned ESM methods for the SISO case are presented. The simulation parameters are chosen based on the WiMAX standard, as a typical example of OFDM system with a variety of operational modes (modulation, coding rate, number of subchannels). For the selected schemes, 100 different channel realizations with normalized total power (over the used subchannels) were produced.

The simulation parameters are:

– channel type: scenario definition in the timing definition language (TDL), pederastian B for 100 different channel realizations;

– subcarrier allocation method: full usage of subchannels (FUSC);

– code type: convolutional turbo code;

– chosen modulation/coding schemes: 4QAM with rate 0.75 for 1 subchannel per coded block (12 bytes); 4QAM with rate 0.75 for 6 subchannels per coded block (72 bytes); 64QAM with rate 0.5 for 3 subchannels per coded block (108 bytes).

In Figs. 1–4 with points are the additive white Gaussian noise (AWGN) link performance curves, while with solid line are the different realizations of the frequency selective case. The figures display the BLER based on the chosen ESNR metric per different channel realization.

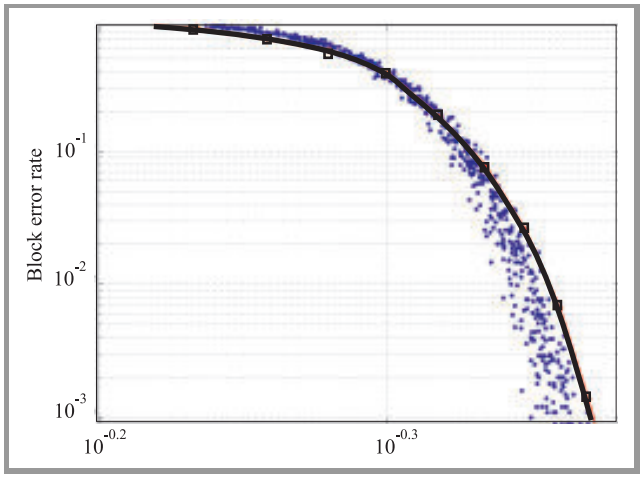


Fig. 1. The BLER based on the RBIR metric, 4-QAM, rate = 0.75, per 1 subchannel.

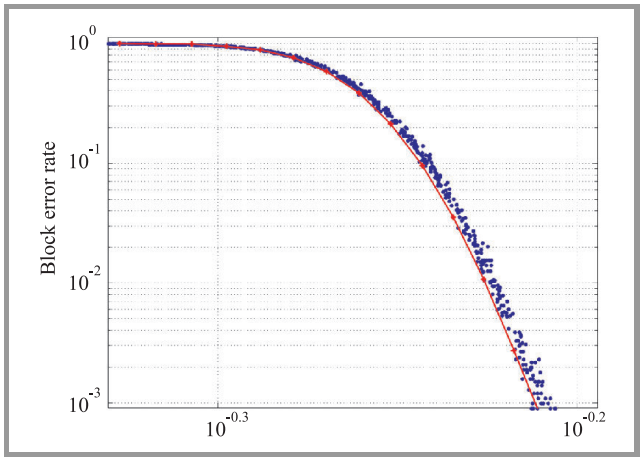


Fig. 2. The BLER based on the MMIB metric, 64-QAM, rate = 0.5, per 3 subchannels.

From the system level simulation perspective, the performance curves for all the methods exhibits low dispersion and good prediction since they are close to the AWGN reference curves. In the EESM case a calibration factor is used (β) for each MCS. The performance is plotted for two

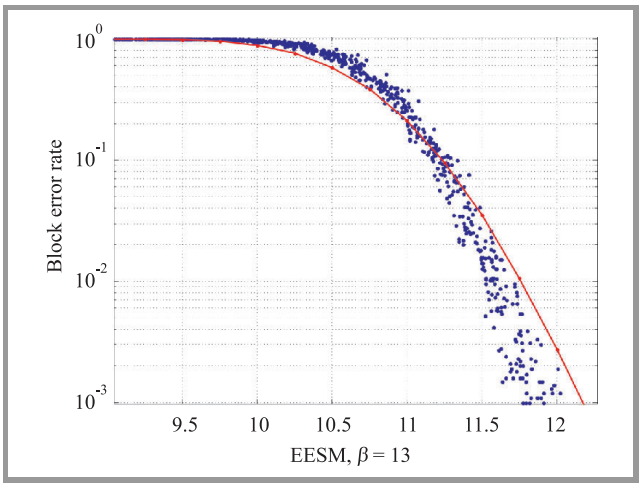


Fig. 3. The BLER based on the EESM metric, 64-QAM, rate = 0.5, per 3 subchannels, $\beta = 13$.

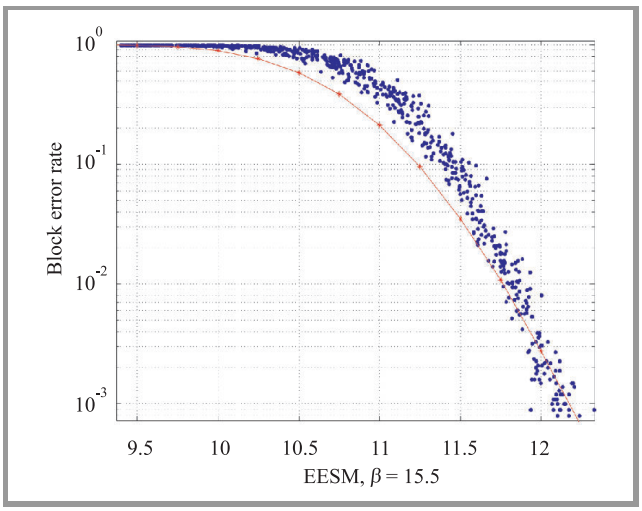


Fig. 4. The BLER based on the EESM metric, 64-QAM, rate = 0.5, per 3 subchannels, $\beta = 15.5$.

different selections of β. This demonstrates that the prediction accuracy in different BLER areas can be controlled via proper calibration, and this property can be exploited in the design of adaptive modulation and coding (AMC) algorithms.

3. PHY Abstraction Proposal for NOFDM Systems – Theoretical Analysis

In this section the main features of NOFDM systems are presented, which are followed by the preliminary modification proposals of PHY abstraction methods for NOFDM systems. Let us stress that the presented proposals need to be intensively tested by the means of computer simulation, which will be the next step of common investigation in this area.

3.1. NOFDM Signal Description

The non-orthogonal multicarrier signal belongs to particular subclass of all multicarrier signals (known as GMC signals) for which the neighboring information-bearing-pulses are not orthogonal. Such approach is in opposition to the well-known OFDM signals, where the pulses transmitted on adjacent subcarriers are mutually orthogonal. From the mathematical point of view, the transmit signal $s(t)$ (one NOFDM frame of N subcarriers and L time slots) is represented as the superposition of the translated and modulated elementary functions $g(t)$ multiplied by the weighting coefficients $c_{l,n}$:

$$\begin{aligned} s(t) &= \sum_{l=0}^{L-1}\sum_{n=0}^{N-1} c_{l,n} g(t-lT)\mathrm{e}^{2\pi jnFt} \\ &= \sum_{l=0}^{L-1}\sum_{n=0}^{N-1} c_{l,n}\cdot g_{l,n}(t). \end{aligned} \tag{6}$$

The above equation describes the inverse Gabor transform, and the weighting coefficients $c_{l,n}$ are the so-called Gabor coefficients, which are obtained by the means of Gabor transform [12]–[14]:

$$c_{l,n} = \int_{-\infty}^{\infty} s(t)\cdot\gamma_{l,n}^{*}(t)dt, \tag{7}$$

where: $\gamma_{l,n}(t)$ represents the window function (localized at time-frequency point (l,n) on time-frequency grid and dual to $g_{l,n}(t)$) used at the receiver to recover the transmit data and $(^{*})$ denotes conjugation.

One can observe, that each coefficient $c_{l,n}$ can be treated as the transmit data symbol located at the nth subcarrier in lth time period carried by the pulse $g_{l,n}(t)$. It is worth mentioning that in order to represent any signal by the means of the sum of the elementary pulses $g_{l,n}(t)$ (called also atoms), these pulses have to create the basis (denoted hereafter as $\boldsymbol{G}$) which spans the considered signal space. In other words, the forward and inverse Gabor transform switch the spaces from one-dimensional (time domain) to two-dimensional (time-frequency plane) and backwards, respectively. At the receiver, the dual set of elementary functions $\gamma_{l,n}(t)$ has to be applied, which constitute the basis $\boldsymbol{\Gamma}$. In order to ensure the perfect reconstruction requirement, the biorthogonality condition between the elementary pulses used on the transmitter and on the receiver side has to be fulfilled [15]. The two abovementioned sets of pulses will be biorthogonal, only if the following relation is true:

$$\sum_{l,n} g_{l,n}(t)\gamma_{l,n}(t') = \delta(t-t'), \tag{8}$$

where $\delta(x)$ is the well-know Dirac delta function, which is non-zero only for $x=0$. Let us stress, that the set of elementary functions constitutes also the so-called "frame" [13], [14] – in such a case, there exist real numbers A and B, $0 \le A \le B < \infty$, for which the following relation holds:

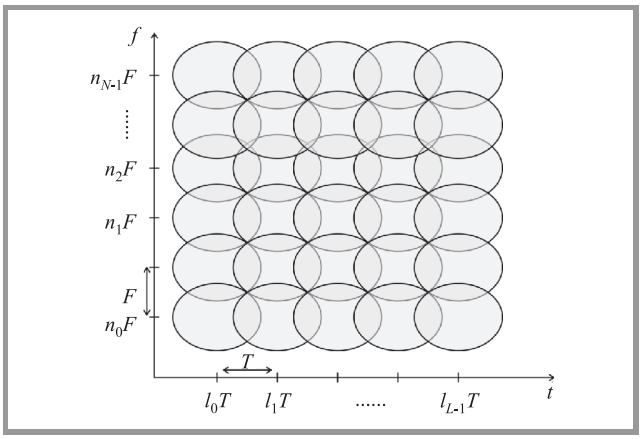


Fig. 5. Time-frequency representation of one GMC frame.

$$A\|s(t)\|^2 \le \sum_{l,n} |\langle s(t), g_{l,n}(t)\rangle| \le B\|s(t)\|^2, \tag{9}$$

where $\|\cdot\|$ and $\langle\cdot\rangle$ are the norm and the inner product, respectively, and $s(t)$ is the data signal.

The exemplary GMC signal frame is depicted in Fig. 5, where one circle represents one waveform. One can observe, that in general the neighboring pulses can overlap each other both in time and in frequency domain.

3.2. Proposals of Modification of PHY Abstraction Methods for NOFDM Systems

When referring to the NOFDM systems, the following aspects have to be considered: first, the NOFDM signal (frame) is represented on time-frequency plane, thus the algorithms shall be two-dimensional, second – the neighboring atoms overlap each other in both domains. These two phenomena have a significant impact on the definition of PHY abstraction methods. In the following the proposals of modification of selected PHY abstraction methods for NOFDM systems will be shortly presented and explained.

3.3. General Considerations

As highlighted in the previous section, besides the straightforward change from one-dimensional to two-dimensional processing, the overlapping between neighboring pulses has to be considered. Thus, in the first step, let take into account all separated values of $SINR_{l,n}$ (related to one atom on time-frequency plane) in the calculation of the effective $SINR_{eff}$:

$$SINR_{eff} = \Phi^{-1}\left(\frac{1}{LN}\sum_{n=0}^{N-1}\sum_{l=0}^{L-1}\Phi\left(SINR_{l,n}\right)\right). \tag{10}$$

Obviously, the definition of the $SINR_{l,n}$ has to be adjusted to NOFDM case, since the power of residual interferences (i.e., the total amount of power coming from the adjacent atoms and affecting the considered pulse at time-frequency

point (l,n)) has to be considered. Let us start from deriving the expression of SINR value for orthogonal systems, in which the cyclic prefix is added at the beginning of every OFDM transmit symbol. Based on [8], one can define the SINR value for nth subcarrier as

$$SINR_n = P_n \overline{G} \left(\frac{N}{N+N_P} \right) \left(\frac{R_D}{N_{sd}/N_{st}} \right), \qquad (11)$$

where: P_n is the frequency-selective fading power value for nth subcarrier, N_p is the length of the cyclic prefix, R_D denotes the power allocated to the data subcarriers, N_{sd} is the number of data subcarriers and N_{st} is the number of total useful subcarriers.

In other words, the factor within the first brackets describes the power loss due to the cyclic prefix removal at the receiver, whereas the term in second brackets indicates the pilots/signaling overhead (the percentage of power allocated to data subchannels referred to the relative numbers of data subcarriers to total useful subcarriers). The so-called geometry $\overline{G}$ (which includes all other factors that affect the received SINR, such path loss, shadowing, interference form base stations, etc.) can be defined as

$$\overline{G} = \frac{I_{or}}{I_{oc}+N_0}, \qquad (12)$$

where: I_{or} is the total received signal power prior to any receiver processing, I_{oc} denotes the total interference power and N_0 is the thermal noise power measured across the noise bandwidth.

In order to adjust the SINR definition to the NOFDM case, the above equations have to be rewritten as follows. First, let define the SINR value related to one waveform localized at (l,n) point on time-frequency grid as

$$SINR_{l,n} = P_{l,n} \overline{G_{ln}} \left(\frac{N_s}{N_s+N_P} \right) \left(\frac{R_D}{N_{sd}/N_{st}} \right), \qquad (13)$$

where all of the parameters should be interpreted on time-frequency (two-dimension) plane. That is, N_s denotes the number of samples of transmit signal (one NOFDM frame) in time-domain without the cyclic prefix of the length of N_P samples, R_D is the total power allocated to data pulses, whereas N_{sd} denotes the number of data-bearing pulses and N_{st} is the total number of useful waveforms on time and frequency (TF) plane. Moreover, one of the reasons for applying of NOFDM is to remove the cyclic prefix used in OFDM to mitigate the effect of inter-symbol interference (ISI). In such a case, above relation can be simplified:

$$SINR_{l,n} = P_{l,n} \overline{G_{ln}} \left(\frac{R_D}{N_{sd}/N_{st}} \right). \qquad (14)$$

Moreover, the geometry $\overline{G_{l,n}}$ can be defined as

$$\overline{G_{l,n}} = \frac{I_{or}}{I_{oc}+I_{int}^{(l,n)}+N_0}, \qquad (15)$$

where $I_{int}^{(l,n)}$ describes the amount of power that comes from the neighboring pulses and affect the transmit data atom at (l,n) point of TF grid.

3.4. Received Bit Mutual Information Rate (RBIR) ESM

In Subsection 2.1, the RBIR ESM method has been introduced. For NOFDM case the RBIR metric definition has to be adjusted to two-dimensional signal representation. Thus, instead of one sum operation over N subcarriers, two sum operations over whole time-frequency plane have to be calculated. In such a case, the values of $SINR_{l,n}$ (see Eq. (14)) and number of bits m assigned to each pulse on time-frequency plane have to be computed. The RBIR metric in NOFDM scenario can be expressed as

$$RBIR = \frac{\sum_{n=0}^{N-1} \sum_{l=0}^{L-1} SI\left(SINR_{l,n}, m(l,n)\right)}{\sum_{n=0}^{N-1} \sum_{l=0}^{L-1} m(l,n)}. \qquad (16)$$

In the above equation, the number of bits carried by the pulse localized at the (l,n) point on time-frequency grid is denoted by $m(l,n)$.

3.5. Exponential-Effective SINR Mapping (ESM)

Similar conclusions as for RBIR ESM method can be drawn for EESM PHY abstraction method. In such a case, the effective SINR can be computed as

$$SINR_{eff} = -\beta \ln \left(\frac{1}{LN} \sum_{n=0}^{N-1} \sum_{l-0}^{L-1} \mathrm{e}^{-\frac{SINR_{l,n}}{\beta}} \right). \qquad (17)$$

The simulation have to be carried out to define the values of the unknown parameter β.

3.6. Mean Mutual Information per Bit (MMIB) ESM

In Subsection 2.1, the metric called mean mutual information per bit ESM has been also defined and shortly described. In such a case, the definition of mean mutual information MI should be modified for NOFDM systems as follows:

$$MI = \frac{1}{mLN} \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \sum_{i=1}^{m} I\left(b_i^{(l,n)}, LLR\left(b_i^{(l,n)}\right) \right), \qquad (18)$$

where $b_i^{(l,n)}$ and $LLR(b_i^{(l,n)})$ denote the ith bit in one tuple (block of m bits mapped to one constellation point) carried on the nth subcarrier and lth time slot in one NOFDM frame and the log-likelihood ratio computed for this particular bit, respectively. The mutual information function is assumed to be a function of the QAM symbol SINR, thus the mean mutual information MI may be alternatively written as

$$MI = \frac{1}{mLN} \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \sum_{m=1}^{m} I\left(SINR_i^{(l,n)}\right). \qquad (19)$$

However, the definition of LLR for NOFDM case has to be derived, which takes into account the overlapping phenomenon of neighboring pulses in time and frequency domain. These derivations are out of scope of this paper.

4. Conclusions

In this paper new efficient PHY abstraction methodologies were described, which allow the system designers to reduce the amount of information sent in the reverse channel from the receiver to the transmitter in order to select the appropriate (in terms of error probability) modulation and coding scheme. On the other side, these signal-to-noise ratio (SNR) mapping techniques give the possibility of proper prediction of the BLER value (needed for system-level simulation) without implementing the particular decoding stages. Provided simulation results show the correctness of such approach, and confirm that SNR mapping is the promising technique for carrying out the system-level simulations. Moreover, it is shown, that these methods, originally proposed for OFDM case, can be also adjusted, extended or even generalized for non-orthogonal multicarrier systems. In order to do this, the overlapping phenomena between neighboring pulses has to be taken into account.

Acknowledgement

This work was supported by the EC in the framework of the FP7 Network of Excellence in Wireless Communications NEWCOM++ (contract no. 216715).

References

[1] IEEE 802.16m-08/004r1, "Evaluation methodology document", http://ieee802.org/16/tgm/

[2] Sony, Intel, "TGn Sync TGn proposal MAC simulation methodology", IEEE 802.11.

[3] ST Micro-Electronics "Time correlated packet errors in MAC simulations", IEEE Contribution, 802.11-04-0064-00-000n, Jan. 2004.

[4] Atheros, Mitsubishi, ST Micro-Electronics and Marvell Semiconductors, "Unified black box PHY abstraction methodology", IEEE Contribution, 802.11-04/0218r1, March 2004.

[5] 3GPP TR 25.892, "Feasibility study for OFDM for UTRAN enhancement (release 6)", ver. 1.1.0, March 2004.

[6] WG5 Evaluation Ad-Hoc Group, "1x EV-DV evaluation methodology – addendum (V6)", July 2001.

[7] Ericsson, "System level evaluation of OFDM – further considerations", TSGRAN WG1, no. 35, R1-03-1303, Nov. 2003.

[8] Nortel, "Effective SIR computation for OFDM system-level simulations", TSG-RAN WG1, no. 35, R03-1370, Nov. 2003.

[9] K. Brueninghaus *et al.*, "Link performance models for system level simulations of broadband radio access systems", in *IEEE 16th Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC*, Berlin, Germany, 2005.

[10] L. Wan *et al.*, "A fading insensitive performance metric for a unified link quality model", in *Proc. IEEE WCNC'06 Conf.*, Las Vegas, USA, 2006.

[11] I. Dagres and A. Polydoros, "Dynamic transceivers: adaptivity and reconfigurability at the signal-design level", in *SDR Forum Tech. Conf.*, Orlando, USA, 2003.

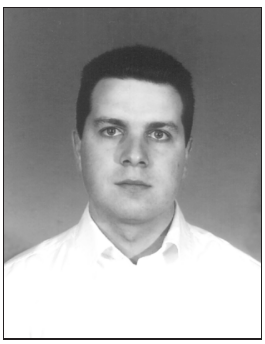
[12] W. Kozek and A. F. Molisch, "Nonorthogonal pulseshapes for multicarrier communications in doubly dispersive channels", *IEEE J. Selec. Areas Commun.*, vol. 16, no. 8, pp. 1579–1589, 1998.

[13] H. Feichtinger and T. Strohmer, *Gabor Analysis and Algorithms. Theory and Applications*. Berlin: Birkhäuser, 1998.

[14] S. Qian and D. Chen, "Understanding the nature of signals whose power spectrum change with time. Joint analysis", *IEEE Sig. Proc. Mag.*, vol. 16, iss. 2, pp. 52–67, March 1999.

[15] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood: Prentice Hall, 1993.

Adrian Kliks received his M.Sc. degree in telecommunication from the Poznań University of Technology, Poland, in 2005. Since 2005 he has been employed at the Institute of Electronics and Telecommunications, and since 2007 as the senior researcher at the Chair of Wireless Communication in the Faculty of Electronics and Telecommunication. Since 2005 he is the Ph.D. student, and starting form the 1 March 2009 he has been working as the Assistent. His research interests cover the wide spectrum of wireless communications. In particular he is interested in multicarrier (both orthogonal an non-orthogonal) systems, in the area of software defined, adaptive and cognitive radios, in hardware implementation on DSP/FPGA and in radio-planning.
e-mail: akliks@et.put.poznan.pl
Chair of Wireless Communications
Poznań University of Technology
Polanka st 3
60-965 Poznań, Poland



Andreas Zalonis received the B.Sc. degree in physics and the M.Sc. degree in telecommunication and electronics, both from the National Kapodistrian University of Athens, Greece, in 2000 and 2002, respectively. He is currently working towards his Ph.D. degree at the same university. Since 2001 he works as a Research Associate in the University of Athens. He has participated in several EU and Greek research projects. His main research interests are on adaptive modulation and coding for multicarrier schemes in next generation digital communications systems, including WiMax and LTE.
e-mail: azalonis@phys.uoa.gr
Department of Physics
Electronics Laboratory
University of Athens
Building: Physics 5, second floor
Panepistimiopolis 15784, Greece

Ioannis Dagres received the B.Sc. degree in computer engineering and the M.Sc. degree in signal processing from the Technical University of Patras, Greece, in 1997 and 1999, respectively. He is currently working towards his Ph.D. degree at the same university. Since 1999 he has participated in several EU and Greek research projects. His research interests include the topics of signal processing for communications, and in particular adaptive signal design and synchronization problems in for multicarrier communication systems.

e-mail: jdagres@phys.uoa.gr
Department of Physics
Electronics Laboratory
University of Athens
Building: Physics 5, second floor
Panepistimiopolis 15784, Greece

Andreas Polydoros was born in Athens, Greece, in 1954. He was educated at the National Technical University of Athens (Diploma in EE, 1977), State University of New York at Buffalo (MSEE, 1979) and the University of Southern California (Ph.D., EE, 1982). He was a faculty member at USC in the Electrical Engineering Department/Systems and the Communication Sciences Institute (CSI) in 1982–1997, a Professor since 1992. He co-directed CSI in 1991–1993. Since 1997 he has been Professor and Director of the Electronics and Systems Laboratory, Division of Applied Physics, Department of Physics, University of Athens.

e-mail: polydoros@phys.uoa.gr
Department of Physics
Electronics Laboratory
University of Athens
Building: Physics 5, second floor
Panepistimiopolis 15784, Greece

Hanna Bogucka received the M.Sc. and the Ph.D. degrees in telecommunications from the Poznań University of Technology (PUT), Poland, in 1988 and 1995, respectively. Since 1988 she has been employed at PUT, currently at the Chair of Wireless Communications as a Professor. In 1994, she received the Young Outstanding Scientists Reward from the Polish Science Foundation. She is involved in the research activities in the area of wireless communications, in particular system design and algorithms to provide reliable transmission of digital signals over dispersive radio channels. She is the author of more than 70 papers, which have been presented at national and international conferences, and published in leading communication journals. She has published 3 handbooks in the area of radio communications and digital signal processing.

e-mail: hbogucka@et.put.poznan.pl
Chair of Wireless Communications
Poznań University of Technology
Polanka st 3
60-965 Poznań, Poland

Paper

Technical and Regulatory Issues of Emergency Call Handling

Wojciech Michalski

Abstract—The paper presents selected technical and regulatory aspects of emergency call handling in communication between citizens and authorities in case of distress. Among the most important technical aspects of emergency call handling are recognition and treatment of emergency call by originating network, routing of such call to the appropriate public safety answering point (PSAP), delivering call-related information to the PSAP as well as architecture and organization of PSAPs. From the legal point of view, of importance are the obligations for the Member States and stakeholders involved in the E112 project included in the EU directives, actions of European Commission related to providing access to the location information as well as obligations concerning emergency call handling included in Polish national law.

Keywords—calling line identity, emergency call, location information, public safety answering point.

1. Introduction

Every year in the European Union several millions of citizens dial an emergency number to access emergency services. It is observed that due to rising penetration of mobile telephony, the share of emergency calls originating from mobile networks grows continuously. Unfortunately, many mobile callers in an emergency situation are not able to indicate their location today.

Such situation makes the work of emergency services extremely difficult since their efficiency, and in particular their response time, depend on knowledge of the caller's location.

Taking into account that an emergency can be anything, from every day incidents like traffic accidents or assault, to major incidents like aeroplane crashes or forest fires, to major disasters such as earthquakes or large-scale terrorist attacks, the emergency communications (EMTEL) elaborated on a broad spectrum of issues related to use of telecommunication services in emergency situations, addressing European Telecommunications Standard Institute (ETSI) members and all stakeholders involved in E112 project.

Currently, ETSI works on defining the user requirements for the four main areas of emergency communications. One of them is emergency call handling – communication from citizens to authorities/organizations. The second comprises public safety communications between authorities/organizations. The next one regards warning systems – communication from authorities/organizations to citizens. The last one concerns the communications amongst citizens during emergencies.

The article presents technical problems of emergency call handling and regulatory issues related to introduction of E112 in public networks in accordance with current telecommunication standards as well as European Union and Polish regulations.

2. Definitions

Emergency call is the call originating from a user to an emergency control centre (ECC), where ECC means the facilities used by emergency response organizations like police, fire brigade and emergency medical services to handle emergency calls.

Emergency call is forwarded to the emergency control centre by public safety answering point (PSAP). The PSAP is treated as physical location where emergency calls are received under the responsibility of a public authority.

Emergency calls are handled under E112 (enhanced 112) defined as emergency communications service using the single European emergency call number 112, which is enhanced with location information of the calling user.

3. Speech Quality and Priority of Emergency Calls

Emergency calls should have priority over all other calls and the priority should be ensured across public networks.

In the fixed network priority should be given from the network access point (associated with the emergency call originated from this point) to the network termination point or PSAP to which appropriate emergency control centre is connected.

In the mobile network priority should be given from the mobile switching centre (MSC), including the air interface, to the network termination point or public safety answering point to which an appropriate emergency control centre is connected.

If the network is not operating under abnormal conditions as a result of a disaster, the speech quality of emergency calls should not be worse than for basic telephone service handled in these conditions.

Otherwise, if the network is operating under abnormal conditions and a trade-off exists between speech quality and connectivity, connectivity should have priority over speech quality.

4. General Provisions

General provisions related to access to emergency call handling are included in directive 2002/22/EC [1] and documents associated with this directive [2] and [3]. Some of them are described in the following sections of this article.

4.1. Ability of Network Resources to Fulfill User'S Needs

According to the directive mentioned above, in addition to any other national emergency call numbers specified by the national authorities, all end users of publicly available telephone services should have the possibility to call the emergency services free of charge by using the single European emergency call number 112. They should be able to do so without any modifications to terminals, networks and devices on the emergency services provider side.

It should be possible to make emergency calls from public and private payphones at any time without the assistance of an operator, on the same principles like from normal phones. Location information within a private network should be available when possible and comply with the requirements of relevant emergency authorities in the area.

Directive 2002/22/EC [1] requires also that emergency calls should be possible even if a voice communication terminal equipment has a PIN-coded lock of the keypad.

The probability that user will be able to make a basic telephone call to appropriate emergency service should be maximized as well.

4.2. Ability of the Public Network Access Point to Enable Emergency Calls

The public network access point should enable emergency calls in each situation, even when normal originating calls have been barred (e.g., because of non-payment of bills) or mobile phone is protected by an identification/authentication procedure. It should enable emergency call originating in a visited network if the mobile phone is technically compatible with the alternate network.

4.3. Recognition and Treatment of Emergency Calls by the Originating Network

Originating network should recognize emergency calls by means of the emergency call number 112 in addition to the local national emergency numbers valid in the originating network.

For each emergency call the originating network should generate emergency call-related information (e.g., location of the caller and calling line identification) and deliver this information to the PSAP or to the corresponding emergency control centre. This information may either arrive at the destination point at the same time as emergency call or be available for retrieval on demand from this point during the call. Handling of emergency call-related information should not delay answering of emergency call.

4.4. Delivering Call-Related Information Concerning User Location

Caller location information may be a geographical address or a set of geographical coordinates. This information enables the emergency control centre to determine the caller's location at the time of calling. The information should be accessible for as long as the emergency lasts via standardized interface after the initial contact is made.

According to the directive 2003/58/EC on privacy and electronic communications [4], public telephone network operator should forward to PSAP the best information available as to the location of the caller, to the extent technically feasible. For each emergency call for which the subscriber number has been identified, public telephone network operator should enable the PSAP to renew the location information through a call back functionality for the purpose of handling the emergency.

In the first case the location information is transferred to the destination point in the push mode because the information is automatically pushed with the initial call together with information contained in the calling line identity (CLI). In the second one, the information is transmitted in the pull mode on demand, using the CLI and preferably the emergency location protocol (ELP) [5].

Generally, location information is based on the calling line number. In a wireline network this information is received together with emergency call. When emergency call is made from mobile phone operated without a SIM card (subscriber identity module card), originating network cannot transmit CLI information to the PSAP.

4.5. Delivering Call-Related Information Concerning User Identification

For every emergency call made to the number 112 originating network should transmit to the PSAP the calling line number of the access (CLI). The emergency control centre should be equipped in functionality available to return a call to the number in the CLI.

If emergency calls are made from mobile phone operated without a SIM card, originating network cannot transmit CLI information to the PSAP because the CLI cannot be determined. In the countries where this is authorized, as an alternative solution, the equipment identity number like international mobile equipment identity (IMEI) may be transmitted by the originating network.

4.6. Handling of Emergency Calls Between Networks

Originating networks should transmit their network identification to the emergency control centre according to the directive [4] which requires that all location information provided to the PSAP is accompanied by an identification of the network from which the call has originated.

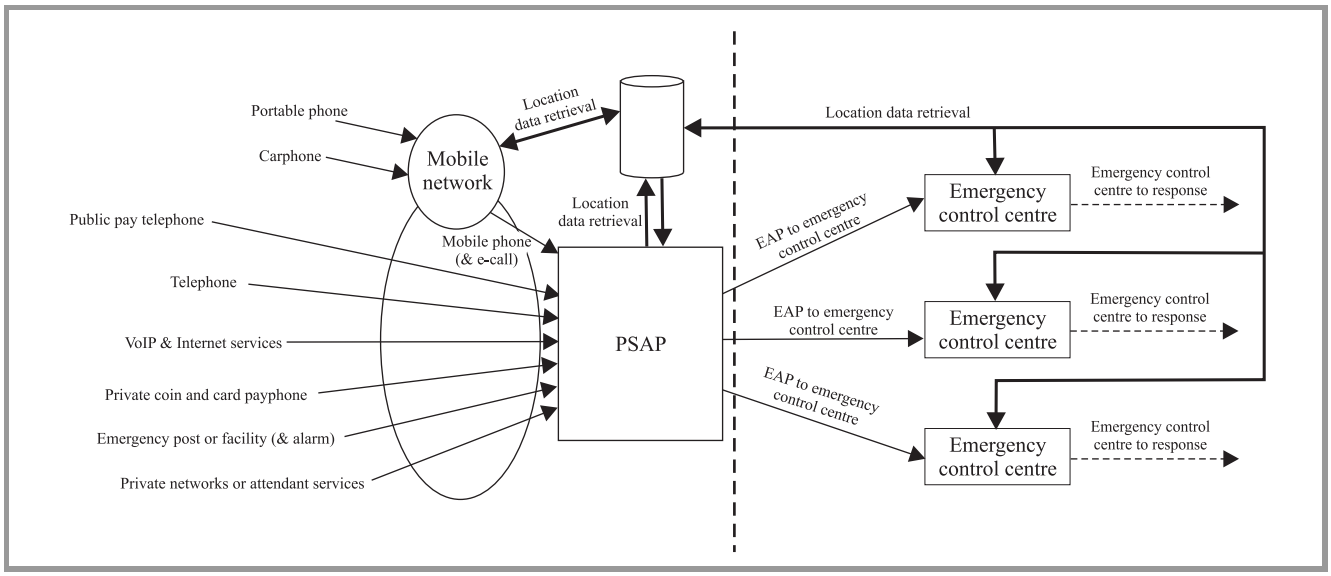


Fig. 1. Basic functional architecture [2].

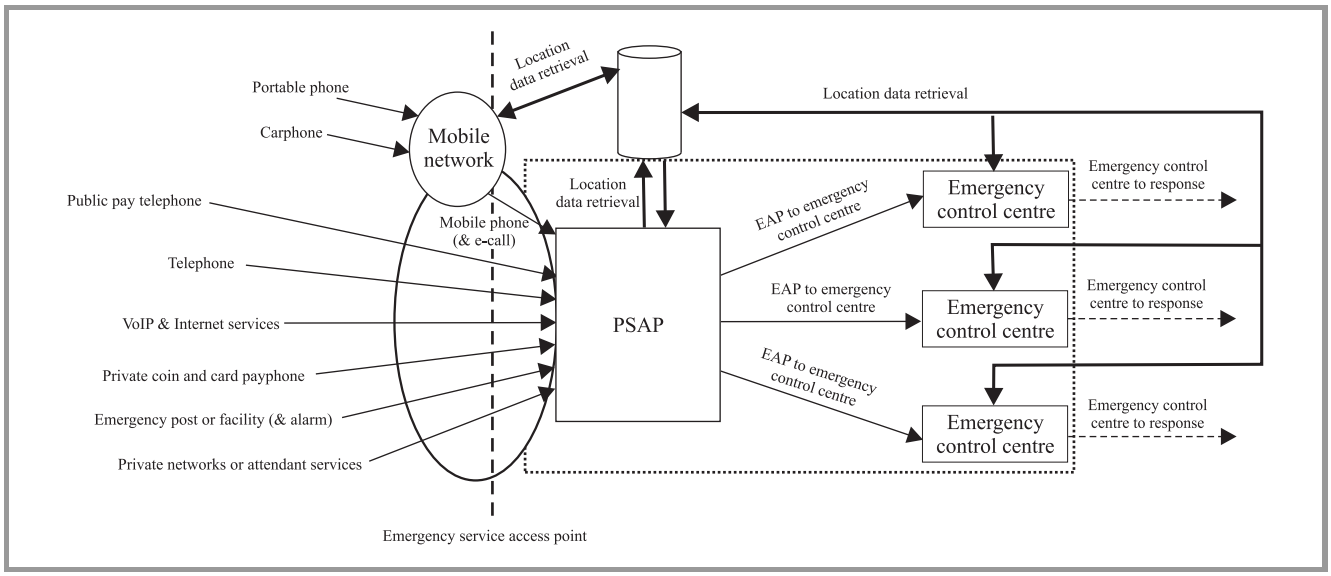


Fig. 2. Integrated PSAP and emergency control center [2].

If the originating network is not connected directly to the PSAP, a transit network is used for transfer of emergency call-related information to the destination point as well as specific routing number (destination number) for identifying responsible emergency service for a specific area.
The transit network should forward this information to the PSAP immediately and in transparent mode, without modification.

4.7. Providing Termination of Emergency Calls to the PSAP

The network should deliver the emergency calls together with any related data, without delay and modification to the PSAP which is directly connected this network. If delivering an emergency calls to the appropriate PSAP is not possible, it must be forwarded to the alternative PSAP.

The PSAP should be provided with access to all of the CLI information. PSAP should be able to release or block repeated nuisance call attempts to the emergency numbers. Only PSAP should be responsible for release of emergency call.

5. The PSAP's Architecture and Organization of the Emergency Control Centres

Basic functional PSAP architecture is illustrated in Fig. 1. Depending on PSAP and emergency control centre physical locations this logical architecture can be mapped into two physical solutions shown in Figs. 2 and 3.

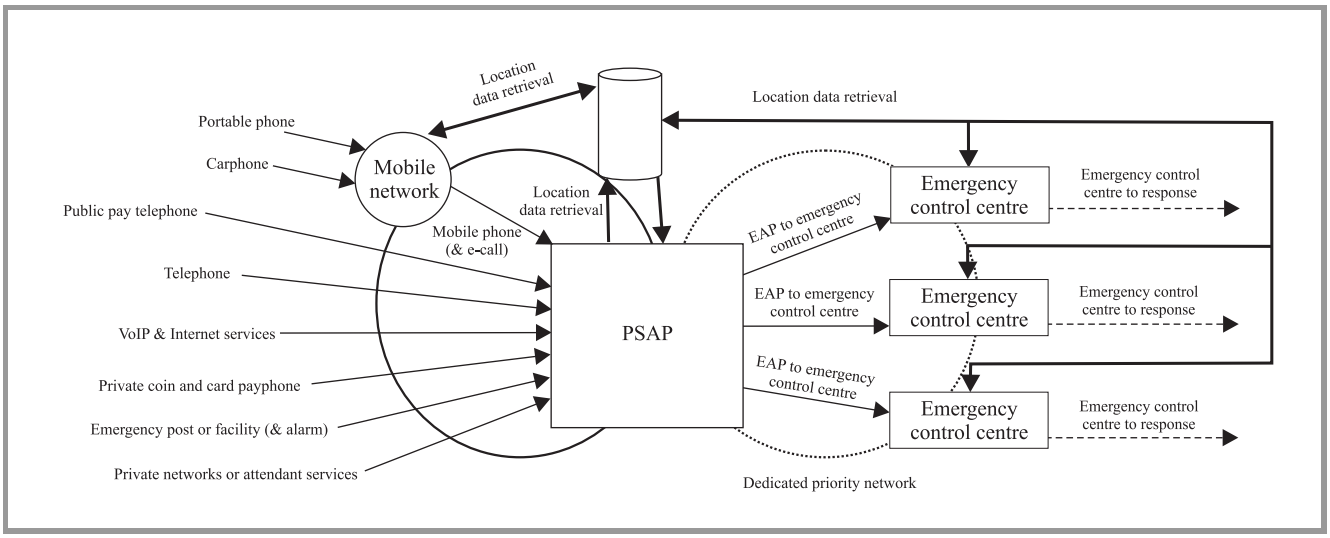


Fig. 3. PSAP on edge of the public network [2].

In the first case, illustrated in Fig. 2, the PSAP and emergency control centre functionalities are integrated into the same physical entity.

In the second solution, shown in Fig. 3, the PSAP sits at the edge of the public network and its functionality is distributed and separated from functionality of the emergency control centre. In this case the network between PSAP and emergency control centre is a dedicated priority network, built using leased lines or secure virtual private network (VPN).

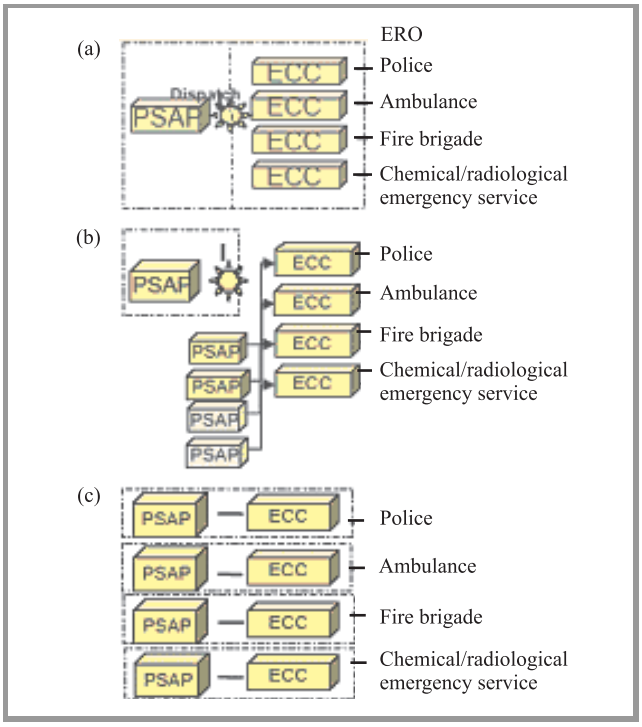


Fig. 4. Basic types of organizational setup [3]: (a) single level 1 PSAP dealing with all emergencies; (b) single level 1 PSAP dealing with all emergencies + directly reachable ERO's having their own answering points; (c) separate PSAPs dealing with emergencies.

Three types of organizational setup of PSAP, ECC and emergency response operations (ERO) recommended by Expert Group on Emergency Access (AGEA), which is the subgroup under the Communication Committee (COCOM) as well as the Technical Group chaired by European Commission (EC), are presented by Fig. 4.

6. Evolution of Regulations Concerning Emergency Calls

6.1. Previous Requirements

Obligations related to emergency calls were defined for the first time in the council decision 91/396/EEC [6]. This document required, in addition to other national emergency numbers, the Member States to ensure that the number 112 was introduced in the public telephone networks, as the number preferred by EC, by 31 December 1992, with a possibility for derogation until 31 December 1996 under certain conditions, e.g., due to high implementation costs. Moreover, the Member States should ensure that emergency calls are correctly received and routed to the appropriate emergency control centre according to technical capabilities existing in the public networks.

Next obligations were defined in the directive 97/13/EC [7], requiring to enable emergency calls even when the normal originating telecommunications services have been barred, e.g., due to of non-payment of bills.

Then, the obligations were included in directive 98/10/EC [8] which more precisely determined the requirements for the emergency calls handling given in using the single European emergency call number the council decision 91/396/EEC [6]. The directive required that it was possible to make emergency calls from public telephones 112 and other national emergency numbers, free of charge and without having to use any means of payment.

6.2. Main Regulations Included in the EU Directives

In the next edition of Community regulations called the "packet of directives of 2002", the requirements for emergency calls made to number 112 was defined in 26 article of universal service directive (2002/22/EC) [1]. In this document European Commission retained all important obligations included in the mentioned above directives and council decision.

First of all, the universal service directive requires it is possible to make emergency calls from public telephones using the single European emergency call number 112 and other national emergency numbers, free of charge and without having to use any means of payment. It is also required that emergency calls can be routed to, and handled within, the appropriate emergency control centre. The users shall be able to make a basic telephone call to an emergency service from any terminal that supports outgoing calls to publicly available telephone services. In particular, this requirement regards the case where emergency calls are made from mobile phones operated without a SIM card.

Moreover, this directive contains new obligations added by the EC. In particular, it is required that public network operators make caller location information available to authorities handling emergencies, to the extent technically feasible, for all calls made to the single European emergency number 112. The directive requires Member States to inform the users about E112 services. The Member States should provide adequate information to their citizens about the existence, use and benefits of E112 services. Citizens should be informed that 112 can connect them to emergency services all across the European Union and that their location will be forwarded. They should also be informed about the identity of the emergency services that will receive their location information and of other necessary details to guarantee fair processing of their personal data.

The universal service directive requires Member States to ensure that the obligations on the processing of caller personal data are respected. However, the directive 2002/58/EC [4; art. 10] permits the network operators to override the restrictions on calling line identification for emergency calls. So, this regulation (possibly only for emergency call purposes) eliminates the protections which users can use in case of other services.

The above mentioned elements form the set of regulations and conditions for their implementation in the Member States, which will be implemented on the principles that are appropriate for such low acts like directives.

6.3. Actions of European Commission Related to Providing Access to Location Information

The next step in creation of regulations regarding user location was Commission recommendation [9]. This recommendation was based on art. 19 of the framework directive 2002/21/EC, so this document has a high law status. Regulations included in this recommendation should be implemented unless serious obstacles exist.

This recommendation determines a number of important elements concerning the scope and mode of caller location (push/pull). This document confirms that location-enhanced emergency call services comprise emergency calls made to number 112 and other national emergency numbers. It means that the Member States with multiple national emergency numbers will have implementation cost higher than others using only one European emergency call number 112.

This document recommends that the "best effort" principle is applied for delivery of location information to PSAPs and that this information is transmitted in a "push" mode. Moreover, the document requires that location information is provided in a non-discriminatory way. In particular, public telephone network operators should not discriminate between the quality of information provided about their own subscribers and other users.

6.4. Proposals of Changes to EU Regulations

European Commission currently considers changing certain regulations concerning emergency call handling.

One of them regards the obligation for public telephone network operators to make caller location information available to authorities handling emergencies to the extent technically feasible. European Commission wants to delete condition "to extent technically feasible".

The second one regards delivery of location information only in the push mode. Moreover, EC recommends that the network operators were debited a cost of data transfer in this mode.

The proposals mentioned above are related to eCall project having a high priority among EC initiatives. The possibility to provide accurate location information for each call in push mode is fundamental to implementation of this project.

6.5. Regulations Included in Polish Law

6.5.1. Obligations Concerning Emergency Call Handling

Obligation for public telephone network operators to make caller information available to authorities handling emergencies is the key regulation concerning emergency calls handling. Set of obligations directly concerning emergency calls is included in the telecommunication law [10] and in other acts having obligatory status. The current status of Polish law concerning E112 determines the elements described below.

6.5.2. Obligations Regarding National Emergency Number

Fundamental obligation included in [10; art. 77] requires providers of publicly available telecommunication services to ensure that emergency calls made from any terminal to emergency numbers are free of charge. Emergency number

is defined in article 2, point 21 as a number available to responsible nominated emergency services specified in the national numbering plan [11]. The emergency numbers are associated with emergency services, including police, fire service, emergency medical services, emergency water services, emergency mountain rescue services and emergency gas services.

6.5.3. Obligations Regarding Assignment of Emergency Calls to the Appropriate Emergency Control Centre

Proper regulation is given in [10; art. 77, par. 2]. According to this obligation, emergency call should be routed to, and handled within, the appropriate emergency control centre. Nominated emergency control centres of the emergency organizations deal with emergency calls from defined geographical areas. Emergency calls are routed to appropriate destination point according to mapping between the location of the caller and the emergency control centre.

6.5.4. Data Available to Responsible Emergency Services

According to [10; art. 78, par. 1], public telephone network operators should make caller location information available to nominated emergency services to the extent technically feasible. The key meaning for this purpose have data regarding caller identification. Appropriate regulation is given in [10; art. 78, par. 1], which requires that public telephone network operators should make calling line identification available to responsible emergency service for this area. According to article 78, par. 1, section 1 [10], user identification should be possible also when the calling line identification is restricted. In the light of the article 78, par. 8 [10], location data should be provided without caller permission if necessary for handling of emergency calls.

6.5.5. Requirements Concerning the Real Time Transferring Data

The article 78, par. 1 [10] requires providers of publicly available telecommunication services to make caller location information available to nominated emergency services in real time. Delivering location information in real time means that delay between receiving of data request message and setting of requested data should not be significant.

National regulations do not precisely define this parameter because it is difficult due to different technologies used in Polish network. Location data based on cell ID methods can be made available faster than when GPS technology is used.

In the CGALIES report [12] it is assumed that coarse location information should be available in 7 s and accurate information should be available in 30 s.

6.5.6. The Changes Required in Polish Regulations

It is necessary to make some changes in [10; art. 77] which obliges providers of publicly available telecommunication services to ensure that emergency calls are free of charge.

First of all this obligation should be limited to providers of publicly available telephone services, not telecommunication services providers. Moreover, obligation related to providing location information should extend to all telephone services providers, because this obligation should be directly associated with providing telephone services.

The number of emergency numbers for which network operators are obliged to provide full functionality should be limited in accordance with concept of PSAPs organization.

In the light of the recommendation [9], the Member States should define precise obligations for telephone service providers related to providing caller location information, as existing general requirements will not suffice in the future.

7. Conclusion

Caller location information is critical for efficiency of emergency services, requiring that mobile positioning function is available anywhere in the network coverage area, anytime. Unfortunately, no method available today meets all requirements of emergency services.

Cell identity (cell ID) and its variants cannot meet accuracy requirements, although only these technologies are able to operate over 100% of the area covered by a network. They can be used only in dense urban environment.

Estimated observed time difference (E-OTD) can meet requirements in all environments except in the rural areas where the network may not provide sufficient number of cells to enable triangulation. This technology works where three or more base transceiver stations (BTS) are visible and the results obtained by E-OTD degrade to cell-ID when only one BTS is visible.

Assisted global positioning system (A-GPS) can fulfill the accuracy requirement in all environments, but performance of this technology in certain indoor environments may be problematic, although there are techniques available to increase the sensitivity of A-GPS receivers and hence improve the probability of a location fix as well as the resulting accuracy indoors.

So, from a purely technical point of view the best solution would be a combination of different technologies; introduction of hybrid technologies would improve the chance of meeting accuracy requirements.

In summary, regarding the works conducted in the technical as well as regulatory areas related to emergency calls handling, one sees recent progress in several fields. However, some problems related to emergency calls aren't solved yet. The most important problem now regards interoperability. It should be ensured that different manufacture's equipments are able to interoperate with each other correctly and without modifications. It should be noted that new standards must be defined to ensure progress in the development of E112.

References

[1] "Directive 2002/22/EC of the European Parliament and of the Council of 7 March 2002 on universal service and users' rights relating to electronic communications networks and services (Universal service directive)", http://www.mi.gov.pl

[2] ETSI SR 002 180: "Requirements for communication of citizens with authorities/organizations in case distress (emergency call handling)", http://portal.etsi.org

[3] "3GPP TR 26.967: eCall data transfer – in band modem solution", http://portal.etsi.org

[4] "Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)", http://eur-lex.europa.eu

[5] ETSI TS 102 164: "Services and protocols for advanced networks (SPAN); Emergency Location Protocols", http://portal.etsi.org

[6] "The Council Decision of 29 July 1991 on the introduction of the single European emergency call number (91/396/EEC)", http://eur-lex.europa.eu

[7] "Directive 97/13/EC of the European Parliament and of the Council of 10 April 1997 on a common framework for general authorizations and individual licences in the field of telecommunications services", http://ec.europa.eu

[8] "Directive 98/10/ EC of the European Parliament and of the Council of 26 February 1998 on the application of open network provision (ONP) to voice telephony and on universal service for telecommunications in a competitive environment", http://eur-lex.europa.eu

[9] "Commission Recommendation of 25 July 2003 on the processing of caller location information in electronic communication networks for the purpose of location-enhanced emergency call services", http://eur-lex.europa.eu

[10] "Telecommunication Law of 16 July 2004" (OJ no. 71 poz. 1800 with changes), http://www.en.uke.gov.pl

[11] "National Numbering Plan for the Public Telephone Network of the Republic of Poland" (lately modified by the Odinance of the Minister of Infrastructure in April 2008), http://www.en.uke.gov.pl

[12] "Report on implementation issues related to access to location information by emergency services (E112): Coordination Group on Access to Location Information for Emergency Services (CGALIES) of 28 February 2002", http://www.telematica.de/cgalies

Wojciech Michalski was born in Bogate, in Poland, in 1952. He received the M.Sc. degree in telecommunications engineering from the Technical University in Warsaw in 1977. He has been with the Switching Systems Department of National Institute of Telecommunications (NIT) since 1977, currently as a senior specialist. His research interests and work are related to PSTN backbone and access networks, GSM networks and IP networks. He is an author and co-author of technical requirements and many documents concerning telecommunication networks, services and protocols.

e-mail: W.Michalski@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland