JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

3/2004

Decision support for telecommunications and information society

Special issue edited by Wiesław Traczyk

Editorial Board

Editorial Advisory Board

ISSN 1509-4553

Circulation: 300 copies

Sowa - Druk na życzenie, www.sowadruk.pl, tel. 022 431-81-40

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

Preface

The set of 21 papers presented during The Third International Conference on *Decision Support for Telecommunications and Information Society DSTIS-2003*, organized by the National Institute of Telecommunications in Warsaw (4–6 September 2003), has been divided into two groups. Papers relevant to operational research have been selected for the special issue of *European Journal of Operational Research*, and remaining papers, devoted to various problems of decision support, are presented here.

Some old methods of hierarchical nonlinear optimization, when combined together and properly used, can give good results with moderate computational effort. Efficient algorithm is based on Kelley's cutting planes, Benders decomposition and ellipsoid methods.

To reduce computational power, needed for networks simulation, one can use Java-based library for distributed simulation, described here. The focus is on the effectiveness of different synchronization protocols and case study results.

Modeling decision processes can be simplified when flow graphs are used to describe a decision algorithm. Branches of the graph are interpreted as decision rules, associated with the certainty and coverage factors.

Fault localization is the core of fault diagnosis in computer networks. Probes technique for locating failures is promising but complicated. Partitions and logic can help in obtaining simpler algorithms.

Multiple criteria decision analysis is presented in two papers: as a tool for evaluation of geographical information (spatial data) and as a method for selection of clients segments with similar behaviors. Different applications show broad possibilities of multicriteria approach.

Uncertainty in multiple criteria decision making is usually analyzed with formal methods, but intuitive methodology, presented here, is also possible.

In the last paper a novel agent-based intelligent communication and decision support system for providing wireless services is described.

Optimization, simulation, modeling, decision analysis – the papers show different methods with a common goal – decision support.

Wiesław Traczyk
Guest Editor

Paper

Direct method of hierarchical nonlinear optimization – reassessment after 30 years

Andrzej Karbowski

Abstract—We consider the optimization problems which may be solved by the direct decomposition method. It is possible when the performance index is a monotone function of other performance indices, which depend on two subsets of decision variables: an individual for every inner performance index and a common one for all. Such problems may be treated as a generalization of separable problems with the additive cost and constraints functions. In the paper both the underlying theory and the basic numerical techniques are presented and compared. A special attention is paid to the guarantees of convergence in different classes of problems and to the effectiveness of calculations.

Keywords—hierarchical optimization, decomposition, the direct method, Benders method, cutting plane method, distributed computations.

1. Introduction

We consider the following optimization problem:

$$\min_{x_1,x_2,\ldots,x_{p-1},v} \psi\Big(f_1(x_1,v), f_2(x_2,v),\ldots \\ \ldots, f_{p-1}(x_{p-1},v), f_p(v)\Big), \tag{1}$$

$$v \in V \subseteq \mathbb{R}^{n_v},\ x_i \in X_i \subseteq \mathbb{R}^{n_i},\ i=1,\ldots,p-1, \tag{2}$$

$$(x_i,v) \in XV_i = \big\{(x_i,v) : g_{ij}(x_i,v) \le 0, \\ j=1,\ldots,m_i\big\},\ i=1,\ldots,p-1, \tag{3}$$

where $\psi : \mathbb{R}^p \to \mathbb{R}$ is an order preserving (i.e., monotonically increasing with all its arguments), continuous function and all functions f_i, g_{ij} are convex and differentiable. We want to solve this problem applying hierarchical two-level approach with the decomposition of (1)–(3) in the direct way (so-called direct method). That is, we would like to apply the following computational scheme:

coordination problem (*CP*):

$$\min_{v\in V\cap V_0} \psi\Big(f_1(\hat{x}_1(v),v), f_2(\hat{x}_2(v),v),\ldots \\ \ldots, f_{p-1}(\hat{x}_{p-1}(v),v), f_p(v)\Big), \tag{4}$$

$$V_0 = \{v | \forall i \in \{1,\ldots,p-1\}\ \exists x_i \in X_i : \\ g_{ij}(x_i,v) \le 0 \quad \forall j=1,\ldots,m_i\big\}, \tag{5}$$

***i*th local problem** (LP_i), $i=1,\ldots,p-1$:

$$\hat{x}_i(v) = \arg\min_{x_i\in X_i} f_i(x_i,v), \tag{6}$$

$$g_{ij}(x_i,v) \le 0, \quad j=1,\ldots,m_i. \tag{7}$$

We will call the variables forming vector v coordinating or complicating variables (the last name stems from the observation, that when they are temporarily fixed the remaining optimization problem is considerably more tractable). They have to belong to a given explicitly set V and to an unknown set V_0, which is the set of admissible values of these variables from the point of view of the local problems. The set V_0 is called solvability set.

Such problems have been considered for more than 30 years in more [10, 17] or less [2] general statement. Surprisingly, they are often treated in some isolation from other problems, which are, in the autor's opinion, very close to them [1, 3, 5, 11]. The latter works were devoted to general problems with two (or more) sets of variables and the possibilities to iterate them in Gauss-Seidel manner to obtain the global optimum. There were no assumptions concerning specific structural properties of the performance index and the constraints' functions. Even terminology is different in these two types of problems. In the first case the variables forming the v vector are called the coordinating variables, while in the second–complicating variables.

The methods proposed depend on the presence of mixed constraints defining sets XV_i (3). If there are no such constraints, the theory considerably simplifies. It will be shown later on, that in this case the coordinating variables v stop to be complicating and there is no need to treat them in a different way than the others. It leads to plane (one level) decision structure, that is without the coordination level, even with some possibilities of desynchronization of calculations between different local units. When such constraints are present, the situation is more complicated and the coordination level is necessary, where the unknown set V_0 has to be taken into account when calculating new values of the coordinating vector v. In the article it will be shown, that actually, it is not necessary to look for a general method of determining the set V_0, and an efficient algorithm based on Kelley's cutting plane method [14], Benders decomposition [1] and ellipsoid method [15, 16] will be proposed.

2. The case of independent constraints on local and coordinating variables

If there are no mixed constraints on local and coordinating variables (7), that is in the definitions (3), (5) of sets XV_i and V_0 $m_i = 0\ \forall i$ we may take as these sets full domains, and as the consequence

$$V \cap V_0 = V \cap \mathbb{R}^{n_v} = V\,. \tag{8}$$

In such circumstances the coordination problem takes the form:

coordination problem for independent sets $(CP-I)$:

$$\min_{v\in V} \psi\Big(f_1(\hat{x}_1(v),v), f_2(\hat{x}_2(v),v),\ldots \\ \ldots, f_{p-1}(\hat{x}_{p-1}(v),v), f_p(v)\Big) \tag{9}$$

and the local problem

***i*th local problem for independent sets** (LP_i-I), $i = 1,\ldots,p-1$:

$$\hat{x}_i(v) = \arg\min_{x_i\in X_i} f_i(x_i,v)\,. \tag{10}$$

Such problem for additive cost function ψ was considered, e.g., in [2, p. 270]. However it seems, that there are possibilities to solve this and a more general problem with (1) performance index more effectively. First of all, let us take that the coordinating vector does not differ qualitatively from the other vectors x_i and denote it by x_p, and its set by X_p that is:

$$x_p = v, \qquad X_p = V, \qquad n_p = n_v\,. \tag{11}$$

Now denoting

$$n = \sum_{i=1}^{p} n_i \tag{12}$$

we will define the performance index $f : \mathbb{R}^n \mapsto \mathbb{R}$ hiding the structure of the function ψ, as:

$$f(x_1,x_2,\ldots,x_p) = \psi\Big(f_1(x_1,x_p), f_2(x_2,x_p),\ldots \\ \ldots, f_{p-1}(x_{p-1},x_p), f_p(x_p)\Big)\,. \tag{13}$$

For typographical convenience the partitioned column vectors:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

will be written in the form $(x_1,x_2,\ldots,x_p)$.

In this notation we deal with the following optimization problem:

$$\min_{x\in X} f(x)\,, \tag{14}$$

where

$$X = X_1 \times X_2 \times \ldots \times X_p\,, \tag{15}$$

$$x = (x_1,x_2,\ldots,x_p) \tag{16}$$

and $x_i \in \mathbb{R}^{n_i}$, $i = 1,\ldots,p$.

For problems with such general structure it is possible to propose two types of optimization algorithms:

- Jacobi algorithm:

$$x_i^{k+1} = \arg\min_{x_i\in X_i} f\Big(x_1^k,\ldots,x_{i-1}^k,x_i,x_{i+1}^k, \\ \ldots,x_p^k\Big),\ i = 1,\ldots,p\,, \tag{17}$$

- Gauss-Seidel algorithm:

$$x_i^{k+1} = \arg\min_{x_i\in X_i} f\Big(x_1^{k+1},\ldots,x_{i-1}^{k+1},x_i,x_{i+1}^k, \\ \ldots,x_p^k\Big),\ i = 1,\ldots,p\,, \tag{18}$$

where k denotes subsequent iterations.

So, in Jacobi algorithm the new values of subvector x_i, that is x_i^{k+1} for every i are obtained on the basis of the same information, that is they may be determined independently of each other. In Gauss-Seidel algorithm to determine new x_i the previous values of subvectors $x_{i+1},\ldots,x_p$ are used, but already new values of subvectors $x_1,\ldots,x_{i-1}$. We may say, that although both these algorithms use decomposition, Jacobi algorithm is parallel, while Gauss-Seidel sequential from its nature.

The following theorems concerning the convergence of these two algorithms have been formulated:

Proposition 1 [3, Prop. 3.9, p. 219]: Suppose that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a continuously differentiable and convex function on the set X. Furthermore, suppose that for each i f is strictly convex function of x_i, when the values of the other components of x are held constant. Let $\{x^k\}$ be the sequence generated by the nonlinear Gauss-Seidel algorithm (18), assumed to be well defined. Then every limit point of $\{x^k\}$ minimizes f over X.

Proposition 2 [7]: Let $\{x^k\}$ be the sequence generated by the proximal Gauss-Seidel method:

$$x_i^{k+1} = \arg\min_{x_i\in X_i}\Big[f\Big(x_1^{k+1},\ldots,x_{i-1}^{k+1},x_i,x_{i+1}^k,\ldots,x_p^k\Big) \\ +\frac{1}{2}\tau_i||x_i - x_i^k||^2\Big],\quad i = 1,\ldots,p\,, \tag{19}$$

where $\tau_i > 0$, $i = 1,\ldots,p$. Then, if f is pseudoconvex on X, every limit point of $\{x^k\}$ is a global minimizer of problem (14).

Proposition 3 [3, Prop. 3.10, p. 221]: Let $f:\mathbb{R}^n \mapsto \mathbb{R}$ be a continuously differentiable function, let γ be a positive scalar, and suppose that the mapping $h: X \mapsto \mathbb{R}^n$, defined by

$$h(x) = x - \gamma \cdot \nabla f(x) \tag{20}$$

is a contraction with respect to the block-maximum norm $||x|| = ||(x_1, x_2, \ldots, x_p)|| = \max_i \frac{||x_i||_i}{w_i}$, where each $||.||_i$ is the Euclidean norm on $\mathbb{R}^{n_i}$ and each w_i is a positive scalar. Then there exists a unique vector $\hat{x}$ which minimizes f over X. Furthermore, the nonlinear Jacobi and Gauss-Seidel algorithms are well defined, that is, a minimizing x_i in Eqs. (17) and (18) always exists. Finally, the sequence $\{x^k\}$ generated by either of these algorithms converges to $\hat{x}$ geometrically.

The first two propositions concern Gauss-Seidel algorithm. Although, as it was written earlier, it is sequential from its nature, in the case of specific structural properties of the optimized functions, like in our case (13), it may be used to obtain the solution of the optimization problem. Owing to monotonicity of the function ψ, when it is strictly convex, the block coordinate problems (18) for $i = 1, \ldots, p-1$ may be simplified to:

$$x_i^{k+1} = \arg\min_{x_i \in X_i} f_i(x_i, x_p^k) \tag{21}$$

and solved independently. Only the last coordinate x_p has to be modified according to formula (18), for new optimal values of $x_1, x_2, \ldots, x_{p-1}$. Strict convexity is necessary to get from the local problems unique solutions. When this function is not strictly convex, but convex or pseudoconvex, according to Proposition 2, we may force the uniqueness of local solutions by adding quadratic proximal terms. Unfortunately, Grippo and Sciandrone theory [7] allows for independent, parallel solutions of block coordinate problems for $i = 1, \ldots, p-1$ only in the case of additive functions ψ.

The third proposition concerns a specific subclass of convex problems. The contraction condition for the mapping h (20) is satisfied for example (for functions from the $C_2(\mathbb{R}^n)$ class) when the Hessian of the function f is constrained, that is there exists such a constant K, that:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \leq K \quad \forall x \in \mathbb{R}^n, \forall i, j \tag{22}$$

and the domination of the main diagonal condition is fulfilled for a positive weights vector $[w_1, w_2, \ldots, w_n]$ (usually we take $w_i = 1, \forall i$):

$$w_i \cdot \frac{\partial^2 f}{\partial x_i^2} > \sum_{j \neq i} w_j \cdot \left| \frac{\partial^2 f}{\partial x_i \partial x_j} \right|. \tag{23}$$

In such conditions, if we take a sufficiently small coefficient γ, more precisely:

$$0 < \gamma < \frac{1}{K} \tag{24}$$

then the mapping h is a contraction in the maksimum norm. The functions whose Hessian is diagonally dominated are a subclass of the set of all convex functions. It results from the Gershgorin's circle theorem (saying that all eigenvalues of the matrix are contained within the union of n disks $K(a_{ii}, \sum_{j \neq i} |a_{ij}|)$, with each disk centered at a diagonal entry of the matrix and having radius equal to the sum of absolute values of off-diagonal entries in that row) and the equivalence of the positive signs of eigenvalues and positive definiteness in the class of symmetric matrices [6].

Unfortunately, not all convex functions have diagonally dominated Hessian. For example a quadratic form $f(x) = \frac{1}{2}x'Ax$ with the matrix:

$$A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix} \tag{25}$$

is convex, but the diagonal dominance condition will never take place, i.e., there are no positive weights w_1, w_2, w_3 for which the condition (23) will be satisfied (it is easy to prove it by a contradiction).

Let us return now to our hierarchical algorithm and state the conclusions from the Proposition 3. We may say, that in the case of functions of the class $C_2(X)$ whose Hessian satisfies conditions (22), (23), and when there are no mixed constraints on local and coordinating variables, it is not necessary to realize the hybrid version of calculations: Gauss-Seidel iterations between coordination and local level and Jacobi iteration between different units of the local level. It is possible and should be useful to treat the coordination problem in the same way as the local problems. Due to the structural properties of the function f–see Eq. (13)–(that it grows monotonically with all functions f_i) the iterations (17) for $i = 1, \ldots, p-1$ will be equivalent to LP_i (6), that is:

$$\begin{aligned} x_i^{k+1} &= \arg\min_{x_i \in X_i} f\left(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_p^k\right) \\ &= \arg\min_{x_i \in X_i} \psi\left(f_1(x_1^k, x_p^k), \ldots, f_i(x_i, x_p^k), \ldots,\right. \\ &\qquad \left. \ldots, f_{p-1}(x_{p-1}^k, x_p^k), f_p(x_p^k)\right) \\ &= \arg\min_{x_i} f_i(x_i, x_p^k), \qquad i = 1, \ldots, p-1 \end{aligned} \tag{26}$$

while for $i = p$ (earlier it was problem CP-I)

$$\begin{aligned} x_p^{k+1} &= \arg\min_{x_p \in X_p} f\left(x_1^k, \ldots, x_{p-1}^k, x_p\right) \\ &= \arg\min_{x_p \in X_p} \psi\left(f_1(x_1^k, x_p), \ldots, f_{p-1}(x_{p-1}^k, x_p), f_p(x_p)\right). \end{aligned} \tag{27}$$

Until now nothing was said about the numerical optimization algorithm solving local problems. Since all local decision variables x_i have to belong to given sets X_i, they have to be constrained optimization procedures. The simplest way is to apply directly the steepest descent algorithm adding to it, to take into account the constraints, the orthogonal projection (with respect to Euclidean norm) of a vector

onto the convex set X_i. Let us define the projection operator $[y]_Z^+$ by:

$$[y]_Z^+ = \arg\min_{z\in Z} ||z-y||\,, \tag{28}$$

where $||.||$ is the Euclidean norm. The simplest constrained optimization algorithm implementing Jacobi iterations (17) will be then:

$$x_i := [h_i(x)]_{X_i}^+ = [x_i - \gamma\nabla_i f(x)]_{X_i}^+,\ \ i=1,\ldots,p\,. \tag{29}$$

Since the projection does not change nonexpansive property [3], this mapping will be a contraction when the mapping h is a contraction. Moreover, different x_i may be calculated totally asynchronously [3], that is without the need to make a new calculation or communication in any finite window.

But it was all about such convex problems with independent admissible sets, where the mapping h defined in (20) was contractive in the maximum block norm. What about these situations, rather more common, where this feature does not take place? Surprisingly, the last algorithm (29) is still valid. The only differences are in the restriction on γ coefficient and in the time dependencies between subsequent iterations of the ith local subvector x_i and in the exchange of information between different local units. If we denote by B the window (measured in the number of iterations of the whole algorithm) in which at least one iteration of each local units and the communication updating their values in the buffers of other units should take place, and by K_1 the Lipschitz constant for the gradient of a convex, nonnegative function f:

$$||\nabla f(x) - \nabla f(y)|| \le K_1 \cdot ||x-y||\ \ \forall x,y\in\mathbb{R}^n \tag{30}$$

the assessment on γ will be as follows [3]:

$$\gamma < \gamma_0(B) = \frac{1}{K_1(1+B+nB)}\,. \tag{31}$$

In this case we deal with so-called partially asynchronous implementation of the algorithm, where B is the measure of asynchronism. For functions f belonging to class $C_2(X)$ the constant K_1 equals K from the assessment (22).

It means, that in the case when the admissible sets are independent, it may be useful to abandon the hierarchical manner of solving the problem (1). In the "peer-to-peer" (Jacobi) version of the algorithm it might be possible to find the solutions faster and even in an asynchronous implementation.

3. The case of mixed constraints on local and coordinating variables

In this case the biggest problem with the above two-level algorithm (4)–(5), (6)–(7), which seems to be quite natural and promising, is that it is very difficult to calculate two things: the set V_0 and the functions $\hat{x}_i(v)$. Because of that the algorithm (4)–(7) is completely impractical–it cannot be directly applied. First of all, solving CP involves the reactivation of all local problems $LP_i, i=1,\ldots,p-1$ after every change of the v vector, that is after every movement in its optimization. It is so, because only in this way we may guarantee the proper first arguments of functions $f_i(\hat{x}_i(v),v)$. It is fast only in these rare cases when we may solve analytically local problems. Yet more difficult situation is with the solvability set V_0. This set is not given explicitly. The direct formula to calculate it was presented by Geoffrion [5] and is the following:

$$V_0 = \left\{ v\in\mathbb{R}^{n_v} : \max_{\lambda\in\Lambda} \min_{x_i\in X_i\, i=1,\ldots,p-1} \sum_{i=1}^{p-1}\sum_{j=1}^{m_i} \lambda_{ij}\cdot g_{ij}(x_i,v) \le 0 \right\}, \tag{32}$$

where

$$\Lambda = \left\{ \lambda\in\mathbb{R}^{m_1+m_2+\ldots+m_{p-1}} : \lambda \ge 0 \sum_{i=1}^{m_1+m_2+\ldots+m_{p-1}} \lambda_i = 1 \right\}. \tag{33}$$

So, it is rather difficult to estimate it and the computational effort to assess whether a given v belongs to this set is comparable with that of solving the whole optimization problem. It would be better to estimate this set by some additional constraints, possibly simple. In the book [10, p. 87] it is written, that: "In general the problem of defining inequalities and equations describing the set V_0 is unsolved" and as the only remedy the penalty function method is suggested:

coordination problem for penalty function method (CP-PFM):

$$\min_{v\in V} \psi\Big(f_1(\hat{x}_1(v),\hat{v}_1(v)) + \rho_{1k}||v-\hat{v}_1(v)||^2, \ldots$$
$$\ldots, f_{p-1}(\hat{x}_{p-1}(v),\hat{v}_{p-1}(v)) + \rho_{(p-1)k}||v-\hat{v}_{p-1}(v)||^2, f_p(v)\Big) \tag{34}$$

***i*th local problem for penalty function method** (LP_i-PFM) $i=1,\ldots,p-1$:

$$[\hat{x}_i(v),\hat{v}_i(v)] = \arg\min_{x_i\in X_i, v_i} \left[f_i(x_i,v_i) + \rho_{ik}||v-v_i||^2 \right]$$
$$g_{ij}(x_i,v_i) \le 0,\ j=1,\ldots,m_i\,. \tag{35}$$

However there is a possibility to estimate both the set V_0 and the function

$$\varphi(v) = \psi\Big(f_1(\hat{x}_1(v),v), f_2(\hat{x}_2(v),v), \ldots$$
$$\ldots, f_{p-1}(\hat{x}_{p-1}(v),v), f_p(v)\Big) \tag{36}$$

by a set of inequalities, growing as the computation progresses. This is a decomposition method proposed by Benders in early sixties [1]. He considered problems (called by him "mixed-variables programming" problems) where both the performance index and the constraints were sums of two components: one linear depending on one set of variables

and one nonlinear (they were called complicating variables; also because in many practical problems, e.g. [12], they are discrete). He proposed an iterative procedure for solving this problem by optimization with respect to either the first or the second group of variables in some auxiliary problems, related to dual representation of the initial problem and to optimality conditions. In the latter–the outer–the number of constraints on the variables corresponding to nonlinear part of the problem was gradually growing. They were delivered by the other–the inner–problem in the way dependent on the existence or not of the feasible solutions in the space of variables corresponding to linear components. Hence, in the decision space of nonlinear part variables either an "optimality cut" or "feasibility cut" was made.

In seventies the procedure proposed by Benders was generalized by Geoffrion [5] to the case of continuous nonlinear problems with performance indices and constraint functions being convex functions for fixed values of complicating variables. In later works Floudas *et al.* [12, 13] presented methods of transformation of many practical nonconvex and mixed continuous-discrete problems to apply this theory. A good review of these methods and well presentation of the algorithms may be found in [11]. We will present the basic procedure on the general problem:

$$\min_{x\in X, v\in V} f(x,v)\,, \tag{37}$$

$$g_j(x,v)\le 0,\ \ j=1,\ldots,m. \tag{38}$$

The solution algorithm is an iterative procedure where every iteration (let us say kth) consists of two parts:

1. Solving the primal problem for the current value of coordinating/complicating variables:

$$\min_{x\in X} f(x,v^k)\,, \tag{39}$$

$$g_j(x,v^k)\le 0,\ \ j=1,\ldots,m. \tag{40}$$

If the problem is feasible (i.e., there exists at least one point $x\in X$ for which all constraints (40) are satisfied) the optimal values of decision variables x^k and Lagrange multipliers λ_o^k are memorized (to be used in optimality cut later on). If not, the following problem assessing the departure from feasibility is solved:

$$\min_{x\in X,\alpha} \sum_{j=1}^{m} \alpha_j\,, \tag{41}$$

$$g_j(x,v^k)\le \alpha_j,\ \ j=1,\ldots,m, \tag{42}$$

$$\alpha_j\ge 0,\ \ j=1,\ldots,m. \tag{43}$$

The optimal values of decision variables $x\in X$ and the Lagrange multipliers in this problem λ_f^k are also memorized (to be used in feasibility cut in the next phase).

2. Solving the relaxed master problem:

$$\min_{\mu, v\in V} \mu\,, \tag{44}$$

$$L_o(x^k,v,\lambda_o^k)\le \mu,\ \ k\in K_o\,, \tag{45}$$

$$L_f(x^k,v,\lambda_f^k)\le 0,\ \ k\in K_f\,, \tag{46}$$

where

$$L_o(x^k,v,\lambda_o^k) = f(x^k,v)+{\lambda_o^k}^T g(x^k,v)\,, \tag{47}$$

$$L_f(x^k,v,\lambda_f^k) = {\lambda_f^k}^T g(x^k,v)\,. \tag{48}$$

Symbols K_o and K_f denote the sets of indices of iterations in which, respectively, the optimal solution of the primal problem existed or not. Functions L_o and L_f are Lagrange functions for primal (39)–(40) and feasibility (41)–(43) problems (the latter restricted to admissible solutions that is for $\hat{\alpha}_j = 0, \forall j$). The assessments on $\varphi(v)$ then result directly from the duality theory.

In the terms of the direct method of hierarchical optimization the first set of inequalities delivers the assessment of the function (36), while the second–the assessment of the set V_0 (32).

The most important classes of problems where this algorithm is proved to converge to the optimum are [5, 11] variable factor programming problems and problems with f, g_j, $j=1,\ldots,m$ linearly separable and convex in x and y, where X is a polyhedron.

The basic drawback of this method is the growing number of constraints of nonlinear type. In the next section we will show how to cope with it.

4. Combining Benders decomposition and Kelley's cutting plane method

Even in Benders' article at the end [1, p. 250] there is a remark on the solution of the relaxed master problem (44)–(46), that if the complicating variables (i.e., nonlinear) components are "convex and differentiable functions (...) problem becomes a convex programming problem that can be solved by-well known methods, e.g., by Kelley cutting plane technique...". It seemed attractive, because in the case when the set V is a polyhedron, if this method is used we actually deal with a linear programming problem.

Let us define:

$$\varphi(v) = L_o\big(\hat{x}_o(v), v, \hat{\lambda}_o(v)\big), \tag{49}$$

$$\xi(v) = L_f\big(\hat{x}_f(v), v, \hat{\lambda}_f(v)\big)\,, \tag{50}$$

where $\hat{x}_o(v)$ is solution of the primal problem (39)–(40), $\hat{x}_f(v)$ is the solution of the feasibility problem (41)–(43), and $\hat{\lambda}_o(v)$, $\hat{\lambda}_f(v)$ are Lagrange multipliers corresponding to

them for given complicating vector. While linearizing the constraints the following expressions may be used [9]:

$$\frac{\partial \varphi(v)}{\partial v} = \frac{\partial f}{\partial v} + \lambda_o^T \frac{\partial g}{\partial v}, \tag{51}$$

$$\frac{\partial \xi(v)}{\partial v} = \lambda_f^T \frac{\partial g}{\partial v}. \tag{52}$$

When we apply Kelley's cutting plane method together with the Benders decomposition, the relaxed master problem (44)–(46) will be replaced by:

$$\min_{\mu, v \in V} \mu, \tag{53}$$

$$\varphi(v^k) + \frac{\partial \varphi^T}{\partial v}(v^k)(v - v^k) \le \mu, \;\; k \in K_o, \tag{54}$$

$$\xi(v^k) + \frac{\partial \xi^T}{\partial v}(v^k)(v - v^k) \le 0, \;\; k \in K_f. \tag{55}$$

The problem is, that at this point Benders was wrong. This algorithm may fail and end in nonoptimal points even in convex problems. The counterexample was shown in Grothey *et al*. [8]. The convex NLP there was:

$$\min_{x_1, x_2, v} v^2 - x_2, \tag{56}$$

$$(x_1 - 1)^2 + x_2^2 \le \ln v, \tag{57}$$

$$(x_1 + 1)^2 + x_2^2 \le \ln v, \tag{58}$$

$$v \ge 1. \tag{59}$$

The optimal solution of this problem is $[x_1, x_2, v] \simeq [0, 0.0337568, 2.721381]$. Starting with the feasible $v = e^2$ we obtain in the first step $\hat{x}_o(e^2) = [0,1]$ and the optimality cut:

$$(e^4 - 1) + \left(2e^2 - \frac{1}{2e^2}\right)(v - e^2) \le \mu. \tag{60}$$

From the relaxed master problem we obtain the new optimal $v = 1 < e$. For this value, however, the primal problem is infeasible and we will get from the feasibility problem $\hat{x}_f = [0,0]$. In general, if $v^k < e$, the following feasibility cut is generated and added to the master problem:

$$(2 - 2\ln v^k) + \left(-\frac{2}{v^k}\right)(v - v^k) \le 0 \;\Leftrightarrow\; v \ge (2 - \ln v^k) v^k.$$

The next values of v from the master problem will be calculated according to the formula:

$$v^{k+1} = (2 - \ln v^k) v^k.$$

They all will be from the interval $(1, e)$ giving the whole time infeasibility and the same optimal values in feasibility problems (actually the sequence v^k will approach e from the left hand side). The authors explain that "the failure of Benders decomposition to converge is due to the fact that the Benders cuts only approach feasibility in the limit and never collect subgradient information from the objective" function of the problem $\varphi(v)$. As the remedy they propose, as they call, "feasibility restoration algorithm", where in the case of infeasibility, after solving feasibility problem, the modified primal problem is solved again with the modified inequalities (40) in such a way, that on the right hand side of them there are positive numbers being values of constraints in the feasibility problem multiplied by some coefficient bigger than 1. Then both the previously obtained feasibility as well as the optimality cuts from this relaxed problem are added to the master problem constraints.

This procedure overcomes the basic disadvantage of the Benders method combined with Kelley's cutting plane algorithm–converging to nonoptimal points, but still has one drawback: the growing number of constraints in master problems as the calculations proceed. One has to wait longer and longer for new values of complicating variables v. How to overcome this difficulty and even to replace the optimization on the upper level with calculation of values of two simple analytic expressions will be shown in the next section.

5. Integration of Benders decomposition with cutting plane and ellipsoid algorithms

The main idea lies in the application (instead of solving relaxed master problem as the optimization problem) one of the simplest algorithms of nondifferentiable optimization, which was proposed by Shor [16], Nemirovski and Yudin [15], namely the ellipsoid algorithm. It will deliver in subsequent iterations the centers of the smallest volume ellipsoids, containing smaller and smaller sets of admissible points, in which the performance index may have a better value than in points it was calculated so far. It is obtained by cutting off the halfspaces of points in which, owing to convexity, for sure the value of performance index is worse than in the current point (if it is feasible) or the value of functions defining constraints is worse than the present one (if the current point is infeasible).

What concerns optimality cuts, we use the same formulae for derivatives as before, that is, it is defined by:

$$\left\langle \frac{\partial \varphi}{\partial v}(v^k), v - v^k \right\rangle \le 0. \tag{61}$$

The calculation of feasibility cuts may be simplified by making individual cuts for the most violated constraint. It is described in the next subsection.

5.1. Feasibility cuts

We will perform for every query point $v^k \in V$ (that is the current value of coordinating variables) verification of the feasibility from the point of view of the constraints (7), independently for all local problems LP_i, $i = 1, \ldots, p-1$, and adding the corresponding linear constraints to coordination problem in the case of a failure.

The verification of feasibility consists in calculation for every LP_i the "constraint index" $g_i(v^k)$, by solving a preliminary optimization problem:

$$g_i(v^k) = \min_{x_i \in X_i} \max_{j=1,\ldots,m_i} g_{ij}(x_i, v^k) \tag{62}$$

before the principal optimization problem.

In the case when $g_i(v^k) > 0$, we draw a conclusion, that for the query point v^k there are no admissible points $x_i \in X_i$ and the rational thing is cutting off a halfspace containing inadmissible values of coordinating variables. This will be obtained by the condition:

$$g_{ij^*}(x_{i_{\min}}, v^k) + \left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v - v^k \right\rangle \leq 0\,, \tag{63}$$

where $x_{i_{\min}}, j^*, v^k$ are such that

$$g_{ij^*}(x_{i_{\min}}, v^k) = g_i(v^k) > 0\,. \tag{64}$$

Proposition 4: Condition (63) assures the elimination from the admissible set V points not belonging to the solvability set V_0.

Proof: To prove the proposition we have to show that:

$$\forall v^* \in V \quad g_{ij^*}(x_{i_{\min}}, v^k) + \left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v^* - v^k \right\rangle > 0$$
$$\Rightarrow \forall x_i \in \mathbb{R}^{n_i} \quad g_{ij^*}(x_i, v^*) > 0\,. \tag{65}$$

From the convexity and smoothness of functions g_{ij} we have for any given pair $(\tilde{x}_i, v^k)$ and all x_i,v:

$$g_{ij}(x_i, v) \geq$$
$$g_{ij}(\tilde{x}_i, v^k) + \left\langle \frac{\partial g_{ij}}{\partial x_i}(\tilde{x}_i, v^k), x_i - \tilde{x}_i \right\rangle + \left\langle \frac{\partial g_{ij}}{\partial v}(\tilde{x}_i, v^k), v - v^k \right\rangle. \tag{66}$$

Setting in (66) $j = j^*$, $\tilde{x}_i = x_{i_{\min}}$ and $v = v^*$ we will get $\forall x_i \in \mathbb{R}^{n_i}$

$$g_{ij^*}(x_i, v^*) \geq g_{ij^*}(x_{i_{\min}}, v^k) + \left\langle \frac{\partial g_{ij^*}}{\partial x_i}(x_{i_{\min}}, v^k), x_i - x_{i_{\min}} \right\rangle$$
$$+ \left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v^* - v^k \right\rangle. \tag{67}$$

That is

$$g_{ij^*}(x_i, v^*) \geq \left[g_{ij^*}(x_{i_{\min}}, v^k) + \left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v^* - v^k \right\rangle \right]$$
$$+ \left\langle \frac{\partial g_{ij^*}}{\partial x_i}(x_{i_{\min}}, v^k), x_i - x_{i_{\min}} \right\rangle. \tag{68}$$

Let us notice, that from the assumption, the term in square brackets is positive. The second component is nonnegative, because we assumed, that $x_{i_{\min}}$ is the solution of the minimax problem. This means that

$$g_{ij^*}(x_i, v^*) > 0 \quad \forall x_i \in \mathbb{R}^{n_i} \tag{69}$$

what completes the proof. □

The interpretation of the Proposition 4 is, that by cutting off from the set V more and more points, we get a better estimate of the set $V \cap V_0$, that is the admissible set in the CP problem (4)–(5).

So, if we restrict our attention to these points of the decision space in which the value of the most violated constraint function g_{ij^*} is better than in the current point, it is sufficient to add a constraint:

$$\left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v^* - v^k \right\rangle \leq 0\,. \tag{70}$$

5.2. Ellipsoid algorithm

The presented algorithm was proposed by Shor [16], Nemirovski and Yudin [15]. At every step we obtain an ellipsoid

$$E_k = \left\{ v | (v - v^k)^T W_k^{-1} (v - v^k) \leq 1 \right\}. \tag{71}$$

It is characterized by two parameters: a matrix W_k and a center v^k. It is assumed, that we start from an ellipsoid E_0 containing the admissible set V. The subsequent ellipsoids E_k are such that E_{k+1} is the minimum volume ellipsoid containing $E_k \cap \{v | \langle h_k, v - v^k \rangle \leq 0\}$. It is defined by:

$$v^{k+1} = v^k - \frac{1}{n_v + 1} \frac{W_k h_k}{\sqrt{h_k^T W_k h_k}}\,, \tag{72}$$

$$W_{k+1} = \frac{n_v^2}{n_v^2 - 1} \left(W_k - \frac{2}{n_v + 1} \frac{W_k h_k h_k^T W_k}{h_k^T W_K h_k} \right), \tag{73}$$

where n_v is the dimension of v. It can be shown, that the volume of E_{k+1} equals the volume of E_k reduced by the factor $(1 - 1/(n_v + 1)^2)$.

5.3. Integration

If we use as the vector h_k in expressions modifying ellipsoids (72), (73) the gradient $\frac{\partial \varphi}{\partial v}(v^k)$ from optimality cut expressions (61), (51) or the gradient $\frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k)$ from feasibility cut expression (70), we will have what we need–a very simple and fast metod of delivering subsequent values of coordinating (complicating) variables with the convergence guarantee.

This approach seems to be the most promising among all, because the calculations on the coordination level are the simplest one can imagine: only two direct formulas without any optimization, iterative process, etc. There are other techniques from cutting plane family generating queries

at new points inside the admissible area (e.g., center of gravity, largest inscribed sphere, volumetric, analytic center methods–see [4]), which prevents the algorithm against blocking, but none of them has so simple and fast master problem iteration.

6. Conclusions

In the paper the basic approaches to solving optimization problems with generalized separable structure, where the performance index is a monotone function of other performance indices depending on individual and common subvectors of decision variables, were presented and compared. It was shown, that in the case when the admissible set is a Cartesian product of individual domains and a domain of common variables (that is coordinating or complicating variables), the problem may be solved by application of hybrid Gauss-Seidel (between coordination and local level) and Jacobi (between different units of the local level) algorithms or in a completely symmetric (Jacobi) version, even with asynchronous iterations. The degree of asynchronism depends on the features of the overall performance index. If its Hessian is restricted and diagonally dominated the steepest descent type iterations and the exchange of information may be totally asynchronous, otherwise they may be partially asynchronous, that is with iterations and communication between local units in a given finite window, dependent on the length of the step in optimization iterations, the dimension of the problem and the assessment on the Hessian elements.

The situation is much more complicated when the admissible set is not a Cartesian product of local and common variables domains. The most natural seems to be the Benders decomposition, where so-called optimality and feasibility cuts obtained after, respectively, admissible or inadmissible queries of complicating/coordinating variables are used to estimate the value function (i.e., the function whose value is the optimal value of the original problem for fixed values of coordinating variables) and the solvability set (i.e., the set of complicating variables for which all mixed constraints can be satisfied for at least one combination of the primal variables). This approach, based on duality relations, although very general and elegant, has one serious drawback–since the estimates of both the value function and the solvability set have to be more accurate as the computations progress, the number of constraints defining them systematically grows. It means, that the problems solved in subsequent iterations are more are more complicated and the time needed for one iteration of master problem is longer and longer. An attempt to simplify calculations by combining Benders decomposition with Kelley's cutting plane method and transform the master problem to LP problem is not a good idea, because, as it was shown in an example, the optimization process even in the convex case may converge to a nonoptimal point. It is possible to avoid it by either so-called feasibility restoration algorithm, which adds an additional optimality cut in an extended domain, or the application on the master (i.e., coordination) level an algorithm which delivers query points lying inside the admissible area, e.g., center of gravity method, the largest inscribed sphere or ellipsoid method, volumetric center method, analytic center method (ACCPM) or the smallest circumscribing ellipsoid method. The latter approach seems to be the most attractive due to its simplicity, noniterative character (that is, the new values of complicating variables are not determined, as for example in optimization, via an iterative process, but directly from two simple formulas) and converges to optimal solution with the geometric rate.

References

[1] J. F. Benders, "Partitioning procedures for solving mixed-variables programming problems", *Numer. Math.*, vol. 4, pp. 238–252, 1962.

[2] D. P. Bertsekas, *Nonlinear Programming.* 2nd ed. Belmont: Athena Scientific, 1999.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs: Prentice Hall, 1989.

[4] S. Elhedhli, J. L. Goffin, and J.-P. Vial, "Nondifferentiable optimization: cutting plane methods", in *Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Dordrecht: Kluwer, 2001, vol. 4, pp. 40–45.

[5] A. M. Geoffrion, "Generalized Benders decomposition", *J. Opt. Theory Appl.*, vol. 10, pp. 237–260, 1972.

[6] G. Golub and J. M. Ortega, *Scientific Computing: an Introduction with Parallel Computing.* San Diego: Academic Press, 1993.

[7] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss-Seidel method under convex constraints", Report R. 467, Istituto di Analisi dei Sistemi ed Informatica, CNDR, Settembre 1998.

[8] A. Grothey, S. Leyffer, and K. I. M. McKinnon, "A note on feasibility in Benders decomposition", Numerical Analysis Report NA/188, Dundee University, 1999.

[9] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Mathematics in Science and Engineering. New York: Academic Press, 1983, vol. 165.

[10] W. Findeisen, F. N. Bailey , M. Brdyś, K. Malinowski, P. Tatjewski, and A. Woźniak, *Control and Coordination in Hierarchical Systems*. Chichester: Wiley, 1980.

[11] C. A. Floudas, "Generalized Benders decomposition, GBD", in *Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Dordrecht: Kluwer, 2001, vol. 2, pp. 207–218.

[12] C. A. Floudas, A. Aggarwal, and A. R. Ciric, "Global optimum search for nonconvex NLP and MINLP problems", *Comput. Chem. Eng.*, vol. 13, no. 10, pp. 1117–1132, 1989.

[13] C. A. Floudas and V. Visweswaran, "A primal-relaxed dual global optimization approach", *J. Opt. Theory Appl.*, vol. 78, no. 2, pp. 187–225, 1993.

[14] J. E. Kelley, "The cutting-plane method for solving convex programs", *J. Soc. Indust. Appl. Math.*, vol. 8, pp. 703–712, 1960.

[15] A. Nemirovski and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*. Chichester: Wiley, 1983.

[16] N. Z. Shor, *Minimization Methods for Non-differentiable Functions*. Berlin: Springer Verlag, 1985.

[17] *Optimization Methods for Large-Scale Systems with Applications*, D. A. Wismer, Ed. New York: McGraw-Hill, 1971.

Andrzej Karbowski received his M.Sc. degree in electronic engineering (specialization automatic control) from Warsaw University of Technology (Faculty of Electronics) in 1983. He received the Ph.D. in 1990 in automatic control and robotics. He works as adjunct both at Research and Academic Computer Network (NASK) and at the Faculty of Electronics and Information Technology (at the Institute of Control and Computation Engineering) of Warsaw University of Technology. His research interests concentrate on data networks management, optimal control in risk conditions, decomposition and parallel implementation of numerical algorithms.
e-mail: A.Karbowski@ia.pw.edu.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

ASimJava: a Java-based library for distributed simulation

Ewa Niewiadomska-Szynkiewicz and Andrzej Sikora

Abstract—The paper describes the design, performance and applications of ASimJava, a Java-based library for distributed simulation of large networks. The important issues associated with the implementation of parallel and distributed simulation are discussed. The focus is on the effectiveness of different synchronization protocols implemented in ASimJava. The practical example–computer network simulation–is provided to illustrate the operation of the presented software tool.

Keywords—parallel computations, parallel asynchronous simulation, computer networks simulation.

1. Introduction

The main difficulty in networks simulation is the enormous computational power needed to execute all events involved by packets transmission through the network. Recently computer network simulation has been an active research area. Numerous software systems have been engineered to aid researchers. The popular commercial and publicly released packet-level simulators like OPNET, NS [10] or OMNeT [11] require costly shared-memory supercomputers to run even medium size network simulation. Since parallel and distributed simulation is fast becoming the dominant form of model execution, the focus is on experiments carried on parallel and distributed software platforms. High level architecture (HLA) [2] standard for distributed discrete-event simulation was defined by the United States Department of Defense. During last years numerous integrated environments for parallel and distributed processing were developed [13]. These software tools apply different techniques for synchronization and memory management, and focus on different aspects of distributed implementation. Many of them are built in Java. SimJava [6] was among the first publicly released simulators written in Java.

The paper deals with the description of an integrated framework for distributed simulation. Asynchronous Simulations Java (ASimJava) can be used to perform simulation experiments carried out on parallel computers or computer networks. It is general purpose environment that can support the researchers of different types of complex systems, but the focus is on communication and computer networks. It targets a variety of potential simulation problems. Two types of networks simulators can be developed using ASimJava:

- connection-level (involving no packet-level operation) simulator for supporting service level agreement (SLA) negotiation process and resource management,
- detailed, involving packet-level operation simulator for testing and analyzing all proposed decision mechanisms.

The paper describes the organization, implementation and usage of ASimJava. Presented practical examples show the range of applications of the discussed software tool.

2. Description of ASimJava

2.1. General description

The ASimJava general structure enables to do parallel and distributed discrete-event simulations, [1] that can be described in terms of logical processes (LPs) and communicate with each other through message-passing. LPs simulate the real life physical processes PPs. Each logical process starts processing as a result of event occurrence (from the event list or having received a new message). It performs some calculations and generates one or more messages to other processes. The calculation tasks executed in parallel require explicit schemes for synchronization. Two simulation techniques are considered [5]: synchronous and asynchronous. Synchronous simulation is implemented by maintaining a global clock (global virtual time–GVT). Events with the smallest time-stamp are removed from the event lists of all LPs, for parallel execution. Parallelism of this technique is limited, because only events with time-stamps equal to that of the global clock can be executed during an event cycle. Asynchronous simulation is much more effective due to its potentially high performance on a parallel platform. In asynchronous simulation each logical process maintains its own local clock (local virtual time–LVT). Local times of different processes may advance asynchronously. Events arriving at the local input message queue of a logical process are executed according to the local clock and the local schedule scheme. Synchronization mechanisms fall into two categories: conservative and optimistic. They differ in their approach to time management. Conservative schemes avoid the possibility of causality error occurring. These protocols determine safe events, which can be executed. Optimistic schemes allow occurrence of causality errors. They detect such error and provide mechanisms for its removal. The calculations are rolled back to a consistent state by sending out antimessages. It is obvious that in order to allow rollback all results of previous calculations have to be recorded.

2.2. System architecture

One of the principle goals of the ASimJava was portability and usage in heterogeneous computing environments. Two versions of ASimJava are implemented: parallel and distributed. It is possible to join both of them in one simulator (Fig. 1). The Java messaging service (JMS) API provided by Sun Microsystems is used for interprocess communication in the case of distributed version. The asynchronous version of distributed simulation is applied in ASimJava.

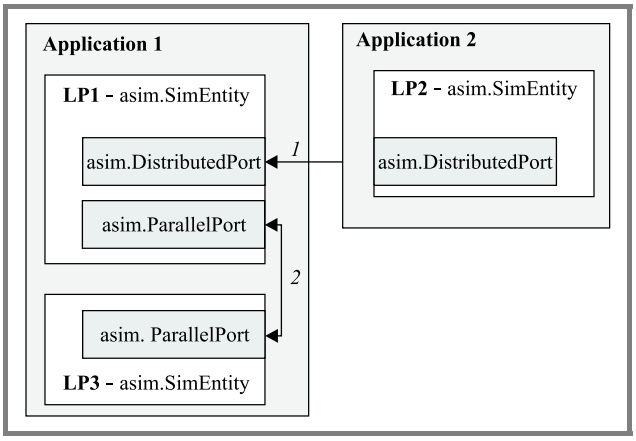


Fig. 1. Combined parallel and distributed simulation. Explanations: *1*–distributed connection, *2*–parallel connection.

Four synchronization protocols are provided:

– conservative protocol with null messages (CMB),

– window conservative protocol (WIN),

– Time Warp (TW),

– Moving Time Window protocol (MTW).

The short description of these protocols is presented in the Appendix.

2.3. Design overview

Current version of ASimJava is composed of five components:

- Graphical user interface (GUI)–responsible for user-system interactions.
- Basic library–a collection of classes implementing basic elements of a simulator, such as: logical processes, events, event lists, messages passing, etc.
- Synchronization protocols library–the library of classes implementing four synchronization algorithms (CMB, WIN, TW and MTW).
- Communication library–the library of classes that provides communication between the user interface and the simulator.
- Toolboxes–the collections of classes implementing basic elements of different systems. Currently available: computer network toolbox that is the collection of classes implementing elements of computer networks, such as router, hub, switch, etc. The package is flexible and can be easily extended by other toolboxes of classes, which are specific to a chosen case study.

The simulator built upon ASimJava classes has hierarchical structure. The simulated system is partitioned into several subsystems (subtasks), with respect to functionality and data requirements. They are implemented as LPs. Next, each LP can be divided into smaller LPs. Hence, the logical processes are nested (Fig. 2). Calculation processes belonging to the same level of hierarchy are synchronized.

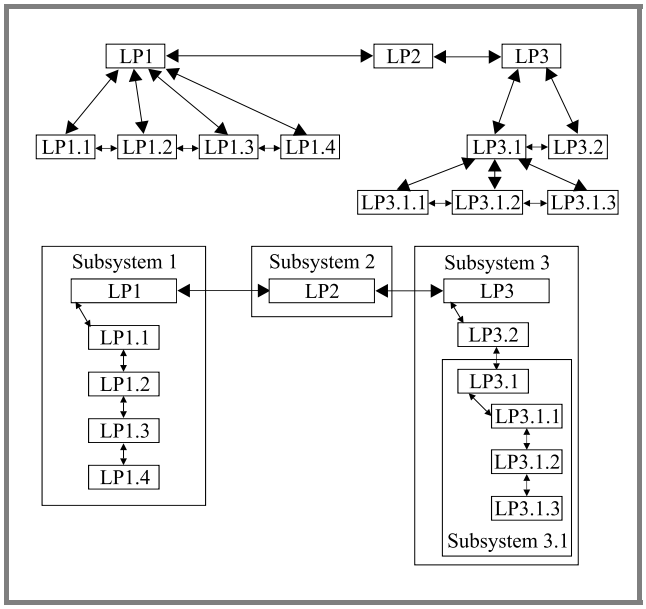


Fig. 2. Simulator structure (example).

Two types of simulators can be distinguished:

1. The simulator consists only of classes provided in ASimJava. The structure of the simulated system together with all model parameters is created using ASimJava graphical interface or may be read from an XML file.
2. New simulator. The user's task is to implement the subsystems' simulators responsible for adequate physical systems simulation. He can create his application applying adequate classes from the ASimJava libraries and including his own code–numerical part of the application.

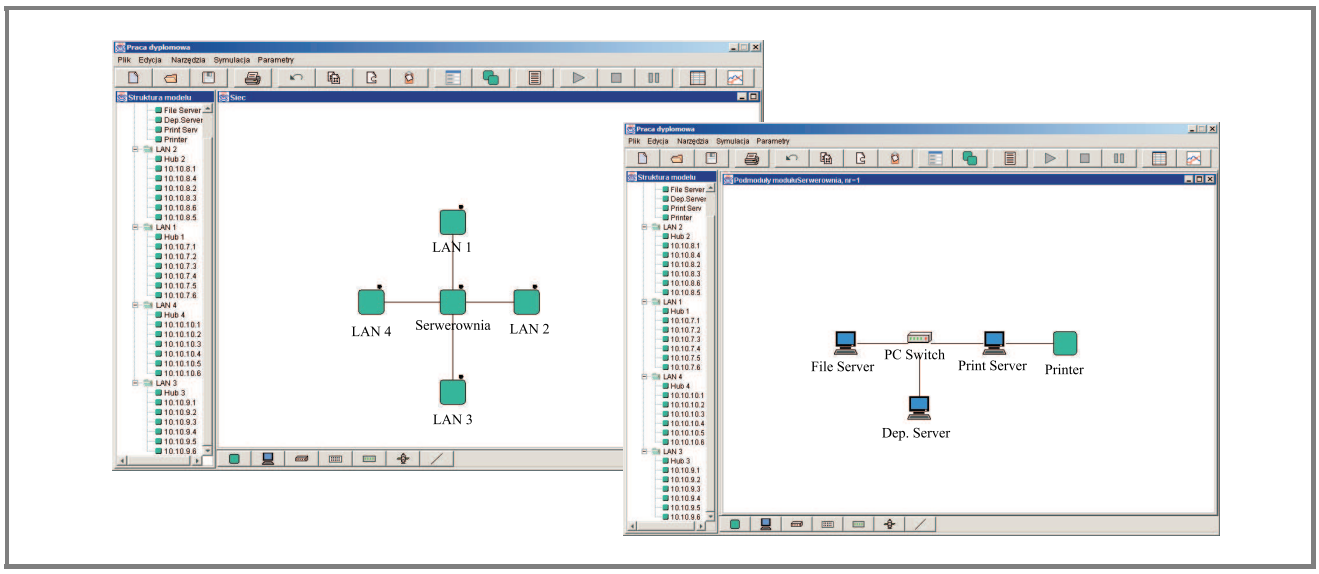


Fig. 3. The simulated system scheme inserting.

2.4. User interface

ASimJava provides the graphical environment (shell) for supporting the considered case study implementation. The most important tasks of the user interface are as follows:

- supporting the process of defining a considered application,
- presenting of the calculation results,
- providing the communication with the user (during design and experimental phases).

The main element of the interface is the graph editor–the graphical tool for inserting the scheme of the simulated system. It is organized in nested manner, too. The user can start from dividing the considered system into several subsystems and inserting the scheme of it. Next, he can divide each subsystem into smaller ones (Fig. 3). For example, in the case of computer networks we can start from the decomposition of our application into local area networks (LANs), and next we can insert all elements that form each LAN (work stations, routers, switches, etc.).
Within the next step the user is asked to provide some information related to the nodes and arcs of the inserted graph. In the case of nodes the required data are: parameters specific to the edited element and event list, in the case of arcs–maximal and minimal flows. The created graph together with all inserted data can be saved into the XML file.

2.5. Simulation under ASimJava

The experimental phase begins when all decisions regarding the simulated system are made. The simulation starts. The adequate programs corresponding to the nodes of the system graph are executed. The results of the calculations are displayed. The user employs monitoring and analysis of the current situation. All results may be recorded into the disc file during the experiment.

3. Case study results

The AsimJava library was used to perform simulation of several systems. In this paper the applications of a simple manufacturing system and a computer network are presented. The objective of all tests was to compare the effectiveness of considered synchronization protocols.

3.1. Manufacturing system

The first case study was related to simulation of a simple distributed manufacturing system, as presented in Fig. 4. A considered system consists of two sources P1 and P2, eight work stations P3–P10 and one sink P11. Jobs enter the manufacturing system at work stations P3 or P4. When a job has been serviced at the work station Pi it proceeds to the next work station. Service times at different work stations are different. Jobs may be queued at a station

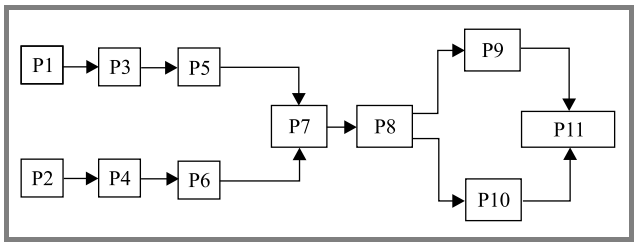


Fig. 4. Example 1: manufacturing system.

awaiting service. A work station takes one job from its input queue when it is free, services that job, and then sends it to the queue of the following work station. All work stations service the jobs in a first come, first served

basis. The job leaves the system after being serviced at work station P9 or P10. It is collected at the sink P11. Simulation experiments were performed under following assumptions:

- Two variants of application were considered:
  - variant *A*: each source generated 2 jobs;
  - variant *B*: each source generated 25 jobs.
- The time intervals in which the sources generate jobs and the service times at different work stations are given in the Table 1.
- The experiments were performed in the network of four Celeron 433 computers. The allocation of LPs simulating adequate physical processes to the computers was as follows: computer *1*: LP1, LP3, LP5; computer *2*: LP2, LP4, LP6; computer *3*: LP7, LP8; computer *4*: LP9, LP10, LP11.

Table 1
Service times (two variants of application)

Variant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
A	2	3	2	3	3	3	7	5	3	3
B	2	5	7	2	2	8	9	5	7	1

The sources, the sink and each work station were simulated by 11 logical processes. If a job j arrived at Pi at time t then its service begins either immediately (at time t), if Pi is idle, or it begins right after the departure of the $(j-1)$ job from this Pi. Let t_{A_j} be the time of arrival of job j at Pi, t_{D_j} the time of departure of job j from Pi and Δt_j the service time at this Pi. Then we obtain:

$$t_{D_j} = \max(t_{A_j}, t_{D_{j-1}}) + \Delta t_j \,. \tag{1}$$

The results of numerical experiments are presented in Tables 2 and 3. Different aspects were considered with respect to (w.r.t.) applied synchronization protocols:

- time of simulation;
- number of additional messages sent by calculation processes (CMB–null messages, WIN–global messages used for lengths of time windows calculation, TW–antimessages, MTW–antimessages and additional messages used to synchronization);
- number of rollbacks at TW;
- different lengths of time window at MTW.

It can be observed (Table 2) that the speed of simulation strongly depends on the applied protocol. The best results were obtained for hybrid approach MTW, the worst for conservative CMB. This was connected with different degrees of parallelism in the case of conservative and optimistic approaches, and overheads (additional messages, rollbacks). The hybrid techniques seem to be promising w.r.t. optimistic TW. It may be profitable to decrease degree of available parallelism, increase the number of additional messages but reduce the number of rollbacks. In the case of combining TW and window techniques, the main problem is to estimate the proper length of the window. The simulation results strongly depend on this parameter (Table 3).

Table 2
Simulation times

Variant	CMB [s]	WIN [s]	TW [s]	MTW [s]
A	247.83	25.81	34.55	20.93
B	316.86	88.43	52.34	41.25

Table 3
Simulation results for different lengths of time window (MTW algorithm, variant *B*)

Time window size	Simulation time [s]	Additional messages	Number of rollbacks
5	54.87	134	0
10	45.09	71	1
20	41.25	36	1
30	44.87	33	2
40	63.82	34	2
50	59.87	23	1

It seems that the length of each window should be calculated adaptively, taking into account the considered application and available hardware platform. So, the degree of parallelism reduction remains a topic of hot debate.

3.2. Computer network

The second case study was related to simulation of an IP network. The library of classes implementing network hardware (host, hub, switch, router, etc.) was developed. The experiments were performed for the network consisting of five LANs and one server room (13 servers), as presented in Fig. 5. Two scenarios of traffic with small packets of transmitted data (variant *A*) and large packets (variant *B*) were considered.

Table 4
Simulation times–computer network

Variant	WIN [s]	MTW-1 [ms]	MTW-10 [ms]
A	18 186	18 090	15 742
B	72 072	71 097	63 996

The experiments were performed in the network of four Celeron 433 computers. The allocation of logical processes simulating the considered computer network to the com-

puters was as follows: computer *1*: LAN1, LAN2; computer *2*: LAN3, LAN4; computer *3*: LAN5; computer *4*: server room.
The results obtained for two synchronization protocols–conservative WIN and hybrid MTW are presented in Table 4. Two lengths of the window in MTW (1 time unit and 10 time units) were taken into account. The application of CMB and TW synchronization protocols brought much worse results–the computation time increased seriously.

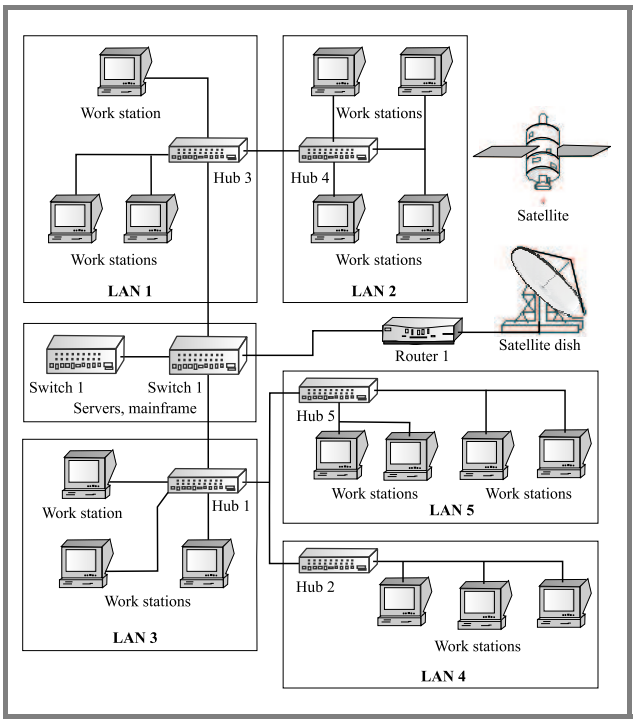


Fig. 5. Example 2: computer network.

Similarly to the previous example the best results were obtained for hybrid protocol.

4. Conclusions

Simulation plays an important role in the computer-aided analysis and design of large computer networks. The ASimJava software framework is suitable to solve many small and large scale problems, based on simulation. The package is flexible and can be easily extended by software modules, which are specific to a chosen application.

Appendix

Synchronization protocols

A.1. CMB protocols

In the case of the CMB scheme [7] each process executes events only if it is certain that no event with the earlier time stamp can arrive. At current time t the ith logical process LPi computes the minimum time $LVT_i = \min_{j=1,\ldots,p_i} t_{ij}$, where t_{ij} denotes the time stamp of the last message received from process LPj and p_i is a number of processes transmitting data to LPi. Next, every LPi simulates all events with time stamps less than its LVT_i. The CMB scheme in its basic form may lead to deadlock. Several approaches were proposed to resolution of deadlock:

- **Null messages**. The additional messages (null messages) [7], are sent by each process to the others, after completing the iteration. They are used to announce absence of messages. The information about the earliest possible time of the next event execution may be appended to the null messages. The evident drawback is the high volume of null messages, especially when we consider the high cost of message-passing in distributed memory machines.
- **Carrier null messages**. The additional messages for processes synchronization propagated through a system carry more information: the list of visited logical processes and pending event time [9].

A.2. Window protocols

Some of conservative algorithms assume constrain concurrent simulation activity to be within a window of global synchronization [1, 8, 9]. A minimum time window is identified for all logical processes, which calculations may be carried out concurrently. Typically, this calculation involves lookahead of some kind. At current time t each logical process LPi is asked to compute the time $T_i(t)$ of the next message it will send, based on its event list. The global time window $[t, T(t))$ is defined; where $T(t) = \min_{i=1,\ldots,p} T_i(t)$, and p is a number of logical processes. Every process can simulate all events with time stamps within this window. Next, $t = T$ and new window is calculated. The evident constraint on this scheme is that only the events within the same window can be executed concurrently. The events performed at the different windows are executed sequentially. The efficiency of this algorithm strongly depends on the calculations decomposition and allocation to the processors.

A.3. Time Warp protocols

The optimistic scheme–Time Warp protocol [3, 4] allows to each LPi keeps calculations in its local simulated time (LVT_i) under the assumption that message communication between processors arrives at proper time. In the case when causality error occurs, e.g., because LPi receives a message with a time stamp smaller than LVT_i, the calculations are rolled back. The state of LPi is restored to that, which existed prior to processing the event. Additional messages, i.e., antimessages are sent to cancel the previously sent messages. After receiving the antimessage the receiver is rolled back, possibly generating additional antimessages. In order to allow rollback, all results of previous calculations has to be recorded, which is connected with requirements of extra resources for states of all processes storing during calculation.

A.4. Moving Time Window protocol

The principal advantage of Time Warp over presented conservative schemes is that it offers the potential for greater exploitation of parallelism to the programmer. Despite this advantage Time Warp has some drawbacks: an excessive amount of wasted, rolled back computations and inefficient use of memory. The danger is greatest when interaction between processors is light and processors loads are uneven. It may result cascading rollbacks. Because of that, the issue of combining conservative and optimistic approaches in a hybrid protocols has received considerable attention in recent years. MTW belongs to the group of protocols, which reduce degree of available parallelism. It limits execution of Time Warp to the defined time window $[t, t+\Delta t)$ [12]. Events with time stamps greater than or equal $t+\Delta t$ are not executed. All processes are synchronized at time $t+\Delta t$ and a new window is simulated. The rollbacks may occur only within the current time window. The problem is to calculate adequate Δt.

References

[1] *Handbook of Simulation*, J. Banks, Ed. New York: Wiley, 1998.

[2] "HLA (high level architecture)", 1998, http://www.dmso.mil/public/transition/hla/

[3] D. A. Jefferson, "Virtual time", *ACM Trans. Programm. Lang. Syst.*, vol. 7, no. 3, pp. 404–425, 1985.

[4] D. A. Jefferson, "Virtual time II: the cancelback protocol for storage management in Time Warp", in *Proc. 9th Ann. ACM Symp. Princ. Distr. Comput.*, New York, USA, 1990, pp. 75–90.

[5] *Systems Modeling and Computer Simulation*, N. A. Kheir, Ed. New York: Marcel Dekker, 1996.

[6] W. Kreutzer, J. Hopkins, and van M. Mierlo, "SimJava–a framework for modeling queueing networks in Java", in *Proc. 1997 Winter Simul. Conf.*, Atlanta, USA, 1997, pp. 483–488.

[7] J. Misra, "Distributed discrete-event simulation", *Comput. Surv.*, vol. 18, no. 1, 1986.

[8] D. M. Nicol, "Performance bounds on parallel self-initiating discrete event simulations", *ACM Trans. Model. Comput. Simul.*, vol. 1, no. 1, pp. 24–50, 1991.

[9] D. M. Nicol and R. Fujimoto, "Parallel simulation today", *Ann. Oper. Res.*, vol. 53, pp. 249–285, 1994.

[10] "NS-2 (network simulator)", 1995, http://www.isi.edu/nsnam/ns/ns-documentation.html

[11] "OMNeT++ (objective modular network testbed in C++)", 1992, http://www.hit.bme.hu/phd/vargaa/omnetpp.htm

[12] L. M. Sokol, D. P. Briscoe, and A. P. Wieland, "MTW: a strategy for scheduling discrete simulation events for concurrent execution", in *Proc. SCS Multiconf. Distr. Simul.*, San Diego, USA, 1988, pp. 34–42.

[13] B. Szymański, A. Saifee, A. Sastry, Y. Liu, and M. Kiran, "Genesis: a system for large-scale parallel network simulation", in *Proc. Parall. Distr. Simul. Conf. PADS 2002*. Washington: IEEE CS Press, 2002.

Ewa Niewiadomska-Szynkiewicz received her M.Sc. in 1986 and Ph.D. in 1996 in control and computer engineering. Since 1987 she is with Warsaw University of Technology and since 2000 with Research and Academic Computer Network (NASK). She is a lecturer (adiunkt) on simulation technologies and optimisation techniques in Institute of Control and Computation Engineering. She was involved in a number of research projects concerned with development of simulation software and techniques for design of on-line operational control. She participated in three European projects within TEMPUS Programme and QoSIPS of 5th Framework Programme project. Her research interests concentrate on computer simulation of large scale systems, data network management and simulation, hierarchical and global optimization, parallel calculations.
e-mail: ewan@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland
e-mail: E.Szynkiewicz@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Andrzej Sikora received his M.Sc. in 2002 from the Warsaw University of Technology. Currently he is a Ph.D. candidate in computer science at the Warsaw University of Technology (Institute of Control and Computation Engineering), Poland. He is a specialist in the application of control and simulation of distributed systems. He has experience in the use of parallel asynchronous computation, modelling, computer simulation, computer networks, Internet techniques, optimisation and decision support, workflow application (Lotus Notes&Domino). His computer skills include programming using Java, C, C++, SQL, HTML, XML, Perl, RMI, JMS.
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

Paper

Data analysis and flow graphs

Zdzisław Pawlak

Abstract—In this paper we present a new approach to data analysis based on flow distribution study in a flow network. Branches of the flow graph are interpreted as decision rules, whereas the flow graph is supposed to describe a decision algorithm. We propose to model decision processes as flow graphs and analyze decisions in terms of flow spreading in the graph.

Keywords—data mining, data independence, flow graph, Bayes' rule.

1. Introduction

In this paper we present a new approach to data analysis based on flow distribution study in a flow network, different to that proposed by Ford and Fulkerson [2], called here a flow graph. Branches of the flow graph are interpreted as decision rules, whereas the flow graph is supposed to describe a decision algorithm. Thus we propose to model decision processes as flow graphs and analyze decisions in terms of flow spreading in the graph.

With every decision rule three coefficients are associated, called strength, certainty and the coverage factors. These coefficients have a probabilistic flavor, but it will be shown in the paper that they can be also interpreted in a deterministic way, describing flow distribution in the flow graph. Moreover, it is shown that these coefficients satisfy Bayes' rule. Thus, in the presented approach Bayes' rule has entirely deterministic interpretation, without reference to its probabilistic nature, inherently associated with classical Bayesian philosophy. This leads to new philosophical and practical consequences. A simple example, of a telecom customer, which is a slight modification of example given in [4], will be used to illustrate ideas presented in the paper.

This paper is a continuation of ideas given in [4–7] and refers to some thoughts presented in [3].

2. Flow graphs

A flow graph is a *directed, acyclic, finite* graph $G = (N, \mathscr{B}, \varphi)$, where N is a set of *nodes*, $\mathscr{B} \subseteq N \times N$ is a set of *directed branches*, $\varphi : \mathscr{B} \to R^+$ is a *flow function* and R^+ is the set of non-negative reals.

If $(x, y) \in \mathscr{B}$ then x is an *input* of y and y is an *output* of x.

If $x \in N$ then $I(x)$ is the set of all inputs of x and $O(x)$ is the set of all outputs of x.

Input and *output* of a graph G *are* defined
$I(G) = \{x \in N : I(x) = \varnothing\}$, $O(G) = \{x \in N : O(x) = \varnothing\}$.

Inputs and outputs of G are *external nodes* of G; other nodes are *internal nodes* of G.

If $(x, y) \in \mathscr{B}$ then $\varphi(x, y)$ is a *troughflow* from x to y. We will assume in what follows that $\varphi(x, y) \neq 0$ for every $(x, y) \in \mathscr{B}$.

With every node x of a flow graph G we associate its *inflow*:

$$\varphi_+(x) = \sum_{y \in I(x)} \varphi(y, x) \tag{1}$$

and *outflow*

$$\varphi_-(x) = \sum_{y \in O(x)} \varphi(x, y). \tag{2}$$

Similarly, we define an inflow and an outflow for the whole flow graph G, which are defined as

$$\varphi_+(G) = \sum_{x \in I(G)} \varphi_-(x), \tag{3}$$

$$\varphi_-(G) = \sum_{x \in O(G)} \varphi_+(x). \tag{4}$$

We assume that for any internal node x, $\varphi_+(x) = \varphi_-(x) = \varphi(x)$, where $\varphi(x)$is a *troughflow* of node x.

Obviously, $\varphi_+(G) = \varphi_-(G) = \varphi(G)$, where $\varphi(G)$ is a *troughflow* of graph G.

The above formulas can be considered as *flow conservation equations* [2]. We will define now a *normalized flow graph*.

A normalized flow graph is a *directed, acyclic, finite* graph $G = (N, \mathscr{B}, \sigma)$, where N is a set of *nodes*, $\mathscr{B} \subseteq N \times N$ is a set of *directed branches* and $\sigma : \mathscr{B} \to <0, 1>$ is a *normalized flow* of (x, y) and

$$\sigma(x, y) = \frac{\varphi(x, y)}{\varphi(G)} \tag{5}$$

is *strength* of (x, y). Obviously, $0 \leq \sigma(x, y) \leq 1$. The strength of the branch expresses simply the percentage of a total flow through the branch.

In what follows we will use normalized flow graphs only, therefore by a flow graphs we will understand normalized flow graphs, unless stated otherwise.

With every node x of a flow graph G we associate its normalized *inflow* and *outflow* defined as

$$\sigma_+(x) = \frac{\varphi_+(x)}{\varphi(G)} = \sum_{y \in I(x)} \sigma(y, x), \tag{6}$$

$$\sigma_-(x) = \frac{\varphi_-(x)}{\varphi(G)} = \sum_{y \in O(x)} \sigma(x, y). \tag{7}$$

Obviously for any internal node x, we have $\sigma_+(x) = \sigma_-(x) = \sigma(x)$, where $\sigma(x)$ is a *normalized troughflow* of x.
Moreover, let

$$\sigma_+(G) = \frac{\varphi_+(G)}{\varphi(G)} = \sum_{x \in I(G)} \sigma_-(x), \tag{8}$$

$$\sigma_-(G) = \frac{\varphi_-(G)}{\varphi(G)} = \sum_{x \in O(G)} \sigma_+(x). \tag{9}$$

Obviously, $\sigma_+(G) = \sigma_-(G) = \sigma(G) = 1$.

3. Certainty and coverage factors

With every branch (x, y) of a flow graph G we associate the *certainty* and the *coverage factors*.
The *certainty* and the *coverage* of (x, y) are defined as

$$cer(x, y) = \frac{\sigma(x, y)}{\sigma(x)} \tag{10}$$

and

$$cov(x, y) = \frac{\sigma(x, y)}{\sigma(y)}, \tag{11}$$

respectively, where $\sigma(x) \neq 0$ and $\sigma(y) \neq 0$.
Below some properties, which are immediate consequences of definitions given above are presented:

$$\sum_{y \in O(x)} cer(x, y) = 1, \tag{12}$$

$$\sum_{x \in I(y)} cov(x, y) = 1, \tag{13}$$

$$\sigma(x) = \sum_{y \in O(x)} cer(x, y)\,\sigma(x) = \sum_{y \in O(x)} \sigma(x, y), \tag{14}$$

$$\sigma(y) = \sum_{x \in I(y)} cov(x, y)\,\sigma(y) = \sum_{x \in I(y)} \sigma(x, y), \tag{15}$$

$$cer(x, y) = \frac{cov(x, y)\,\sigma(y)}{\sigma(x)}, \tag{16}$$

$$cov(x, y) = \frac{cer(x, y)\,\sigma(x)}{\sigma(y)}. \tag{17}$$

Obviously the above properties have a probabilistic flavor, e.g., Eqs. (14) and (15) have a form of total probability theorem, whereas formulas (16) and (17) are Bayes' rules. However, these properties in our approach are interpreted in a deterministic way and they describe flow distribution among branches in the network.
A (*directed*) *path* from x to y, $x \neq y$ in G is a sequence of nodes $x_1, \ldots, x_n$ such that $x_1 = x$, $x_n = y$ and $(x_i, x_{i+1}) \in \mathscr{B}$ for every i, $1 \leq i \leq n-1$. A path from x to y is denoted by $[x \ldots y]$.

The *certainty*, the *coverage* and the *strength* of the path $[x_1 \ldots x_n]$ are defined as

$$cer[x_1 \ldots x_n] = \prod_{i=1}^{n-1} cer(x_i, x_{i+1}), \tag{18}$$

$$cov[x_1 \ldots x_n] = \prod_{i=1}^{n-1} cov(x_i, x_{i+1}), \tag{19}$$

$$\sigma[x \ldots y] = \sigma(x)\,cer[x \ldots y] = \sigma(y)\,cov[x \ldots y], \tag{20}$$

respectively.
The set of all paths from x to $y (x \neq y)$ in G denoted $< x, y >$, will be called a *connection* from x to y in G. In other words, connection $< x, y >$ is a sub-graph of G determined by nodes x and y.
For every connection $< x, y >$ we define its certainty, coverage and strength as shown below:

$$cer < x, y > = \sum_{[x \ldots y] \in < x, y >} cer[x \ldots y], \tag{21}$$

the *coverage* of the connection $< x, y >$ is

$$cov < x, y > = \sum_{[x \ldots y] \in < x, y >} cov[x \ldots y], \tag{22}$$

and the *strength* of the connection $< x, y >$ is

$$\sigma < x, y > = \sum_{[x \ldots y] \in < x, y >} \sigma[x \ldots y] = \sigma(x)\,cer < x, y > = \sigma(y)\,cov < x, y > . \tag{23}$$

Let $[x \ldots y]$ be a path such that x and y are input and output of the graph G, respectively. Such a *path* will be referred to as *complete*.
The set of all complete paths from x to y will be called a *complete connection* from x to y in G. In what follows we will consider complete paths and connections only, unless stated otherwise.
Let x and y be an input and output of a graph G respectively. If we substitute for every complete connection $< x, y >$ in G a single branch (x, y) such $\sigma(x, y) = \sigma < x, y >$, $cer(x, y) = cer < x, y >$, $cov(x, y) = cov < x, y >$ then we obtain a new flow graph G' such that $\sigma(G) = \sigma(G')$. The new flow graph will be called a *combined* flow graph. The combined flow graph for a given flow graph represents a relationship between its inputs and outputs.

4. Flow graph and decision algorithms

Flow graphs can be interpreted as decision algorithm.
Let us assume that the set of node of a graph is interpreted as a set of formulas, denoted Φ, Ψ, etc. The formulas are understood as propositional functions.
Then every branch (Φ, Ψ) can be understood as a decision rule $\Phi \to \Psi$, read if Φ *then* Ψ; Φ will be referred to as a *condition*, whereas Ψ—*decision* of the rule. Such a rule is characterized by three numbers, $\sigma(\Phi, \Psi)$, $cer(\Phi, \Psi)$ and $cov(\Phi, \Psi)$.

Thus every path $[\Phi_1 \ldots \Phi_n]$ determines a sequence of decision rules $\Phi_1 \to \Phi_2$, $\Phi_2 \to \Phi_3, \ldots, \Phi_{n-1} \to \Phi_n$.
From previous considerations it follows that this sequence of decision rules can be interpreted as a single decision rule $\Phi_1 \Phi_2 \ldots \Phi_{n-1} \to \Phi_n$, in short $\Phi^* \to \Phi_n$, where $\Phi^* = \Phi_1 \wedge \Phi_2 \wedge \ldots \wedge \Phi_{n-1}$, characterized by

$$cer(\Phi^*, \Phi_n) = cer[\Phi_1 \ldots \Phi_n], \tag{24}$$

$$cov(\Phi^*, \Phi_n) = cov[\Phi_1 \ldots \Phi_n], \tag{25}$$

and

$$\sigma(\Phi^*, \Phi_n) = \sigma(\Phi_1)\, cer[\Phi_1 \ldots \Phi_n] = \\ = \sigma(\Phi_n)\, cov[\Phi_1 \ldots \Phi_n], \tag{26}$$

where $\sigma(\Phi)$ is truth value of the formula Φ and $\sigma(\Phi, \Psi)$ in the strength of the decision rule $\Phi \to \Psi$.
Similarly, every connection $< \Phi, \Psi >$ can be interpreted as a single *decision rule* $\Phi \to \Psi$ such that:

$$cer(\Phi, \Psi) = cer < \Phi, \Psi >, \tag{27}$$

$$cov(\Phi, \Psi) = cov < \Phi, \Psi >, \tag{28}$$

and

$$\sigma(\Phi, \Psi) = \sigma(\Phi)\, cer < \Phi, \Psi > = \\ = \sigma(\Psi)\, cov < \Phi, \Psi >, \tag{29}$$

Let $[\Phi_1 \ldots \Phi_n]$ be a path such that Φ_1 is an input and Φ_n an output of the flow graph G, respectively. Such a path and the corresponding connection $< \Phi_1, \Phi_n >$ will be called *complete*.
The set of all decision rules $\Phi_{i_1} \Phi_{i_1} \ldots \Phi_{i_{n-1}} \to \Phi_{i_n}$ associated with all complete paths $\Phi_{i_1} \ldots \Phi_{i_n}$ will be called a *decision algorithm* induced by the flow graph.

5. Dependencies in flow graphs

Let $(x, y) \in \mathscr{B}$. Nodes x and y are independent on each other if

$$\sigma(x, y) = \sigma(x)\, \sigma(y). \tag{30}$$

Consequently

$$\frac{\sigma(x, y)}{\sigma(x)} = cer(x, y) = \sigma(y) \tag{31}$$

and

$$\frac{\sigma(x, y)}{\sigma(y)} = cov(x, y) = \sigma(x). \tag{32}$$

If

$$cer(x, y) > \sigma(y) \tag{33}$$

or

$$cov(x, y) > \sigma(x), \tag{34}$$

then x and y *depend positively* on each other.

Similarly, if

$$cer(x, y) < \sigma(y) \tag{35}$$

or

$$cov(x, y) < \sigma(x) \tag{36}$$

then x and y *depend negatively* on each other.
Let us observe that relations of independency and dependences are symmetric ones, and are analogous to that used in statistics.
For every $(x, y) \in \mathscr{B}$ we define a *dependency factor* $\eta(x, y)$ defined as

$$\eta(x, y) = \frac{cer(x, y) - \sigma(y)}{cer(x, y) + \sigma(y)} = \frac{cov(x, y) - \sigma(x)}{cov(x, y) + \sigma(x)}. \tag{37}$$

It is easy to check that if $\eta(x, y) = 0$, then x and y are independent on each other, if $-1 < \eta(x, y) < 0$, then x and y are negatively dependent and if $0 < \eta(x, y) < 1$ then x and y are positively dependent on each other.

6. Illustrative example

We will illustrate the above ideas by means of a simple tutorial example concerning a telecom provider. This example is a modification of the example given in [4].
Suppose we have three groups of telecom customers classified with respect to age: *young* (*students*), *middle aged* (*workers*) and *old* (*pensioners*). Moreover, suppose we have data concerning place of residence of customers: *town, village* and *country*.
Let us assume that the customers are buying a telecom service in vacation promotion and some of the customers are leaving the telecom provider in 3 months.
The initial data are presented in Fig. 1.

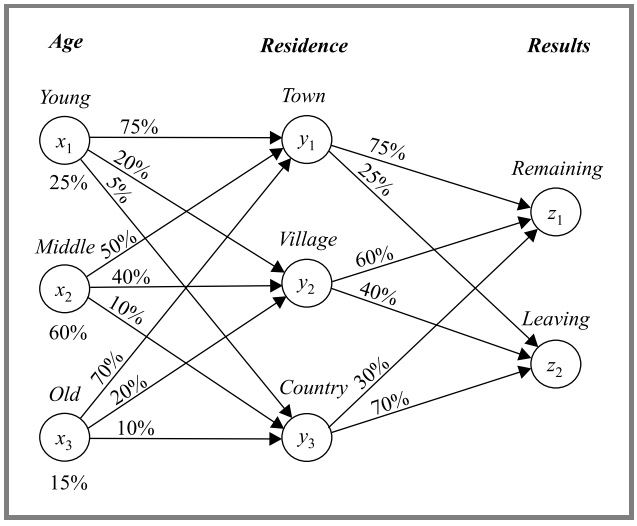


Fig. 1. Initial data.

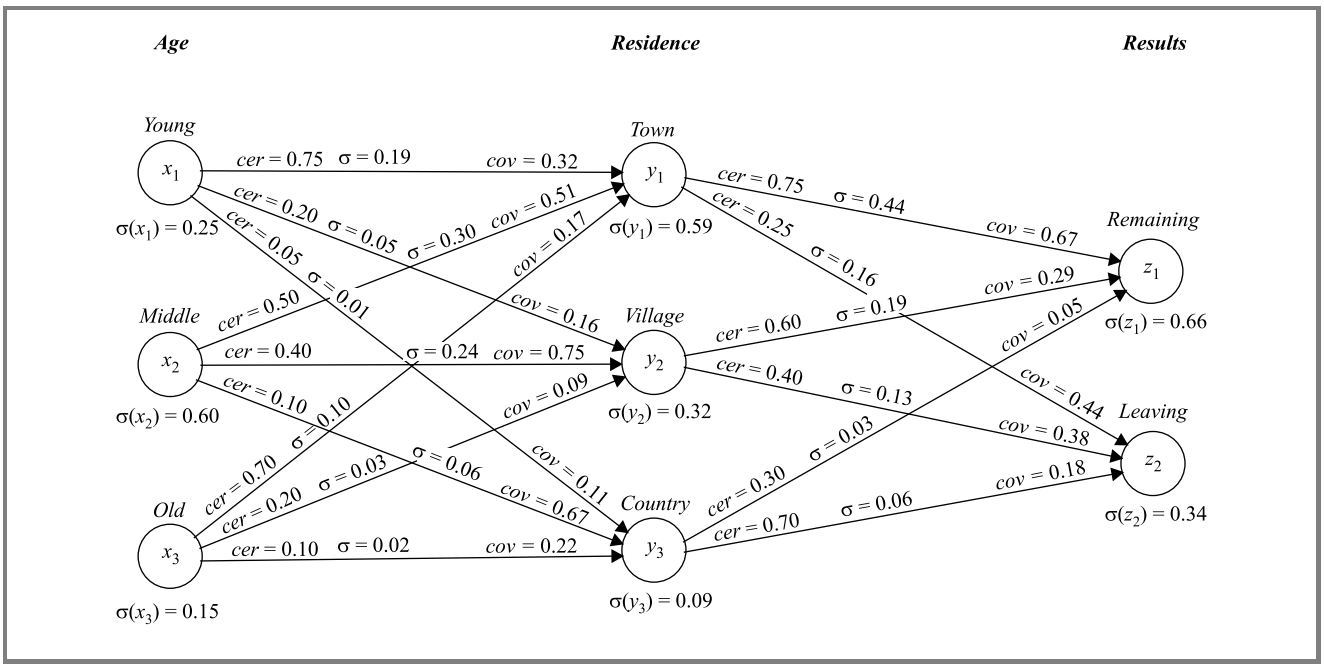


Fig. 2. Final results.

That means that these are 25% young customers, 60%—middle aged and 15% old—in the data base. Moreover, we know that 75% of young customers are living in towns, 20%—in villages and 5%—in the country, etc. We also have from the database that 75% town customers are not leaving the provider, whereas 25% are leaving the provider after 3 month, etc.

We want to find a relationship between the customer's group and the final result of the promotion after three month.

Applying the ideas presented in previous sections we get the results presented in Fig. 2.

Figure 2 shows general structure of patterns between customers and promotion results. Many interesting conclusions can be drawn from the picture, but we leave them for the interested reader.

We might be also interested in finding the relationship between age group and final result of the promotion. To this end we have to eliminate from the flow graph *residence*. In other words we have to compute all connections between

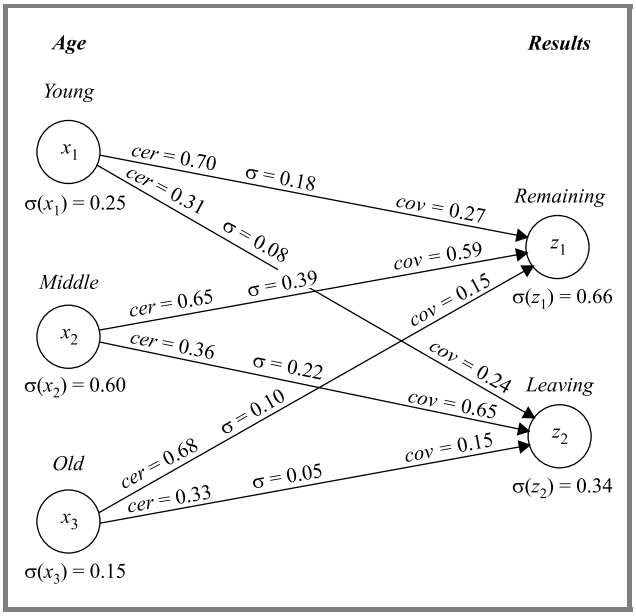


Fig. 3. Simplified flow graph.

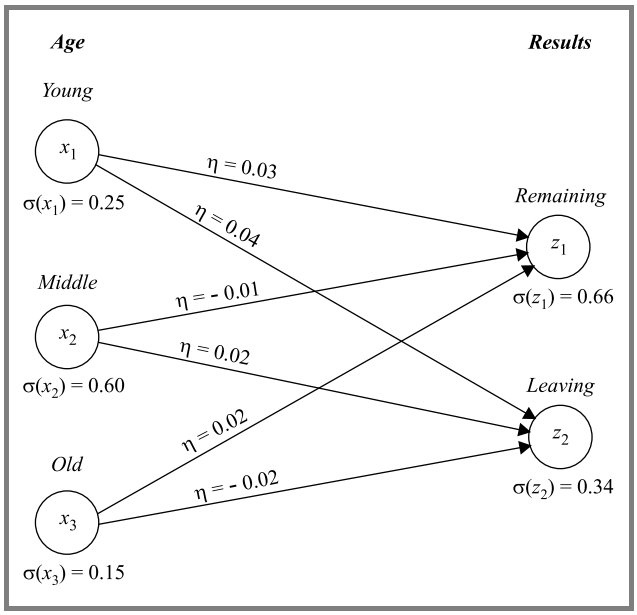


Fig. 4. Dependency coefficient.

age groups and the results, or—the relationship between input and output of the flow graph. The result is shown in Fig. 3.

Many interesting decision rules can be obtained from Figs. 2 and 3. Again we leave the task for the interested reader.

Dependences in flow graph presented in Fig. 3 are shown in Fig. 4.

It can be seen from the flow graph that all the dependency factors are very low and almost close to zero. That means, that in view of the data, practically, there is no relationship between groups of customers and the final result.

7. Conclusions

The paper presents a new approach to decision algorithm analysis. It is revealed that certain classes of decision algorithms can be represented as flow networks, and basic properties of such algorithms can be expressed in terms of flow distribution in a corresponding flow network. A method of simplification of such algorithms is presented.

References

[1] M. Berthold and D. J. Hand, *Intelligent Data Analysis—an Introduction*. Berlin [etc.]: Springer-Verlag, 1999.

[2] L. R. Ford and D. R. Fulkerson, *Flows in Networks*. Princeton [etc.]: Princeton University Press, 1962.

[3] J. Łukasiewicz, "Die logishen Grundlagen der Wahrscheinilchkeitsrechnung" (Kraków, 1913), in *Jan Łukasiewicz—Selected Works*, L. Borkowski, Ed. Amsterdam [etc.], North Holland Publ., Warsaw: Polish Scientific Publ., 1970.

[4] M. Kryszkiewicz, H. Rybiński, and M. Muraszkiewicz, "Data mining methods for telecom providers", Institute of Computer Sciences, Warsaw University of Technology, Research Report 28/02 and "MOST—mobile open society thorough wirless telecommunications technology for mobile society", M. Muraszkiewicz, Ed., Warsaw University of Technology, 2003, pp. 210–220.

[5] Z. Pawlak, "Probability, truth and flow graphs", in *RSKD—Rough Sets in Knowledge Discovery, Proc.*, A. Skowron and M. Szczuka, Eds., Warsaw, Poland, 2003, pp. 1–9.

[6] Z. Pawlak, "Flow graphs and decision algorithms", in *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, G. Wang, Y. Yao, and A. Skowron, Eds., *Lecture Notes in Artificial Intelligence*. Berlin: Springer, 2003, vol. 2639, pp. 1–10.

[7] *Bayes's Theorem*, Proceedings of the British Academy, R. Swinburne, Ed. Oxford University Press, 2002, vol. 113.

Zdzisław Pawlak was born in Łódź (Poland), in 1926. He obtained his M.Sc. in 1951 in electronics from Warsaw University of Technology, Ph.D. in 1958 and D.Sc. in 1963 in the theory of computation from the Polish Academy of Sciences. He is a Professor of the Institute of Theoretical and Applied Informatics, Polish Academy of Sciences and the University of Information Technology and Management and Member of the Polish Academy of Sciences. His current research interests include intelligent systems and cognitive sciences, in particular, decision support systems, knowledge representation, reasoning about knowledge, machine learning, inductive reasoning, vagueness, uncertainty and decision support. He is an author of a new mathematical tool, called rough set theory, intended to deal with vagueness and uncertainty. About two thousand papers have been published by now on rough sets and their applications world wide. Several international workshops and conferences on rough sets have been heldin recent years. He is a recipient of many awards among others the State Award in Computer Science in 1978, the Hugo Steinhaus Award for achievements in applied mathematics in 1989. Doctor honoris causa of Poznań University of Technology in 2002. Member of editorial boards of several dozens international journals. Deputy Editor-in-Chief of the *Bulletin of the Polish Academy of Sciences*. Program committee member of many international conferences on computer science. Over forty visiting university appointments in Europe, USA and Canada, about fifty invited international conference talks, and over one hundred seminar talks given in about fifty universities in Europe, USA, Canada, China, India, Japan, Korea, Taiwan, Australia and Israel. About two hundred articles in international journals and several books on various aspects on computer science and application of mathematics. Supervisor of thirty Ph.D. theses in computer science and applied mathematics.

e-mail: zpw@ii.pw.edu.pl
Institute for Theoretical and Applied Informatics
Polish Academy of Sciences
Bałtycka st 5
44-100 Gliwice, Poland

University of Information Technology and Management
Newelska st 6
01-447 Warsaw, Poland

Paper

Probes for fault localization in computer networks

Wiesław Traczyk

Abstract—Fault localization is a process of isolating faults responsible for the observable malfunctioning of the managed system. This paper reviews some existing approaches of this process and improves one of described techniques—the probing. Probes are test transactions that can be actively selected and sent through the network. Suggested innovations include: mixed (passive and active) probing, partitioning used for probe selection, logical detection of probing results, and adaptive, sequential probing.

Keywords—fault localization, probes, computer networks, partitions, logic design.

1. Introduction

As computer networks increase in size, heterogeneity and complexity, effective management of such networks becomes more important and more difficult. Network management is essential to ensure the good functioning of these networks.

The International Standards Organization has divided network management tasks into six categories, as part of their Open System Interconnection Model. One of these categories—the fault management—can be characterized as detecting when network behavior deviates from normal and formulating a corrective course of action. Fault management deals with [1]:

- *fault detection*, to know whether there is a failure or not in the network;
- *fault localization*, to know which is(are) the component(s) that has/have failed and caused the received alarms;
- *fault isolation* so that the network can continue to operate, which is the fast and automated way to restore interrupted connections;
- *network (re-)configuration* that minimizes the impact of a fault by restoring the interrupted connections using spare equipment;
- *replacement* of the failing component(s).

Fault localization is the core of fault diagnosis and means a process of analyzing external symptoms of network disorder to isolate possibly unobservable faults responsible for the symptoms' occurrences. Traditionally, fault localization has been performed manually by experts but, as systems grew larger and more complex, automated fault localization techniques became critical.

The terms used so far (and in the future) require more precise definitions [1, 2].

Object is a part of the network that has separate and distinct existence. An object can be a node, a layer in a protocol stack, a software process, a virtual link, a hardware component, etc. Objects in a communication system consist of other objects, down to the level of smallest objects which are considered indivisible and called *elements*.

Fault (also referred to as *root problem*) can be defined as an unpermitted deviation of at least one characteristic parameter or variable of a network object from acceptable or usual or standard values. Faults may be classified according to their duration time as *permanent, intermittent* and *transient*.

Error, a consequence of fault, is defined as a discrepancy between observed and correct value. Fault may cause one or more errors.

Failure is an error that is visible to the outside world. Errors may propagate within the network causing failures of faultless hardware or software.

Symptoms are external manifestation of failures. They are observed (and send to the network manager) as *alarms*.

Communication networks are built on several layers, performing each fault management functions independently. When a failure occurs, several symptoms are issued to the network manager from the different management layers, and fault management functions start in parallel. Research is carried out to allow interoperability between different layers, to avoid task duplication and increase efficiency.

This paper discusses some approaches to automated fault localization and presents the new method based on probes.

2. Fault localization techniques

All techniques performing fault diagnosis rely on analysis of symptoms and events (such as warnings and parameters of the network elements) that are generated or detected during the occurrence of the fault. One can divide them in two main categories. The first ones are *passive approaches*, which compute fault location hypotheses on the basis of signals, generated by network elements by oneself and sent to management centers. The second ones are *active approaches*, which periodically check the state of the network elements, whether they are correct or not.

Among **passive fault localization techniques** two families of methods are of special interest: artificial intelligent (AI) methods and fault propagation methods.

Artificial intelligent techniques for fault localization. This the most widely used family contains a lot of methods and appropriate systems [1–4].

Model-based systems construct an abstract model of the network. The model represents the network topology and is able to generate predictions of the normal behavior of the system. This predictions are compared with network observations and used for obtaining fault hypotheses. Depending of the kind of model, different approaches can be used: deterministic, probabilistic, temporal, finite state machines, etc. The advantages of these systems are that they are able to cope with incomplete information and with unforeseen failures. The drawback is the difficulty of developing good model for large networks and computation complexity.

Rule-based systems describe human expert knowledge in the form of decision rules, linking logical description of the network state (rules conditions) with partial or final localization hypotheses (rules conclusions). These systems do not require profound understanding of the architectural and operational principles of the network, and can effectively take human expertise into account. The disadvantages of rule-based systems are: the translation of human expertise into the set of rules, which cover all cases in an exhaustive manner is hard, and the need to search for all possible fault hypotheses slows down the global functioning of the system.

Case-based systems make their decisions based on experience and past situations. They try to acquire relevant knowledge of past cases and previously used solutions to propose solutions for new problems. If these solutions can not be taken directly from the case-base and need special reasoning on the base of closely matched situations, case-based systems are computationally complex. Their advantages are efficiency and speed when the submitted problem was previously solved, and on-line learning that allows storing newly solved cases.

All three described systems are the special cases of *expert systems* or *knowledge-based systems* but since the most popular expert systems are rule-based, sometimes only these systems are known as expert systems.

Some other AI techniques (neural networks, decision trees, etc.) are rarely used in these applications.

Fault propagation methods [2, 5, 6]. This family of techniques require a priori specification of how a failure condition in one object is relevant to failure condition in other object. It is important because some errors can propagate failures through the network, generating many different alarms.

Code-based techniques use causality graph model to describe the cause-and-effect relationships between network events. For each problem and each symptom a unique binary code is assigned, and fault propagation patterns are represented by a codebook. Fault localization is performed by finding a fault whose code is the closest match to the code of symptoms. For small systems this technique is very effective.

Bayesian networks take into account uncertainty about dependencies within the managed network and about the set of observed symptoms. Uncertainty is represented by probabilities in a believe (Bayesian) network. The best symptom explanation is a result of Bayesian inference. The method is computationally complex and needs many values of events probability.

Dependency graph is a directed graph whose nodes correspond to objects and whose edges denote the fact that a fault in starting object may cause a fault in ending object. Probabilities may be assigned to nodes and edges, describing uncertain relationships and events. Comparing a state of the graph with known state of the network one can find the source of fault symptoms.

Active fault localization techniques construct managing tools which, instead of waiting for symptoms from the network, ask objects about their state and parameters. These techniques are not so popular as passive approaches but in some cases they may be very useful and therefore deserve attention.

Intelligent agents [8] are simply software processes that live on every managed node, collecting, forwarding and setting management information, either at predefined intervals or when requested to by management station.

Monitoring technique [9] locate in some network nodes the computers *(monitors)* which are guaranteed by self-testing. Each monitor tests the adjacent nodes and links, and sends results of testing to the management station. Proper number of monitors can cover all nodes and links in the network. More advanced technique starts from only one monitor. Its adjacent nodes that pass the tests can became new monitors, then test their non tested adjacent nodes and connected links, and so on.

Probing technique [10, 11] use an active measurement approach, called *probing*. A probe is a program that executes on a particular machine (called a *probe station*) by sending a command or transaction to a server or network element, and measuring the response. The objects represented by nodes may be physical entities such as routers, servers and links, or logical entities such as software components, database tables, etc. It is assumed that each node of the tested network can be either "up", functioning correctly, or "down", not functioning correctly. A probe either succeeds or fails: if it succeeds, then every object it tests is up; it fails if any of the objects it tests are down.

Fault localization attempts to determine the state of the system from the probe results, so effectiveness of localization depends on the number of probes and their paths. For practical networks the problem of achieving the minimal set of effective probes is solved only approximately. In the next section known probing technique will be modified, with the goal to simplify its practical applications.

Each of presented above (and many others) approaches has some advantages and drawbacks thus further research, improving existing methods, is still needed.

3. Probes for locating failures

Probes technique is already used by IBM's EPP technology and seems to be promising. The main problem with it is the need for effective algorithm of probes generation. Minimizing the number of probes is important because probing increases network overhead, probe results must be stored and analyzed, and modifications enforced by changes in network configuration are simpler for smaller set of probes. Active probe selection [12] gives some positive results but is based on the prior probability distribution over system states, difficult to achieve.

Approach proposed here tries to improve probing by:

1) application of mixed (passive and active) technique;
2) partitioning used for probe selection;
3) logical detection of probing results;
4) adaptive probing;

it is assumed that symptoms received by managing center refer to high level of the network topology, signaling defects of the whole path, with many nodes (objects).

The set of A tested nodes will be denoted by $\mathbf{N} = \{N_1, N_2, \ldots, N_A\}$ with node name N_i from the set of natural numbers (for simplicity). Binary state $n_i \in \{0,1\}$ of each node N_i equals 1 if node N_i is correct and equals 0—if it is not. The state of the whole network is described by the binary vector $\mathbf{n} = \langle n_1, n_2, \ldots, n_A \rangle$.

A probe S_j is represented by the set of tested nodes $S_j = \{N_{j_a}, N_{j_b}, \ldots, N_{j_z}\}$, and the set $\mathbf{S} = \{S_1, S_2, \ldots, S_B\}$ designates all probes used for tests.

Mixed technique. In the conventional method the set **N** consists of all nodes in the network, what makes construction of the probes set **S** very difficult and prolongs the time of testing. Instead, we can wait for symptoms generated by network equipment (passive part) and on this base to fix the set of nodes, suspected for malfunctioning. Usually it is not difficult, especially if symptoms concern communication paths or channels. Suspected set of nodes is much smaller then the whole set of nodes, even if it is defined approximately and in excess. Only this smaller set is traversed by probes (active part of the technique).

Partitions for probe selection. To describe probes needed for a fault localization one can use calculus of partitions.

Partition $\pi(X)$ of a set X is a family of subsets X_k (*blocs*), such that $\pi(X) = \{X_1, X_2, \ldots X_K\}$ and for each i, j there is $X_i \cap X_j = \emptyset$ and $X_1 \cup X_2 \cup \ldots \cup X_K = X$. A block of partition π_l of the set X is denoted as X_{π_l}. *Product* of two partitions π_1 and π_2 of the same set X is a partition $\pi(X) = \pi_1(X).\pi_2(X)$ such that for all blocks X_{π_1}, X_{π_2} there exists a block X_π such that $X_\pi = X_{\pi_1} \cap X_{\pi_2}$. Partition with K blocks of the set with K elements is marked as π_0.

Each probe may be considered as partition of the set of nodes from **N**, with two blocks: the first block consists of all nodes tested by this probe and the second block contains all remaining nodes from **N**. If $\pi_j(\mathbf{N})$ refers to a probe S_j then $\pi_j(\mathbf{N}) = \{N_{j_a}, N_{j_b}, \ldots, N_{j_z}; \mathbf{M}_j\}$, where $\mathbf{M}_j$ is a set of remaining nodes.

The results of B probes have to separate one faulty node from A nodes. It can be done if the product of partitions related to probes gives the partition π_0, i.e., $\pi_1.\pi_2 \ldots \pi_B = \pi_0$. It is important advice for probe selection.

Usually managers are able to define the set of probes easily generated by probe stations. From this set a special algorytm should select these probes which contain nodes from **N**. Desirable probes contain the number of nodes nearing the value A/2, because in this case their informative power is the highest. Since B probes can distinguish 2^B nodes, in the optimal case $A \approx 2^B$. When the product of partitions describing the best probes is not equal π_0, elements of blocks with more then one node should be separated by additional probes with appropriate partitions.

For example if $\mathbf{N} = \{1,2,3,4,5\}$ and the two primarily selected probes have partitions $\pi_1 = \{1,2,3; \mathbf{M}_1\}$ and $\pi_2 = \{3,4,5; \mathbf{M}_2\}$, then

$$\pi = \pi_1.\pi_2 = \{1,2;\ 3;\ 4,5\} \neq \pi_0.$$

Two probes can separate node 1 from 2 and node 4 from 5. Choosing $S_3 = \{1,4\}$, i.e., $\pi_3 = \{1,4;\mathbf{M}_3\}$ we have

$$\pi_1.\pi_2.\pi_3 = \{1;\ 2;\ 3;\ 4;\ 5\} = \pi_0.$$

It means that probes S_1, S_2, S_3 create the minimal set localizing faults in 5-object network.

Detection of probing results. Each probe S_j from the set **S** may give positive or negative result of testing. Having all these results we should compute the number(s) of node(s) with a fault. Logical functions will help in this task.

A probe $S_j = \{N_{j_a}, N_{j_b}, \ldots, N_{j_z}\}$ will be described by conjunction of logical variables v^α:

$$\eta(S_j) = V_j = v_{j_a}^{\alpha_a}.v_{j_b}^{\alpha_b}.\ldots.v_{j_z}^{\alpha_z}.$$

Here $\alpha \in \{0,1\}$, $v^0 = \overline{v}$, $v^1 = v$, and v_x^1 means that a node N_x is correct and v_x^0 means that it is not. Similarly V_j is used if the result of probe S_j is positive and $\overline{V_j}$—if it is negative.

Total result of testing with the set of probes $\mathbf{S} = \{S_1, S_2, \ldots, S_B\}$ will be described by conjunction $\mathbf{V} = \eta(\mathbf{S})$, with differentiated formulas **V**, depending on the result of probing:

$\mathbf{V}_0 = V_1.V_2.\ldots.V_B$—if all probes gave positive result,
$\mathbf{V}_1 = \overline{V_1}.V_2.\ldots.V_B$—if only first probe gave negative result,
$\mathbf{V}_{1,2} = \overline{V_1}.\overline{V_2}.V_3.\ldots.V_B$—if two first probes gave negative result,

and so on.

Decimal pointers can be obtained by the set $\Gamma(\alpha) = \{i|\alpha_i = 0\}$, taking values of α from notations V^{α}:

$$\mathbf{V}_{\Gamma(\alpha)} = V_1^{\alpha_1}.V_2^{\alpha_2}.\ldots.V_B^{\alpha_B}.$$

When all symbols V^{α} are substituted by the appropriate conjunctions and reformulated, final formula shows the nodes with a fault.
Continuing the example—if probes

$$S_1 = \{1,2,3\} \qquad S_2 = \{3,4,5\} \qquad S_3 = \{1,4\}$$

are used for testing, results can be computed from the following equations:

$$\begin{aligned}
\mathbf{V}_0 &= V_1.V_2.V_3 = v_1.v_2.v_3.v_3.v_4.v_5.v_1.v_4 = \\
&= v_1.v_2.v_3.v_4.v_5, \\
\mathbf{V}_1 &= \overline{V}_1.V_2.V_3 = (\overline{v_1} \vee \overline{v_2} \vee \overline{v_3}).v_3.v_4.v_5.v_1.v_4 = \\
&= \overline{v_2}.v_1.v_3.v_4.v_5, \\
\mathbf{V}_2 &= V_1.\overline{V_2}.V_3 = v_1.v_2.v_3.(\overline{v_3} \vee \overline{v_4} \vee \overline{v_5}).v_1.v_4 = \\
&= \overline{v_5}.v_1.v_2.v_3.v_4, \\
\mathbf{V}_3 &= V_1.V_2.\overline{V_3} = v_1.v_2.v_3.v_3.v_4.v_5.(\overline{v_1} \vee \overline{v_4}) = 0, \\
\mathbf{V}_{1,2} &= \overline{V_1}.\overline{V_2}.V_3 = (\overline{v_1} \vee \overline{v_2} \vee \overline{v_3}).(\overline{v_3} \vee \overline{v_4} \vee \overline{v_5}).v_1.v_4 = \\
&= \overline{v_3}.v_1.v_4, \\
\mathbf{V}_{1,3} &= \overline{V_1}.V_2.\overline{V_3} = (\overline{v_1} \vee \overline{v_2} \vee \overline{v_3}).v_3.v_4.v_5.(\overline{v_1} \vee \overline{v_4}) = \\
&= \overline{v_1}.v_3.v_4.v_5, \\
\mathbf{V}_{2,3} &= V_1.\overline{V_2}.\overline{V_3} = v_1.v_2.v_3.(\overline{v_3} \vee \overline{v_4} \vee \overline{v_5}).(\overline{v_1} \vee \overline{v_4}) = \\
&= \overline{v_4}.v_1.v_2.v_3, \\
\mathbf{V}_{1,2,3} &= \overline{V_1}.\overline{V_2}.\overline{V_3} = \\
&= (\overline{v_1} \vee \overline{v_2} \vee \overline{v_3}).(\overline{v_3} \vee \overline{v_4} \vee \overline{v_5}).(\overline{v_1} \vee \overline{v_4}) = \\
&= \overline{v_1}.\overline{v_3} \vee \overline{v_1}.\overline{v_4} \vee \overline{v_1}.\overline{v_5} \vee \overline{v_2}.\overline{v_4} \vee \overline{v_3}.\overline{v_4}.
\end{aligned}$$

From these formulas we can conclude: negative result from first probe means that node 2 is not correct, negative result from second probe means that node 5 is not correct, etc. Negative result from probe 3 is impossible, because probes 1 and 2 gave positive result.
Table 1 summarizes the results.

Table 1
Tests results

Tests	Probes				
Negative probes	1	2	1, 2	1, 3	2, 3
Incorrect nodes	2	5	3	1	4

Adaptive probing. In all approaches described above the set of probes **S** was defined on the basis of the whole set of nodes **N**. But the probes can also be defined sequentially:

- Step 1—probe S_1 is fixed for the set $\mathbf{N_1} = \mathbf{N}$.
- Step 2—probe S_2 is fixed for the set $\mathbf{N_2} = \mathbf{M_1}$ if the result of S_1 is positive and for the set $\mathbf{N_2} = \mathbf{N_1} \setminus \mathbf{M_1}$, if it is negative.
- Step 3, 4, ... as above, until all nodes are tested.

Such adaptive probing can be useful if the algorithm defining probes is fast enough.

4. Conclusion

Four suggestions presented in this paper can improve a procedure of probing, but some further research is needed. For the total diagnose of the large set of nodes it will be required to have:

– automatic generation of a set of probes from the network topology,

– additional probe selection for the case with more than one fault,

– additional probe selection for the case with dynamic routing.

References

[1] C. Mas and P. Thiran, "A review on fault location methods and their application to optical networks", *Opt. Netw. Mag.*, vol. 2, no. 4, 2001.

[2] M. Steinder and A. S. Sethi, "The present and future of event correlation: a need for end-to-end service fault localization", in *World Multi-Conf. Syst., Cyber. Inform.*, N. Callaos *et al.*, Ed., Orlando, USA, 2001.

[3] K. Hashimoto *et al.*, "A new diagnostic method using probabilistic temporal fault models", *IEICE Trans. Inform. Syst.*, no. 3, 2002.

[4] S. Bibas *et al.*, "Alarm driven supervision for telecommunication network: i-off-line scenerios generation", *Ann. Telcommun.*, vol. 51, no. 9–10, pp. 493–500, 1996.

[5] I. Katzela and M. Schwartz, "Schemes for fault identification in communication networks", *IEEE Trans. Netw.*, no. 3(6), 1995.

[6] M. Steinder and S. Sethi, "End-to-end failure diagnosis using belief networks", in *Proc. Netw. Oper. Manag. Symp.*, Florence, Italy, 2002.

[7] Chi-Chun Lo *et al.*, "Coding-based schemes for fault identification in communication networks", *J. Netw. Manag.*, no. 10, 2000.

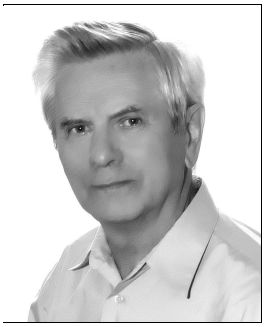
[8] E. U. Ekaette and B. H. Far, "A framework for distributed fault management using intelligent software agents", in *CCEC*, Monteral, Canada, 2003.

[9] H. Masuyama *et al.*, "A diagnosis method of computer networks", http://mylab.tottori-u.ac.jp

[10] M. Brodie *et al.*, "Optimizing probe selection for fault localization", in *Distr. Syst. Oper. Manag.*, Nancy, France, 2001.

[11] M. Brodie *et al.*, "Intelligent probing: a cost-effective approach to fault diagnosis in computer networks", *IBM Syst. J.*, vol. 41, no. 3, 2002.

[12] M. Brodi *et al.*, "Active probing strategies for problem diagnosis in distributed systems", http://research.ibm.com/piple/r/rish



Wiesław Traczyk is a Professor of the National Institute of Telecommunications and also of the Warsaw University of Technology, Institute of Control and Computation Engineering. His research interests include expert systems, approximate reasoning, failures in computer networks and data mining.

e-mail: W.Traczyk@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland
e-mail: traczyk@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

Site selection for waste disposal through spatial multiple criteria decision analysis

Mohammed A. Sharifi and Vasilios Retsios

Abstract—This article deals with the application of spatial multiple criteria evaluation (SMCE) concepts and methods to support identification and selection of proper sites for waste disposal. The process makes use of a recently developed SMCE module, integrated into ITC's[1] existing geographic information system called ILWIS. This module supports application of SMCE in planning and decision making processes through several compensatory and non-compensatory approaches, allowing inclusion of the spatial and thematic priority of decision makers. To demonstrate the process, a landfill site selection problem around the town of Chinchina, in Colombia, is used as an example. Based on different objectives, a spatial data set consisting of several map layers, e.g., land use, geological, landslide distribution, etc., is made available and used for modeling the site selection process.

Keywords—SMCE, geographic information systems, planning, decision-making, site selection.

1. Introduction

There are four analytical function groups present in most geographic information systems (GIS) models: selection, manipulation, exploration and confirmation. Selection involves the query or extraction of data from the thematic or spatial databases. Manipulation entails transformation, partitioning, generalization, aggregation, overlay and interpolation procedures. Selection and manipulation in combination with visualization can be powerful analysis tools. Data exploration encompasses those methods which try to obtain insight into trends, spatial outliers, patterns and associations in data without having a preconceived theoretical notion about which relations are to be expected [1, 2]. The data driven approach, sometimes called data mining, is considered very promising, due to the fact that theory in general in many disciplines is poor and moreover, spatial data is becoming increasingly available (rapid move from a data poor environment to a data rich environment).

Confirmative analysis, however is based on a priori hypothesis of spatial relations, which are expected and formulated in theories, models and statistical relations (technique driven). Confirmative spatial methods and techniques originate from different disciplines like operation research, social geography, economic models and environmental sciences. The four analytical functions can be considered as a logical sequence of spatial analysis. Further integration of the maps/results from spatial analysis is an important next step to support decision-making, which is called evaluation [2]. The lack of enough functionality especially in exploitative and confirmative analysis and evaluation in GIS packages has been the topic of many debates in the scientific communities and as a result techniques to support these steps have gained more attention. In this context several studies have demonstrated the usefulness of integrating multi-criteria decision analysis techniques with GIS in a user-friendly environment. However, there is a trade-off between efficiency, ease of use, and flexibility of the system. The more options are predetermined and available from the menu of choices, the more defaults are provided, the easier it becomes to use a system for a progressively small set of tasks.

In this context, a spatial multiple criteria evaluation (SMCE) module has been developed and integrated in a user-friendly environment into ITC's existing geographic information system called ILWIS. This module supports the implementation of framework for the planning and the decision making process as described by [3] and includes several compensatory and non-compensatory approaches, enhancing the spatial data analysis capability of GIS to support planning and decision-making processes. This article tries to demonstrate this capability in a site selection process for waste disposal in Chinchina, located in the Central Cordillera of the Andes in Colombia (South America).

2. Theoretical framework

2.1. Spatial multiple criteria evaluation

Conventional multi-criteria decision making (MCDM) techniques have largely been non-spatial. They use average or total impacts that are deemed appropriate for the entire area under consideration [4]. The assumption that the study area is spatially homogenous is rather unrealistic because in many cases evaluation criteria vary across space. The most significant difference between spatial multi-criteria decision analysis and the conventional multi-criteria decision analysis is the explicit presence of a spatial component. Spatial multi-criteria decision analysis therefore requires data on the geographical locations of alternatives and/or geographical data on criterion values. To obtain information for the decision making process the data are processed using both MCDM and GIS techniques.

[1]International Institute for Geo-Information Science and Earth Observation, Enschede, The Netherlands.

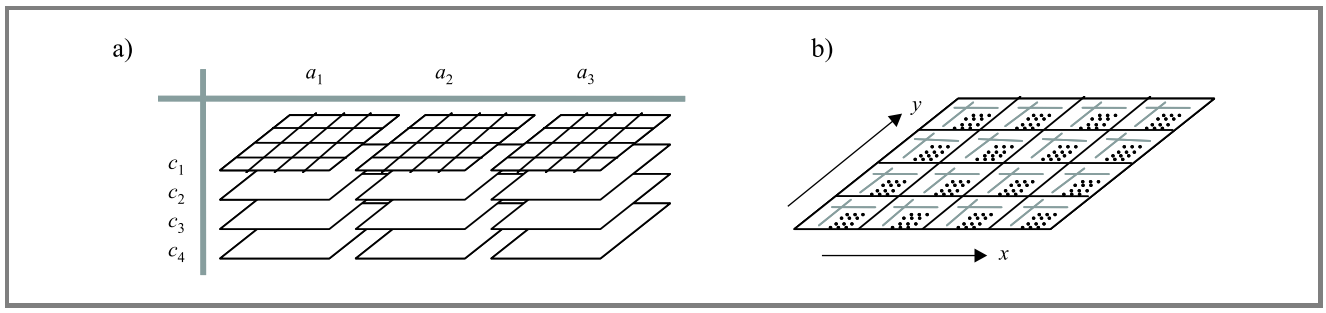


Fig. 1. Two interpretations of a 2-dimensional decision problem (a) table of maps; (b) map of tables.

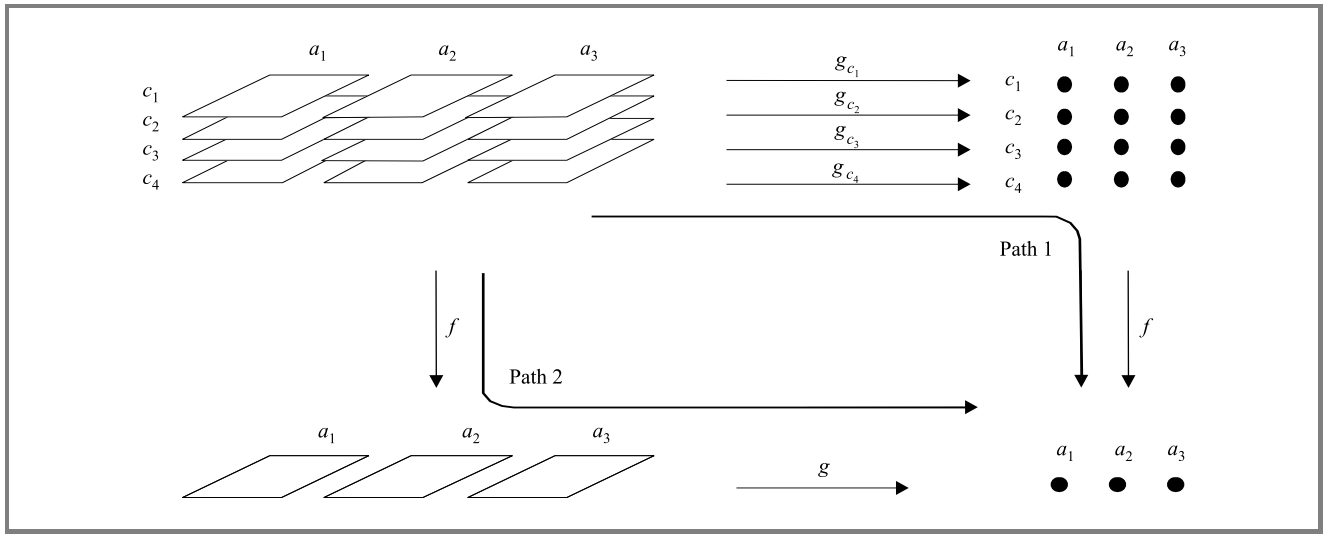


Fig. 2. Two paths of spatial multi-criteria evaluation.

Spatial multi-criteria decision analysis is a process that combines and transforms geographical data (the input) into a decision (the output). This process consists of procedures that involve the utilization of geographical data, the decision maker's preferences and the manipulation of the data and preferences according to specified decision rules. In this process multidimensional geographical data and information can be aggregated into one-dimensional values for the alternatives. The difference with conventional multi-criteria decision analysis is the large number of factors necessary to identify and consider, and, the extent of the interrelationships among these factors. These factors make spatial multi-criteria decision analysis much more complex and difficult [5]. GIS and MCDM are tools that can support the decision makers in achieving greater effectiveness and efficiency in the spatial decision-making process. The combination of multi-criteria evaluation methods and spatial analysis is referred as spatial multiple criteria evaluation. SMCE is an important way to produce policy relevant information about spatial decision problems to decision makers.

An SMCE problem can be visualized as an evaluation table of maps or as a map of evaluation tables as shown in Fig. 1 [6].

If the objective of the evaluation is to rank all alternatives, the evaluation table of maps has to be transformed into a single final ranking of alternatives. Actually, the function has to aggregate not only the effects but also the spatial component. To define such a function is rather complicated. Therefore, the function is simplified by dividing it into two operations: 1) aggregation of the spatial component and 2) aggregation of the criteria. These two operations can be carried out in different orders, which are visualized in Fig. 2 as Path 1 and Path 2. The distinguishing feature of these two paths is the order in which aggregation takes place. In the first path, the first step is aggregation across spatial units (here spatial analysis is the principal tool); the second step is aggregation across criteria, with multi-criteria analysis playing the main role. In the second path these steps are taken in reverse order. In the first case, the effect of one alternative for one criterion is a map. This case can be used when evaluating the spatial evaluation problem using so called Path 1. In the second case, every location has its own 0-dimensional problem and can best be used when evaluating the spatial problem using the so called Path 2 (Fig. 2).

2.2. SMCE and integrated planning and decision support system

Advances in information technology and remote sensing have provided extensive information on processes that are

taking place on the earth's surface, much of which is organized in computer systems, some is freely available and other is accessible at an affordable price. Research in disciplinary sciences has also produced insight into many physical and socio-economic processes. Yet much of the existing information and knowledge is not used to support better management of our resources. Geo-information technology has offered appropriate technology for data collection from the earth's surface, information extraction, data management, routine manipulation and visualization, but it lacks well-developed, analytical capabilities to support decision-making processes. For improved decision-making, all these techniques, models and decision-making procedures have to become integrated in an information processing system called integrated planning and decision support system (IPDSS). The heart of an IPDSS as defined by [7] is model based management, which includes quantitative and qualitative models that support resource analysis, assessment of potential and capacity of resources at different levels of management. This is the most important component of the system, which forms the foundation of model-based planning support [8]. It includes three classes of models, which make use of existing data, information and knowledge for identification of a problem, formulation, evaluation and selection of a proper solution. These models are:

- A process/behavioural model describing the existing functional and structural relationships among elements of the planning environment to help analysing and assessing the actual state of the system and identify the existing problems or opportunities. This also supports "resource analysis", which clarifies the fundamental characteristics of land/resources and helps understanding the process through which they are allocated and utilized [8, 9].
- A planning model, which integrates potential and capacity of resources (biophysical), socio-economic information, goals, objectives, and concerns of the different stakeholders to simulate the behaviour of the system. Conducting experimentation with such a model helps understanding the behaviour of the system and allows generation of alternative options/solutions to address the existing problems.
- An evaluation model, which allows evaluation of impacts of various options/solutions and supports selection of the most acceptable solution, which is acceptable to all stakeholders, and improves the management and operation of the system.

Spatial multiple criteria evaluation can play a very important role in the development and application of above models. In the process/behavioural model it will help to assess the current state of the system. Today, sustainability assessment of the resource management is one of the very critical issues in the management science. There is great interest to assess sustainability of agricultural development, sustainability of forest management, sustainability of cities, etc. What is sustainable management and how it can be assessed and improved is, however a very important research question in many cases.

Spatial multiple criteria evaluation can also be applied in the evaluation and planning model. In the evaluation it will allow assessment and multiple criteria evaluation of several options/alternatives in order to help understanding their impacts, pros and cons, their related trade-offs and the overall attractiveness of each option or alternative. Here the alternatives have specified locations (boundaries) and their performance on each criterion can be represented by a separate map (more than one-dimensional table of maps). This type of analysis is based on the multiple attribute decision analysis techniques [6]. In the planning model, it can help to formulate/develop alternative options. Here, in the planning process alternatives are formed out of pixels of one map. The types of analysis that are applied here, are based on the multiple objective decision analysis techniques [6]. In this process, the whole decision space is divided in two sets, mainly the efficient and non-efficient ones, which are then used for proper design of alternatives. A good example of SMCE application in planning and decision models is site selection, which will be demonstrated through a case study explained in the following sections of this article.

3. Case study

In this chapter, a case study on selection of a waste disposal site is carried out in order to demonstrate some of the capabilities of SMCE as implemented in the ILWIS GIS.

3.1. Problem definition

The municipality of the town of Chinchina, located in the Central Cordillera of the Andes in Colombia (South America), wants to investigate areas suitable for waste disposal. Up till today all the garbage from the city of 150000 inhabitants is dumped in a river. However, due to an increase in environmental awareness, the municipality of Chinchina has decided to construct a proper waste disposal site. For this purpose, assistance from the regional planning department has been requested. The planning department forms a team, consisting of a geologist, a geomorphologist, a hydrologist and an engineer.

After a one-month period in which field studies were conducted and multidisciplinary plenary meetings were held, the team submitted a report to the municipality, in which the following criteria in selecting areas suitable for waste disposal were considered:

Biophysical criteria

Constraints[2]

- The waste disposal site cannot be built on landslides which are active or may become active in the future.

[2]Constraints are binding criteria (no compensation is allowed).

- The waste disposal site can only be constructed in areas which do not have an important economic or ecological value.
- Areas should have sufficient size/capacity (at least 1 hectare) to be used as a waste disposal site for a prolonged time.

Factors[3]

- The waste disposal site should preferably be constructed on areas with no landslides.
- The waste disposal site should preferably be constructed on areas with the least important economic or ecological value.
- The waste disposal site should preferably be located on a terrain with a slope less than 20 degrees.
- The waste disposal site should preferably be located within 2 km from the city limits of Chinchina.
- The waste disposal site should preferably be located at least 300 meters from any existing built-up area.
- The waste disposal site should be constructed on clay-rich soils (preferably more than 50% clay).
- The waste disposal site should have a high soil thickness.
- The waste disposal site's soil should have a very low permeability (preferably 0.05 meters per day or less).

Socio-economic criteria (*factors*)

- The overall site transportation costs should be as low as possible.
- Once a waste disposal site is introduced, the land value of the surroundings and other locations will change. The negative effect on the land value should, if possible, be minimized for land that currently has a significant value.
- Once a waste disposal site is introduced, the pollution of the surroundings and other locations will change. The effect on the pollution should be as low as possible to locations that are sensitive to it.

The following digital raster maps were made available to be used for analysis of the biophysical criteria:

- "Slide": a map whereby each pixel is classified in one of four classes: 'no landslide', 'stable', 'dormant' and 'active'.
- "Landuse": a map whereby each pixel is classified in one of eight classes: 'built-up area', 'coffee', 'shrubs', 'forest', 'pasture', 'bare', 'riverbed', 'lake'.
- "Slope": a map whereby for each pixel the average slope of the corresponding area is stored, as a numerical value (in degrees).
- "Geol": a geological map of the area with several attributes, one of them being the geological class, another the average clay thickness, another the average clay percentage, and another the average permeability.
- "Distance_from_city": a map whereby for each pixel its distance from the nearest point of the city of Chinchina is stored, as a numerical value (in meters).
- "Distance_from_built_up": a map whereby for each pixel its distance from the nearest built-up area is stored, as a numerical value (in meters).

Maps for the socio-economic criteria are not yet available. They can only be produced for a potential site.

3.2. Site selection process

The site selection process is carried out in two phases: in phase one, SMCE is applied in order to identify (design) potential areas, which are biophysically suitable for waste disposal. In the next phase, SMCE is applied to compare/evaluate potential sites considering their socio-economic and biophysical characteristics in order to make the final recommendation (choice of a solution). The socio-economic characteristics reflect the impact of a site on several spatial (and sometimes non-spatial) aspects. They can only be assessed for a potential site, which is why they cannot be used as a criterion in the design phase. In the choice phase of the site selection process, the suitability of each site, which is identified as a potential site in the first phase, will be assessed by means of SMCE, considering socio-economic factors. Figure 3 presents the site selection process.

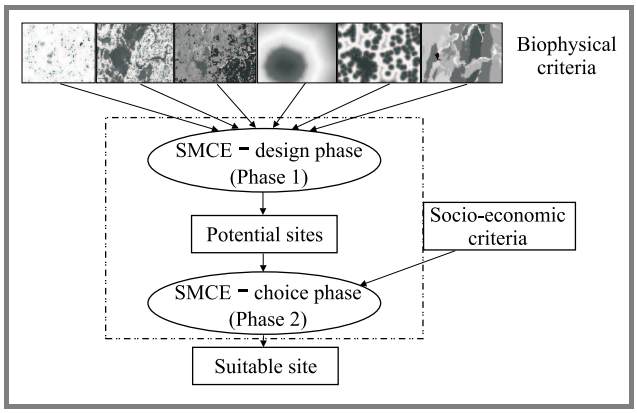


Fig. 3. Flowchart for the entire site selection process. Due to printing limitations, original color maps had to be converted to black and white, with loss of some detail.

[3]Factors are non-binding and are considered as preferred situations that can compensate each other.

3.3. SMCE application in identifying potential sites (Phase 1—design)

In this phase, SMCE is used as a basis for a planning model, which can support development/design of alternative solutions. Here each point/pixel in the map (area of interest) is considered as a potential element of a site. Therefore their related quality and characteristics are evaluated through SMCE (map of tables). This process is implemented through the following main steps (Fig. 4):

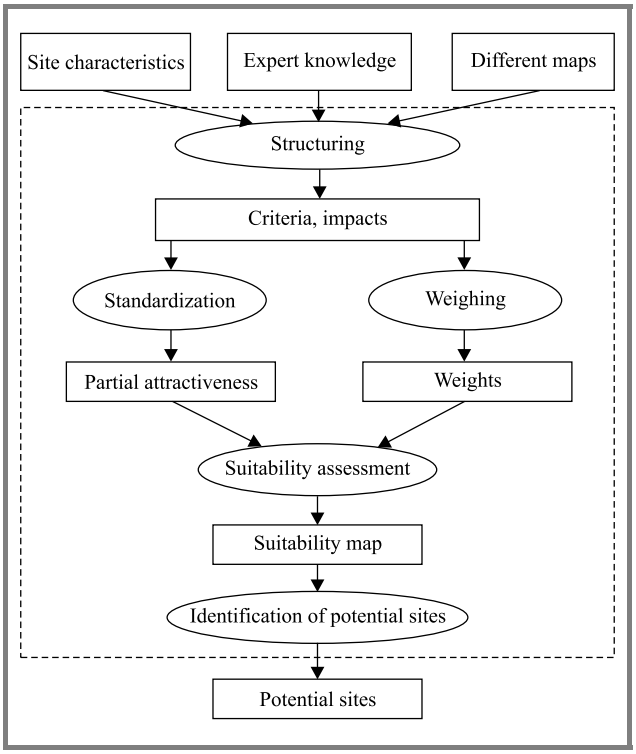


Fig. 4. Flowchart for the SMCE design phase, which results in selection of potential sites.

1. Problem structuring, which leads to identification of the main criteria that should be considered absolutely necessary as well as those that are preferable. Naturally, the information related to those criteria has to be collected and presented in the proper format.
2. Identification of the relevant transformation functions that convert the facts (data) related to each selected criterion to a value judgment, the so-called "utility". This process identifies the partial attractiveness of the region of interest for a site with respect to each criterion.
3. Identification of the relative importance of each criterion with respect to the others, in order to find the level of contribution of each criterion into the achievement of its related objectives (weight assessment).
4. Assessment of the overall attractiveness of every point in the map (pixel) applying the proper decision rule.
5. Formation of the potential sites by connecting the suitable points (pixels) in order to design potential sites with the required size and capacity.

Above steps are explained in the following paragraphs.

3.3.1. Structuring

Structuring in SMCE refers to the identification of alternatives, criteria that are used for their assessment, together with measurement or assessment of the performance of each alternative with respect to each criterion "impact" or "effect". In the same way here structuring refers to identification of the biophysical quality and quantity of site-characteristics, and their relationships, which should be considered in the determination of sites for suitable waste disposal. The relationships between the site characteristics/criteria are established by development of a so-called "criteria tree", which considers all the relevant criteria and groups them in clusters of comparable criteria that are forming a specific quality of the potential sites. Next, a map representing land quality in the area of interest is prepared.

In the SMCE module implementation in ILWIS this process is greatly facilitated through development of the criteria tree structure. The leaves of the tree are indicators that are represented by separate maps. The related map will eventually be assigned to each leaf in the tree. As was mentioned earlier, some of the criteria are binding and act as constraints (can not be compensated) and some act as factors that can be compensated. These are presented in Fig. 5, which presents the criteria tree of the case.

3.3.2. Partial valuation (standardization)

In Fig. 5, at the leaves of the criteria tree, each criterion is represented by a map of a different type, such as a classified map (forest, agriculture, etc.) or a value map (slope, elevation, etc.). For decision analysis the values and classes of all the maps should be converted into a common scale, which is called "utility". Utility is a measure of appreciation of the decision maker with respect to a particular criterion, and relates to its value/worth (measured in a scale 0 to 1). Such a transformation is commonly referred as "standardization".

Different standardization is applied to each different type of maps:

- For "value maps", standardization is done by choosing the proper transformation function from a set of linear and nonlinear functions. The outcome of the function is always a value between 0 and 1. The function is chosen in such a way that pixels in a map that are highly suitable for achieving the objective result in high standardized values, and unsuitable pixels receive low values. ILWIS' SMCE module provides a number of linear and nonlinear functions. Possible standardization methods for value maps in the

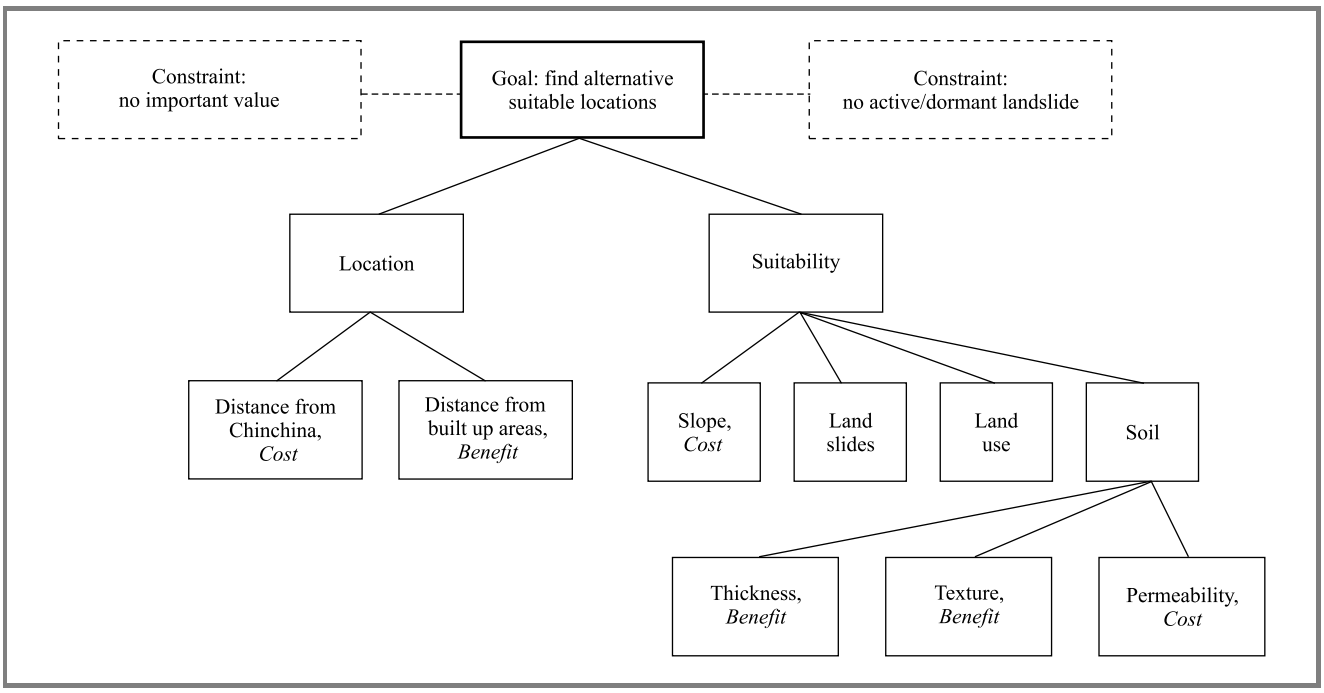


Fig. 5. Criteria tree for identifying the potential site for waste disposal.

developed SMCE module are, e.g., "Maximum", "Interval" and "Goal". Together with the "cost/benefit" property of the criterion, this information is sufficient for applying the selected standardization method in the correct way.

- For "classified maps", standardization is done by matching a value between 0 and 1 to each class in the map. This can be done directly, but also by pair wise comparing or rank-ordering the classes.

3.3.3. Weighing

The next step in SMCE is the identification of the relative importance of each indicator, the so-called weights. ILWIS' SMCE module provides support for a number of techniques (direct, pair wise comparison and rank-ordering) that allow elicitation of weights in a user-friendly fashion, at any level and for every group in the criteria tree. The criteria tree designed in the first step enables giving weights to a few factors at a time, as the branches of one group only are compared to each other. Starting, e.g., with the group "Soil", the factors "Thickness", "Texture" and "Permeability" are compared to each other and a weight is assigned to them. Factors are always weighed, but for constraints there is no weight involved, because they simply mask out the areas which are not interesting.

3.3.4. Suitability assessment/derivation of overall attractiveness

After partial valuation and identification of the relative importance of each criterion in the site selection process, the next step is to obtain the overall attractiveness (suitability) of each point (pixel) in the map (composite index map) for waste disposal. For this process, ILWIS' SMCE module supports several techniques. One of the most transparent and understandable techniques is the weighted summation that is implemented in a user-friendly fashion at each level, for every group of indicators. For the waste disposal criteria tree, starting at the beginning of the tree, a weighted sum formula is written out based on the two first level groups:

$\texttt{suitability\_map} = w_1\texttt{*Location} + w_2\texttt{*Suitability}$

Here w_1 and w_2 are the weights that were produced in the weighing process.

Then, recursively, the groups are substituted by the formula that will generate them from their components, which results in the following:

$\texttt{suitability\_map} = w_1\texttt{*(}w_{11}\texttt{*Distance\_from\_Chinchina} + w_{12}\texttt{*Distance\_from\_builtup\_areas)} + w_2\texttt{*(}w_{21}\texttt{*Slope} + w_{22}\texttt{*Land\_slides} + w_{23}\texttt{*Land\_use} + w_{24}\texttt{*Soil)}$

Here, `Distance_from_Chinchina`, `Distance_from_builtup_areas`, `Slope`, etc. represent the "standardized" version of the corresponding maps.

Substituting the group "Soil" will make the formula even longer.

At the end, the "standardized" maps are written in terms of the original maps and the corresponding value function that will standardize them.

In the developed SMCE module, it is a one step process (single mouse-click) to produce the formula that corresponds to the criteria tree and execute it in order to generate the composite index map named `suitability_map`. Although not explicitly mentioned, the constraints are also taken care of in this formula.

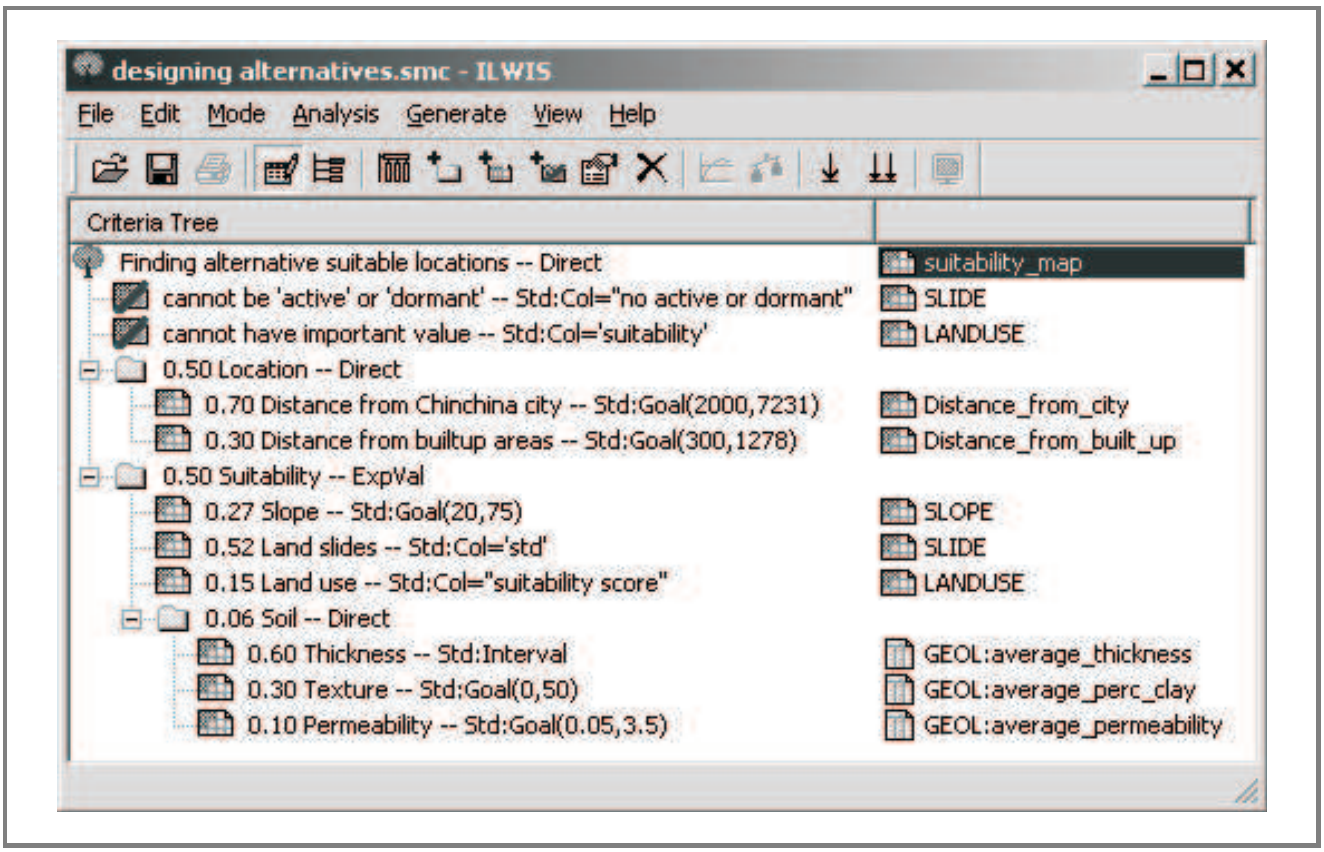


Fig. 6. Criteria tree for identifying suitable waste disposal sites in the SMCE module.

3.3.5. Identification of potential sites

The resulting suitability map for waste disposal is showing the overall attractiveness of each point (pixel) presented in the scale between 0 and 1 for the whole area of interest. In this map each map element (pixel) is 156 m^2 (12.5×12.5 m) with a composite index between 0 and 1. The higher the index, the more suitable the land is. Based on expert knowledge the potential site should have an area of at least 10 000 m^2, corresponding to at least 64 connected pixels. To identify the most suitable locations with sufficient capacity (size) the following steps are implemented:

1. By setting a threshold on the suitability index, the whole area is classified to the classes "suitable" and "unsuitable". This will generate a map with several "suitable" sites.

2. From the "suitable" sites, the ones with sufficient capacity are identified. The "minimum area" (in m^2) required for a site is considered here.

The threshold on the suitability index mentioned in Step 1 is found by trial and error, so that a reasonable number of candidate sites can be designed.
The result of this process is the final output of the design phase: the "potential sites" which satisfy the biophysical factors as good as possible, and have sufficient capacity for being used as a waste disposal site for a longer period.

3.3.6. Practical use of the developed SMCE module in the design phase

Structuring the criteria to determine their impact by setting the relation between factors, constraints and the objective, standardizing and weighing and finally performing the weighted summation is integrated into a few easy steps with the SMCE module developed. Figure 6 shows the module's window at the moment when the waste disposal criteria tree has been fully defined, all criteria standardized and all groups weighed.

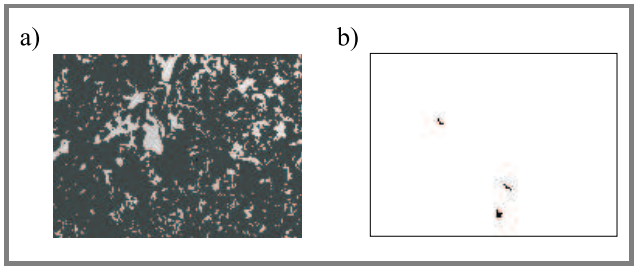


Fig. 7. The suitability map on the left (a) gives the three potential sites on the right (b). Due to printing limitations, the maps are printed as black and white, and the red–yellow–green gradation was converted to black–grey–white.

Generation of the composite index map (the "suitability_map") is now single mouse-click away. Unsuitable areas, i.e., areas with suitability value 0, are denoted with the

red color. When suitability increases, the color gradually transits to yellow, and then to green as suitability gets closer to 1. With a few more steps, the suitability map translates to a map indicating the potential sites for waste disposal. Three sites are identified to have both high suitability and sufficient capacity (Fig. 7).

3.4. SMCE application for site selection (Phase 2—choice)

In the previous phase SMCE was used to identify potential sites (planning mode). In this phase, SMCE will be used

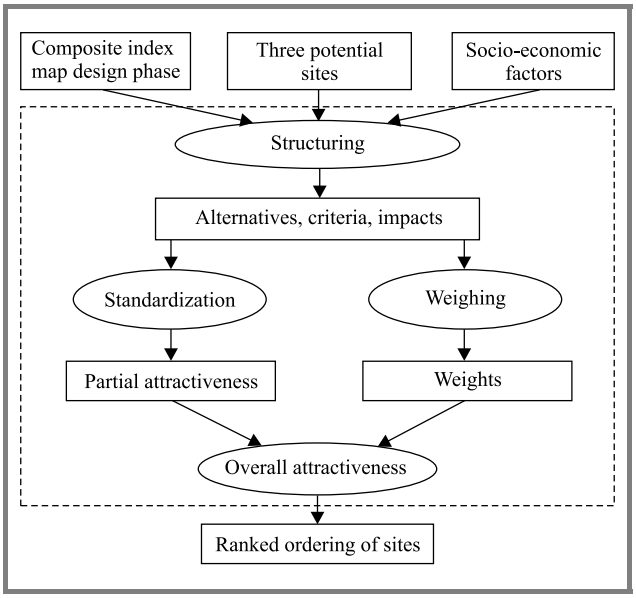


Fig. 8. Flowchart for the SMCE choice phase, which results in ranking of potential sites.

to rank them and choose the most attractive site (choice mode). In the same way as was presented in Fig. 4, this phase includes the following steps (Fig. 8):

1. Problem structuring: identification of alternatives, criteria and their impacts. Here, each of the potential sites from the previous phase is an alternative from which a final choice has to be made. One of the criteria considered in this phase is the suitability of each of the potential sites. The other are the socio-economic criteria.

2. Partial valuations of all alternatives on each criterion. This is carried out through a value function that is based on the attractiveness of each criterion. In this way all criteria are standardized and will represent the level of appreciation and contribution of each indicator to the overall attractiveness of each site.

3. Identification of the relative importance of each criterion in the overall attractiveness of the site, which leads to elicitation of weights for the socio-economic and biophysical factors.

4. Identification of the overall attractiveness of each alternative (each of the potential sites) and ranking and recommendation of the most suitable site.

The most important difference between this phase and the design phase is that here several data sets, one set for each potential site, go through the same SMCE process as in the design phase and result in one composite index map for each potential site. The data sets are not handled independently. The same criteria tree is used and the same weights are used for all, and the standardization step gets an extra dimension.
The identified steps are explained in the next sections.

3.4.1. Structuring

In this step, the problem is structured, by identifying which are the alternatives, on which criteria the decision should be based, and what is their impact. In the design phase, three sites were identified as being the potential sites. Those are then the three alternatives from which a choice is made in this phase.
The criteria on which the decision is based are the site suitability calculated in the design phase, and the socio-economic criteria "transportation cost", "land value reduction" and "effect on pollution". Those are grouped and inserted into a criteria tree in order to determine their impact (Fig. 9).

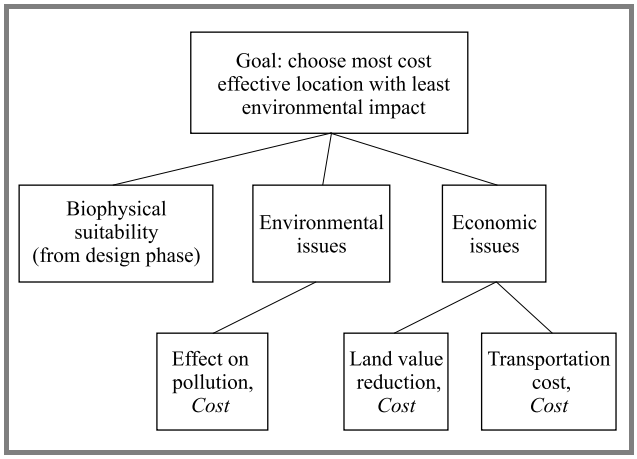


Fig. 9. Criteria tree for the SMCE choice phase of the selection of a waste disposal site.

This criteria tree shows that the impact of the biophysical suitability is on the same level as the environmental and economic criteria of the sites. All environmental and economic criteria are costs to the objective: "most economic location with least environmental impact", but the site suitability is a benefit. The maps for these criteria can only be generated now that the potential sites are known, as they depend on knowledge of the potential site.
For the site suitability criterion, one suitability map per potential site is produced, based on the composite index map from the design phase. This is a simple step, where the suitability of the areas in the composite index map that don't

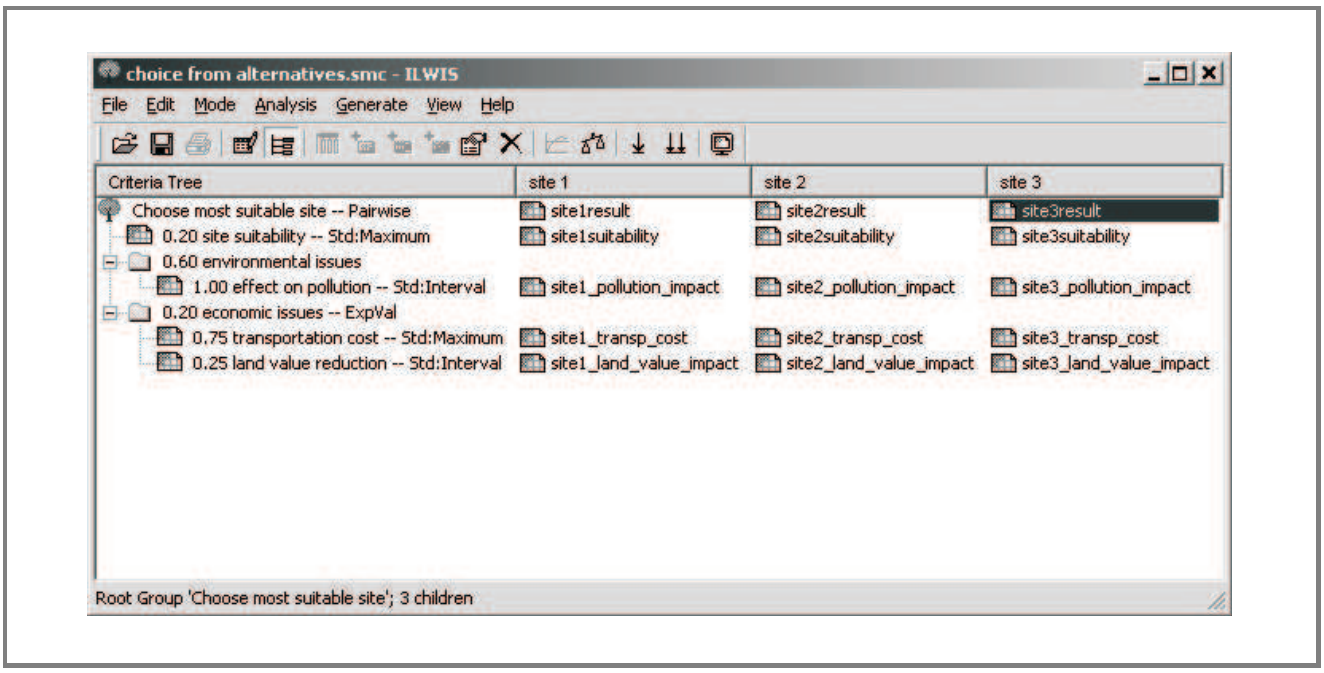


Fig. 10. The SMCE module in the choice phase; deciding on the best from three sites.

belong to the potential site is set to 0. Instead of having the suitability as one value (by taking, e.g., the average for each site), the spatial aspect is preserved.
The maps for the socio-economic criteria are based among others on the location of the potential sites identified in the design phase. Each of the three criteria is handled in its own way:

1. "Transportation cost": The overall city to site transportation costs should be minimized. For each site, a map is calculated that indicates the cost for transporting garbage to the waste disposal site for each point in Chinchina city.

2. "Land value reduction": The impact on the land value should be as little as possible. To calculate this, a map with the original land value is used in order to calculate a map for each site with the change in land value.

3. "Effect on pollution": The effect on the pollution should be as little as possible. To calculate this, a map with the originally polluted areas around the river is used together with a map indicating the sensitivity of different areas to pollution. The result is a map with the effect on the pollution for each site.

Calculating the required maps is done with functionality of the GIS into which the SMCE module is integrated.

3.4.2. Standardization

As in the design phase, the "utility" must be determined for each criterion, i.e., the function that converts the pixels of the three corresponding maps (one for each site) to a value between 0 and 1. In this phase, an extra dimension is given to this process by making sure that the range is the same for all three sites per criterion. Only then it is meaningful to compare maps of the three sites to each other.
This changes the way in which histogram values used in standardization functions are calculated. Where the maximum in the design phase was simply the maximum value of one map, here it is the maximum value of all the alternative maps for one criterion. The same goes for the minimum.

3.4.3. Weighing

As in the design phase, every group in the criteria tree has to be weighed. The same weigh methods are available.

3.4.4. Assessment of the overall attractiveness of the sites

As in the design phase, a weighted summation formula is generated for the criteria tree. The difference is that it is applied once for each potential site. The maps calculated for each site are used as input. The result is one composite index map for each potential site.

3.4.5. Final site selection

The composite index maps can be compared to each other in several ways, in order to rank-order the sites. The most common way is to aggregate the composite indexes of each site through their histogram values (e.g., maximum, average, sum, connectivity index) and rank-order the sites accordingly. The one with the most favorable selected histogram value becomes the site recommended by the SMCE process.

3.4.6. Practical use of the developed SMCE module in the choice phase

As in the design phase, development of the criteria tree, standardization, weighing and performing the weighted summation is a matter of few easy steps with the SMCE module developed. In the window of Fig. 10 the complete criteria tree for choosing one of the three potential waste disposal sites is shown.
Equivalent to the composite index map generated in the design phase, when selection of a site has an unacceptable effect to an area, i.e., the composite index value is 0, this is denoted with the red color. As the effect becomes more acceptable, the color gradually transits to yellow, and finally to green to denote a satisfactory effect with composite index value 1. In this way, the composite index maps indicate not only how much more attractive a site is compared to another, but have this attractiveness distributed spatially (Fig. 11).

Fig. 11. The three composite index maps that correspond to the three sites. The maps are printed as black and white, and the red–yellow–green gradation was converted to black–grey–white.

If at this point the location that becomes better or worse is not interesting, other values could give an outcome. As an example, the sites are rank-ordered as follows: first preference is given to the site with the largest area with high values. In case of equality, preference is given to the site with the smallest area with low values.
For the three composite index maps, the aggregated information in Table 1 helps the rank-ordering.

Table 1
Values taken from the histogram of the three composite index maps

Area [m^2]	Site 1	Site 2	Site 3
With composite index >= 0.58	5 469	20 781	5 469
With composite index <= 0.45	178 594	78 906	170 469

This results in the rank-ordering site 2, site 3 and site 1. The final choice according to the criteria used is thus site 2.

4. Concluding remarks

With the development of GIS, environmental and natural resource managers increasingly have information systems at their disposal in which data are more readily accessible, more easily combined and more flexibly modified to meet the needs of environmental and natural resource decision making. It is thus reasonable to expect a better informed, more explicitly reasoned decision-making process. But despite the proliferation of GIS software systems and the surge of public interest in the application of a system to resolve real world problems, the technology is commonly seen as complex, inaccessible, and alienating to the decision makers [10]. The reasons for this estrangement are varied. In part the early development and commercial success of GIS was fuelled more by the need for efficient spatial inventory rather than decision support systems. As a result, few systems yet provide any explicit decision analysis tools. To alleviate above problems, enough analytical capability should be integrated/connected to GIS in order to provide DSS functionality in a user friendly environment.

One of the very important analytical capabilities is spatial multi-criteria evaluation which together with the analytical functionality of GIS, supports producing decision and policy relevant information about spatial decision problems to decision makers. GIS and MCDM can support decision makers in achieving greater effectiveness and efficiency in the spatial decision-making process, therewith enhancing the use of geo-information. In this context a user friendly SMCE module has been developed and integrated into ITC's geographic information system called ILWIS. This module, which is based on the framework for the planning and decision making process as developed at ITC, is designed and implemented in such a way that can help the integration of information from a variety of sources (spatial, non-spatial) to support planning and decision making processes.

A good example of ILWIS' SMCE module in planning and decision making modes is site selection, which has been demonstrated through the case study in this paper. The case shows how effectively and efficiently SMCE can be applied in the process of designing and ranking of alternative sites for waste disposal.

References

[1] J. W. Tukey, *Exploratory Spatial Data Analysis*. Boston: Adison Wesley, 1977.

[2] L. Anselin and A. Getis, "Spatial statistical analysis and geographic information systems", in *Geographic Information Systems, Spatial Modelling and Policy Evaluation*, M. Fisher and P. Nijkamp, Eds. Berlin: Springer-Verlag, 1992.

[3] M. A. Sharifi, W. van der Toorn, A. Rico, and E. M. Emmanuel, "Application of GIS and multicriteria evaluation in locating a sustainable boundary between the Tunari national park and Cochabamba city (Bolivia)", *J. Multi-Crit. Decis. Anal.*, vol. 11, no. 3, pp. 109–164, 2002.

[4] R. J. Tkach and S. P. Simonovic, "A new approach to multi-criteria decision making in water resources", *J. Geogr. Inform. Decis. Anal.*, vol. 1, no. 1, pp. 25–43, 1997.

[5] J. Malczewski, *GIS and Multicriteria Decision Analysis*. New York: Wiley, 1999.

[6] M. A. Sharifi and M. van Herwijnen, *Spatial Decision Support System: Theory and Practice*. Enschede: ITC Lecture Serie, 2003.

[7] M. A. Sharifi, "Integrated planning and decision support systems for sustainable watershed development", in *Sem. Sustain. Watersh. Develop.*, Tehran, Iran, 2002.

[8] M. A. Sharifi, "Planning supports to enhance land utilisation systems", in *Sem. Imp. Land Utiliz. Syst. Agricult. Product.*, Tokyo, Japan, 2003.

[9] M. A. Sharifi and H. van Keulen, "A decision support system for land use planning", *Agricult. Syst.*, vol. 45, pp. 239–257, 1994.

[10] K. Fedra, "GIS and environmental modeling", in *Environmental Modelling with GIS*, M. F. Goodchild, B. O. Park, and L. T. Steyaert, Eds. New York: Oxford University Press, 1993.

Mohammed Ali Sharifi is born in Teheran, Iran, on 30 December 1944. In 1967 he obtained an M.Sc. degree in agricultural engineering from the University of Teheran. After that, he moved to the Netherlands, where in 1973 he obtained an M.Sc. in photogrammetric engineering at the International Institute for Geo-Information Science and Earth Observation (ITC). In 1992 he obtained a Ph.D. degree in agricultural and environmental sciences from the Agricultural University of Wageningen in the Netherlands. Since then he has worked at ITC. His current position is Associate Professor in decision support systems and land use planning, in the Department of Urban Regional Planning and Geo-Information Management.

e-mail: alisharifi@itc.nl
Department of Urban and Regional Planning
and Geo-Information Management
International Institute for Geo-Information Science
and Earth Observation (ITC)
P.O. Box 6, 7500AA
Enschede, The Netherlands

Vasilios Retsios is born in Athens, Greece, on 20 April 1972. In 1990 he moved to the Netherlands for studying computer science at the University Twente. In 1996 he acquired an M.Sc. with specialization in system architecture. In 1997 he started working at the International Institute for Geo-Information Science and Earth Observation (ITC), developing distance learning GIS and Remote Sensing courses. In 2000 he started working as a software developer in the Department of Geo-Software Development at ITC, making major contributions to ITC's geographic information system called ILWIS. The most recent contribution is a module that performs "spatial multiple criteria evaluation".

e-mail: retsios@itc.nl
Department of Geo-Software Development
International Institute for Geo-Information Science
and Earth Observation (ITC)
P.O. Box 6, 7500AA
Enschede, The Netherlands

Paper

Multicriteria analysis for behavioral segmentation

Janusz Granat

Abstract—Behavioral segmentation is a process of finding the groups of clients with similar behavioral patterns. The basic tool for segmentation is a clustering algorithm. However, the clusters generated by the algorithm depend on the preprocessing steps as well as parameters of the algorithm. Therefore, there are many possibilities of dividing the clients into segments and it is a subjective process. In this paper we will focus on application on multicriteria analysis for selecting the best partition of clients into segments.

Keywords—segmentation, clustering, multicritera analysis, telecommunications.

1. Introduction

The knowledge about clients becomes one of the most important assets in making various business decisions. In this paper we will focus on telecommunication industry. However, the methods that we will present are also valid for other industries. The only requirement is that they have operational databases that are sources for building behavioral features of the clients. Behavioral segmentation is a process of finding the groups of clients with similar behavioral patterns. It is one of the key data mining tasks for marketing departments of telecommunication operators [1, 6, 9]. In this paper we will focus on multicriteria application in segmentation process.

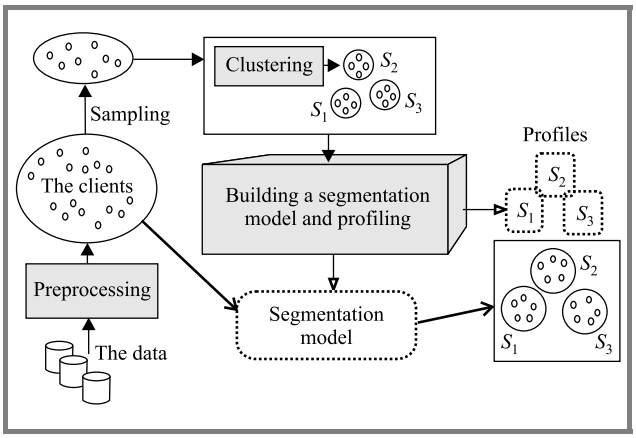


Fig. 1. The segmentation process–analyst view.

We can distinguish two views on segmentation process. The first, analyst view (see Fig. 1), focuses on preprocessing, selection of samples, clustering, building the segmentation model and profiling. Preprocessing consist of loading the data from various sources and transforming it into a table where each row corresponds to a client and columns correspond to the selected features of clients. In classical segmentation the features might be divided into the following groups: geographic (e.g., area, region, city, size),

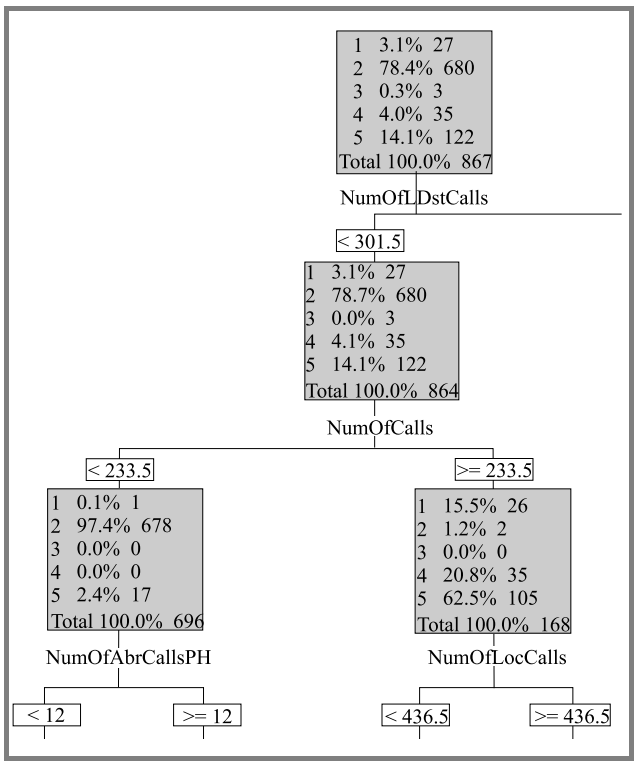


Fig. 2. Model as the cluster profile tree (SAS system).

```
IF      31.5 <= NumOfLDstCalls
THEN
 NODE:    3
 N    :   3
 1    :   0.0%
 2    :   0.0%
 3    : 100.0%
 4    :   0.0%
 5    :   0.0%

IF      12 <= NumOfIntCalls
AND NumOfCalls < 23.5
AND NumOfLDstCalls < 31.5
THEN
 NODE:    7
 N    :   1
 1    : 100.0%
 2    :   0.0%
 3    :   0.0%
 4    :   0.0%
 5    :   0.0%
```

Fig. 3. Model as the set of rules (SAS system).

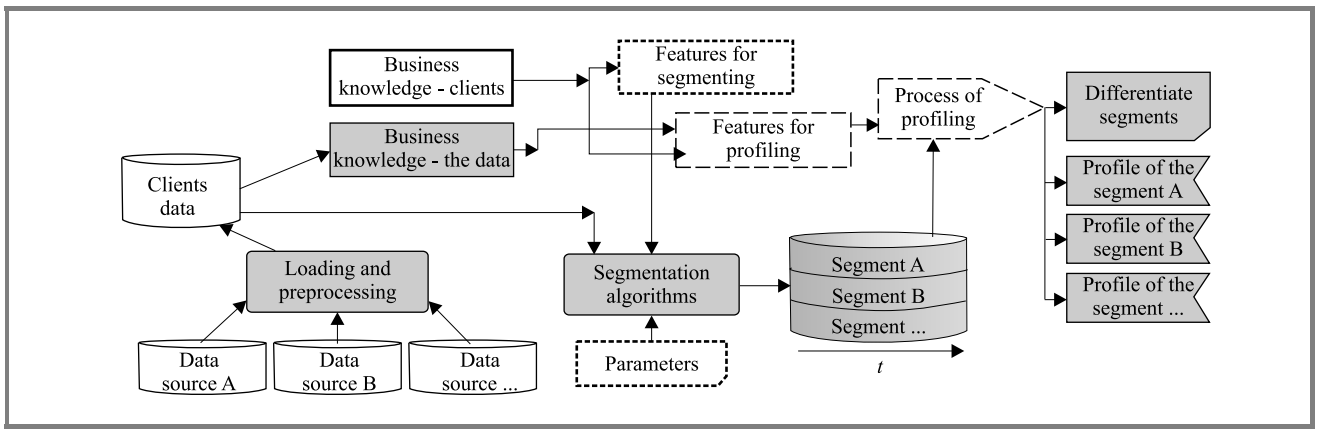


Fig. 4. The segmentation process–business view.

demographic (age, life stage, marital status), socioeconomic (e.g., income, education), psychographic (e.g., personality, lifestyle) [7]. The organizations must buy this type of data or gather it by questionnaires [10].

Telecommunication operators have the data stored in billing databases and others dedicated to specific services. The data about each call is stored in call detail records (CDR). These records contain the following data: caller number, called number, timestamp, the length of the call, tariff units per call, etc. It should be stressed that this data has incomparable quality in comparison to demographic, socioeconomic data, etc. This leads to much better modeling results. CDR records are transformed into client's behavioral features. The examples of behavioral features are the following: number of calls, number of calls within peak hours, number of calls within off-peak hours, number of different called parties, number of specific type connections (local, long distance, international, mobile, premium rate numbers with prefix 0700, toll-free numbers with prefix 0800, internet provider), total length of calls, total length of calls within peak hours, total length of calls within off-peak hours, number of not answered calls, etc. In practical modeling there are more then 1000 such features. If we have, e.g., 5 million of clients the resulting behavioral segmentation table is huge. Therefore, we choose a sample of clients for building the segmentation model. Next, we apply segmentation algorithms several times and generate various instances of the segmentation model. The model might have various views, e.g., in SAS system the model is represented as:

- the cluster profile tree (Fig. 2),
- the set of rules (Fig. 3),
- the SAS 4GL code,
- the C language code.

A cluster profile tree has the percentages and numbers of cases assigned to each cluster and the threshold values of each input variable is displayed as a hierarchical tree. The set of rules is a text file that lists all the rules used to create the profile tree.

In fact, segmentation model belongs to the class of the classification models. We generate several models. Then we usually choose subjectively one of them. This paper shows how this phase might be supported by multicriteria analysis. The model is then applied to whole population of clients. After that, profiles for all segments are built.

The segmentation process in the company has to take into account knowledge of business people about the data, the features that might be used for clustering and the features of clients that might be used for profiling (see Fig. 4).

The business people set up goals of segmentation and then actively participate in the segmentation process. We have quite different results of segmentation, even if it is based on the same data, by setting different goals. As a result of segmentation in a business environment there are profiles of each segment (which describe the features that are common for the segments), as well as so called differentiate profiles (which describe the features that differentiate the segments).

2. The formal model

Modeling is based on the dynamic information system [2, 8] defined as follows:

$$IT = (X, F, V, \rho, T, R), \tag{1}$$

where: T is a nonempty set whose elements are called moments of time, R is a order on the set T (here we assume a linear order), X is the finite and nonempty set of n objects or observations, F is finite and nonempty set of features of the objects, $V = \bigcup_{f \in F} V_f$, V_f is a set values of feature $f \in F$, called the domain of f, ρ is an information function: $\rho : F \times X \times T \to V$.

The dynamic information system considers explicitly time. Segmentation is a dynamic process because clients change their behavior. In this paper we are focusing on multicriteria selection of partition of segments at a specific moment of time t so the time might not be considered.

There is a set of objects X of a dynamic information system IT and the similarity measure between objects $x_i, x_j \in X^{id}$, $i \neq j$:

$$\varphi(x_i, x_j).$$

If there are linguistic, nominal, boolean, and interval-type of features, along with quantitative attributes, the symbolic similarity between the objects is applied [5].
Moreover, clustering depends on parameters of the preprocessing steps Γ and parameters of the algorithms Ω. Let us Δ ($\Delta = \Gamma \cup \Omega$).
Clustering algorithms divide the set of objects into m subsets of similar objects (based on the similarity measure):

$$X \mapsto_{\varphi(x_i,x_j),\Delta} \{X_{S_{1,\Delta}}, X_{S_{2,\Delta}}, \ldots \cup X_{S_{m,\Delta}}\},$$

$$X_{S_{i,\Delta}} \subset X, X_{S_{i,\Delta}} \cap X_{S_{j,\Delta}} = \emptyset \text{ for } i \neq j, \bigcup_i X_{S_{i,\Delta}} = X.$$

For the set of parameters Δ, we have the corresponding identifiers of the clusters:

$$\mathrm{Seg}_\Delta = \{S_{1,\Delta}, S_{2,\Delta}, \ldots, S_{m,\Delta}\}.$$

For a huge data set we have to find the clusters of objects for a training set $X^{Train} \subset X$ and then we build a classification model that can be applied to X set. Let us consider the selected f_S feature of the object (where $S \in FS$, FS is the index of cluster feature), called cluster feature, and the subsets of input features f_I ($I \in FI$, $FI = F \setminus FC$, FI is the index set of input features, F is the index set of all object features).
The model is defined as follows:

$$\rho(f_S, x_i) = M_S\big(\rho(f_{k1}, x_i), \rho(f_{k2}, x_i), \ldots, \rho(f_{kk}, x_i)\big),$$

where: x_i is an object identifier, $k1, k2, \ldots, kk \in FI$.

3. The multicriteria analysis

The analyst divides the set of clients into various segments by setting various sets of parameters Δ. The multicriteria analysis might be applied to final selection of the segments. For each of the set of parameters Δ_i we have a partition $\mathrm{Seg}_{\Delta i}$ and a vector of criteria that evaluate this segmentation $\mathbf{q}_i = (q_{i,1}, q_{i,2}, \ldots q_{i,r})$ (see Table 1). Let the set of segmentation results $\mathrm{Seg} = \{\mathrm{Seg}_{\Delta_1}, \mathrm{Seg}_{\Delta_2}, \ldots, \mathrm{Seg}_{\Delta_n}\}$. The multicriteria problem is defined as follows:

$$\min_{\mathbf{Seg}}, \max, \mathrm{stab}\,\mathbf{q}.$$

Table 1
Table of evaluations of segmentations

Partition	q_1	q_2	$\ldots$	q_r
Seg_{Δ_1}	$q_{1,1}$	$q_{1,2}$	$\ldots$	$q_{1,r}$
Seg_{Δ_2}	$q_{2,1}$	$q_{2,2}$	$\ldots$	$q_{2,r}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Seg_{Δ_n}	$q_{n,1}$	$q_{n,2}$	$\ldots$	$q_{n,r}$

The following examples of criteria might be considered:

- *Number of segments*:

$$q_{i,1} = N\mathrm{Seg}_{\Delta_i}.$$

- *Outliers frequency*:

$$q_{i,2} = OF_{\Delta_i} = \frac{n_{\mathrm{Seg}_{\Delta_i}} - no_{\mathrm{Seg}_{\Delta_i}}}{n_{\mathrm{Seg}_{\Delta_i}}},$$

  where: $n_{\mathrm{Seg}_{\Delta_i}}$ is the number of objects in all segments for Δ_i, $no_{\mathrm{Seg}_{\Delta_i}}$ is the number of deleted outliers.

- *Segments frequency*:

$$q_{i,3} = SF_{\Delta_i} = \frac{\sum_{l=1}^{m} \frac{n_{S_{l,\Delta_i}}}{\max_{\mathrm{Seg}_{\Delta_i}}}}{N\mathrm{Seg}_{\Delta_i}},$$

  where: $n_{S_{l,\Delta_i}}$ is the number of objects in the segment, $\max_{\mathrm{Seg}_{\Delta_i}}$ is the maximal numer of objects.

- *Segments compactness* [4], that evaluates how well the segments are redistributed by the clustering algorithm, compared to the whole set X:

$$q_{i,4} = \frac{1}{m} \sum_{l=1}^{m} \frac{v(X_{S_{l,\Delta_i}})}{v(X)},$$

  where:

$$v(X) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} d^2(x_i, \bar{x})},$$

  $d(x_i, x_j)$ is a distance metric between two objects x_i and x_j, n is the number of objects in X, m is a number of segments, $\bar{x}$ is the mean of X, $v(X_{S_{l,\Delta_i}})$ is calculated for objects in the segment.
  A smaller $q_{i,4}$ indicates a higher homogeneity of the objects in the segments.

- *Separation of segments* [4]:

$$q_{i,5} = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1, j\neq k}^{m} \exp\left(-\frac{d^2(x^c_{S_{i,\Delta}}, x^c_{S_{j,\Delta}})}{2\sigma}\right),$$

  where: σ is a Gaussian constant, m is the number of clusters, $x^c_{S_{i,\Delta}}$ is the centroid of the cluster $S_{i,\Delta}$, $d()$ is the distance metric used by the clustering algorithm, and $d(x^c_{S_{i,\Delta}}, x^c_{S_{j,\Delta}})$ is the distance between the centroid of segments.
  A smaller segment separation criterion indicates a larger overall similarity among the segments.

- $\ldots$ $q_{i,r}$ (several other criteria might be considered).

Table 2
Example of table of evaluations of segmentations

Partition	*NC*	*Outliers*	*FC*
$\text{Seg}_1 = \text{Seg}_{\Delta_1}$	10	1.0	0.115
$\text{Seg}_2 = \text{Seg}_{\Delta_2}$	6	1.0	0.171
$\text{Seg}_3 = \text{Seg}_{\Delta_3}$	5	1.0	0.208
$\text{Seg}_4 = \text{Seg}_{\Delta_4}$	5	0.62	0.208
$\text{Seg}_5 = \text{Seg}_{\Delta_5}$	15	0.98	0.082
$\text{Seg}_6 = \text{Seg}_{\Delta_6}$	15	0.87	0.119

Illustrative example (Table 2) has been prepared on sample of telecommunications data. We have generated six partitions of clients by the SAS system. Then, we calculated the values of three criteria $N\text{Seg}_{\Delta_i}$, OF_{Δ_i} SF_{Δ_i}. Then the multicriteria analysis has been applied.
For specification of the preferences, we have applied the ISAAP software, developed by Granat and Makowski [3]. The user specifies preferences, including values of criteria that he/she wants to achieve and to avoid. Those values are called *aspiration* and *reservation* levels, respectively. The graphs of the so called component achievement functions and the related solution are presented in Fig. 5. The solution, marked by rectangle, is projected into two-dimensional space, in which, for each criterion, its values (the x axis) and the degree of satisfaction (the y axis) of meeting preferences, expressed by aspiration and reservation levels, are reported. In the next step, an interactive procedure is used to assist the user in selecting a segment that best corresponds to his/her preferences. In our example, the user changes reservation point for *SF* criterion to value equal to 0.15. The new solution is presented in Fig. 6. The user can continue the interactive process in order to find another segment.

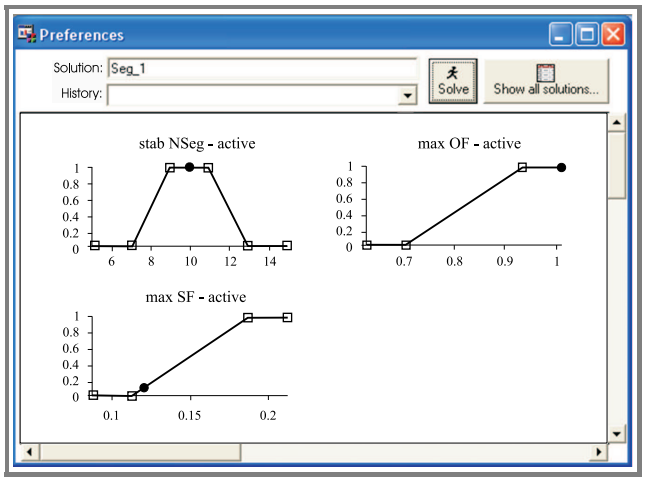


Fig. 5. The ISAAP screen.

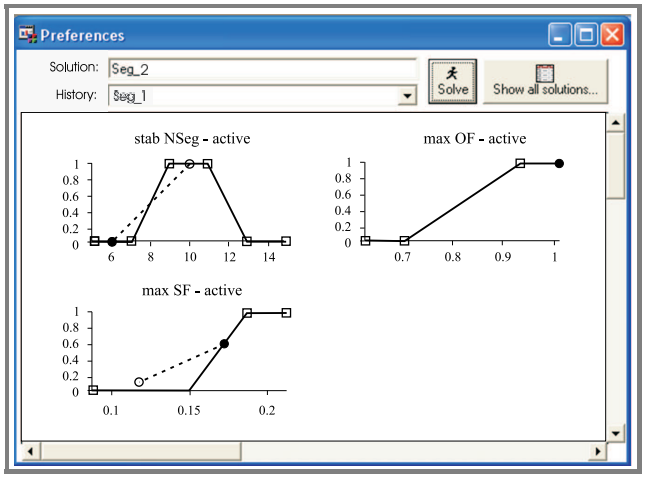


Fig. 6. The ISAAP screen–new preferences.

4. Conlusions

The quality of the behavioral segmentation process significantly influences the clients relationship managament. As we have shown, the process of behavior in business environment is very complex and requires various interactions of analysts and business representatives. Selection of the final partition of clients into segments that is well attuned to the business goals is usually a subjective task. There is a need for development of analytical tools that improve objective comparison of partition of clients into segments. This paper shows how to apply a multicriteria analysis in this process and it is a step into development of such tools.

References

[1] J.-L. Amat, "Using reporting and data mining techniques to improve knowledge of subscribers; applications to customer profiling and fraud management", *J. Telecommun. Inform. Technol.*, no. 3, pp. 11–16, 2002.

[2] J. Granat, "Data mining and complex telecommunications problems modeling", *J. Telecommun. Inform. Technol.*, no. 3, pp. 115–120, 2003.

[3] J. Granat and M. Makowski, "Interactive specification and analysis of aspiration-based preferences", *Eur. J. Oper. Res.*, vol. 122, no. 3, pp. 469–485, 2000.

[4] J. He, A.-H. Tan, C.-L. Tan, and S.-Y. Sung, "On quantitative evaluation of clustering algorithms", in *Clustering and Information Retrieval*, W. Wu, H. Xiong, and S. Shekhar, Eds. Kluwer, 2003.

[5] K. Mali, "Clustering and its validation in a symbolic framework", *Patt. Recogn. Lett.*, vol. 24, pp. 2367–2376, 2003.

[6] R. Mattison, *Data Warehousing and Data Mining for Telecommunications*. Boston, London: Artech House, 1997.

[7] M. McDonald and I. Dunbar, "Market segmentation. How to do it, how to profit from it". Palgrave Publ., 1998.

[8] E. Orłowska, "Dynamic information systems", in *Annales Societatis Mathematicae Polonae*, Series IV: *Fundamenta Informaticae*, vol. 5, no. 1, pp. 101–118, 1982.

[9] M. Shawa, C. Subramaniama, G. Tana, and M. Welgeb, "Knowledge management and data mining for marketing", *Decis. Supp. Syst.*, vol. 31, no. 1, pp. 127–137, 2001.

[10] P. Verhoef, P. Spring, J. Hoekstra, and P. Leeflang, "The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands", *Decis. Supp. Syst.*, vol. 34, pp. 471–481, 2002.

Janusz Granat received his M.Sc. in control engineering (1996) and his Ph.D. (1997) in computer science from the Warsaw University of Technology. He holds a position as an Assistant Professor at the Warsaw University of Technology, and is the leader of a research group on applications of decision support systems at the National Institute of Telecommunications in Warsaw. He lectured decision support systems and various subjects in computer science. His scientific interests include data mining, modeling and decision support systems, information systems for IT management. Since 1988 he has been cooperating with IIASA. He contributed to the development of decision support systems of DIDAS family and the ISAAP module for specifying user preferences. He has been involved in various projects related to data warehousing and data mining for telecommunication operators. He was also involved in EU MiningMart project.

e-mail: J.Granat@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland
Institute of Control and Computation Engineering
Warsaw University Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

A new methodology of accounting for uncertainty factors in multiple criteria decision making problems

Vladimir I. Kalika

Abstract—A new approach is proposed to select a predetermined number of "reasonable" (the best in a certain sense) alternatives from the considerable (maybe a vast) set of initial alternatives according to an arbitrary number of optimization criteria and accounting for uncertainty factors. The approach is based on using a special intuitive methodology, developed to account for uncertainty factors when solving such multiple criteria decision making (MCDM) problems. This methodology is based on performing multi-variant computations (MVC) and finding their "stable-optimal" solutions, and it's realized as a multi-level hierarchical system of MVC series. It's possible to use this methodology for solving various real problems.

Keywords—multiple criteria decision making, uncertainty factors, "reasonable" solutions, multi-level hierarchical system of multi-variant computation series, "stable-optimal" solutions, scenarios, Monte Carlo simulations.

1. Introduction

An approach [1–6] is proposed to select *a predetermined number of "reasonable" alternatives* (their set is named RAS) from their *considerable* (maybe vast) *initial set* (ISA) according to multiple criteria, presented by *an arbitrary number of optimization criteria*. This selection is performed as accounting for *uncertainty factors*, inherent in both the considered problem and its solution process.

The solution process considered should include the following basic stages: (1/2) creating the ISA/ISCAV, where the ISCAV, interrelated with the ISA, should reflect the problem solution objectives, expressed by multiple criteria; (3) multi-criteria optimization in the ISCAV/ISA space to reach the RAS by decreasing ISCAV/ISA and accounting for uncertainty factors. The methodology to perform the stages (1/2) is specific to each considered problem, but for the stage (3) a quite *universal intuitive methodology* was developed, based on accounting for uncertainty by performing *multi-variant computations* (MVC) and finding their *"stable-optimal" solutions*. This MVC process is realized as *a multi-level hierarchical system of MVC series*, where each its level includes a totality of *scenarios*, having the same specific nature for this level. These scenarios reflect varying problem conditions and parameters. Such varying allows to account for uncertainty factors using procedures specific to each level of the multi-level hierarchical system considered. So, one type of procedures to account for uncertainty consists of varying the assigned (a priori) different versions of ISA and ISCAV as well as multi-criteria optimization techniques, used in computations for the upper (exterior) levels of the hierarchical system. Other types of such procedures involve some directed or random sorting out of possible values of problem parameters. This corresponds to the system intermediate and interior levels, where these parameters might be considered as any intervals of random variables. In this case a Monte Carlo simulation is used. Another aspect of accounting for uncertainty is using only such versions of multi-criteria optimization techniques, which were especially modified (by us) to account for uncertainty factors.

Generally, the main aspect of accounting for uncertainty in the proposed methodology is use of the multi-level hierarchical system of MVC series itself to reach the required solutions. It is what precisely allows to account for uncertainty factors, having different nature, by a way of finding "stable-optimal" solutions of MVC series, formed for each level of scenarios. Moving this way, subsets of "stable-optimal" alternatives are formed for each level of the system, based on analysis of such "stable-optimal" subsets, obtained before for the preceding (lower) level. In the end of this moving "from bottom to top" of the system, the required RAS is reached as a "stable-optimal" subset for the first (top) level of this hierarchical system. For each suitable real problem, based on processing vast amount of initial information, the *final solution* may be found by a way of the RAS (or the RAS series) analysis (usually, non-formal) using additional qualitative and quantitative criteria and estimates.

Various investigations have been performed to apply the proposed approach to solving several real problems, actual for conditions prevailing in Israel [1–6].

Although several methods were developed to solve various multiple criteria decision making problems, our intuitive approach might be considered as an original one. In our opinion, this approach might be used to solve problems for which other methods are not convenient. With this point of view, we will compare it with the well-known

AHP methodology of Prof. T. Saaty [8, 9]. Concerning this comparison, it's possible, on our opinion, to say the following:

1. At first glance both these methods use the same multi-level structures to solve the appropriate problems. However, we see *different essence of the levels in these structures*:
   - in the AHP, such levels represent only objects (criteria and alternatives), considered as obvious parameters in the solution process;
   - in our approach, these levels reflect the series of multi-variant computations, which are performed to account for uncertainty factors, specific only to the considered level objects, having various nature.

   This is the main difference between both methods and their solution methodologies, leading to the discrepancy between their fields of application.
2. The AHP methods consider, mainly, the MCDM problems with small number of initial alternatives, but our approach is oriented to solve *the MCDM problems with a considerable number of initial alternatives*. The areas of MCDM problems suitable to application of these methods are very different.
3. The AHP is based on *use of expert estimates of objects' priorities, but our approach can practically avoid use of such estimates*. Use of our approach expands possibilities for solving various MCDM problems, but AHP methods (when they can be practically used) can give more reliable results.
4. The AHP considers one top goal and small number of objects (criteria or alternatives) on other hierarchy levels; *in our approach the number of versions (scenarios) on each hierarchy level may be arbitrary*.
5. *It's possible to include the criteria multi-level hierarchy system, used in the AHP, in our solution methodology as well*.

Let's consider further some basic features of the proposed methodology.

2. Calculation process peculiarities

In accordance with an available uncertain situation, the process to solve the considered problem is treated as *a two-step* one. In its *first step*, a "reasonable" alternatives set (RAS) should be selected from all initial alternatives in accordance with joint accounting for multiple criteria, assigned a priori. This first step might be completed also by finding a totality of such RAS, including several ones, that depends mainly on organization of the second step of solving the whole problem. In this *second step, the final solution of the whole problem*, ready to be used in a practical decision making process is found, basing on the RAS analysis obtained, performed basically in a non-formal way using additional qualitative (including subjective) and quantitative information, criteria and procedures.

Thus, the proposed approach allows to *sharply decrease* the amount of information needed for decision making.

This first step includes the following basic stages: (1-2) constructing *the initial sets of alternatives* (ISA) and *criteria assessment vectors* (ISCAV); (3) decreasing the considerable (maybe vast) ISA/ISCAV to the required (usually small) RAS.

The calculation methodologies used to construct the ISA/ISCAV should be specific to each MCDM problem considered, elaborated especially to account for specific features of this problem. These elaborations can have various basic directions and "bottlenecks". In our experience of solving the appropriate problems, we have encountered situations, when the basic information and calculation difficulties were related to the ISA construction as well as when the ISA was formed in obvious and easy way, but the ISCAV construction required considerable efforts.

In our experience with the ISA construction process [1–6], we had very difficult case of forming the vast initial set of alternatives for the problem of power generation system expansion (PGSE) planning [1–3, 5], where each such alternative reflected the dynamic PGSE strategy. The case of implicit assignment of initial alternatives is linked with the problem of stock buying on the stock market [6], where each initial alternative reflects the natural operation of a stock buying.

The ISCAV construction process consists of the following: (a) *assignment of the criteria totality*; (b) *development of the criteria calculation models*, allowing to determine *criteria assessment vector* for each considered alternative, where this vector is represented for this alternative by one numerical value for each alternative from the ISA; (c) *forming the initial set of such vectors* (ISCAV), interrelated with the ISA. For this ISCAV construction process, we can have the opposite situations: (a) there are the natural criteria (economic, technical, reliability, others), where the criteria calculation models are developed, mainly, by using the existing models, methods and procedures (e.g., [1–3, 5]); (b) the necessity is raised to create principally new models to form the ISCAV (e.g., for the above considered stock market problem [6]). We will consider in detail (Section 4) the latter situation "(b)" for the same stock market problem, where a new approach, different from one developed early (see [6]), is described.

To implement stage (3) of the problem solution process, we have developed a quite universal *intuitive solution methodology* [4–6] to reach the RAS by decreasing the ISA. This methodology of *accounting for uncertainty*, applicable to various MCDM problems, is based on performing multi-variant computations (MVC) and finding their "stable-optimal" solutions.

The developed *methodology of accounting for uncertainty* is implemented as *a multi-level hierarchical system of MVC series*. Each *l-level* ($l = 1,\ldots,L$) of this system includes *a totality of l-scenarios*, having the same nature, specific only to this l-level. Such l-scenarios reflect the possible variations of parameters and conditions, corresponding to this l-level. *Each $l^\wedge$-MVC series*, corresponding to a fixed $l^\wedge$-level, reflects a combination of $l^\wedge$-scenarios from their totality and generates an appropriate *subset of "stable-optimal" alternatives*. Forming a combination of $l^\wedge$-MVC series allows to find the appropriate set of "stable-optimal" subsets, corresponding to this $l^\wedge$-level ($l^\wedge = 1,\ldots,L$). On a basis of this set processing, a "stable-optimal" subset of the next upper ($l^\wedge$–1)-level might be determined using *a special procedure*. Its "key" operations are based on calculating the highest frequencies of entering into this full set for the alternatives from the "stable-optimal" subsets of $l^\wedge$-level, forming this set.

Thus, the multi-level hierarchical system of MVC series performance, realizing the calculation stage (3) to reach the resulting RAS, consists of a successive forming of sets of "stable-optimal" subsets for all l-levels ($l = 1,\ldots,L$), beginning with the lowest L-level ($l = L$) and ending with the top l-level ($l = 1$). The resulting (one or several) RAS should be also found as subset (subsets) of "stable-optimal" alternatives, where this finding represents a final operation in the calculation process considered. In a case when several RAS are derived, their non-formal analysis is performed in the second step of the whole solving process in order to find a final solution of the problem.

The *multi-level hierarchical system of MVC series* can have various structures and content, differing by the number of l-levels as well as the accepted *system of l-scenarios*. We have already considered *nine-* and *six-level hierarchical systems* with some variations in their totalities of l-scenarios [4–6].

At present, we use *a six-level hierarchical systems* [4–6] based on *the multi-criteria optimization technique* TOPSIS [7], modified to consider *the criteria weights as random variables* which are presented by the intervals of their *possible values* [1–6]. These values are determined inside these intervals by *Monte Carlo simulations.*

3. Illustration of the six-level hierarchical system performance on the sample

The process of this six-level hierarchical system performance for the simple sample is illustrated in Fig. 1. This process consists of successive forming of all possible l-MVC series ($l = 1,\ldots,6$), from the lowest 6-level ($l = 6$) to the top 1-level ($l = 1$), and determination of the "stable-optimal" subset for each such l-MVC series, reflecting a combination of full Scenarios. Each *full Scenario* is a combination of l-scenarios, taken one at a time for each of all l-levels ($l = 1,\ldots,6$). We will present this calculation process to reach the RAS, performed from "bottom" ($l = 6$) of the hierarchical system to its "top" ($l = 1$), for the conditions of the considered sample.

The initial data of the considered simple sample are presented in Fig. 2, where 40 points reflect all initial alternatives, i.e., the ISA includes 40 i-alternatives (points $\{i = 1-40\}$). Each such i-point (i-alternative) has two coordinates (the criteria values $\{C_{ij},\ j = 1,2\}$), for example in Fig. 2 we can see that 1-point (1-alternative) has the coordinate (criteria) values $\{C_{11} = 1.0,\ C_{12} = 13.5\}$.

We will consider this MCDM problem according to the above mentioned principle, where multi-criteria optimization (MCO) on all l-levels of this six-level hierarchical system is based on finding the following minimal (for all i) scalar sums with the random j-criteria weights $\{W_j,\ j = 1,\ldots,J\}$:

$$\min_{\{i=1,\ldots,I\}} \{C_{i1}W_1 + \ldots + C_{ij}W_j + C_{iJ}W_J\}. \qquad (1)$$

Here, these random variables are presented by the intervals $\{[w_j^{\min}, w_j^{\max}],\ j = 1,\ldots,J\}$ of their possible values $\{w_j,\ j = 1,\ldots,J\}$, where each such value is chosen within its interval using a Monte Carlo simulation.

According to our sample conditions, we have $\{i = 1,\ldots,40;\ I = 40\}$ and $\{j = 1,2;\ J = 2\}$, that leads to the necessity to solve the following minimization problem:

$$\min_{\{i=1,\ldots,40\}} \{C_{i1}W_1 + C_{i2}W_2\}. \qquad (1')$$

The considered six-level hierarchical system of MVC series, applicable to the sample conditions, should include the following l-scenarios for all l-levels (l = 1–6) of this system:

- $l = 1$. Single 1-scenario, representing a version of MCO technique TOPSIS (see [1–3]), modified to consider the criteria weights as random variables and to use Monte Carlo simulations (the appropriate MCDM problem was reflected above by formula (1′).
- l = 2–3. Single 2-, 3-scenarios, reflecting single versions for the ISA/ISCAV, both corresponding to data represented in Fig. 2 and reflecting 40 initial alternatives, having the numbers {points $i = 1,\ldots,40$}, as well as 2 criteria with the numbers $\{j = 1,2\}$.
- l = 4–5. The totalities of 4-, 5-scenarios reflect the accepted (see [6]) 9 versions of possible values' intervals for the random criteria weights W1,W2:
  (1) $\{[0.475, 0.525\}, [0.475, 0.525\}\}$;
  (2) $\{[0.45, 0.55\}, [0.45, 0.55\}\}$; …
  (4) $\{[0.6175, 0.6825\}, [0.3325, 0.3675\}\}$; …
  (9) $\{[0.2975, 0.4025\}, [0.5525, 0.7475\}\}$.
- $l = 6$. The totalities of 6-scenarios represent 90 combinations of possible values of random criteria weights, obtained within the above 9 intervals using the 10 assigned series of Monte Carlo simulations. Each series of 2 Monte Carlo simulations generates

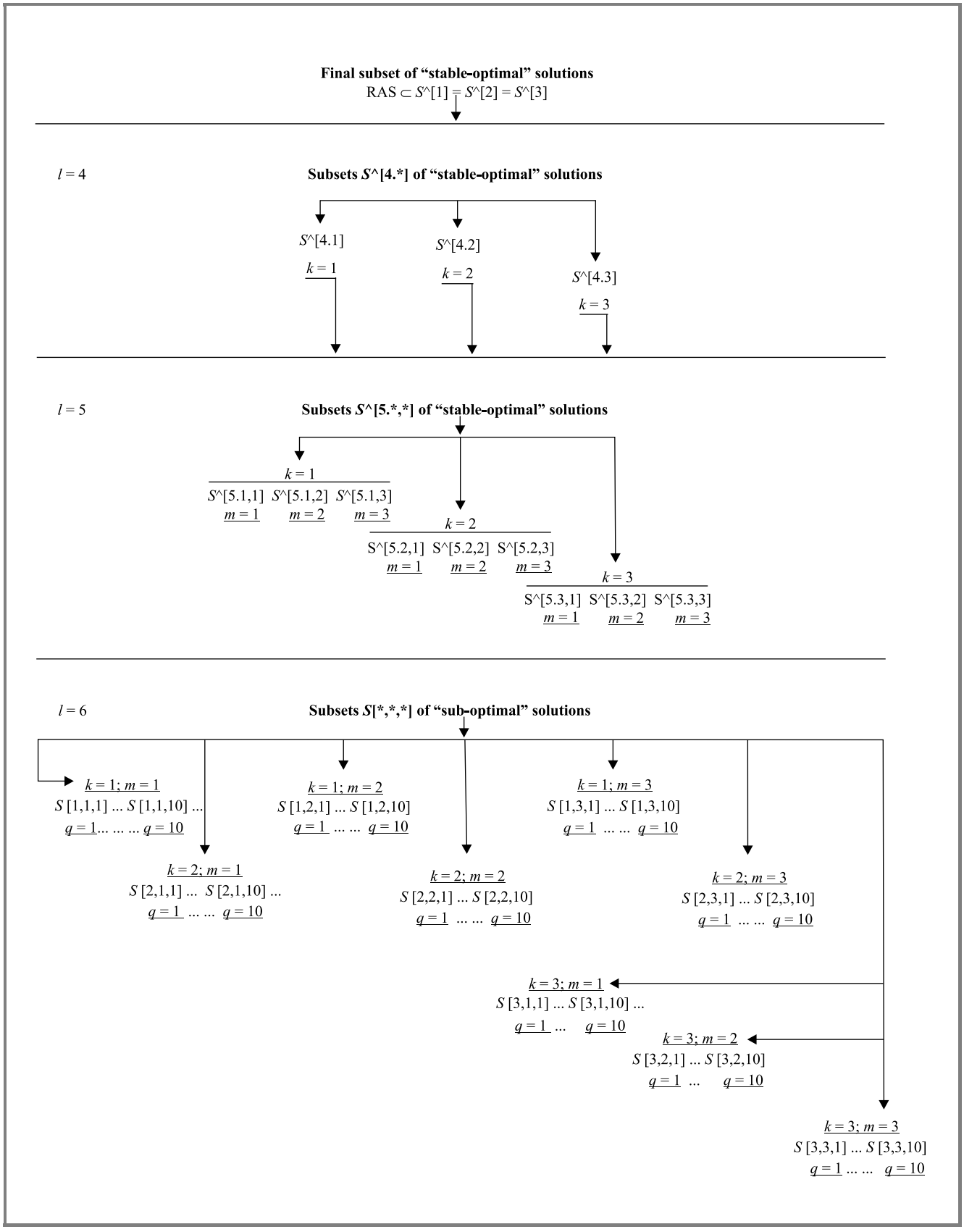


Fig. 1. Results of the calculation process for the considered sample: ninety (90) of "sub-optimal" subsets ($l = 6$), deriving nine (9) "stable-optimal" subsets ($l = 5$), from them–three (3) "stable-optimal" subsets ($l = 4$), and finally–the resulting subset RAS = $S^{\wedge}[1] = S^{\wedge}[2] = S^{\wedge}[3]$.

one combination of 2 weight values $\{w_1, w_2\}$ inside the appropriate intervals $\{[w_1^{\min}, w_1^{\max}], [w_2^{\min}, w_2^{\max}]\}$ (e.g., we have such intervals (4) $\{[w_1^{\min} = 0.6175, w_1^{\max} = 0.6825], [w_2^{\min} = 0.3325, w_2^{\max} = 0.3675]\}$ and their inside values $\{w_1 = 0.6643, w_2 = 0.3412\}$).

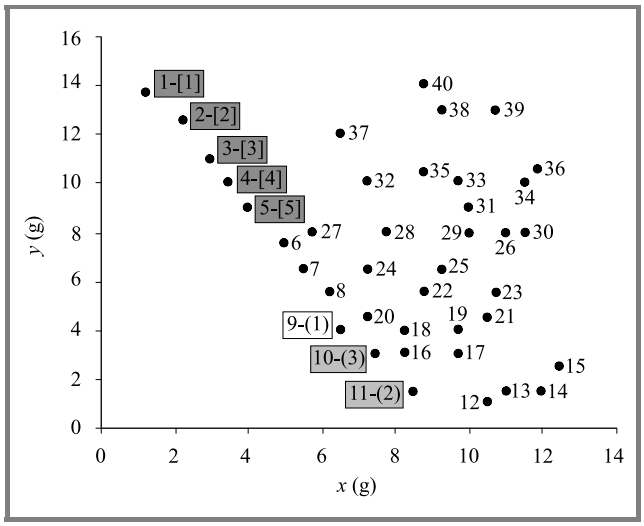


Fig. 2. Points and their coordinates reflecting ISA and ISCAV.

Consideration of all l-scenarios mentioned above allows to form the totality of 90 full Scenarios $\{1,1,1,k,m,q\}$, corresponding to single l-scenarios for three of the first l-levels (l = 1–3), K ($k = 1,\ldots,3$; $K = 3$) 4-scenarios, M ($m = 1,\ldots,3$; $M = 3$) 5-scenarios and Q ($q = 1,\ldots,10$; $Q = 10$) 6-scenarios.

Each full Scenario defines a *mono-optimization problem* (1′) and its solution–a subset, including a predetermined number of "sub-optimal" alternatives. Analysis of sets of such subsets allows to find the "*stable-optimal*" *subsets*. All this forms the solution process to reach the RAS, implementing the considered six-level hierarchical system performance for this sample. This solution process is presented in Fig. 1.

According to Fig. 1, the above mentioned solution process includes the following *operations*:

a) Choosing the first full Scenario–combination $\{1,1,1,1,1,1\}$, reflecting 1-scenarios taken one at a time for each of all l-levels (l = 1–6) $\{k = 1, m = 1, q = 1\}$; we form a *mono-optimization sub-problem* (1′) using 2 Monte Carlo simulations to determine the values of scalar sums for 40 (the number of alternatives) *criteria assessment vectors*.

b) A predetermined number (e.g., $N = 12$) of minimal values are selected among these scalar sums, that allows to form the subset $S[1,1,1,1,1,1]$ or $S[1,1,1]$ (in Fig. 1) of *sub-optimal alternatives*, including only their numbers (e.g., $\{11,9,\ldots\}$).

c) In the same way, by varying all ten 6-scenarios and leaving unchanged l-scenarios for all upper five l-levels (l = 1–5), we determine the set of "sub-optimal" subsets $\{S[1,1,1],\ldots,S[1,1,10],\ k = 1, m = 1,\ q = 1,\ldots,10\}$. It defines a MVC series, for which a *subset* $S^{\wedge}[5.1,1]\{l = 5,\ k = 1,\ m = 1\}$ of "stable-optimal" alternatives is determined using the special procedure.

d) Repeating the preceding operations, while varying all three 5-scenarios $\{l = 5,\ m = 1,2,3\}$ and leaving unchanged l-scenarios for all upper four l-levels (l = 1–4), a set of "stable-optimal" subsets $\{S^{\wedge}[5.1,1], S^{\wedge}[5.1,2], S^{\wedge}[5.1,3],\ l = 5,\ k = 1,\ m = 1,2,3\}$ is determined. This allows to find (on a basis of this set analysis) the "stable-optimal" subset $S^{\wedge}[4.1]$ $\{l = 4,\ k = 1\}$, corresponding to the next 4-level.

e) Continuing this process, we can define the set of "stable-optimal" subsets $\{S^{\wedge}[4.1], S^{\wedge}[4.2], S^{\wedge}[4.3],\ l = 4,\ k = 1,2,3\}$ and for it–the "stable-optimal" subset $S^{\wedge}[3]$ $\{l = 3\}$, corresponding to the next 3-level. Since we have only single version for each of three upper l-levels (l = 1–3), their "stable-optimal" subsets are the same ones: $S^{\wedge}[3] = S^{\wedge}[2] = S^{\wedge}[1]$. This is *the resulting RAS*, including 12 ($S = 12$) *i*-points/"reasonable" alternatives $\{9, 11, 10, 8, 20, 7, 16, 12, 5, 6, 13, 18\}$, which are picked out in Fig. 2.

We would like to underline that in this RAS, all selected "stable-optimal" alternatives gain their priorities on a basis of *frequency of their presence* in all "stable-optimal" subsets, obtained for the preceding l-level, as well as accounting for the *sum of places*, which they have in these subsets.

Thus, the result we obtain for this sample is not an obvious one (with general positions). It becomes more illustrative if we locate all 40 alternatives-points in the Euclid space and estimate their distances to the Origin of the coordinates. We see in Fig. 2 that the best alternatives–points 9, 11, 10 are the nearest to the Origin of the coordinates.

4. The implementation for a problem of buying on the stock market

This implementation has two main purposes: (1) to apply our approach to the problem, where good statistic data allowing to reach the "reasonable" solutions using the proposed methodology are available; (2) to demonstrate the possibility of overcoming the difficulties of the ISCAV modeling in case of a real problem, where it's required to apply analytical methods of such modeling since it isn't possible to construct any natural (implicit) optimization criteria. Such an attempt was made early [6], but in this paper a new approach is demonstrated, where another totality of such criteria is considered, linked with other approach to construct a greater part of them.

4.1. Statement of the problem

The following problem is considered: *to select* (in the current T-day) a holding of stock, including a predetermined number of stocks, proper for buying on the stock market for the next prognosis $(T+1)$-day. Such selection is performed on a basis of processing the appropriate statistical data, where such data are considered for the period $[1,\ldots,T]$ of T days, as well as for expert estimates' use. All this concerns two parameters of stock market process: *deal sums, stock prices*. Thus, the problem solution purpose is to determine a proper quantity of each type of stock to be bought for the prognosis $(T+1)$-day.

These resulting proper stocks might be selected in accordance with one of two *purposes*: (a) to be sold on the days $(t = T+1, T+2, \ldots)$, nearest to the prognosis $(T+1)$-day (this is the *speculative Model A*); (b) to be kept for a long period as a part of decision makers' available capital (the *keeping Model B*).

The accent on this statistical data processing is more convenient in a framework of the *Model A* use; applying the *Model B* should be based, first of all, on using the expert estimates of production conditions for the enterprises, whose stocks are bought. However, the principal peculiarity of the considered problem, connected with the competition of great quantity of stocks, seriously limits the possibility to take such expert estimates for all considered stocks. Accounting for it, we are more closely focused on using the *Model A*, considering quite long period of the appropriate statistical data, especially if this period is characterized by "a stable behaviour" of the stock market considered. In this situation it's possible to expect that these observed statistical data are a sum of various aspects, affecting the "stock market behaviour of each considered stock", like the status of production, psychology, interaction of stock market buyers and sellers, etc.

The choice of this problem to apply the proposed approach of MCDM accounting for uncertainty is caused, first of all, by availability of required initial (statistical) data as well as of specific methodological difficulties to apply such approach to this real situation. Such difficulties were related mainly to criteria modeling and ISCAV construction.

4.2. Methodological peculiarities of ISA construction

For the problem considered, the operations necessary to construct the ISA have not been of our main methodological interest, since: (a) an initial alternative concept is very implicit–a stock itself is such alternative; (b) it was not needed to develop the special calculation procedures to construct a vast ISA. The latter (case (b)) is explainable by the measurable quantity of stocks, which can be considered in each existing stock market. In our opinion, this situation is very different from the one we met solving the problem [1–3, 5], with a practically non-measurable quantity of initial alternatives.

Thus, in the considered problem of stock buying the ISA (each its version) may be presented as the set $\{i = 1,\ldots,I\}$ of i-alternative's numbers.

4.3. Methodological peculiarities of criteria modeling (ISCAV construction)

The main methodological peculiarity of ISCAV construction in the considered problem is related to *non-implicit character of criteria*, which should express the optimization process in the problem. This aspect differentiates this problem from others (e.g., see [1–3, 5]), where there is a full possibility to assign (a priori) such natural criteria (e.g., economic, reflecting profit maximization or expenditure minimization; environmental–pollution minimization, etc.). Thus, the modeling process for the present problem is linked with necessity to perform some analytical research to express the required optimization criteria. Such approach allowed to define some basic concepts for modeling of various types of such criteria, providing the choice of "reasonable" stocks for buying on the prognosis $(T+1)$-day in order to sell them in the future in a short/long time (*Model A* and *Model B*).

4.3.1. Constructing the criteria, related to estimates of stock deal sums (DS) values [6]

The i-stocks, having the greatest expected *prognosis* (for $(T+1)$-day) *absolute values* $\{A^{pr}(i, T+1), i = 1,\ldots,I\}$ for *deal sums* (DS), are preferable in both *Model A* and *Model B*. Such confirmation is based on the following sentence: the stocks with the greatest values $A^{pr}(i, T+1)$ might be in great demand on $(T+1)$-day and few following days. Expected values $\{A^{pr}(i, T+1), i = 1,\ldots,I\}$ are determined on a basis of the appropriate *statistical data processing* to define their trends as well as by an implicit assigning of expert estimates. The first way corresponds to *Model A*, since it allows to estimate the conditions for selling the stocks on the days nearest to the $(T+1)$-day; the second expert way is more convenient to *Model B*, since experts might predict more long-term tendencies. However, this expert way is difficult to implement for a large quantity of stocks. In such case, the first trend way might also be used for the *Model B*, if we could determine reliable long-term trends. In both cases (for *Model A* and *Model B*), when the trends might be used to find the required prognosis, various trend types are found, and the required values $\{A^{pr}(i,t), i = 1,\ldots,I;\ t = T+1, T+2,\ldots\}$ are determined as the weighted values of these trends' continuations. Further, we will consider the criteria for *Model A* only.

Thus, we use the *maximization* (on all i-stocks, $i = 1,\ldots,I$) *of absolute prognosis* (on $T+1$-day) DS *values* $A^{pr}(i, T+1)$ as the *Criterion 1*, formulated as follows:

$$\max_{\{i=1,\ldots,I\}} \{C_{i1}\} = \max_{\{i=1,\ldots,I\}} \{A^{pr}(i, T+1)\}, \qquad (2)$$

where these values $\{A^{pr}(i,\ T+1),\ i = 1,\ldots,I\}$ are determined by constructing various types of trends and weight-

ing these trend prognosis (on $T+1$-day) values. Let's consider the appropriate method.

These *trends* should be calculated on a basis of statistical data processing. The period of statistical data, intended for such processing, could be arbitrary, and its optimal duration could be established after many tests.

At present, the following 6 *types of trends* may be used to obtain the following prognosis (for any t-day) values for all i-stocks ($i=1,\ldots,I$), based on statistical data processing for the period $[1,\ldots,t-1]$: (1) *Linear (Lin)* $A^{(1)}(t)$, (2) *Exponential (Exp)* $A^{(2)}(t)$, (3) *Logarithmic (Log)* $A^{(3)}(t)$, (4) *Polynomial (Pol) 3rd (third) order* $A^{(4)}(t)$, (5) *Power (Pow)* $A^{(5)}(t)$, (6) *Hyperbolic (Hpr)* $A^{(6)}(t)$.

When all these trend prognosis values $\{A^{(j)}(t),\ j=1,\ldots,6\}$ are obtained, the required *prognosis* (*for any t-day*) values $A_i^{pr}(t)$ of the *Criterion 1* are calculated for all i-stocks ($i=1,\ldots,I$) as follows:

$$A^{pr}(i,T+1)=A^{(1)}(i,t)w^{(1)}+A^{(2)}(i,t)w^{(2)}+\ldots$$
$$+A^{(5)}(i,t)w^{(5)}+A^{(6)}(i,t)w^{(6)}, \qquad (3)$$

where the weight values $\{w^{(1)},w^{(2)},w^{(3)},w^{(4)},w^{(5)},w^{(6)}\}$ are assigned by experts or calculated using appropriate models. In both cases, we can't consider these weights as random variables.

For the latter case, the *special heuristic multi-step procedure* was developed early [6] to calculate *weights* $\{w^{(1)},w^{(2)},w^{(3)},w^{(4)},w^{(5)},w^{(6)}\}$, using the same statistical data. This procedure is based on finding the values $A^{pr}(i,T+1)$ from formula (3) for the days $\{t=T+1-k,\ 0<k\ll T\}$, preceding the basic $(T+1)$-prognosis day. In this case, the weight values, needed for (3), should reflect the validity of all trend-prognosis values $A^{pr}(i,T+1)$ by the following way of their calculation for all preceding t-days, corresponding to the assigned (a priori or in the calculation process itself) integer numbers k, and of the comparison of these calculated values with the appropriate actual (statistical) data:

Step 1: Stepping k^*-days back from the prognosis $(T+1)$-day (where $0<k^*\ll T$ is a fixed integer number), we determine the DS values $\{A^{(j)}(i,t^*),\ j=1,..,6\}$ from (3) for the chosen prognosis t^*-day ($t^*=T+1-k^*$) and each i-stock ($i=1,\ldots,I$), according to all 6 trends considered.

Step 2: Comparing these values $\{A^{(j)}(i,t^*),\ j=1,\ldots,6\}$ with the appropriate fact (statistical) DS values $A^{fact}(i,t^*)$, we can calculate the *aberration values* $D^{(j)}(i,t^*)$ as follows:

$$D^{(j)}(i,t^*)=\left|[A^{(j)}(i,t^*)-A^{fact}(i,t^*)]\right|/A^{fact}(i,t^*),$$
$$j=1,\ldots,6;\ \ i=1,\ldots,I. \qquad (4)$$

Step 3: Comparing the derived aberration values (4) with the *limits* $L(t^*)$, assigned (a priori) for deal sums (DS) and t^*-day, we select the "good" i-stocks, having these absolute aberration values lower than limits ($D^{(j)}(i,t^*)<L(t^*)$). The *numbers* (*quantity*) *of "good" i-stocks*, corresponding to each considered j-trend and reflecting its validity, are designated as $\{N^{(j)}(t^*),\ j=1,\ldots,6\}$.

Step 4: We determine the weights $\{w^{(j)}(t^*),\ j=1,\ldots,6\}$ from (3) as follows:

$$w^{(j)}(t^*)=N^{(j)}(t^*)/\left[N^{(1)}(t^*)+N^{(2)}(t^*)+\ldots\right.$$
$$\left.+N^{(5)}(t^*)+N^{(6)}(t^*)\right],\quad j=1,\ldots,6. \qquad (5)$$

This four-step procedure might be repeated for several (R) t-days $\{t=t^1,\ldots,t^R\}$, corresponding ($t=T-k$) to the assigned values $k\{k=k^1,\ldots,k^R\}$. Thus, we obtain R totalities of weights (5). On their basis, we can find the following *weighted average weights* to obtain the criterion values $A^{pr}(i,T+1)$, according to (3):

$$w^{(j)}[R]=d(t^1)w^{(j)}(t^1)+d(t^2)w^{(j)}(t^2)+\ldots$$
$$+d(t^R)w^{(j)}(t^R),\quad j=1,\ldots,6, \qquad (6)$$

where $\{d(t^1),\ldots,d(t^R),\ldots\}$ are the assigned (a priori) weights of t-days and sum of these weights' sum for all such days should be equal to 1. This leads to the following formula:

$$w^{(1)}[R]+w^{(2)}[R]+\ldots+w^{(6)}[R]=1. \qquad (7)$$

It's possible to exclude some j-trends from consideration, if it is possible to take a priori their weights $w^{(j)}=0$ in formula (3). Besides this, we could vary the totalities of $\{w^{(j)}[R],\ j=1,\ldots,6\}$ from (6), (7), considering various totalities of j-trends as well as k-days $\{k=k^1,\ldots,k^R\}$.
To estimate the quality of the derived DS prognosis absolute values $\{A^{pr}(i,T+1),\ i=1,\ldots,I\}$, they are compared with the DS fact absolute values $\{A^{fact}(i,T),\ i=1,\ldots,I\}$, taken for the last T-day from the appropriate statistical data. The *minimization of relative estimates* $A^{pr\wedge}(i,T+1)$, reflecting for each i-stock the ratio of absolute value of this pair values' difference to the last (fact) of them, is *Criterion 2*:

$$\min_{\{i=1,\ldots,I\}}\{C_{i2}\}=\min_{\{i=1,\ldots,I\}}\{A^{pr\wedge}(i,T+1)\},$$
$$A^{pr\wedge}(i,T+1)=\left\{\left|(A^{pr}(i,T+1)-A^{fact}(i,T))\right|/A^{fact}(i,T)\right\},$$
$$i=1,\ldots,I. \qquad (8)$$

Use of this *Criterion 2* is based on the following principle: when the value $A^{pr\wedge}(i,T+1)$ is less, the appropriate DS prognosis absolute value $A^{pr}(i^*,T+1)$ may be considered more reliable.

4.3.2. Constructing the criteria, connected with estimates of stock prices values

The approach to form the criteria for stock prices (SP) for *Model A* should be, in principle, different from the one for DS considered above. If we consider *Model A*, oriented on the stock buying for the prognosis $(T+1)$-day and their selling for the next t-days $\{t = T+2, T+3, \ldots\}$, we should account for the sinusoidal character of changing the SP values during the whole period of SP values observation (the days $[1, \ldots, T]$). This sinusoidal pattern is dictated by the nature itself, of stock buying on stock market when the SP falls, volumes of such stocks buying increase, and this tendency remains up to the day of reaching SP minimum. After this, another picture is observed, when the SPs go up, and holders of these stocks begin to sell such stocks, that leads again to SPs fall, and so on. Accounting for it, in a framework of *Model A*, it's expedient to buy those stocks, which is expected when the current minimum of its SP sinusoid is close to the prognosis for $(T+1)$-day. Such aspect might be reflected by *Criteria 3–5*, having very specific methodology of their construction, based on accounting for sinusoidal pattern of the initial data used.

In accordance with the sinusoidal character of changes to SP values, the set of all "*j-sinusoidal Hills*" $\{H[i,j], j = 1, \ldots, J[i]\}$, corresponding to each i-stock, is defined as one, reflecting the SP values, given on a totality of time j-intervals $\{[Tl[i,j], Tr[i,j]],\ j = 1, \ldots, J[i]\}$, measured in integer numbers of days of the period $[1,T]$. Each such j-interval includes t-days from interval $[1,T]$, arranged between the *left/right bounds* $(Tl[i,j] < Tr[i,j])$ of this j-interval. These bounds (for $j = 1, \ldots, J[i]$) corresponding to the neighboring minimal SP values, observed for the period $[1,T]$, are subject to the following conditions:

$$1 \le Tl[i,1] < Tr[i,1] < Tl[i,2], \ldots, Tl[i,j] < Tr[i,j], \ldots$$
$$\ldots, Tl[i,J[i]] < Tr[i,J[i]] \le T. \qquad (9)$$

Let's consider *a heuristic algorithm* of these *j-intervals constructing*, defined for each i-stock and based on considering the appropriate indicators, where $t[j]$ is the number of current day [j-interval] in the observed period $[1, \ldots, T]$ of statistical data, $SP[t]$ is the SP value for this t-day, and $Dif[t] = SP[t] - SP[t-1]$:

1. Establishing the *initial conditions*: $Dif[1] = 0$, $t = 2$, $j = 1$, $Dif[2] = SP[2] - SP[1]$.

2. Going to the next $(t+1)$-day with calculation $Dif[t+1] = SP[t+1] - SP[t]$.

3. Performing the joint analysis of signs for the differences $Dif[t+1], Dif[t]$; the following cases may be:

   3.a) If $Dif[t+1] > 0$, $Dif[t] \le 0$, we fix the upper limiting day $Tc = t+1$ for this *cycle* 3.a of finding the first preceding $(Tc-t1)$-day, when the value $Dif[Tc-t1] < 0$ $(t1 = 1, \ldots, Tc-1)$. If this value is reached, the right bound-day $Tr[i,j] = t$ of j-Hill is established to reflect the current minimal SP value in the sinusoid considered (for i-stock), and we go to its next $(j+1)$-Hill with fixing its left bound-day $Tl[i, j+1] = t+1$. Thus, the j-interval $[Tl[i,j], Tr[i,j]]$ is completed. However, in the process of executing this cycle, we can encounter two particular *cases*:

   3.a1) $Dif[Tc-t1] > 0$ for some value $t1$ $(t1 = 1, \ldots, Tc-2)$;

   3.a2) $Tc - t1 = 1$, i.e., we reached in this cycle the beginning of sinusoid (the 1-day).

   In both these cases (3.a1) and (3.a2), operations of *cycle* 3.a end without completion of the interval, and we go to operation 2 for the next $(t+2)$-day.

   3.b) If $Dif[t+1] < 0$, $Dif[t] \ge 0$, we execute the *cycle* 3.b (it's similar to the *cycle* 3.a) of finding the first preceding $(Tc-t1)$-day, when $Dif[Tc-t1] > 0$ $(t1 = 1, \ldots, Tc-1)$. If this is reached, we fix the day $T\max[i,j] = t$ of maximal SP value for j-Hill with saving j-Hill and its other indicators without changes. In this case we can also encounter two particular *cases*:

   3.b1) $Dif[Tc-t1] < 0$ for some value $t1$ $(t1 = 1, \ldots, Tc-2)$ and execution of *cycle* 3.a is stopped on $(Tc-t1)$-day without establishing $T\max[i,j]$;

   3.b2) *cycle* 3.b ends, reaching (Tc–$t1$=1)-day, and we fix the day $T\max[i, j=1] = t$.

4. After completion of all these 3 operations and for all other cases of relationship between $Dif[t+1]$ and $Dif[t]$, we go to the next $(t+2)$-day of the period $[1, \ldots, T]$ $(t < T)$ to perform the next operation 2.

We bring the following appropriate *parameters* into operations to construct the *Criteria 3–5*:

– The *summary distances* $DsSHw[i]$ or $DsSHp[i]$ (both in days) *of all j-intervals or their part*, corresponding to the full period $[1,T]$ or its continuous part, including the assigned number $K(>1)$ of j-intervals from the fixed j^*-interval to the $J[i]$-interval $(j^* = J[i] - K)$. It's calculated as follows:

$$DsSHw[i] = Tr[i,J[i]] - Tl[i,1] + 1,$$
$$DsSHp[i] = Tr[i,J[i]] - Tl[i,j^*] + 1,$$
$$i = 1, \ldots, I. \qquad (10)$$

– The *average distances* $DsAHw[i]$, $DsAHp[i]$ (both expressed in days, maybe with 0.1 day resolution),

determined as the quotient from division of the summary distances (10) on the appropriate numbers of j-intervals, forming these summary distances:

$$DsAHw[i] = DsSHw[i]/J[i],$$

$$DsAHp[i] = DsSHp[i]/K,$$

$$i = 1,\ldots,I. \tag{11}$$

– The *residual distance* $DsRHl[i]$ (in days) reflects the *last non-completed* $(J[i]+1)$*-Hill* (*-interval*), including all t-days after $Tr[i,J[i]]$ (the $J[i]$-day of the last minimal SP value) up to the final T-day of the period $[1,T]$, where:

$$DsRHl[i] = T - Tr\big[i,J[i]\big], \quad i = 1,\ldots,I. \tag{12}$$

Using these parameters, we can define *Criteria 3–5* as follows:

1. According to the formulas (9)–(11), the *average distances* $DsAHw[i]$, $DsAHp[i]$ from (11) are calculated for all i-stocks considered. By comparing them, the *relative coefficient* $CfAwp[i]$, characterizing their closeness, is calculated as the quotient from division of the absolute value of these average distances' difference on their maximal value; the appropriate formula is as follows:

$$CfAwAp[i] = \big|\big(DsAHw[i] - DsAHp[i]\big)\big|/$$

$$/\max\big\{DsAHw[i], DsAHp[i]\big\}, \quad i = 1,\ldots,I. \tag{13}$$

2. The *final average distance* $DsAHF[i]$, determined as weighted sum of both average distances from (11), should reflect (as well as the *Criterion 3* itself) the prediction of the expected day $Tr\big[i,J[i]+1\big]$ (maybe with 0.1 day resolution) of current minimal SP value (this prognosis day of $(J[i]+1)$-SP minimum should be the next one after the last fact day of SP minimum). Thus, $DsAHF[i]$ is calculated as:

$$DsAHF[i] = DsAHw[i]\cdot WAw + DsAHp[i]\cdot WAp,$$

$$i = 1,\ldots,I, \tag{14}$$

where the expert weights WAw, WAp $(WAw + WAp = 1)$ are assigned according to the following principle: *when making the prognosis for the nearest future it's more important to take into account what happened in the last fact days*. According to it, we take $WAp > WAw$ (it's possible to take: $WAw = 0.4$, $WAp = 0.6$).

3. We formulate the *Criterion 3*, based on the above considered principles, as *minimization* (on all totality of i-stocks) *of closeness for the i-stock expected* $(J[i]+1)$*-day of its SP sinusoid minimum to the accepted prognosis* $(T+1)$*-day*. This expected $(J[i]+1)$-day is calculated (maybe in with 0.1 day resolution) for each i-stock as the end of its last non-completed $(J[i]+1)$-sinusoidal Hill $H\big[i,J[i]+1\big]$:

$$Tr\big[i,J[i]+1\big] = Tr\big[i,J[i]\big] + DsAHF[i],$$

$$i = 1,\ldots,I, \tag{15}$$

where: $Tr\big[i,J[i]\big]$ is the last fact day of SP minimum, $DsAHF[i]$ is from (14).

Thus, the *Criterion 3* is minimization of the value, calculated as the corrected (using (13)) absolute value of difference between the rated day from (15) and the prognosis $(T+1)$-day:

$$\min_{\{i=1,\ldots,I\}}\{C_{i3}\} = \min_{\{i=1,\ldots,I\}}\{C_3^{\sin}[i,T+1]\},\ C_3^{\sin}[i,T+1]$$

$$= \big|\big(Tr[i,J[i]+1] - T - 1\big)\big|\cdot(1 + CfAwAp[i])\},$$

$$i = 1,\ldots,I. \tag{16}$$

4. The *Criterion 4* should reflect *the accordance between the last fact non-completed* $(J[i]+1)$*-sinusoidal Hill* $H\big[i,J[i]+1\big]$ *and the prognosis of* $(T+1)$*-day*. This accordance may be estimated as follows: we find the *day* $T\max\big[i,J[i]+1\big]$ of maximal value on this last non-completed Hill, and the *Criterion 4* is considered as minimization of relative difference between this maximal day and this *Hill initial day* $Tr\big[i,J[i]\big]$; thus, the *Criterion 4* value is calculated as the quotient from division of this difference on this Hill full distance $DsRHl[i]$ (see (12)):

$$\min_{\{i=1,\ldots,I\}}\{C_{i4}\} = \min_{\{i=1,\ldots,I\}}\{C_4^{\sin}[i,T+1]\},\ C_4^{\sin}[i,T+1]$$

$$= \big(T\max\big[i,J[i]+1\big] - Tr\big[i,J[i]\big]\big)/\big(T - Tr\big[i,J[i]\big]\big),$$

$$i = 1,\ldots,I. \tag{17}$$

Such minimization approach is based on the following principle: if the relative value $C_4^{\sin}[i,T+1]$ is small, it is possible to expect the great distance $(T - T\max\big[i,J[i]+1\big])$, i.e., the next (out $[1,\ldots,T]$) SP minimum value could be expected as one closer to the prognosis of $(T+1)$-day. Thus, such stocks are good for buying (only with these positions) on this $(T+1)$-day. However, if the required value $T\max\big[i,J[i]+1\big]$ isn't found on the Hill $H\big[i,J[i]+1\big]$, we assume the *Criterion 4* value as follows:

$$C_4^{\sin}[i,T+1] = 1 + Eps[i], \quad i = 1,\ldots,I, \tag{18}$$

where the values $\{Eps1[i] > 0, i = 1,\ldots,I\}$ are some expert estimates $Eps[i]$ (or they may be accepted as $\{Eps[i] = 1/\big(T - Tr\big[i,J[i]\big]\big), i = 1,\ldots,I\}$).

5. The *Criterion 5* estimates *the difference between the fact and rated values of SP maximal value on the last*

non-completed Hill $H[i, J[i]+1]$. It may be accepted as an additional estimation of agreement between the fact and rated data. If there is the fact maximum $T\max[i, J[i]+1]$, the *Criterion 5* value is calculated by defining the absolute value of difference between the value $T\max[i, J[i]+1]$ and the rated maximal value of this Hill, derived using a half of distance $DsAHF[i]$ from (14), where this difference is corrected using the value $CfAwAp[i]$ from (13). Thus, we have the following *minimization Criterion 5*:

$$\min_{\{i=1,\ldots,I\}}\{C_{i5}\} = \min_{\{i=1,\ldots,I\}}\{C_5^{\sin}[i, T+1]\}, \qquad (19)$$

where:

$$C_5^{\sin}[i, T+1] = \big|\big(T\max[i, J[i]+1] - Tr[i, J[i]] - 0.5DsAHF[i]\big)\big| \cdot (1 + CfAwAp[i])$$

$$i = 1, \ldots, I. \qquad (19')$$

If $T\max[i, J[i]+1]$ does not exist, the *Criterion 5* value is accepted as in (18):

$$C_5^{\sin}[i, T+1] = \big|\big(T + Eps1[i] - Tr[i, J[i]] - 0.5DsAHF[i]\big)\big| \cdot (1 + CfAwAp[i]),$$

$$i = 1, \ldots, I, \qquad (20)$$

where the values $\{Eps1[i] > 0,\ i = 1, \ldots, I\}$ are some expert estimates.

4.3.3. Illustrative sample of *Criteria 1–5* constructing

We will illustrate the above considered methods of *Criteria 1–5* constructing on the real sample of data for one stock (3-stock) of the holding "Tel-Aviv-100" on the Israeli stock market. The appropriate statistical data for this stock are considered for the period 11/11/01–01/01/02, which included 32 working days ($T = 32$, $t = 1, \ldots, 32$). The prognosis day ($T + 1 = 33$) corresponds to 02/01/02. All these data present the deal sums (DS), shown in Fig. 3, and the stock prices (SP), presented in Table 1.

Construction of *Criterion 1*, reflecting the prognosis (for 33-day) of absolute value $A^{pr}(3, 33)$ for deal sums (DS), is performed according to (3), where only the linear (Lin) $A^{(1)}(3, 33)$, logarithmic (Log) $A^{(3)}(3, 33)$ and polynomial (Pol) 3rd order $A^{(4)}(3, 33)$ trends are taken into account (it corresponds to $w^{(2)} = 0$, $w^{(5)} = 0$, $w^{(6)} = 0$, $w^{(Ex)} = 0$ in (3)). These trends for 3-stock are presented in Fig. 3, where they are shown by bold (Lin), stroke (Log) and dotted (Pol) lines. In this Fig. 3, we can see that their prognosis (for 33-days) values are very close to Log- and Pol-trends ($A^{(3)}(3, 33) = 3757.9$; $A^{(4)}(3, 33) = 3754.4$), but they differ from Lin-trend ($A^{(1)}(3, 33) = 5029.1$). According to it and the closeness of values $A^{(3)}(3, 33)$ and $A^{(4)}(3, 33)$ to the known (statistics for 02/01/02) fact DS values ($y^{fact}(3, 33) = 3840.8$), we can expertly assign, to realize calculations according to formula (3) for this 3-stock, the following weight values: $w^{(1)} = 0.2$, $w^{(3)} = w^{(4)} = 0.4$ (we would like to underline that such assigning is performed conditionally accounting only for this situation and for this 3-stock). Using these weight values and in accordance with (3), we calculate this *Criterion 1* value $C_{31} = A^{pr}(3, 33) = 5029.1 \cdot 0.2 + 3757.9 \cdot 0.4 + 3754.4 \cdot 0.4 = 4010.7$ ($= 1.04\, y^{fact}(3, 33)$ is very closed to the fact).

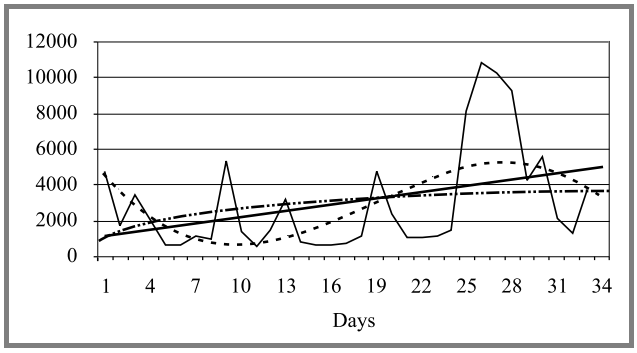


Fig. 3. Lin (bold), Log (stroke) and Pol (dotted line) trends ($j = 1, 3, 4$) and statistic curve for DS of 3-stock.

Using the derived value of *Criterion 1* and the fact DS value for the T-day $A^{fact}(3, 32) = 1293.7$, the *Criterion 2* value is calculated according to formula (8): $C_{31} = A^{pr\wedge}(3, 33) = |(A^{pr}(3, 33) - A^{fact}(3, 32))/A^{fact}(3, 32)| = |(4010.7 - 1293.7)|/1293.7 = 2.10$.

We illustrate construction of *Criteria 3–5* through the analysis of 3-stock price (SP) sinusoidal data, presented in Table 1. Such construction is performed according to (9)–(20) and the special procedure described before. The procedure defines the basic parameters of the criteria C_{33}–C_{35} construction, reflecting the "j-sinusoidal Hills" $\{H[3, j],\ j = 1, \ldots, J[3]\}$ and their j-intervals $\{[Tl[3, j], Tr[3, j]],\ j = 1, \ldots, J[3]\}$. So, according to operation 1 in this procedure, we establish the following initial conditions: $t = 2$, $j = 1$, $Dif[2] = SP[2] - SP[1] = 3283 - 3290 = -7$ (see Table 1). Executing operations 2–3, we find $t = 3$, $Dif[3] = 63$, and the case 3.a ($Dif[3] > 0$, $Dif[2] \le 0$) is established. Performing the appropriate cycle ($t1 = 1$, $Dif[t - t1] = Dif[2] < 0$), we fix the first minimal SP value for $t = 2$ ($Tr[3, 1] = 2$) and go to the next 2-Hill $H[3, 2]$ ($j = 2$, $Tl[3, 2] = 3$). Through the operation 4 going to the next day $t = 4$, we repeat again the operations 2–3, reaching $Dif[4] = -127$ and the case 3.b ($Dif[4] < 0$, $Dif[3] > 0$), where the maximal SP value in 2-interval is fixed for $t = 3$ ($T\max[3, 2] = 3$). Continuing this process, we define the following full totality of j-intervals for this 3-stock: [1, 2], [3, 7], [8, 12], [13, 17], [18, 22], [23, 24], [25, 26], [27, 32], including the following t-days of maximal SP val-

Table 1
SP values for 3-stock and their differences $Dif[t]$ for all t-days ($t = 1,\ldots,32$)

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
SP	3 290	3 283	3 346	3 219	3 144	3 097	3 040	3 050	3 437	3 424	3 417	3 388	3 478	3 427	3 368	3 36
$Dif[t]$		−7	63	−127	−75	−47	−57	10	387	−13	−7	−29	90	−51	−59	−12
t	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
SP	3 310	3 361	3 489	3 482	3 415	3 387	3 404	3 399	3 412	3 226	3 261	3 315	3 412	3 406	3 353	3 330
$Dif[t]$	−46	51	128	−7	−67	−28	17	−5	13	−186	35	54	97	−6	−53	−23

ues: $\{1,3,9,13,19,23,25,29\}$. Thus, we can obtain the values needed for (9)–(20): $J[3]=7$, $Tl[3,1]=1$, $Tr[3,7]=26$, $DsSHw[3]=26-1+1=26$, $K=3$ (it's assigned a priori), $j*=7-3=4$, $DsSHp[3]=Tr[3,7]-Tl[3,4]+1=14$, $DsAHw[3]=DsSHw[3]/J[3]=26/7=3.714$, $DsAHp[3]=DsSHp[3]/K=14/3=4.667$, $DsRHl[3]=T-Tr[3,7]=32-26=6$ (for non-completed 8-interval [27,32]), $CfAwAp[3]=|(DsAHw[3]-DsAHp[3])|/\max\{DsAHw[3],\ DsAHp[3]\}=|3.714-4.667|/4.667=0.204$. According to (11), (14) and the above accepted expert weights ($WAw=0.4, WAp=0.6$), we find the values $DsAHF[3]=3.714\cdot 0.4+4.667\cdot 0.6=4.286$, $Tr[3,8]=26+4.286=30.286$.

Thus, in accordance with the above calculated parameter values using formulas (16)–(20) and accounting for the SP maximum $T\max[3,8]$ availability in the last non-completed 8-interval [27, 32], the following *Criteria 3–5* values are calculated:

$$C_{33} = C_3^{\sin}[3,33] = |(30.286-33)|\cdot 1.204 = 2.714\cdot 1.204 = 3.268;$$

$$C_{34} = C_4^{\sin}[3,33] = (29-26)/(32-26) = 3/6 = 0.5;$$

$$C_{35} = C_5^{\sin}[3,33] = |(29-26-0.5\cdot 4.286)|\cdot 1.204 = 1.032.$$

We could perform some preliminary analysis of these criteria values obtained $\{4010.7, 2.10, 3.268, 0.5, 1.032\}$ with the intent to estimate quality of these results in accordance with these criteria. As is evident from the foregoing, the concordance with the fact data for the prognosis 33-day period is very good (104%) for the *Criterion 1* and very bad (210%) for the *Criterion C2*. The first characterizes the good reflection of general tendencies of DS statistical data by the chosen totality of trends and their weight values, the second–the sharp fall of fact DS on day 32 (see Fig. 3). The good values of *Criteria 3–5* characterize a good estimate of current SP minimum, calculated taking into account the fact 3-sinusoid "behaviour" in the average and in the last t-days.

4.3.4. Peculiarities of the six-level hierarchical system performance for the problem considered

The above considered (Section 3) six-level hierarchical system of MVC series is applied, at present, to reach the required "reasonable" solutions for the problem of stock buying on the stock market. We will not comment here on the aspects of this application contents and results, but we will accent only the methodological aspects of this application. With these positions we will consider the peculiarities of: (a) forming this hierarchical system; (b) performance of the appropriate computations.

To solve the problem considered, we accepted performance, in principle, of the same six-level system that was presented above (Section 3). The totalities of possible scenarios for all 6 levels ($l = 1,\ldots,6$) of this six-level system of MVC series as applied to this problem of stock buying are as follows:

- $l = 1$. Only one 1-scenario is accepted, intended to use the same modified TOPSIS method [1–3], based on considering the following *scalar goal function* (the partial case of the above general function (1))

$$\min_{\{i=1,\ldots,I\}}\{(-C_{i1}W_1)+C_{i2}W_2+C_{i3}W_3+C_{i4}W_4+C_{i5}W_5\}, \tag{21}$$

  where the first component, reflecting the maximizing *Criterion 1*, is considered with the sign "-".

- $l = 2$. Various 2-scenarios correspond to different ISA versions, whose variation might reflect the change of the observed stock groups, where this changes may include: (a) types of stocks; (b) period of observation; (c) number of considered stocks inside the same group, etc.

- $l = 3$. At present, we consider only one 3-scenario, reflecting the *Criteria 1–5*, presented in (21).

- $l = 4-5$. Various combinations of 4- and 5-scenarios should reflect different versions of weights of possible value j-intervals $\{[w_j^{\min}, w_j^{\max}],\ j = 1,\ldots,5\}$, whose construction is linked with assigning (a priori) their "central points" $\{w_j^{\wedge}, j = 1,\ldots,5\}$ (4-scenarios) and their quite standard surroundings by the bounds (5-scenarios). In a framework of multi-variant computations performed, we consider different versions of such "central points", reflecting: (a) expert estimates, where the weight $w_1^{\wedge}$ of *Criterion 1* is more preferable than others ($w_1^{\wedge} > w_j^{\wedge},\ j = 2,\ldots,5$) and

the weight $w_2^{\wedge}$ of *Criterion 2* is the least one; (b) estimates, opposite to the preceding case (a); (c) the uncertain situation ($w_1^{\wedge} = w_j^{\wedge} = w_3^{\wedge} = w_4^{\wedge} = w_5^{\wedge} = 0.2$). The totality of 5-scenarios used presents 4 types of interval bounds (see [6]), surrounding these "central points" on: (1)5%, (2)10%, (3)15%, (4)20% (e.g., we consider the interval $[0.95w_j^{\wedge}, 1.05w_j^{\wedge}]$).

- $l = 6$. The quantity of 6-scenarios depends upon the assigned number of Monte Carlo simulation series (each such series includes 5 Monte Carlo simulations according to the number of criteria). This quantity might be considerable according to our wishes (e.g., it's varied from 10 to 100 to estimate the sensitivity of results to such variations).

Thus, the above considered variations of 2-, 4-, 5- and 6-scenarios lead to a considerable number of multi-variant computations, related to reaching different RAS totalities. Their analysis allows, in principle, to research the influence of varying the problem conditions on the final solutions, but this analysis of varied results have led to the necessity to perform extensive research before the development of pithy (rich in content) methodology for behaviour on stock market. However, the research already performed shows *viability of our MCDM approach as applied to this real problem*.

5. Conclusions

An original, intuitive methodology to solve various real MCDM problems is proposed. This methodology reflects the approach, focused primarily on accounting for uncertainty factors in the process of selecting a predetermined number of "reasonable" alternatives from their considerable (maybe vast) initial set in accordance with an arbitrary number of optimization criteria, considered jointly as a multiple criteria. The methodology allows to take into account the uncertainty factors of different nature in a framework of multi-level hierarchical system of multi-variant computation series.

At present, the main purpose of promotion of this methodology is the research connected with application of this approach to real MCDM problems. The "bottlenecks" of such application may result from creating the initial sets of alternatives or criteria assessment vectors (ISA or ISCAV). From this point of view, the problem considered in this paper is a very suitable one, since it reflects the rich statistical data, which could be used to apply the proposed approach. Besides, this problem is connected with a lack of implicit optimization criteria, creating difficulties in ISCAV construction.

Thus, the main purpose of this paper was to show the possibility of applying the quite general methodology of MCDM accounting for uncertainty to real problem proposed, using the quite reliable initial data regarding the construction of initial alternatives and multiple criteria (the ISA/ISCAV). An important aspect of this application is related to *modeling the totality of non-implicit criteria* (ISCAV), based on reliable statistical data processing.

Other aspects of modeling the problem of stock buying on the stock market are linked to practical usage of the appropriate calculation results. Here, we can suggest two directions of work:

1) developing the methods of successful "behaviour" on stock market in buying the "reasonable" stocks;

2) accumulating the experience in researching such "behaviour" by performance of multi-variant computations to estimate the influence of various factors on the choice of "reasonable" solutions.

At present, we carry out work in the second direction, performing a lot of multi-variant computations while varying the problem conditions and parameters. In this way, we hope to start the "self-education" process, leading in the future to a development of successful "behaviour" on the stock market in buying the "good" stocks.

References

[1] V. I. Kalika and S. Frant, "Multi-criteria analysis accounting for uncertainty factors in electricity generation expansion planning", in *Proc. 12th Power Syst. Computat. Conf.*, Dresden, Germany, 1996, vol. 1, pp. 145–151.

[2] V. I. Kalika and S. Frant, "Environmental aspects of power generation", *Energy Sour.*, vol. 21, pp. 687–704, 1999.

[3] V. I. Kalika and S. Frant, "Multi-criteria optimization accounting for uncertainty in dynamic problem of power generation expansion planning", in *Research and Practice in Multiple Criteria Decision Making*, Y. Y. Haimes and R. E. Steuer, Eds., *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, 2000, vol. 487, pp. 409–420.

[4] V. I. Kalika, "Multi-level hierarchical system to take into account uncertainty in multi-criteria decision making process", in *Proc. Sixth Int. Symp. Anal. Hierar. Process, ISAHP 2001*, Switzerland, 2001, pp. 173–180.

[5] V. I. Kalika and S. Frant, "Multi-criteria approach for power generation expansion planning", in *Multiple Criteria Decision in the New Millennium*, M. Koksalan and S. Zuionts, Eds., *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, 2001, vol. 507, pp. 458–468.

[6] V. I. Kalika and G. Rossinsky, "Methodology of multi-criteria decision making accounting for uncertainty and some applications", *Int. J. Manag. Decis. Mak.*, vol. 4, no. 2/3, pp. 240–271, 2003.

[7] B. N. Massam, "Multi-criteria decision making (MCDM): techniques in planning", *Prog. Plan.*, vol. 30, pp. 1–84, 1988.

[8] T. L. Saaty, *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.

[9] T. L. Saaty, "The seven pillars of the analytic hierarchy process", in *Multiple Criteria Decision in the New Millennium*, M. Koksalan and S. Zuionts, Eds., *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, 2001, vol. 507, pp. 15–37.

[10] R. E. Steuer, *Multiple Criteria Optimization: Theory, Computation and Application*. New York: Wiley, 1986.

Vladimir Isaak Kalika has been a Senior Research fellow at Haifa University since 1992. From 1964, he worked towards a doctorate at the Moscow Central Economic-Mathematical Institute of the USSR Academy of Sciences, having received a Ph.D. in economic-mathematical modeling in 1967. Since 1967, working both in Russia and Israel, he had dealt with theoretical and practical problems of mathematical modeling in the fields of environment protection, electricity generation, oil pipeline systems, gas supply, construction, road transport and finance. Dr. Kalika has more than 150 publications to his name; most recently these (more than 20) have concentrated on the development of a new methodology for multi-criteria decision-making accounting for uncertainty.

e-mail: kalika@econ.haifa.ac.il
Natural Resources and Environmental
Research Center
Haifa University
Mount Carmel
Haifa 31905, Israel

Paper

Exploring agent-based wireless business models and decision support applications in an airport environment

Yapeng Wang, Laurie Cuthbert, Francis J. Mullany, Panagiotis Stathopoulos, Vasilios Tountopoulos, Dimitrios A. Sotiriou, Nikolas Mitrou, and Michael Senis

Abstract—This paper describes an intelligent communication and decision support system for providing wireless services in an airport environment. A novel agent-based business model is proposed and the value chain is analysed for wireless applications. This system is studied and developed within the scope of the IST ADAMANT project, where the Athens International Airport (AIA) is used as the trial environment. First of all, a set of advanced, realistic decision support application scenarios enhancing the airport facilities both for the passengers and for the airport staff is identified. Most of the applications can be summarised as location-based personalised services. They refer both to airport internal users and to passenger users. In order to provide these services, location-sensitive service level agreements (SLAs) and radio resource management (RRM) are introduced. The design of such a system is envisaged based on a generic, multi-agent architecture, which is also presented in this paper.

Keywords—wireless application, multi-agent system, wireless business model, wireless SLA, radio resource management.

1. Introduction

The advance in wireless communication market enables users to experience enhanced delivery of personalized services through the integration of various radio technologies. However, the existing management platforms cannot ensure the scalability and reliability for the interworking of different networks. Therefore, the need for research activities in network management by developing and validating flexible architectures for the support of heterogeneous infrastructures is apparent.

This paper presents an intelligent decision and management system (IDMS) for providing wireless services in an airport environment. This system is studied and developed within the scope of the IST ADAMANT (airport decision and management network) project, where AIA is used as a trial environment. The envisaged system will be capable of handling crisis situations, as its benefits are clearer under such circumstances. The approach adopted is generic covering any mode of trans-port; although within this paper it is limited to an airport environment, as the "hot spots" that are naturally generated there provide some of the most difficult challenges. A system, as generic as this, has not been implemented before and it is technically challenging.

A set of advanced, realistic application scenarios enhancing the airport facilities both for the passengers and for the airport staff is identified, which make obvious the need for a scalable anticipatory environment able to provide roaming/location-based and personalized services based on service level agreements. These scenarios include the internal bus arrival time information estimation, flight information display on demand, mobile video/photo information for security and surveillance, automatic billing application for airport fuelling companies, and passenger support in the area of the main terminal building.

To meet the requirements imposed by the above scenario, an agent-based architecture is proposed. The multi-agent architecture forms a framework for implementing the interactions between the very different types of entities involved in the proposed scenarios, whether at the service level or at the network transport level. This type of architecture also has the advantage of scalability and robustness of operation in congestion and emergency situations.

The SLA management for location-sensitive applications and business models for wireless applications in an airport environment are analysed. A hybrid business model balances the benefits of all business roles, where appropriate value chains ensure the market place will be running smoothly.

2. System overview

This section presents briefly the operational environment for the functionality of the IDMS. More specifically, it describes the airport environment main structure, the airport user groups, and the existing set of services and wireless telecommunication infrastructure.

The airport environment consists of the core building which is the main terminal (MTB) and satellite building, the administration buildings and the related building of all the airport providers (such as handlers, police, fire brigade, the air-traffic control tower), as well as the outdoor environment which includes the access motorway, the new metro station and the parking spaces. Four different categories of potential users, within this operational environment, can be identified: passengers, meters and greeters, the airport staff and other airport service and content providers (i.e., security companies, airlines, duty-free and facility companies).

Different communications services are targeted at the different user categories listed above. For the provision of

these services, a telecommunication infrastructure has been developed. GSM/GPRS, WLAN and TETRA [2] are the wireless technologies used. An IP backbone network is provided for Internet services and the PABX network for conventional telephony services. The airport holds the airport services and operations centre (ASOC), which combines all critical airport operations mechanisms and controls. The airport operational database (AODB) contains real-time information about the arriving and departing flights and other flight related information (gates, stands, etc.). This information is distributed within the airport community via the flight information display system (FIDS). All the airport data is processed through a central security system, which is called the universal flight information system (UFIS). This provides technical and logical functions for an effective and reliable data processing of operational flight information and holds the central UFIS database of the airport.

3. Application scenarios

In this section, the set of applications that the proposed system supports for enhancing the operational environment of the airport is described.

3.1. Internal bus arrival time information

This application aims at developing a real-time component that estimates the waiting time of a passenger at the bus stop, prior to the arrival of the internal bus service of the airport. Towards this end, the data transfer functionality of either the airport TETRA system or the GPRS functionality of the GSM system is exploited. Real-time data regarding the location (from on-board GPS units) and the speed of the buses will be sent to the ASOC, where suitable software will estimate their arrival time at the bus stops, within the airport area. From the ASOC centre, this information will be distributed to every bus stop, where a LED display or a flat screen will be available, or even to individuals' handheld devices.

3.2. Flight information display system on demand

This application will provide on-demand arrival and departure information on portable wireless terminals, based on the existing non-personalized UFIS flight information system of the airport. This service is to be considered as an extension of the existing non-personalized UFIS applications, as far as new interfaces and communication infrastructure are concerned. Target users are the airport staff, ground handlers, airlines, concessionaires and passenger service providers. This service will be fully personalised according to individual customer or group preferences. For example, ground handlers are more interested in the exact time of arrival and the location of the aircraft, whereas concessionaires are more likely to be interested in time delays or predictions of the number of passengers transiting in the next hour. The system interfaces with the network resource management ADAMANT subsystem, providing relevant information to aid the resource management subsystem in its provision of pro-active congestion management mechanisms (e.g., the dropping of less critical user groups). Security issues will also be taken into consideration.

3.3. Mobile video/photo information for security and surveillance

This scenario concerns the development of a mobile photo/video camera system for the real-time transfer of photos or video to the ASOC for security and surveillance purposes and back to the security/emergency staff. Potential users of this system are the airport security personnel, airport police, fire brigade and ambulance services, etc. Transmitting real time photographs and/or video of an emergency situation or accident, by the ASOC intelligent decision and management system, will provide instantly the necessary elements for the immediate evaluation of the situation and the rapid activation of the necessary emergency-response teams. The remote monitoring of the crisis event through instant photos or video will result in more effective crisis evaluation and measures. The images and/or video will be directed to the system users by the ASOC, according to their profile, location, the overall situation and the user's preferences. This application will exploit commercial off-the-shelf equipment enhanced with additional security and authentication mechanisms and the ADAMANT network resource management features, in order to provide connectivity to the security personnel under all network congestion circumstances.

Networks GSM/GPRS and WLAN have the capability to transfer multimedia data in order to provide real-time visual information, according to the users' location and the overall situation. The user terminal comprises a PDA equipped with a GPRS and IEEE 802.11b WLAN PCMCIA cards and a GPS device in order to extract precise location information. The recently deployed multimedia messaging service in GSM/GPRS networks can be an enabling technology for such an application. In the longer term, video streaming may be possible in areas with WLAN coverage or over future 3G networks. This service complements existing networks of fixed cameras that provide surveillance information from certain airport areas.

3.4. Automatic billing application for airport fuelling companies

The aim of this application is the development of a real time component that will automatically charge the amount of fuel loaded onto an aircraft from the respective fuelling company. The IMDS utilises the SDS functionality of the TETRA system. Moreover, it combines the information granted from airport's existing FIDS system with locally stored information from fuelling trucks and drivers.

3.5. Passenger support in the area of the main terminal building

This is the core application of the project, aiming at providing passengers with *personalised* and *location-based* information related to the airport, as well as broadband Internet access during their stay in the main terminal building. Users of this service can access the content through their mobile terminals (PDAs or portable computers), or fixed PCs located inside the airport premises. Services depend on passenger preference profiles, as well as the status (e.g., departing, arriving, or delayed) of the user's flight.
The service will be controlled by a session manager platform able to:

- Provide the passenger with the necessary flight information and according to his/her profile and flight status, guide him/her through the departure procedure of checking in, clearing security, and reporting to the departure gate.
- Allow the airport to track the passenger.
- Provide personalised information with respect to other airport commercial facilities such as restaurants, shopping, etc., especially to passengers who get to the gate early, or to business travellers making use of the airport lounges. Flight delays can increase the demand enormously, as travellers can be encouraged to use these commercial facilities. To this end, the IMDS can exploit knowledge about the user preferences and the type of trip (e.g., regular business travellers or travel for leisure).
- Inform arriving passengers about means of transport (to and from the airport), accommodation, and local tourist information.

This application will help the airport authority to offer a value-added service and increase the passenger satisfaction, and on the other hand allow them to make more efficient use of the time spent at the airport premises and to increase the concessionaires' incomes. Finally in this system, it is possible to extend the coverage area to other "hotspots" like conference halls, hotels, restaurant and bars, etc.

4. The agent approach of system architecture

This section describes the main agent architecture for the IDMS in the airport area. IDMS deals with resource management strategies for GSM/GPRS, WLAN and TETRA and addresses SLA management issues in the context of providing advanced services to the airport users. The IDMS architecture is based on the "one-stop-shop" business concept, which identifies all the processes that should be in place, in order to include a service in the business service portfolio. In that respect, in such a scalable anticipatory environment, the following business entities can be identified: the users, the service providers, the network providers for GSM/GPRS, WLAN and TETRA infrastructures and the content providers.
The development of the IDMS is based on multi-agent systems (MAS) [3]. This section introduces the components that comprise the main agent architecture and provide the framework for the interaction between the entities identified previously. More specifically, the following generic agent components can be defined:

- user agents (UA),
- user resource agents (URA),
- location agent (LA),
- service broker agent (SBA),
- service provider agents (SPA),
- content provider agents (CPA),
- resource broker agent (RBA),
- network provider agents (NPA).

The rest of this section is dedicated to the analysis of these agents and the description of their role in the IDMS.
The UA manages all the information related to the user terminal and behaviour in the airport, such as user preferences, travel information, privacy issues, etc. Every user of the IDMS owns a UA, located inside the user terminal. In cases where a terminal cannot support the UA software, a proxy user agent located at the SBA can be used. The UA is activated any time the user registers to the IDMS. Its main role is to set and update the user profile. Any time the user wants to make use of a specific service, the UA communicates with the SBA and sends an application request. The UA also performs SLA monitoring functionality by monitoring some crucial SLA parameters, such as the received bit rate.
Another important functionality of the UA is to update the user's location at certain intervals, in order to provide location-based services. In that respect, the UA sends location update messages to a LA via the SBA, in order to inform it about the current position of the corresponding user. The LA holds a database, which keeps records of the current position of all the users registered to the IDMS. It can then maintain information about the geographical location of the user at any time.
The SBA plays a key role in the whole architecture, acting as a mediator agent by providing the interface between the UAs and the SP agent components. The SBA performs on behalf of the IDMS, subscription and identity checks for incoming users. Furthermore, the SBA maintains user profiles containing information such as the set of services and applications that the corresponding users are willing to subscribe to and the reference quality level at which a specific service should be provided.
The SBA undertakes the responsibility of prioritizing and delivering the incoming messages to the appropriate agents.

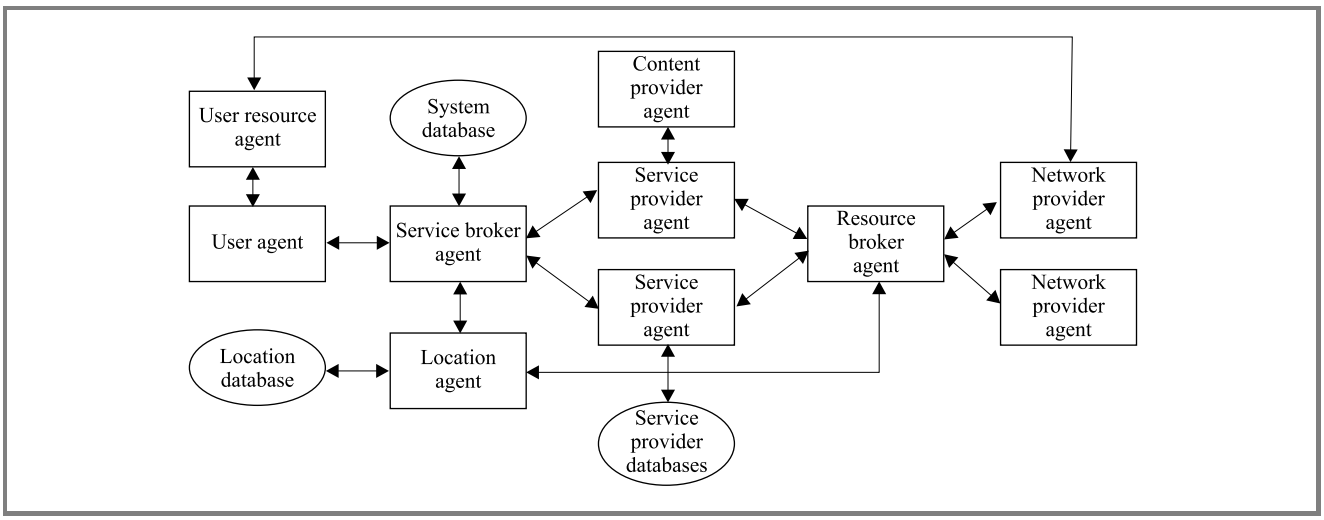


Fig. 1. System agent architecture.

The prioritization strategies can be based on many parameters, such as the type of the request, the SLA management policies and the preference settings of the UA. According to such information, the IDMS can decide about the priority given to multiple incoming service requests, with the aim of maximising the IDMS's profit.

The SBA can, also, act as a facilitator for SLA negotiation and notification functionality between the UAs and the SPAs. It is responsible for transferring and supporting SLA messages both from and to the SPAs. These messages may involve the SLA proposal of an SPA in response to the service request of a UA, the SLA response of the UA to the SPA and other SLA renegotiation messages.

The SPA maps to the existing entities of the airport environment for the provision of the available services. It is mainly responsible for the accomplishment of the role of the service provider to the end users. So, the role of the SPA is to respond to the requests from the SBA and offer the requested services to the UA, according to certain criteria and to retrieve pertinent information from local databases. In that respect, it communicates with the CPA, which provides service content information messages regarding the operation of the available services. The SPA can, also, perform SLA management functionality for the negotiation and monitoring of the SLAs with the UAs, through the SBA, and report any violations of the agreed SLA contracts.

The RBA is the gateway of the network domain with the other entities. Its main task is to find the best NPA to serve the incoming requests from an SPA. In that respect, the RBA should exploit the information gathered from the service domain and report to the most appropriate NPA about the facilitation of the specific service request. The incoming messages may involve the service type and other user related information. The RBA also interacts with the LA, in order to receive the user's location in the system. It can then decide the appropriate NPA with resource availability for location-based services.

The behaviour of the RBA is more important for the IDMS in cases of lack of resources, due to unexpected events, such as emergency cases and flight delays, which lead to local hot spots. In such cases, the challenge for the RBA is to find reliable solutions from the most appropriate NPA. In that respect, it should monitor the performance of the underlying networks, in order to identify the current congestion levels of the available networks and report any degradation in the system performance. So, at certain intervals, the RBA is informed by the NPA about (or alternatively infers itself) the current network status. Subsequently, the RBA can perform adequate functionality for finding the most appropriate NPA and ask it for resources. This functionality is crucial, especially in cases of high priority service requests, such as in emergency cases.

The network provider agent (NPA) is that agent in the system with primary responsibility for provisioning of the transport function for the services supported by the system. Hence, the role of a NPA is twofold:

1. **Negotiation**. NPAs "represent" the networks in the system's efforts to match the required network transport capabilities to the services requested by the user. A resource broker agent will contact one or more NPAs and by some mechanism, come to a contracted arrangement whereby one or more NPAs undertake to reserve and allocate network resources to the transport link needed for an agreed price.
2. **Resource management**. A given NPA organises its own internal resources, either in isolation or with the aid of other agents, to supply the agreed transport links with the contracted QoS arising from the first role.

In the context of the IDMS, there is one NPA for each of the access networks that are associated with the system, i.e., the WLAN system, the various GSM/GPRS operators, the TETRA network, etc. It should be noted that the NPA functionality may be distributed among a number of subagents within the network [1].

The user resource agent is an optional agent in the user terminal, with the role of collaborating with NPAs and/or other URAs, to control the user terminal's use of resources so as to optimise the operation of the user's transport link over the radio interface. The resources over which the URA may have influence include the air interface and usage of battery power.

Based on the description provided above, Fig. 1 summarises the system agent architecture for the accomplishment of the main objectives of the IDMS. As it can be seen, some of the agent components hold interactions with external databases, which then (partially) represent the necessary interface between the agent components of the IDMS and the external world.

5. Service level agreement management framework

As users move through an airport, they are roaming through the network, and the SLA management provided should be sensitive to the local conditions, adjusting SLA guarantees according to the roaming and location conditions as well as promptly notifying the user of these changing conditions. Location dependency of SLA is envisaged when users move between locations with different coverage characteristics or with different congestion situations (e.g., different cells of a GSM/GPRS/UMTS network) or when users move between locations served by different network technologies (e.g., UMTS/WLAN/TETRA).

In addition, in order to manage congestion and crisis scenarios, a service provider should be able to define policies to prioritise the allocation of resources to the most critical services and customers, to deliver the best possible service levels based on combination of SLA agreed with the customers and current conditions of the network.

The SLA management framework defined here is used to develop SLA templates and SLA contracts for service subscribers, in which the service level objectives may change as the user travels to or moves through the airport.

It includes components to: provide resource management with contracted SLA terms in order to allocate network resources (e.g., bandwidth) accordingly; monitor the service level experienced; notify compliance or non-compliance of the service level objectives; provide reporting functions for the detailed analysis of the service level offered by the network; communicate with service components to adjust the service behaviour based upon the management policies set by the provider.

Performing SLA management also implies coordination and exchanging information between providers in the value-chain and between providers and customers. Such coordination and information exchange is formalized in ADAMANT through the definition of business process models and the corresponding business interactions amongst providers, focusing on service management processes (e.g., service subscription, service assurance, service planning).

6. Business models and value chain

The business model of the ADAMANT system involves several different market players and can be characterized by multiple relations which, depending on the provided service, might differ.

In general, the following players are expected to be involved in the ADAMANT business model:

- **Customers/users**. In general, we can distinguish between the user of a service and the customer that subscribes to that service and negotiates a SLA with the service provider. In some cases they might coincide, e.g., in the case of passengers, in other cases they might differ, e.g., in the case in which a company subscribes to a service for a certain number of its employees. In ADAMANT, customers can be passengers, members of the Athens International Airport, ground handlers, etc.

- **Service providers**. A service provider (SP) is a company or organization that provides wireless communication services as a business. SPs may operate networks, or they may simply integrate the services of other providers in order to deliver a total service to their customers. Providing a wireless communication service to any one end customer may involve multiple SPs, where one provider may "sub-contract" to other providers to fulfil the customer's needs. According to the specific nature of the services provided we can distinguish among content providers, application service providers, service integrators/content brokers, Internet service providers, etc. In the ADAMANT context, depending on the specific application, content providers may be the AIA, e.g., providing flight information, or the airport concessionaires, e.g., airlines and travel agencies providing information about flight offers, or shops and duty frees providing information about special offers. The role of application service provider, service integrator and content broker is typically assumed by AIA (or more precisely, by some of its internal departments, e.g., the ASOC or IT&T departments), who acts as a "one-stop-shop" for services offered to the airport community. The role of Internet service providers is again covered by AIA, through its IT&T department. Other ISPs may be involved in the ADAMANT application scenarios, such as OTENet, and Panafonet.

- **Network providers**. These provide wireless communication services on an underlying network infrastructure, where, in this context,the services are basic telecommunication services, such as voice and data channels, IP capacity, etc. To be precise, network providers are actually a sort of service providers, i.e., providers of basic network connectivity services, so they can be considered as network service providers. We can distinguish the following

sub-roles for network providers, according to the network technology used:

- mobile operators provide voice and data services on GSM/GPRS (and UMTS);
- WLAN operators operate WiFi (IEEE 802.11 [4]) wireless LAN networks in hot spots such as the airport main terminal; in the ADAMANT context the WLAN is operated according to the **neutral host model**: in this model the location owner, AIA, owns the infrastructure (access points, antennas, cabling, session management software, firewalls, routers) and offers it for exploitation to different providers, in this case to OTENet;
- specialized network operators provide voice and data services on specific network technologies.

As stated above, the classification of business roles presented above is useful to clearly separate functional roles within the business model, but it might be the case that, in a certain application scenario, the same business entity plays more than one role, or even different business units within the same organization, like in the case of AIA departments. It may also happen that the same business entity plays different roles in different application scenarios.

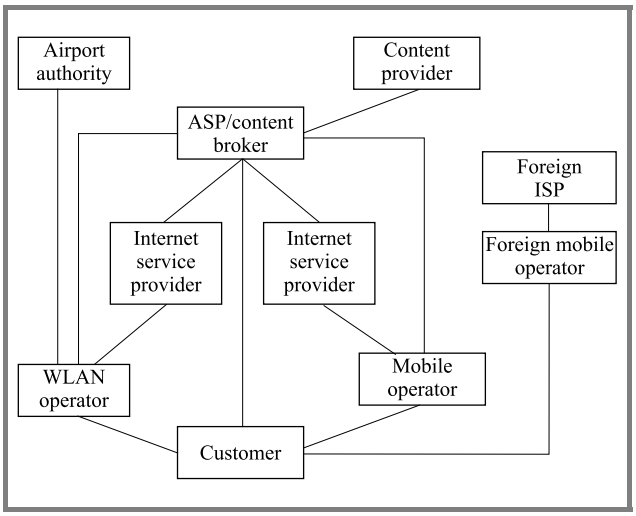


Fig. 2. Business actors and their relationships in the ADAMANT context.

For the generic business roles, the customer (user) subscribes for a service with either a service provider or a network provider, according to the nature of the service requested and the specific business model used for providing that service. In fact, for the most generic case, it is possible that the customer, beyond having a direct relationship with the network provider, would also have a direct business relationship with the service provider. For example, not all services would be billed via the account the user holds with the network provider. Typically, a subscription for a value-added service could be performed with a service provider, while a network access/connectivity service could be subscribed to with a network provider.

The business relationships between roles in the ADAMANT context are depicted in Fig. 2. In the figure, specialized service provider and network provider roles have been added to reflect the actual service and networks typologies that are available in the designed application scenarios.

The customer may have direct relationships with both the service and network providers. Mobile and WLAN operators interact with Internet service providers (ISPs) for accessing the public Internet, and such ISPs in turn may interact with an ASP/content broker, either for providing Internet access to the ASP/broker, or to buy content from it. Mobile and WLAN operators may have relationships in the cases in which an operator may want to offer its customers services using the other operators' networks (e.g., a mobile operator offers download of video-clips to its customers through a WLAN when the customer is in the airport, which would be much faster). For the case of foreign visitors, foreign mobile operators are also represented, since these would have roaming relations with the local operators.

7. Conclusion

This paper has illustrated the on-going work of ADAMANT project, the application scenarios and the IDMS architecture. The paper also introduces the proposed SLA management framework for location-sensitive wireless applications and business models and value chains. Through the multi-agent architecture, the IDMS can provide management of the communication resources and ensure that the service provisioning of the airport users is in line with their SLAs. The new business model and value chain balance the requirements of all the business roles and ensure the smooth operation of the future wireless marketplace. Within this project, the IDMS system and new business models will be tested and validated through experiments, trials and demonstrations based on the defined application scenarios.

Acknowledgement

The authors would like to thank the European Commission for their support of the ADAMANT project.

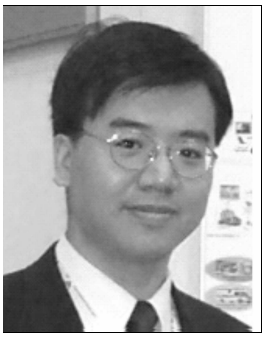
References

[1] "IST Project SHUFFLE (An agent based approach to controlling resources in UMTS networks)", 2001, http://www.ist-shuffle.org

[2] "Terrestrial Trunked Radio (TETRA): Voice plus Data (V+D) Part 1: General network design", ETSI EN 300 392-1 V1.2.1 (2003-01).

[3] W. Brenner, R. Zarnekow, and H. Wittig, *Intelligent Software Agents*. Berlin, Heidelberg: Springer-Verlag, 1998.

[4] "IEEE Standard for Information Technology—LAN/MAN—Specific requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications 1999", ISO/IEC 8802-11: 1999 (802.11, 1999).



Yapeng Wang received two B.E. degrees in telecommunication engineering and in computer and its applications from North China Electric Power University in China, 1998. He received M.Sc. degree in Internet computing from Queen Mary, University of London (QMUL) in 2002. He is now a Research Assistant for IST ADAMANT project and also a Ph.D. student in Department of Electronic Engineering in QMUL. His research interests are in areas of broadband wireless communications, radio resource management and agent-based intelligent applications. He is a student member of IEE.

e-mail: yapeng.wang@elec.qmul.ac.uk
Department of Electronic Engineering
Queen Mary, University of London
327 Mile End Road, London, UK

Laurie Cuthbert is a Professor in the Department of Electronic Engineering at Queen Mary, University of London, where he is currently Head of the department. In the late 1980s he founded the Telecoms Research Group with its main emphasis being on research into ATM, but since then it has broadened its interests to include multimedia, mobile, Internet protocols and applications and intelligent control of networks. He is a Fellow of the Institution of Electrical Engineers and has been active in the Professional Groups of the Institution.

e-mail: laurie.cuthbert@elec.qmul.ac.uk
Department of Electronic Engineering
Queen Mary, University of London
327 Mile End Road, London, UK

Francis J. Mullany received B.E. and Ph.D. degrees in electronic engineering from University College Dublin, National University of Ireland, in 1992 and 1998, respectively. He was a visiting scholar at the Institute of Radio Communications, Helsinki University of Technology, Finland from 1994 to 1995 and an assistant lecturer at the Department of Electronic and Electrical Engineering in University College Dublin from 1996 to 1997. Since 1998, he has been a research engineer at the Wireless Research Laboratory of Bell Labs Research, Lucent Technologies and is based in Swindon, UK. He is a member of the IEEE and the IEE. Dr. Mullany's research interests centre on radio access architectures. Issues of interest to him include radio system design and dimensioning, radio resource management, physical layer algorithms, and hardware architectures.

e-mail: mullany@lucent.com
Wireless Research Laboratory
Bell Labs, Lucent Technologies
Quadrant, Stonehill Green, Westlea
Swindon, Wiltshire, SN5, 7DJ, UK

Panagiotis Stathopoulos was born in Athens, Greece, in 1976. He received the Dipl.-Eng. degree from the School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), Greece, in 1999. His graduate thesis was in the area of multiservice ATM networks management using intelligent agents. Since December 1999, he is a Research Associate with the Computer Network Laboratory of the School of Electrical and Computer Engineering of NTUA, pursuing a Ph.D. degree in the area of broadband communications. He is involved in several IST projects of the European Community. He is a member of the Technical Chamber of Greece and a student member of IEEE.

e-mail: pstath@telecom.ntua.gr
National Technical University of Athens
Department of Electrical and Computer Engineering
Telecommunications Systems Laboratory
9 Heroon Polytecheiou st
Zographou 15773, Athens, Greece

Vasilios Tountopoulos was born in Athens, Greece, in 1977. He received the diploma in electrical and computer engineering from the University of Patras in 2000. His dissertation, which was elaborated at the Wireless Communications Laboratory of the Department of Electrical and Computer Engineering of the University of Patras, dealt with the handover procedures on UMTS and IMT-2000 cellular systems. He was awarded in 2001 Ericsson's awards of excellence in telecommunications. Since February 2001, he is a Research Associate at the Computer Network Laboratory of the School of Electrical and Computer Engineering of the National Technical University of Athens (NTUA), working for his Ph.D. degree in the area of broadband wireless communications. He is involved in many IST projects of the European

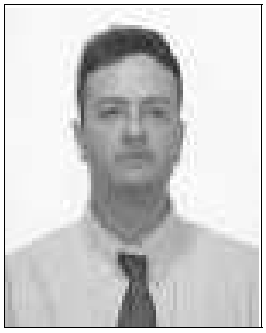
Community. He is a member of the Technical Chamber of Greece.
e-mail: vttounto@telecom.ntua.gr
National Technical University of Athens
Department of Electrical and Computer Engineering
Telecommunications Systems Laboratory
9 Heroon Polytecheiou st
Zographou 15773, Athens, Greece

Dimitrios Athanasios Sotiriou was born in Volos, Greece, in 1978. In September 2000 he received the diploma in electrical and computer engineering from the National Technical University of Athens (NTUA). After a year in London, in 2001 he received a Master's degree in telecommunications and digital signal processing from Imperial College (IC). During the course he worked for PA consulting and completed his thesis on load control algorithms for UMTS. A year later, he obtained his MBA in Business Administration from the Economic University of Athens and NTUA. Since September 2002, he is a Research Associate at the Computer Network Laboratory of the School of Electrical and Computer Engineering of NTUA, working for his Ph.D. degree in the area of mobile networks. He is involved in IST projects of the European Community and is also a member of the Technical Chamber of Greece.

e-mail: dsot@telecom.ntua.gr
National Technical University of Athens
Department of Electrical and Computer Engineering
Telecommunications Systems Laboratory
9 Heroon Polytecheiou st
Zographou 15773, Athens, Greece

Nikolas Mitrou received the diploma degree in electrical engineering from the National Technical University of Athens (NTUA) in 1980, the M.Sc. degree in systems and control from the UMIST, Manchester, in 1982, and the Ph.D. degree in electrical engineering from the NTUA in 1986. He is now a Professor in the School of Electrical and Computer Engineering of NTUA. His research interests are in the areas of digital communications, communication networks and networked multimedia in all range of studies: design, implementation, modelling, performance evaluation and optimisation, applications. He is a member of the IEEE and member of the IFIP (TC6).

e-mail: mitrou@softlab.ntua.gr
National Technical University of Athens
Department of Electrical and Computer Engineering
Telecommunications Systems Laboratory
9 Heroon Polytecheiou st
Zographou 15773, Athens, Greece



Michael Senis received his Dipl.-Ing. in electrical engineering, from the National Technical University of Athens (NTUA), Greece in 1990 and his M.Sc./D.I.C. in Communications and Digital Signal Processing from Imperial College of Science Technology & Medicine, University of London in 1991. During 1995–1999 he worked for Panafon-Vodafone in the Planning and Functionality Group of the technical division. In March 1999 he joined Athens International Airport S.A. (AIA) as Project Manager of wireless telecommunications responsible for the project realisation of the airport wireless infrastructure which is comprised of: trunked radio system (TETRA), paging system, GSM/DCS mobile radio infrastructure, air-ground communications for airlines, wireless LAN, etc. In February 2002 he has been promoted to Head of the telecommunications section within the IT & Telecommunications Department of AIA S.A. Currently, among his extensive responsibilities is the technical project leader of the European ADAMANT project which is in progress. He is also participating in other European projects like IM@GINE IT. He is member of the Technical Chamber of Greece, the IEEE and the Greek Association of Mechanical and Electrical Engineers.

e-mail: msenis@aia.gr
Athens International Airport SA
Information Technology & Telecommunications Depart.
Business Development Division Spata 19019
ATC & CONTROL Tower Building (32), Greece

Regular paper

Towards broadband global optical and wireless networking

Marian Marciniak

Abstract—This paper presents a novel, non-conventional approach to the future optical and wireless hybrid transport network, capable of supporting/dominant kinds of traffic, i.e., voice/real time, wireless and packet data traffic in a single transport network. The proposed model combines different technologies as connection and connectionless networks, optical cable and wireless (microwave/millimetre wave or optical wireless), being suitable for a variety of purposes and services in order to achieve global broadband networking features. Our new networking model contains an extension to wireless world in order to achieve mobility and personalisation of connection. From the networking point of view it consists of an upgrade of real-time traffic with the microwave modulated optical wave, in order to carry out conventional mobile wireless signals via optical fibres over long distances and without significant distortion. The whole available bandwidth can be fully exploited in the hybrid network. In the IP part of the network the quality of service (QoS) can be differentiated for various classes of packets and network reliability/survivability can be categorised for the whole hybrid network. This proposal combines complete and revolutionary shift to packet traffic with mooth evolutionary upgrades. We believe the model presented here is a powerful tool to trace the future evolution of telecommunications worldwide for the next 25 years.

Keywords—optical and wireless networking, radio-over-fibre, optical packet networks, transparent networks.

1. Introduction

Global broadband networking is a crucial requirement for the future information society, which needs a seamless personal access to broadband services, everybody to anybody, from any location to any destination, anytime. The technology of conventional connection-oriented voice networks has been developed for a hundred of years already and is in a mature state. Terabit per second transmission via optical fibre links is a reality with the use of asynchronous transfer mode/synchronous digital hierarchy (ATM/SDH) over wavelength division multiplexing (WDM) technologies. In addition, the transparent optical transmission at distances of thousands of kilometers become a reality with the advent of optical amplifiers.

Although the optical fibre communications still faces a number of technological difficulties and physical constraints with chromatic dispersion (CD), polarisation mode dispersion (PMD) and nonlinear optical interactions as classical examples, it is still the best technology to achieve terrestrial global networking. However, a serious drawback of that technology is the lack of mobility features, so in this paper a suitable combination of optical and wireless technologies is addressed and discussed.

It is worth to underline that actually the basic concept of networking is a subject of revolutionary change from classic circuit switched, connection – oriented networks to modern packet switched, connectionless transmission of data [1]. In fact this is driven by the dramatic expansion of Internet users worldwide. Therefore circuit-switched networks which work excellent for voice traffic and in general for real-time traffic are not at all a suitable solution for packet data traffic.

Packet traffic is the basis of Internet. It has a statistic nature, so at certain moments of time and network points the flow of data might be much higher that the network equipment is able to handle. This means that some data have to be stored in buffers, and if available buffer capacity is insufficient, then some data have to be sacrificed. This affects the quality of service provided by the network. This statistical phenomenon is called "burst of packets" and has provoked an effort to search for suitable means to cope with it [2]. For some applications this is not a critical situation, as the missing data can be retransmitted. But usually retransmission is not allowed for real-time voice or video services. Nevertheless, even when the networks are overprovisioned, i.e., with potential performance characteristics much better that actually needed, one could easily imagine an immense and impossible to handle data flow generated by computer viruses or malicious attacks, and a complete blocking of the network might result from that.

New technologies are being proposed to enable successful real-time service transmission through packet network [3]. However, even generally accepted, IP protocol is not so obvious as the only future platform for a converged voice/data network. In fact the main problem to be resolved is the lack of QoS guarantees for real-time traffic inherent to this approach.

In this paper we propose a combination of different technologies, such as connection and connectionless networks, optical cable and wireless (microwave/millimetre wave or optical wireless), suitable for a variety of purposes and services in order to achieve global broadband networking features. In addition to our recent proposal of a hybrid connection and connectionless networks superimposed on the top of a physical WDM layer, and on different optical wavelengths, our new network model contains an extension to wireless world in order to achieve mobility and personalisation of the connection. From the network point of view it contains an upgrade of real-time traffic with the microwave

modulated optical wave, in order to carry out conventional mobile wireless via optical fibres over long distances and without a significant distortion.

The paper is organised as follows: Section 2 discusses the optical fibre communication technologies, while Section 3 addresses the relevant characteristics of the wireless mobile world. In Section 4 the hybrid optical and wireless networking basis are laid out. Conclusions and future work are summarised in Section 5.

2. Optical fibre communication technology

The transparent features of the network will be discussed in detail, pointing out that that "transparent" and "all-optical" does not mean necessarily the same [4]. The impact of optical transparency on a successful deployment of future optical packet switched networks is discussed. Actually the opacity of a network is considered as resulting from conversion to electronics only. However, an "optical opacity" is inherent to several all-optical solutions, too. Indeed, the optical transparency has to be sacrificed in future optical logic elements, e.g., for signal processing, packet switching, etc. Implementation of networking functions with photonic components is summarised and directions of future development are pointed out. Finally, a novel non-conventional approach to voice + IP network is presented in the concluding part of the paper.

2.1. Optical transparency

The advent of erbium-doped fibre amplifiers which replaced electronic regenerators in fibre transmission links in early 90s resulted in optical transparency of the links [5]. The evolution of actual telecommunication networks towards transparent networks started at that time. Transparent networks at regional and global scale with transmission speed exceeding one terabit-per-second became a reality. Such networks are commonly referred to as "terabit networks" [6].

The notion of transparency of a transmission link (not necessarily optical) is much older than optical fibre communications. Its primary meaning is that the output signal is proportional to the signal at the input of a link. Consequently, the transparency is rather an analogue feature, apparently contrary to modern digital transmission schemes.

An ideal transparency is not realisable in an optical network, since even an ideal glass fibre exhibits attenuation, chromatic dispersion of the first and higher orders, and optical nonlinearities. Moreover real fibres exhibit polarization mode dispersion (PMD), resulting from random local lack of circular symmetry of the fibre due to manufacturing imperfections and local stresses caused by cable structure. Those features of a fibre result in distortion, crosstalk, and noise in the transmitted optical signal.

Transparent photonic network insures the scaleability, i.e., possibility of future upgrades. Wavelength becomes a new degree of freedom (wavelength-switched and routed networks), and can be switched in wavelength routers and converters.

The lack of an ideal transparency requires very high wavelength precision and stability of optical sources in a dense WDM network, which considerably increases the cost of the devices. Therefore, the goal is not to loose the precious wavelength. The solution is to keep the signal in optical domain while it traverses as large part of the network as possible, and this is why transparency is a so important.

Transparent network still includes attenuation and/or amplification, and eventually wavelength conversion of signals. Transparent wavelength conversion assumes conservation of temporal signal shape, superimposed on a different wavelength.

In a way transparent networks go back to the analogue age. Transparent components of the optical network treat the passing signals in an analogue way. The transparency length is a distance over which the signal can be transmitted successfully. Transmission over longer distances requires some form of regeneration. The transparency length depends on number of factors, and it can be increased in the future, when the technology is sufficiently developed.

Moreover, the expected introduction of optically transparent fibre links to subscriber networks will allow to take advantage from WDM technology also in that area. New ways of providing access are emerging to satisfy the need for interactive broadband services. A combination of various signals (i.e., analogue or digital radio and television, interactive broadband services, Internet traffic) could then be transmitted simultaneously. The emphasis is on the possibility to transmit conventional wireless radio signal. What is really very important, is that the transparent optical networks are scaleable, i.e., provide a potential for future upgrades [7].

2.2. Optical switching and routing

Degrees of freedom of an optical network are:

- 3-D space co-ordinates;
- time (and resulting possibility of optical time domain multiplexing – OTDM);
- wavelength (WDM);
- polarization of light.

Those provide opportunities for optical switching in space, time, wavelength, and polarisation domains. In addition to that, logical on/off switching may be implemented in optical logic elements.

Optical routing can be realised as wavelength routing in a transparent way as:

- an analogue and passive solution or
- an analogue and active solution with wavelength conversion.

2.3. All-optical opacity

All-optical packet routing involves some intelligence of the router and decisions based on the information included in the packet. Therefore it is not realisable in transparent way. Even though all-optical routing concept involves optical logics, optical memory, etc., it is not optically transparent and it exploits optically opaque elements. The signal remains in optical domain, but due to digital operations the fundamental transparency condition of proportionality between output and input signals is not satisfied.
Emerging evolution directions of optical network infrastructure and its include migration:

- from circuit switched optical networks, with analogue processing of carrier frequency, i.e., in spectral domain, where the transparency is a positive characteristics;
- to a packet switched network, with direct digital processing of signals in time domain, realizing all-optical switching and in particular exploiting optically opaque all-optical routing devices.

A combination of transparency exploitation in wavelength domain and all-optical logic in time domain seems to be the most justified way of network evolution.

3. Wireless mobile world

Wireless technology has been developed during the last decade for the mobile networks. New frequency bands are exploited in view of transmission capacity and reliability, the 60 GHz band being an example. Sophisticated 3G and beyond broadband and interactive services are foreseen in many countries in the near future. European telecom companies have invested a lot of money in the UMTS licenses. Unfortunately, it is still not obvious when the investement will bring the expected revenues.
While microwave and millimeter wave links have excellent mobility characteristics that is impossible to achieve with other transmission media (wireless optical links have very poor performance compared to microwave ones), they still suffer from a number of constraints, most of them resulting from electromagnetic compatibility (EMC) requirements, in order to avoid interference and crosstalk. Also the wireless links suffer from the attenuation of signal due to air characteristics, weather, smog, and the local shape of terrain or trees and buildings. The line of sight between the transmitter and receiver is usually an essential requirement for reliable transmission. Microwave spectrum is expensive and limited. Fibre optic technology can help to transmit wireless signal superimposed on the optical wavelength, as we have proposed and analysed recently [8].

4. Hybrid optical and wireless networking

Among real time services the voice is still dominant. As videophone is concerned, this technology is not likely to be accepted widely in the near future. In fact people even prefer short message system (SMS) to spare bandwidth, time and money. Television broadcasting will never be done via public telecom networks as well.
Attempts to transmit voice over IP have an inherent difficulty to guarantee quality of service [9] – in fact nobody guarantees anything, as the basis of Internet is a "best-effort" principle. So "real-time" in fact stops to be "real" in packet traffic. Recently we have proposed to stop to care so much about voice traffic which is already an excellent developed technology, and to start to think about separate voice and data networks [10]. We propose a hybrid network in which voice is carried on dynamically allocated wavelengths, according to an instantaneous demand for real-time service traffic. Table 1 shows a comparison of main features of both types of traffic in a hybrid network [11].
Our non-conventional approach consists of voice (and other real-time services) subnetwork implemented within data traffic network. Voice is transmitted via circuit-switched subnetwork, while IP traffic travels in a packet-switched connectionless network. The two kinds of traffic are separated and interleaved in frequency (wavelength) domain, not in time domain.
The conventional mobile microwave/millimetre wave signal transmission can be included in the transparent real-time part of the network by the means of modulating the optical carrier wavelength with the mobile signal [12]. Then it can be transmitted over long distances via fibres before being detected at an optical receiver and processed further.
The network intelligence has to be located at IP routers and has to provide the real-time subnetwork including microwave transmission on a sufficient number of wavelengths [13]. This approach allows to profit fully from both SDH/ATM technology – best suited for real time-circuit switched services, and from IP protocol – developed uniquely for packet-switched traffic. Moreover, the QoS can be differentiated for various classes of services [14].

Table 1
Voice versus Internet traffic

Characteristics	Voice, real-time	Internet, data
Bandwidth	Dedicated on demand	As wide as available
Basic principle	Circuit-switched	Packet-switched
Packet length	Constant	Variable
Quality of service	Guaranteed	Best-effort
Lost packets	No retransmission	Retransmitted
Traffic	Deterministic	Statistic
Other	Instantaneous bandwidth (# of wavelengths) controlled logically in IP routers	Intelligence
	Transparent	Includes all-optical opacity
Access	Conventional twisted-pair access to public exchange offices	Broadband access to servers, e.g., via cable-TV or mobile

5. Conclusions and future work

The novel non-conventional approach to the future hybrid network retains the well-developed voice technology with transparent transmission. Voice traffic is carried via dynamically allocated wavelengths in conventional way as circuit-switched traffic. The number of wavelengths is controlled by IP layer according to the instantaneous demand for real-time traffic. All remaining wavelengths are available for the IP traffic, which becomes free of real-time restrictions and can adopt variable-packet length, no idle bits, and best-effort scheme. As a consequence, the whole available bandwidth can be fully exploited in the hybrid network. In the IP part of the network, the quality of service can be differentiated for various classes of packets and network reliability/survivability can be categorised for the whole hybrid network.

The terms "all-optical network" and "transparent network" are not equivalent. After a decade of triumph of transparent WDM transmission, evolution towards optical packet-switched networks appears to be imminent. All-optical network requires optically opaque (i.e., not transparent) optical solutions. Opaque all-optical elements have to be introduced in the all-optical packet switched network. Those optically opaque elements will perform all-optical signal processing in general, and all-optical routing functions in particular.

Voice and IP traffic have fundamentally different characteristics and requirements which should not be overlooked and have to be taken into account when designing basics of future real-time service and IP converged network. On the other hand, the future development of converged network should not lose the well developed voice traffic technology with ATM and SDH.

The approach presented here is a result of in-depth investigation of different networking principles and traffic schemes, and physical constraints that characterise the classical fixed fibre network and new mobile wireless world. This proposal substitutes the complete and revolutionary shifting to packet traffic that a number of people foresee, with a smooth evolution and network upgrade. What is really worth to note is that real time traffic provides security and network availability, conserving a number of connections even in an malicious attack occurs. So we believe the model presented here is a powerful tool to trace the future evolution of telecommunications worldwide for the next 25 years.

Acknowledgements

Numerous and very profitable interactions with partners from European Projects: the NEXWAY Network of Excellence, *Advanced Infrastructure for Photonic Networks* (COST 266), *Reliability of Optical Components and Devices in Communications Systems and Networks* (COST 270), and *Towards Mobile Broadband Multimedia Networks* (COST 273), as well as with the ITU Study Group 15 *Optical and Other Transport Networks*, and IEC Technical Committee 86 *Fiber Optics* are warmly acknowledged. This research is being supported by the State Committee for Scientific Research under grants 632/E-242/SPB/COST/T-11/DZ198/2003-2005 and 632/E-242/SPB/COST/T-11/DZ 311/2003-2005.

References

[1] M. Marciniak, "From circuit- to packet-switched or to hybrid network?", in *5th Int. Conf. Transp. Opt. Netw. ICTON 2003, Worksh. All-Opt. Rout.*, Warsaw, Poland, 2003, vol. 1, pp. 47–50.

[2] S. Bjørnstad, C. M. Gauger, M. Nord, E. Baert, F. Callegati, D. Careglio, W. Cerroni, J. Cheyns, D. Colle, P. Demeester, C. Develder, D. R. Hjelme, G. Junyent, M. Klinkowski, M. Marciniak, M. Kowalewski, M. Lackovic, M. Pickavet, C. Rafaelli, J. Sole-Pareta, N. Stol, E. Van Breudegem, and P. Zaffoni, "Optical burst switching and optical packet switching", in *COST 266 Final Report*. Croatia: University of Zagreb, 2003, pp. 115–154.

[3] D. Wright, "Voice over MPLS compared to voice over other packet transport technologies", *IEEE Commun. Mag.*, vol. 40, no. 11, pp. 124–132, 2002.

[4] M. Marciniak, "Transparent versus opaque issues in an all-optical packet switched network", in 1*0th Int. Worksh. Opt. Waveg. Theory Numeric. Modell*., Nottingham, UK, 2002.

[5] M. Marciniak, "Optical transparency and optical opacity in future all-optical packet switched network", in *XXVIIth Gener. Assemb. Int. Union of Radio Sci. URSI 2002*, Maastricht, NL, 2002.

[6] M. Marciniak, "The impact of optical transparency on the successful development of all-optical terabit networks", in *5th World Multi-Conf. Syst., Cyber. Inform. SCI 2001*, Orlando, USA, 2001.

[7] M. Marciniak, "Reliability aspects of the future hybrid optical network and quality of service issues for real time and packet traffic", in *Proc. Int. Conf. Advanc. Optoelectron. & Lasers CAOL 2003*, Alushta, Ukraine, 2003, vol. 1, pp. 63–68.

[8] L. Smoczyński and M. Marciniak, "A comparison of different radio over fibre system concepts with regard to applications in mobile Internet and multimedia access", in *XXVIIth Gener. Assemb. Int. Union of Radio Sci. URSI 2002*, Maastricht, NL, 2002.

[9] M. Marciniak, M. Kowalewski, and M. Klinkowski, "Advanced optical infrastructure for the emerging optical internet services", in *SSGRR 2002s–Int. Conf. Advanc. Infrastr. e-Business, e-Education, e-Science, and e-Medicine on the Internet*, L'Aquila, Italy, 2002, book of abstracts, pp. 82–83.

[10] M. Marciniak, "Towards hybrid real-time & photonic packet network", in *Conf. Opt. Internet & Austr. Conf. Opt. Fibre Technol. COIN/ACOFT 2003*, Melbourne, Australia, 2003, pp. 461–464.

[11] M. Marciniak, "Broadband optical and wireless networking for personal information services", in *7th Int. Conf. "Evolution of telecommunication transport networks. Construction, development and management"*, Partenit, Ukraine, 2004

[12] L. Smoczyński and M. Marciniak, "Radio-over-fibre 60 GHz broadband access", in *Int. Topic. Meet. Microw. Photon. MWP 2003*, *Nefert. Worksh. Broadband Opt. Wirel. Access*, Budapest, Hungary, 2003.

[13] N. Geary, A. Antonopoulos, and J. O'Reilly, "Analysis of the potential benefits of OXC-based intelligent optical networks", *Opt. Netw. Mag.*, vol. 4, no. 2, pp. 20–31, 2003.

[14] M. Klinkowski and M. Marciniak, "Services differentiation in MPLS photonic packet networks", in *5th IFIP Work. Conf. Opt. Netw. Des. Modell. ONDM 2003*, Budapest, Hungary, 2003, vol. I, pp. 283–290.

Marian Marciniak Associate Professor has been graduated in solid state physics from Marie-Curie Sklodovska University in Lublin, Poland, in 1977. From 1985 to 1989 he performed Ph.D. studies in electromagnetic wave theory at the Institute of Fundamental Technological Research, Polish Academy of Sciences, followed by Ph.D. degree (with distinction) in optoelectronics received from Military University of Technology in Warsaw. In 1997 he received his Doctor of Sciences (habilitation) degree in physics/optics from Warsaw University of Technology. From 1978 to 1997 he held an academic position in the Military Academy of Telecommunications in Zegrze, Poland. In 1996 he joined the National Institute of Telecommunications in Warsaw where he actually leads the Department of Transmission and Fibre Technology. Previous activities have included extended studies of optical waveguiding linear and nonlinear phenomena with analytic and numerical methods including beam-propagation methods. Actual research interests include photonic crystal technology and phenomena, optical packet-switched networks, and the future global optical and wireless network. Recently he has introduced and developed a concept of a hybrid real-time service end photonic packet network. He is an author or co-author of over 190 technical publications, including a number of conference invited presentations and 13 books authored, co-authored and/or edited by himself. He is a Senior Member of the IEEE – Lasers & Electro-Optics, Communications, and Computer Societies, a member of The New York Academy of Sciences, The Optical Society of America, SPIE – The International Society for Optical Engineering and its Technical Group on Optical Networks, and of the American Association for the Advancement of Science. In early 2001 he originated the IEEE/LEOS Poland Chapter and he has served as the Chairman of that Chapter until July 2003. He is widely involved in the European research for optical telecommunication networks, systems and devices. He was the originator of accession of Poland to European Research Programs in the optical telecommunications domain, in chronological order: COST 240 *Modelling and Measuring of Advanced Photonic Telecommunication Components*, COST P2 *Applications of Nonlinear Optical Phenomena*, COST 266 *Advanced Infrastructure for Photonic Networks*, COST 268 *Wavelength-Scale Photonic Components for Telecommunications*, COST 270 *Reliability of Optical Components and Devices in Communications Systems and Networks*, COST 273 *Towards Mobile Broadband Multimedia Networks*, and very recently two new starting actions COST 288 *Nanoscale and Ultrafast Photonics* and COST P11 *Physics of Linear, Nonlinear and Active Photonic Crystals*. In all but two those projects he acted as one of the originators at the European level. He has been appointed to Management Committees of all those Projects as the Delegate of Poland. In addition, he has been appointed as the Evaluator of the European Union's 5th Framework Program proposals in the Action Line *All-Optical and Terabit Networks*. He is a Delegate to the International Telecommunication Union, Study Group 15: *Optical and Other Transport Networks*, and to the International Electrotechnical Commission, Technical Committee 86 *Fibre Optics* and its two sub-Committees. He served as a member of Polish Delegation

to the World Telecommunication Standards Assembly WTSA 2000. From 2002 he participates in the work of the URSI – *International Union of Radio Science, Commission D – Electronics and Photonics*. In 2000 he originated and actually serves as the Chairman of the Technical Committee 282 on *Fibre Optics* of the National Committee for Standardisation. Since May 2003 he serves as the Vice-President of the Delegation of Poland to the Intergovernmental Ukrainian-Polish Working Group for Cooperation in Telecommunications. He is the originator and the main organiser of the *International Conference on Transparent Optical Networks* ICTON starting in 1999, and a co-located events the *European Symposium on Photonic Crystals* ESPC and *Workshop on All-Optical Routing* WAOR since 2002. He is the Technical Program Committee Co-Chair of the *International Conference on Advanced Optoelectronics and Lasers* CAOL, and he participates in Program Committees of the *Conference on the Optical Internet & Australian Conference on Optical Fibre Technology* COIN/ACOFT, the *International Conference on Mathematical Methods in Electromagnetic Theory* MMET, the *International Workshop on Laser and Fiber-Optical Network Modeling* LFNM, and the *International School for Young Scientists and Students on Optics, Laser Physics and Biophysics/Workshop on Laser Physics and Photonics*. He serves as a reviewer for several international scientific journals, and he is a Member of the Editorial Board of *Microwave & Optoelectronics Technology Letters* journal, Wiley, USA, and the *Journal of Telecommunications and Information Technology*, National Institute of Telecommunications, Poland. Languages spoken: Polish (native), English, French, and Russian. His biography has been cited in Marquis *Who's Who in the World, Who's Who in Science and Engineering*, and in the *International Directory of Distinguished Leadership of the American Biographical Institute*.

e-mail: M.Marciniak@itl.waw.pl
e-mail: marian.marciniak@ieee.org
Department of Transmission and Fibre Technology
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Regular paper

Influence of common path on availability of ring network

Ivan Rados

Abstract—This paper analyses availability of the ring network which uses the path protection switching (sub-network connection protection – SNCP). Influence of the common path on the ring network availability is analyzed. Data regarding failures of optical fibre cables and equipment used for calculations have been obtained during years-long observation of SDH network in HT Mostar as well as from manufacturers.

Keywords—availability, failure rate, protection, ring network, SDH network.

1. Introduction

Modern industrialized societies are considerably dependent on telecommunication services. Interruption of service for any reason, be it equipment failure or human factor, can cause isolation in telecommunications sense as well as great losses for users and network operators. Hence, "survival" of the transmission network in the conditions of failure and mistake, becomes a primary task of network operators. In synchronous digital hierarchy network (SDH), more standardized mechanisms that provide "survival" of transmission network, are predicted [1]. One of these mechanisms, "path protection", will be analyzed in this paper.

Ring structure, which provides two separated paths inside the ring, is particularly interesting. However, in real networks there are cases when paths inside the ring structure are not completely physically separated although they constitute the logical ring. In those cases the failures cannot be considered statistically independent. This is particularly related to the urban environment where it is difficult to obtain permission for the construction of another optical cable duct, so both working and protection paths inside the ring use may fibres in the same cable. This paper will provide a precise analysis of the influence of the common path on availability of 2 Mbit/s channels between two nodes inside the ring structure.

2. On availability in general

The availability $A(t)$ is defined as the probability that the system is operating at a specified point of time t [2]. As a more useful measure, availability A is determined as the ratio between the total time of failure-free operation and the total monitoring time:

$$A = \frac{MTTF}{MTTF + MTTR}\ , \tag{1}$$

where $MTTF$ (mean time to failure) is mean time till the failure occurs and $MTTR$ (mean time to repair) a mean time of repair.

A common unit related to availability and widely used in networking is the FIT (failures in time), where 1 FIT corresponds to 1 failure in 10^9 hours. This measure has some relevance to availability analysis that lies in the fact that the FIT value can be used to calculate the mean time to failure (in hours) of a component as follows:

$$MTTF = \frac{10^9}{FIT}\,. \tag{2}$$

Unavailability U is probability complementary to availability [3], i.e., $U = 1 - A$. When reporting system/network performance, unavailability U is often expressed as MDT (mean down time) in minutes per year [min/year], i.e.,

$$MDT = 365 \cdot 60 \cdot 24 \cdot U\,. \tag{3}$$

As SDH network generally consists of cable sections and nodes, optical fibre failure rate is calculated separately from node failure rate. Optical fibre failure [4] rate λ per km of installed cable per year [1/hkm] is calculated according to the equation

$$\lambda = n / M \cdot T\,, \tag{4}$$

where n is a number of failures over monitoring time, M the length of installed cable in km and T monitoring period in hours. Failure rate of individual modules that make the node (add/drop multiplexer) is received from the manufacturer.

Availability of the series structure which includes n elements (Fig. 1) is simply the product of their individual availabilities:

$$A_s = A_1 \cdot A_2 \cdot \ \ldots\ \cdot A_n = \prod_{i=1}^{n} A_i\,. \tag{5}$$

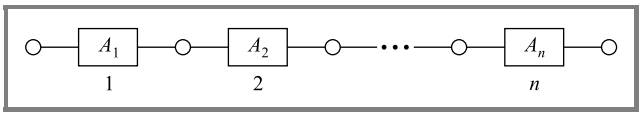


Fig. 1. Series structure from n elements.

As the availability equation shows, the failure of any element causes unavailability of the whole structure.

Generally, for parallel structure of n branches (Fig. 2a) availability is

$$A_p = 1 - \prod_{i=1}^{n} (1 - A_i)\,. \tag{6}$$

Since, two branches (working and protection ones) are needed for the "protection path" mechanism, we are going to analyze parallel structure availability as in Fig. 2b, reflected in the following formula:

$$\begin{aligned} A_P &= 1-[(1-A_a)(1-A_b)] \\ &= 1-(1-A_a-A_b+A_aA_b) \\ &= 1-1+A_a+A_b-A_aA_b \\ &= A_a+A_b-A_aA_b\,. \end{aligned} \tag{7}$$

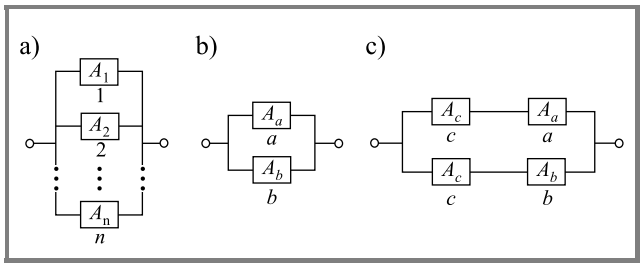


Fig. 2. Parallel structure of (a) n branches; (b) two branches; (c) two branches with common element one.

If we add to each branch the same element c, but in fact two elements of common section, whose failures are mutually completely dependent (Fig. 2c), availability of such a structure is a union probability of non-disjunctive events. Assuming that failures on network sections which are physically separated are mutually independent, we have

$$\begin{aligned} A_{cp} &= p\big[(c\cap a)\cup(c\cap b)\big] \\ &= p\big[c\cap a\big]+p\big[c\cap b\big]-p\big[(c\cap a\cap c\cap b)\big] \\ &= p\big[c\cap a\big]+p\big[c\cap b\big]-p\big[(c\cap a\cap b)\big] \\ &= A_cA_a+A_cA_b-A_cA_aA_b\,. \end{aligned} \tag{8}$$

As the above equation shows, the failure on the c element results in reduction of parallel structure availability, because c is a common element for both branches.

3. Protection path mechanism in the ring network

In a network, using the protection path (SNCP ring) mechanism, signal in the source node is transmitted in both ring directions, and a higher quality signal is chosen on the destination source input. The path the transmission signal passes through during its normal operations is called the working path (P_0), and the path the signal passes in case of failure on its working path is called protection path (P_1) [5].

We assume that the network consists of N nodes and N cable sections linking those nodes. In order to determine availability of signal we will introduce the expressions related to availability of node and availability of cable section between two nodes:

- a_{f_i}, availability of i-cable section of working path;
- a_{n_j}, availability of j-node which belongs to the working path.

Nodes in the ring in which transmission signal is extracted and added are called "termination" nodes (s and t in Fig. 3). Nodes in the ring between termination nodes the signal only passes through from the "east" to the "west" side, are called "through" (transit) nodes. Since transmission signal passes through different components inside those two node types, their availability is different:

- a_{n_jt}, termination node is working;
- a_{n_jp}, transit node is working.

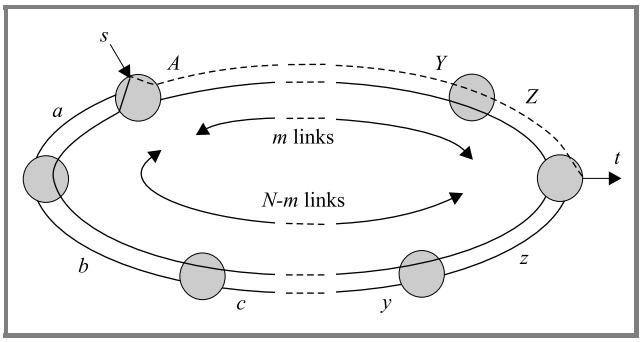


Fig. 3. Path protection in the ring network.

If all nodes of the same type are equal, then

$$a_{n_jt}=a_{nt},\ \forall j\,,$$

$$a_{n_jp}=a_{np},\ \forall j\,.$$

If transmission signal on the working path P_0 passes m cable sections between termination nodes, availability of the working path is equal to the product of all nodes and cable sections availability on that path:

$$\begin{aligned} a_{st}(P_0) &= \prod_{i,\,j\in P_0} a_{f_i}a_{n_j} = \prod_{i\in P_0} a_{f_i}\prod_{j\in P_0} a_{n_j} \\ &= a_{nt}^2\,a_{np}^{m-1}\prod_{i\in P_0}^{m} a_{f_i}. \end{aligned} \tag{9}$$

In case of failure on the working path, transmission signal passes $N-m$ cable sections and $N-m-1$ on protection path P_1 (Fig. 4).
Availability of the protection path is

$$\begin{aligned} a_{st}(P_1) &= \prod_{i,\,j\in P_1} a_{f_i}a_{n_j} = \prod_{i\in P_1} a_{f_i}\prod_{j\in P_1} a_{n_j} \\ &= a_{nt}^2\,a_{np}^{N-m-1}\prod_{i\in P_1}^{N-m} a_{f_i}. \end{aligned} \tag{10}$$

Availability of the transmission signal between s and t nodes is completely defined with these two paths, so the

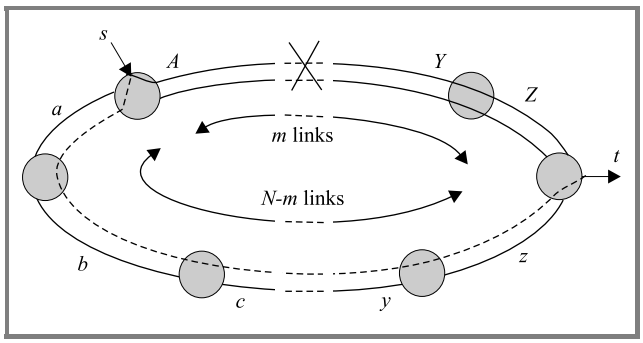


Fig. 4. Path protection in the ring network.

availability of the ring with path protection mechanism A_{st}, with completely independent paths is calculated as the availability of the parallel structure:

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

$$A_{st}(a) = a_{nt}^2\, a_{np}^{m-1} \prod_{i\in P_0}^{m} a_{f_i} + a_{nt}^2\, a_{np}^{N-m-1} \prod_{i\in P_1}^{N-m} a_{f_i}$$

$$-a_{nt}^2\, a_{np}^{m-1} \prod_{i\in P_0}^{m} a_{f_i}\, a_{nt}^2\, a_{np}^{N-m-1} \prod_{i\in P_1}^{N-m} a_{f_i}\,, \qquad (11)$$

where a denote the set of fibre and node availabilities.
Although the equation in the big bracket contains a product of multiplication of two equal parts $(a_{nt})^2 \cdot (a_{nt})^2$, only one is taken for availability $(a_{nt})^2$ because the cause of node failure is the same and we obtain:

$$A_{st}(a) = a_{nt}^2 \left(a_{np}^{m-1} \prod_{i\in P_0}^{m} a_{f_i} + a_{np}^{N-m-1} \right.$$

$$\left. \times \prod_{i\in P_1}^{N-m} a_{f_i} - a_{np}^{N-2} \prod_{i\in P_0,P_1}^{N} a_{f_i} \right). \qquad (12)$$

If we suppose that two cable sections have the same length, then the availability of each is the same, i.e.,

$$a_{f_i} = a_f, \forall i.$$

In this case availability between s and t nodes is:

$$A_{st}(a) = a_{nt}^2 \left(a_{np}^{m-1}\, a_f^m + a_{np}^{N-m-1} \right.$$

$$\left. \times\, a_f^{N-m} - a_{np}^{N-2}\, a_f^N \right). \qquad (13)$$

If there is one common part of the path then the availability equation should include the availability of this part, and hence we have

$$A_{st\,cp}(a) = A_{cp}\, a_{nt}^2 \left(a_{np}^{m-1} \prod_{i\in P_0}^{m} a_{f_i} + a_{np}^{N-m-1} \right.$$

$$\left. \times \prod_{i\in P_1}^{N-m} a_{f_i} - a_{np}^{N-2} \prod_{i\in P_0,P_1}^{N} a_{f_i} \right), \qquad (14)$$

where A_{cp} is availability of the common path.

4. Analysis of the common path influence on availability

Availability analysis will be performed on the ring-shaped SDH network consisting of different numbers of nodes, assuming equal distances between nodes (20 and 30 km). Availability calculation was performed for 2 Mbit/s link between two nodes, assuming that the paths are completely independent and that there is one common path of different length. Availability value of "termination" and "through" (transit) nodes was calculated on the basis of data collected during system operation ($MTTR$ = 6.58 hours) and data on individual equipment module failure rate, provided by the manufacturer (Table 1).

Table 1
Unavailability and *MDT* for termination and transit nodes

Unavailability	U $[\times 10^{-5}]$	*MDT* [min/year]
Termination node	0.113	0.69
Transit node	7.385	38.82

As shown in Table 1, the termination node has much smaller unavailability than transit node because it has redundancy of all cards. With transit node the signal passes through from "east" to the "west" side, resulting in serial structure which is very sensitive to failures of individual elements.
Average failure rate for fibre optic cables is calculated on the basis of the number of failures (29) along the 795.135 km of installed cables during six years' period – see Table 2.

Table 2
Failure rate, unavailability and *MDT* for fiber optic cables

λ [*FIT*/km]	U $[\times 10^{-5}]$	*MDT* [min/year]
693.91	1.30	6.86

We shall first calculate the availability of the ring network consisting of different number of nodes, assuming completely independent paths and equal distances among nodes (20 km). As we can see in Table 3, with lesser number of nodes and consequently shorter total length of working path, the influence of node availability on total availability between source and target nodes is dominant. It results from the fact that in total availability, with, e.g., $N = 6$ nodes

Table 3
Unavailability and *MDT* for ring-shaped SDH network

Nodes	6	8	10	12	14
U $[\times 10^{-6}]$	3.34	3.95	4.75	5.74	6.91
MDT [min/year]	1.76	2.08	2.50	3.01	3.63
Cable [%]	20.64	32.19	42.76	51.73	59.05
Node [%]	74.36	67.81	57.24	48.27.	40.95

the influence of source and target node availability is prevalent with no less than 74.36% share. With increasing number of nodes and total length of the working path, the cable influence becomes dominant ($N = 14$, cables share: 59.05%).
If distance between nodes is increased to 30 km (still with completely independent paths) there is an almost linear increase of mean time to failure ($N = 10$, MDT increases for approximately 30%; $N = 12$, MDT increases for approximately 33%).

Table 4
Unavailability and MDT for ring-shaped SDH network ($d = 30$ km)

Nodes	6	8	10	12	14
U $[\times 10^{-6}]$	4.05	5.23	6.76	8.65	10.90
MDT [min/year]	2.13	2.75	3.56	4.55	5.73
Cable [%]	34.42	48.69	59.77	67.99	74.03
Node [%]	65.58	51.31	40.23	32.01	25.97

The results given in Table 4 show that the influence of node availability has decreased and the influence of cables on total availability has increased, since total length of the working path has increased. If we assume that there is 1 km of the common path, then the availability between source and target nodes is slightly greater, as seen in Table 5.

Table 5
Unavailability and MDT with 1 km of common path ($d = 20$ km)

Nodes	6	8	10	12	14
U $[\times 10^{-5}]$	1.64	1.70	1.78	1.87	1.94
MDT [min/year]	8.62	8.94	9.36	9.88	10.24
Cable [%]	83.80	84.22	84.71	85.25	85.82
Node [%]	16.20	15.78	15.29	14.75	14.18

Existence of only 1 km of the common path (e.g., $N = 8$, $d = 20$ km) increases mean down time by 429.81% or from 2.08 min/year (completely independent paths) to 8.94 min/year. Also, the influence of cable availability in total availability between two nodes becomes dominant (84.22% of cable availability, 15.78% of node availability). If for $d = 30$ km we assume 1 km of the common path (Table 6), the influence of cable availability on total availability becomes even more dominant ($N = 8$, cables 85.32%, nodes 14.68%). The result given in Tables 5 and 6 show that the common part of the path has the greatest impact on total availability between two nodes. For $N = 10$ and $d = 20$, 1% of the common path (1 km) in relation to total length of the working path gives about 73% of total availability. With $N = 10$ and $d = 30$ this impact is slightly less (about 66%), since the share of this part of the path on total length of the working path is smaller (0.66%).

Table 6
Unavailability and MDT with 1 km of common path ($d = 30$ km)

Nodes	6	8	10	12	14
U $[\times 10^{-5}]$	1.71	1.82	1.98	2.17	2.39
MDT [min/year]	8.99	9.61	10.42	11.41	12.59
Cable [%]	84.47	85.32	86.26	87.23	88.18
Node [%]	15.53	14.68	13.74	12.77	11.82

Although there are only two cables in the same trench on the common path, during the calculation we used only the availability of one cable because, on the basis of the analysis of failure situations in HT Mostar that have been done so far, both cables are always cut during failures. The main cause of optical fibre cable cuts is digging.

5. Conclusion

The results obtained in this analysis show that in ring networks with small number of nodes and completely independent paths, the nodes (particularly the termination ones) have a greater impact on total availability than the cables since the total length of working path is small and the availability of parallel structure is very close to unity. The increased distance between the nodes, and consequently the total length of the working path, leads to increased mean time to failure, and thus greater impact of cables on total availability of the ring network.
Further on, the availability of the ring structure mostly depends on the existence of the common path for working and protection directions. Very short common sections significantly reduce availability between two nodes in the ring structure. Negative influence of the common path is clearly shown in the equation for availability of the ring structure $A_{st\,cp}$ because, if $A_{cp} = 0$ (common factor of the equation), then the communication between two nodes is broken. Improved availability can be achieved with total independence of the working and the protection paths, which means that routing of working and protection path in the same cable or trench, must be avoided.

References

[1] C. Coltro, "Evolution of transport network architectures", *Alcatel Telecommun. Rev.*, 1st Quarter 1997, pp. 10–18.

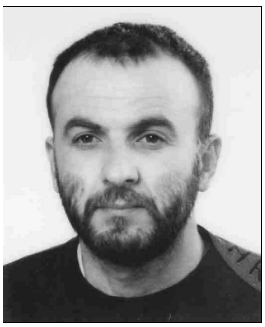
[2] R. Inkret and B. Mikac, "Availability analysis of different WDM network protection scenarios", in *Proc. Conf. WAON'98*, Zagreb, Croatia, 1998, pp. 121–128.

[3] J. Baudron, A. Khadr, and F. Kocsis, "Availability and survivability of SDH networks", *Electr. Commun.*, 4th Quarter 1993.

[4] I. Jurdana and B. Mikac, "An availability analysis of optical cables", in *Proc. Conf. WAON'98*, Zagreb, Croatia, 1998, pp. 153–160.

[5] M. R. Wilson, "The quantitative impact of survivable network architecture on service availability", *IEEE Commun. Mag.*, May 1998, pp. 122–127.

[6] I. Rados, "Availability analysis of synchronous digital hierarchy network". Master thesis, University of Zagreb, 2000 (in Croatian).



Ivan Rados received the B.Sc. degree in electrical engineering from the University of Split, Croatia, in 1983, and M.Sc. degree from the University of Zagreb, in 2000. In 1985 he joined the PTT (Post and Telecommunication) office in Tomislavgrad. Since 1992 he has been working at Department of Transmission Systems of the HT Mostar (Croatian Telecommunication). His research interests include digital transmission systems, optical systems and networks, availability and reliability of telecommunication systems. He has published 8 papers in international conference proceedings.

e-mail: ivan.rados@ht.ba
Croatian Telecommunication d.o.o. Mostar
Kneza Branimira bb
88 000 Mostar, Bosnia and Herzegovina

INFORMATION FOR AUTHORS

The *Journal of Telecommunications and Information Technology* is published quarterly. It comprises original contributions, both regular papers and letters, dealing with a broad range of topics related to telecommunications and information technology. Items included in the journal report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

The *Journal* is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, speech devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology. Suitable research-related manuscripts should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published by invitation only.

Papers published by invitation and regular papers should contain up to 15 and 8 printed pages respectively (one printed page corresponds approximately to 3 double-space pages of manuscript, where one page contains approximately 2000 characters).

Manuscript: An original and two copies of the manuscript must be submitted, each completed with all illustrations and tables attached at the end of the papers. Tables and figures have to be numbered consecutively with Arabic numerals. The manuscript must include an abstract limited to approximately 100 words. The abstract should contain four points: statement of the problem, assumptions and methodology, results and conclusion, or discussion, of the importance of the results. The manuscript should be double-spaced on only one side of each A4 sheet (210 × 297 mm). Computer notation such as Fortran, Matlab, Mathematica etc., for formulae, indices, etc., is not acceptable and will result in automatic rejection of the manuscript. The style of references, abbreviations, etc., should follow the standard IEEE format.

References should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. **The reference entry** (correctly punctuated according to the following rules and examples) **has to contain**.

From journals and other serial publications: initial(s) and second name(s) of the author(s), full title of publication (transliterated into Latin characters in case it is in Russian, possibly preceded by the title in Russian characters), appropriately abbreviated title of periodical, volume number, first and last page number, year. E.g.:

[1] Y. Namihira, "Relationship between nonlinear effective area and modefield diameter for dispersion shifted fibres", *Electron. Lett.*, vol. 30, no. 3, pp. 262–264, 1994.

From non-periodical, collective publications: as above, but after title – the name(s) of editor(s), title of volume and/or edition number, publisher(s) name(s) and place of edition, inclusive pages of article, year. E.g.:

[2] S. Demri, E. Orłowska, "Informational representability: Abstract models versus concrete models" in *Fuzzy Sets, Logics and Reasoning about Knowledge*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314.

From books: initial(s) and name(s) of the author(s), place of edition, title, publisher(s), year. E.g.:

[3] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.

Figure captions should be started on separate sheet of papers and must be double-spaced.

Illustration: Original illustrations should be submitted. All line drawings should be prepared on white drawing paper in black India ink. Drawings in Corel Draw and Postscript formats are preferred. Colour illustrations are accepted only in exceptional circumstances. Lettering should be large enough to be readily legible when drawing is reduced to two- or one-column width – as much as 4:1 reduction from the original. Photographs should be used sparingly. All photographs must be gloss prints. All materials, including drawings and photographs, should be no larger than 175 × 260 mm.

Page number: Number all pages, including tables and illustrations (which should be grouped at the end), in a single series, with no omitted numbers.

Electronic form: A floppy disk together with the hard copy of the manuscript should be submitted. It is important to ensure that the diskette version and the printed version are identical. The diskette should be labelled with the following information: a) the operating system and word-processing software used, b) in case of UNIX media, the method of extraction (i.e. tar) applied, c) file name(s) related to manuscript. The diskette should be properly packed in order to avoid possible damage during transit.

Among various acceptable word processor formats, TEX and LATEX are preferable. The *Journal's* style file is available to authors.

Galley proofs: Proofs should be returned by authors as soon as possible. In other cases, the article will be proof-read against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within two weeks after receiving the offprints.

The copy of the "Journal" shall be provided to each author of papers.

Copyright: Manuscript submitted to this journal may not have been published and will not be simultaneously submitted or published elsewhere. Submitting a manuscript, the authors agree to automatically transfer the copyright for their article to the publisher if and when the article is accepted for publication. The copyright comprises the exclusive rights to reproduce and distribute the article, including reprints and also all translation rights. No part of the present journal may be reproduced in any form nor transmitted or translated into a machine language without permission in written form from the publisher.

Biographies and photographs of authors are printed with each paper. Send a brief professional biography not exceeding 100 words and a gloss photo of each author with the manuscript.

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Editorial Office

tel. +48(22) 512 81 83
tel./fax: +48(22) 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.itl.waw.pl/jtit