JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

SPECIAL ISSUE 2024

(*Contents continued on back cover*)

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

25 years

2000-2024

Editor-in-Chief

Steering Editor

Editorial Advisory Board

Editorial Team

eISSN 1899-8852 Special Issue 2024
Page numbers in brackets refer to the original editions

JOURNAL OF TELECOMMUNICATIONS
AND INFORMATION TECHNOLOGY

JTIT – past and present

Time flies and Journal of Telecommunications and Information Technology (JTIT) is already 25 years old! To celebrate this occasion, we have prepared a special issue presenting 25 top articles, i.e. those that were most frequently cited in other publications according to Web of Science and Scopus data. We hope this will be a great opportunity to take a look back into the past. We also hope that all authors who have ever published their papers in JTIT will take this opportunity to check whether their piece has made it to the "Top 25 of Most Cited Publications".

A quarter of a century marks an entire era in the advancement of today's technology. Rather naturally, older publications differ, both visually and thematically, from current ones. Is proves, however, that the journal is evolving and developing on a continuous basis.

The very first issue of JTIT appeared in the middle of the year 2000. It was a *Special Issue on Applications of Nonlinear Optical Phenomena* edited by our guest editors – Marian Marciniak and Werner Blau. It comprised contributions to the 1st Workshop of COST (*European Cooperation in Science and Technology*) Action P2 "*Applications of Nonlinear Optical Phenomena*", held on 12-13 June 1998 in Limerick, Ireland. The issue contained Preface, Introductory Remarks, and 18 technical papers, all published on 90 pages. The preface contained a statement by prof. Andrzej P. Wierzbicki, director of the *National Institute of Telecommunications* and the author of the idea to create the JTIT. The introductory remarks to the issue were extended jointly by local workshop organizer prof. Werner Blau from *Trinity College Dublin*, and COST P2 Action Chairman prof. Yves Lion from *Université de Liège*. The topics discussed in the paper focused on the importance of photonic technologies in a number of technology-intensive fields, such as telecommunications, IT, diagnostics, quality control, etc. The contributions originated from eight countries: France, Hungary, Ireland, Italy, Russia, the Netherlands, Ukraine, and Poland.

Over the years, the topics that the publication was dealing with became more aligned with the current profile and title of the Journal. The number of foreign authors increased and special editions were slowly phased out and replaced by their regular counterparts. The Journal's program council, made up of highly renowned scholars from all over the world, also helped achieve wide-scale acceptance of the publication within the scientific community.

The Journal has become widely recognizable, both in Poland and abroad. Its reach has increased considerably, as has the group of authors publishing in JTIT on a regular basis for a number of years.

The first editor-in-chief, prof. Paweł Szczepański (2000-2019), represented JTIT's published – the *National Institute of Telecommunications*. Since mid-2019, the position of the editor-in-chief has been held by prof. Adrian Kliks (*Poznań University of Technology*). Since 2021, JTIT has been published exclusively in electronic form, and the repository of its publications contains over 1,200 articles.

I am very pleased to play my part in this project and I hope that the next 25 years will be remembered by the entire scientific community as a very productive period.

Ewa Kapuściarek, Head of the Scientific Information Center

JOURNAL OF TELECOMMUNICATIONS
AND INFORMATION TECHNOLOGY

1,2/2000

Special issue edited by Werner J. Blau and Marian Marciniak

Rough set theory and its applications

Zdzisław Pawlak

http://doi.org/10.26636/jtit.2002.140

Abstract — In this paper rudiments of the theory will be outlined, and basic concepts of the theory will be illustrated by a simple tutorial example, concerning churn modeling in telecommunications. Real life applications require more advanced extensions of the theory but we will not discuss these extensions here. Rough set theory has an overlap with many other theories dealing with imperfect knowledge, e.g., evidence theory, fuzzy sets, Bayesian inference and others. Nevertheless, the theory can be regarded as an independent, complementary, not competing, discipline in its own rights.

***Keywords* — rough set, decision rules, churn modeling.**

1. Introduction

Rough set theory can be regarded as a new mathematical tool for imperfect data analysis. The theory has found applications in many domains, such as decision support, engineering, environment, banking, medicine and others.

This paper presents basis of the theory which will be illustrated by a simple example of churn modeling in telecommunications.

Rough set philosophy is founded on the assumption that with every object of the universe of discourse some information (data, knowledge) is associated. Objects characterized by the same information are *indiscernible (similar)* in view of the available information about them. The *indiscernibility relation* generated in this way is the mathematical basis of rough set theory. Any set of all indiscernible (similar) objects is called an *elementary set*, and forms a basic *granule (atom)* of knowledge about the universe. Any union of some elementary sets is referred to as a *crisp (precise)* set – otherwise the set is *rough (imprecise, vague)*. Each rough set has boundary-line cases, i.e., objects which cannot be with certainty classified, by employing the available knowledge, as members of the set or its complement. Obviously rough sets, in contrast to precise sets, cannot be characterized in terms of information about their elements. With any rough set a pair of precise sets, called the *lower* and the *upper approximation* of the rough set, is associated. The lower approximation consists of all objects which *surely* belong to the set and the upper approximation contains all objects which *possibly* belong to the set. The difference between the upper and the lower approximation constitutes the *boundary region* of the rough set. Approximations are fundamental concepts of rough set theory.

Rough set based data analysis starts from a data table called a *decision table*, columns of which are labeled by *attributes*, rows – by *objects* of interest and entries of the table are *attribute values*. Attributes of the decision table are divided into two disjoint groups called *condition* and *decision* attributes, respectively. Each row of a decision table induces a *decision rule*, which specifies decision (action, results, outcome, etc.) if some conditions are satisfied. If a decision rule uniquely determines decision in terms of conditions – the decision rule is *certain*. Otherwise the decision rule is *uncertain*. Decision rules are closely connected with approximations. Roughly speaking, certain decision rules describe lower approximation of decisions in terms of conditions, whereas uncertain decision rules refer to the boundary region of decisions.

With every decision rule two conditional probabilities, called the *certainty* and the *coverage* coefficient, are associated. The certainty coefficient expresses the conditional probability that an object belongs to the decision class specified by the decision rule, given it satisfies conditions of the rule. The coverage coefficient gives the conditional probability of reasons for a given decision.

It turns out that the certainty and coverage coefficients satisfy Bayes' theorem. That gives a new look into the interpretation of Bayes' theorem, and offers a new method data to draw conclusions from data.

In the paper rudiments of the theory will be outlined, and basic concepts of the theory will be illustrated by a simple tutorial example of churn modeling. Real life applications require more advanced extensions of the theory but we will not discuss these extensions in this paper.

Rough set theory has an overlap with many other theories dealing with imperfect knowledge, e.g., evidence theory, fuzzy sets, Bayesian inference and others. Nevertheless, the theory can be regarded as an independent, complementary – not competing discipline, in its own rights.

More information about rough sets and their applications can be found in the references and the Web.

2. Illustrative example

Let us start our considerations from a very simple tutorial example concerning churn modeling in telecommunications, which is a simplified version of an example given in [1]. In Table 1, six facts concerning six client segments are presented.

In the table condition attributes describing client profile are: *In* – incoming calls, *Out* – outgoing calls within the same operator, *Change* – outgoing calls to other mobile operator, the decision attribute describing the consequence is *Churn* and *N* is the number of similar cases.

Each row in the table determine a decision rule. E.g., row 2 determines the following decision rule: *"if the number of incoming calls is high and the number of outgoing calls is high and the number of outgoing calls to the mobile operator is low then these is no churn".*

According to [1]: *"One of the main problem that have to be solved by marketing departments of wireless operators is to find the way of convincing current clients that they continue to use the services. In solving this problems can help churn modeling. Churn model in telecommunications industry predicts customers who are going to leave the current operator".*

Table 1
Client segments

Segment	*In*	*Out*	*Change*	*Churn*	*N*
1	medium	medium	low	no	200
2	high	high	low	no	100
3	low	low	low	no	300
4	low	low	high	yes	150
5	medium	medium	low	yes	220
6	medium	low	low	yes	30

In other words we want to explain churn in terms of clients profile, i.e., to describe market segments {4, 5, 6} (or {1, 2, 3}) in terms of condition attributes *In, Out* and *Change*.

The problem cannot be solved uniquely because the data set is *inconsistent*, i.e., segments 1 and 5 have the same profile but different consequences.

Let us observe that:

- segments 2 and 3 (4 and 6) can be classified as sets of clients who *certainly* do not churn (churn),
- segments 1, 2, 3 and 5 (1, 4, 5 and 6) can be classified as sets of clients who *possibly* do not churn (churn),
- segments 1 and 5 are *undecidable* sets of clients.

This leads us to the following notions:

- the set {2,3} ({4,6}) is the *lower approximation* of the set {1,2,3}({4,5,6}),
- the set {1,2,3,5} ({1,4,5,6}) is *the lower approximation* of the set {1,2,3} ({4,5,6}),
- the set {1,5} is the *boundary region* of the set {1,2,3}({4,5,6}),

which will be discussed in the next paragraph more exactly.

3. Information systems and approximations

In this section we will examine approximations more exactly. First we define a data set, called an information system.

An *information system* is a pair $S = (U, A)$, where U and A, are finite, nonempty sets called the *universe*, and the set of *attributes*, respectively. With every attribute $a \in A$ we associate a set V_a, of its *values*, called the *domain* of a. Any subset B of A determines a binary relation $I(B)$ on U, which will be called an *indiscernibility relation*, and defined as follows: $(x,y) \in I(B)$ if and only if $a(x) = a(y)$ for every $a \in A$, where $a(x)$ denotes the value of attribute a for element x. Obviously $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$, i.e., a partition determined by B, will be denoted by $U/I(B)$, or simply by U/B; an equivalence class of $I(B)$, i.e., block of the partition U/B, containing x will be denoted by $B(x)$. If (x,y) belongs to $I(B)$ we will say that x and y are *B-indiscernible (indiscernible with respect to B).* Equivalence classes of the relation $I(B)$ (or blocks of the partition U/B) are referred to as *B-elementary sets* or *B-granules*.

Suppose we are given an information system $S = (U, A)$, $X \subseteq U$, and $B \subseteq A$. Let us define two operations assigning to every $X \subseteq U$ two sets $B_*(X)$ and $B^*(X)$, called the *B-lower* and the *B-upper approximation* of X, respectively, and defined as follows:

$$B_*(X) = \bigcup_{x \in U} \left\{ B(x) : B(x) \subseteq X \right\},$$

$$B^*(X) = \bigcup_{x \in U} \left\{ B(x) : B(x) \cap X \neq \emptyset \right\}.$$

Hence, the *B-lower approximation* of a set is the union of all *B-granules* that are included in the set, whereas the *B-upper approximation* of a set is the union of all *B-granules* that have a nonempty intersection with the set. The set

$$BN_B(X) = B^*(X) - B_*(X)$$

will be referred to as the *B-boundary region* of X.

If the boundary region of X is the empty set, i.e., $BN_B(X) = \emptyset$, then X is *crisp (exact)* with respect to B; in the opposite case, i.e., if $BN_B(X) \neq \emptyset, X$ is referred to as *rough (inexact)* with respect to B.

Thus, the set of elements is rough (inexact) if it cannot be defined in terms of the data, i.e. it has some elements that can be classified neither as member of the set nor its complement in view of the data.

4. Decision tables and decision rules

If we distinguish in an information system two disjoint classes of attributes, called *condition* and *decision attributes*, respectively, then the system will be called a *decision table* and will be denoted by $S = (U, C, D)$, where C and D are disjoint sets of condition and decision attributes, respectively.

Let $S = (U, C, D)$ be a decision table. Every $x \in U$ determines a sequence $c_1(x), \ldots, c_n(x)$, $d_1(x), \ldots, d_m(x)$, where $\{c_1, \ldots, c_n\} = C$ and $\{d_1, \ldots, d_m\} = D$.

The sequence will be called a *decision rule induced by x* (in S) and will be denoted by $c_1(x),\dots,c_n(x) \rightarrow d_1(x),\dots,d_m(x)$ or in short $C \rightarrow_x D$.

The number $supp_x(C,D) = |A(x)| = |C(x) \cap D(x)|$ will be called a *support* of the decision rule $C \rightarrow_x D$ and the number

$$\sigma_x(C,D) = \frac{supp_x(C,D)}{|U|},$$

will be referred to as the *strength* of the decision rule $C \rightarrow_x D$, where $|X|$ denotes the cardinality of X.

With every decision rule $C \rightarrow_x D$ we associate the *certainty factor* of the decision rule, denoted $cer_x(C,D)$ and defined as follows:

$$cer_x(C,D) = \frac{|C(x) \cap D(x)|}{|C(x)|} = \frac{supp_x(C,D)}{|C(x)|} = \frac{\sigma_x(C,D)}{\pi\big(C(x)\big)},$$

where $\pi\big(C(x)\big) = \frac{|C(x)|}{|U|}$.

The certainty factor may be interpreted as a conditional probability that y belongs to $D(x)$ given y belongs to $C(x)$, symbolically $\pi_x(D|C)$.

If $cer_x(C,D) = 1$, then $C \rightarrow_x D$ will be called a *certain decision* rule; if $0 < cer_x(C,D) < 1$ the decision rule will be referred to as an *uncertain decision rule*.

Besides, we will also use a *coverage factor* of the decision rule, denoted $cov_x(C,D)$ and defined as

$$cov_x(C,D) = \frac{|C(x) \cap D(x)|}{|D(x)|} = \frac{supp_x(C,D)}{|D(x)|} = \frac{\sigma_x(C,D)}{\pi\big(D(x)\big)},$$

where $\pi\big(C(x)\big) = \frac{|D(x)|}{|U|}$.

Similarly

$$cov_x(C,D) = \pi_x(C|D).$$

If $C \rightarrow_x D$ is a decision rule then $D \rightarrow_x C$ will be called an *inverse decision rule*. The inverse decision rules can be used to give *explanations (reasons)* for a decision.

For Table 1 we have the certainty and coverage factors are as shown in Table 2.

Table 2
Parameters of the decision rules

Decision rule	Strength	Certainty	Coverage
1	0.20	0.48	0.33
2	0.10	1.00	0.17
3	0.30	1.00	0.50
4	0.15	1.00	0.38
5	0.22	0.52	0.55
6	0.03	1.00	0.07

Let us observe that if $C \rightarrow_x D$ is a decision rule then

$$\bigcup_{y \in D(x)} \{C(y) : C(y) \subseteq D(x)\}$$

is the lower approximation of the decision class $D(x)$, by condition classes $C(y)$, whereas the set

$$\bigcup_{y \in D(x)} \{C(y) : C(y) \cap D(x) \neq \emptyset\}$$

is the upper approximation of the decision class by condition classes $C(y)$.

Approximations and decision rules are two different methods to express properties of data. Approximations suit better to express topological properties of data, whereas decision rules describe in a simple way hidden patterns in data.

5. Probabilistic properties of decision tables

Decision tables (and decision algorithms) have important probabilistic properties which are discussed next.

Let $C \rightarrow_x D$ be a decision rule and let $\Gamma = C(x)$ and $\Delta = D(x)$. Then the following properties are valid:

$$\sum_{y \in \Gamma} cer_y(C,D) = 1, \tag{1}$$

$$\sum_{y \in \Delta} cov_y(C,D) = 1, \tag{2}$$

$$\pi\big(D(x)\big) = \sum_{y \in \Gamma} cer_y(C,D) \cdot \pi\big(C(y)\big) = \\ = \sum_{y \in \Gamma} \sigma_y(C,D), \tag{3}$$

$$\pi\big(C(x)\big) = \sum_{y \in \Delta} cov_y(C,D) \cdot \pi\big(D(y)\big) = \\ = \sum_{y \in \Delta} \sigma_y(C,D), \tag{4}$$

$$cer_x(C,D) = \frac{cov_x(C,D) \cdot \pi\big(D(x)\big)}{\sum_{y \in \Delta} cov_y(C,D) \cdot \pi\big(D(y)\big)} = \frac{\sigma_x(C,D)}{\pi\big(C(x)\big)}, \tag{5}$$

$$cov_x(C,D) = \frac{cer_x(C,D) \cdot \pi\big(C(x)\big)}{\sum_{y \in \Gamma} cer_y(C,D) \cdot \pi\big(C(y)\big)} = \frac{\sigma_x(C,D)}{\pi\big(D(x)\big)}. \tag{6}$$

That is, any decision table satisfies Eqs.(1)–(6). Observe that formulae (3) and (4) refer to the well known *total probability theorem*, whereas (5) and (6) refer to *Bayes' theorem*.

Thus in order to compute the certainty and coverage factors of decision rules according to formula (5) and (6) it is enough to know the strength (support) of all decision rules only. The strength of decision rules can be computed from data or can be a subjective assessment.

6. Decision algorithm

Any decision table induces a set of *"if ... then"* decision rules.

Any set of mutually, exclusive and exhaustive decision rules, that covers all facts in S and preserves the indiscernibility relation included by S will be called a decision algorithm in S.

An example of decision algorithm in the decision Table 1 is given below:

		cer.
1)	*if (In, high) then (Churn, no)*	1.00
2)	*if (In, low) and (Change, low) then (Churn, no)*	1.00
3)	*if (In, med.) and (Out, med.) then (Churn, no)*	0.48
4)	*if (Change, high) then (Churn, yes)*	1.00
5)	*if (In, med.) and (Out, low) then (Churn, yes)*	1.00
6)	*if (In, med.) and (Out, med.) then (Churn, yes)*	0.52

Finding a minimal decision algorithm associated with a given decision table is rather complex. Many methods have been proposed to solve this problem, but we will not consider this problem here.

If we are interested in *explanation* of decisions in terms of conditions we need an *inverse* decision algorithm which is obtained by replacing mutually conditions and decisions in every decision rule in the decision algorithm.

For example, the following inverse decision algorithm can be understood as explanation of churn (no churn) in terms of client profile:

		cer.
1')	*if (Churn, no) then (In, high) and (Out, med.)*	0.33
2')	*if (Churn, no) then (In, high)*	0.17
3')	*if (Churn, no) then (In, low) and (Change, low)*	0.50
4')	*if (Churn, yes) then (Change, yes)*	0.38
5')	*if (Churn, yes) then (In, med.) and (Out, med.)*	0.55
6')	*if (Churn, yes) then (In, med.) and (Out, low)*	0.07

Observe that certainty factor for inverse decision rules are coverage factors for the original decision rules.

7. What the data are telling us

The above properties of decision tables (algorithms) give a simple method of drawing *conclusions* from the data and giving *explanation* of obtained results.

From the decision algorithm and the certainty factors we can draw the following conclusions.

- No churn is implied with *certainty* by:
  - high number of incoming calls,
  - low number of incoming calls and low number of outgoing calls to other mobile operator.
- Churn is implied with *certainty* by:
  - high number of outgoing calls to other mobile operator,
  - medium number of incoming calls and low number of outgoing calls.
- Clients with medium number of incoming calls and low number of outgoing calls within the same operator are *undecided* (no churn, cer. = 0.48; churn, cer. = 0.52).

From the inverse decision algorithm and the coverage factors we get the following explanations:

- the *most probable* reason for no churn is low general activity of a client,
- the *most probable* reason for churn is medium number of incoming calls and medium number of outgoing calls within the same operator.

8. Summary

In this paper the basic concepts of rough set theory and its application to drawing conclusions from data are discussed. For the sake of illustration an example of churn modeling in telecommunications is presented.

References

[1] J. Grant, "Churn modeling by rough set approach", manuscript, 2001.

[2] S. K. Pal and A. Skowron, Eds., *Rough Fuzzy Hybridization*. Springer, 1999.

[3] Z. Pawlak, *Rough Sets – Theoretical Aspects of Reasoning about Data.* Boston, London, Dordrecht: Kluwer, 1991.

[4] Z. Pawlak, "Decision rules, Bayes' rule and rough sets", in *New Direction in Rough Sets, Data Mining, and Granular-Soft Computing*, N. Zhong, A. Skowron, and S. Ohsuga, Eds. Springer, 1999, pp. 1–9.

[5] Z. Pawlak, "New look Bayes' theorem – the rough set outlook", in *Proc. Int. RSTGC-2001*, Matsue Shimane, Japan, May 2001, pp. 1–8; *Bull. Int. Rough Set Soc.*, vol. 5, no. 1/2, 2001.

[6] L. Polkowski and A. Skowron, Eds., *Rough Sets and Current Trends in Computing*. Lecture Notes in Artificial Intelligence 1424, Springer, 1998.

[7] L. Polkowski and A. Skowron, Eds., *Rough Sets in Knowledge Discovery*. Vol. 1–2, Springer, 1998.

[8] L. Polkowski, S. Tsumoto, and T. Y. Lin, Eds., *Rough Set Methods and Applications – New Developments in Knowledge Discovery in Information Systems*. Springer, 2000, to appear.

[9] N. Zhong, A. Skowron, and S. Ohsuga, Eds., *New Direction in Rough Sets, Data Mining, and Granular-Soft Computing*. Springer, 1999.

More info about rough sets can be found at:

http://www.roughsets.org
http://www.cs.uregina.ca/∼roughset
http://www.infj.ulst.ac.uk /staff/I.Duentsch
http://www-idss.cs.put.poznan.pl/staff/slowinski/
http://alfa/mimuw.edu.pl
http://www.idi.ntnu.no/∼aleks/rosetta/
http://www.infj.ulst.ac.uk/∼cccz23/grobian/grobian.html

Zdzisław Pawlak
Institute of Theoretical and Applied Informatics
Polish Academy of Sciences
Bałtycka st 5
44-000 Gliwice, Poland

WEALTHY – a wearable healthcare system: new frontier on e-textile

Rita Paradiso, Giannicola Loriga, Nicola Taccini, Angelo Gemignani, and Brunello Ghelarducci

http://doi.org/10.26636/jtit.2005.4.339

Abstract— A comfortable health monitoring system named WEALTHY is presented. The system is based on a wearable interface implemented by integrating fabric sensors, advanced signal processing techniques and modern telecommunication systems, on a textile platform. Conducting and piezoresistive materials in form of fibre and yarn are integrated in a garment and used as sensors, connectors and electrode elements. Simultaneous recording of vital signs allows extrapolation of more complex parameters and inter-signal elaboration that contribute to produce alert messages and patient table. The purpose of this publication is to evaluate the performance of the textile platform and the possibility of the simultaneous acquisition of several biomedical signals.

Keywords— fabric sensors, fabric electrodes, physiological signs.

1. Introduction

One of the emerging new tendencies for healthcare monitoring systems is rising from areas relatively far away from the traditionally involved technologies.

During the last decade we have assisted at a revolution in telecommunication domain, while during 80's the electronic devices scale has shifted from a micro to a nano dimension. Nowadays, a new generation of monitoring devices based on the growth of the knowledge derived from the past research experience and on the use of textile multi sensing interfaces is rising.

The systems have to combine the advances of telecommunication, microelectronics and material science to guarantee a continuously remote monitoring of multiple physiological functions, as well as comfort and wearability. The spotlight is shifting from external environment control to human oriented systems, where the subject-actor is constantly virtually linked and interactive.

This tendency is changing dramatically the common life style, as well as the needs of people. Citizens are becoming more and more used in telecommunicating and in managing information, and the idea of a surrounding virtual world is no more an alien concept. New tools are being developed to be used every where, during normal life, capable to help people to increase their health status awareness, to train them to act at a preventive level by modifying their life style, to give them the feeling of a reassuring link. The interaction between physician and patient is growing in quality and the contribution is coming from both sides.

New systems designed to be minimally invasive, based on flexible and smart technologies conformable to the human body are conceived to improve the autonomy and the quality of life of patients. They are also cost-effective in providing around-the-clock assistance, in helping physicians to monitor cardiac patients during rehabilitation phase, in decreasing hospitalization time.

The system can also assist professional workers subject to considerable physical and psychological stress and/or environmental and professional health risks.

The aim of the work presented is to set up a fully integrated garment system, able to acquire simultaneously, in a "natural" environment a set of physiological parameters. The system is designed to be minimally invasive, comfortable and wearable, to this aim conductive and piezoresistive materials in form of fibre and yarn are used to realize clothes where knitted fabric sensors and electrodes are distributed and connected to an electronic portable unit, the acquired signals can then be transmitted to a remote monitoring system.

The simultaneous recording of vital signs allows parameters extrapolation and inter-signal elaboration [1, 2] that contribute to produce alert messages and personalized tables of user's health.

2. The WEALTHY system

Strain fabric sensors based on piezoresistive yarns, and fabric electrodes realized with metal based yarns, enable the realization of wearable and wireless instrumented garments capable of recording physiological signals and to be used by the patient during everyday activity. Breathing pattern, electrocardiogram, electromyogram, activity pattern or behaviour, temperature, can be listed as physiological variables to be monitored through the proposed system. A miniaturized short-range wireless system can be integrated in the sensitive garment and used to transfer the signals to the WEALTHY box/PCs, PDA and mobile phones. An "intelligent" system for the alert functions, able to create an "intelligent environment" by delivering the appropriate information for the target professional is the complementary function to be implemented. The system is targeting the monitoring of patients suffering from heart diseases during and after their rehabilitation.

3. WEALTHY functions

The WEALTHY system has been developed as the integration of several functional modules. The main functions are shown in Fig. 1, namely: sensing, pre-processing, transmission, processing and data management.

The garment interface is connected with the portable WEALTHY device where the local processing as well as the communication with the network is performed. A knitted fabric platform containing insulated conductive tracks connected with sensors and electrodes has been implemented to make the cloth. Most signals are transmitted unprocessed to the monitoring system where they can be analyzed off-line. In order to reduce the needed data capacity of the wireless link to the central monitoring system, some sensor signals are processed by the portable patient unit (PPU) to extract essential parameters. Local pre-processing of signals has to be decided in a trade-off between the gain in term of wireless link occupancy and the increase of needed local processing power.

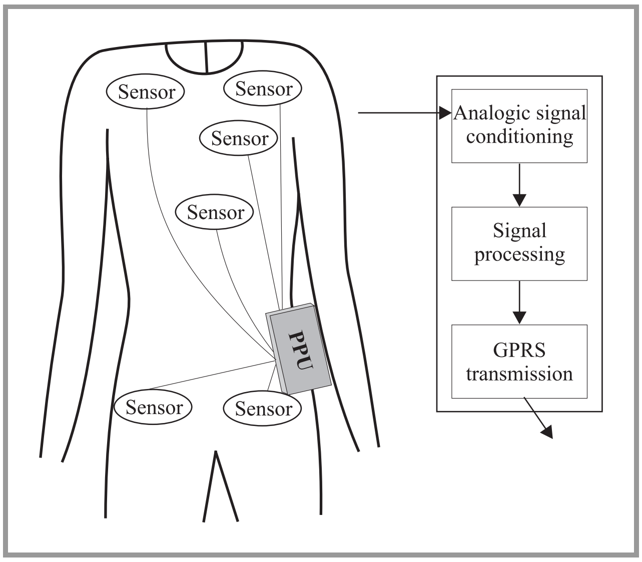


Fig. 1. Overall WEALTHY function.

The ECG leads that can be gathered are:

- precordial V2 and V5;
- einthoven D1, D2, D3.

ECG signals are sampled on the PPU at 250 Hz, a local processing is applied in order to extract parameters with a higher sampling rate, so that ECG parameters, such as heart rate (HR) value and QRS duration can be computed with a significant number of samples.

In order to decrease the amount of data transmitted by GPRS, the ECG signal is decimated to obtain a sampling rate of 100 Hz, and the operator at the monitoring centre can view and record only one ECG signal by selecting the desired one.

Respiration and movement activity come from piezoresistive sensors, sampled at 16 Hz. The signals from these sensors are transmitted without local processing.

The PPU is designed to have a simple user interface, a few LEDs and a buzzer for user warning purpose and a button to let him manually trigger an alarm. The PPU electronics is built on an "Europe" form factor board (first prototype dimension: 160 mm × 100 mm, first prototype weight: 400 g) and packaged in a metallic enclosure. It contains the necessary functions to condition physiological signals, such as filtering, digital analysis and to perform specific higher level processing like HR extraction, run the application, as well as communicate over GPRS with a monitoring centre. All the circuits, sensors and communication module are powered by a 1100 mAh/3.6 V lithium battery. The battery autonomy ranges between a few hours and eight hours, depending on the level of use of the GPRS link. It can be recharged using a dedicated front panel connector.

The WEALTHY central monitoring system is a s/w module interpreting physical sensor data received from the PPU and representing them in simple, graphical forms. It will be used by the proper staff in order to judge the automatically generated alerts and forward only the critical alerts to the doctors and the patients.

The central monitoring system performs the following tasks:

- coordinates and controls the data flow between the different actors;
- collects and stores the data transmitted by the sensors integrated in the WEALTHY garment through the portable patient unit;
- continuously monitors vital health parameters of the patients;
- generates alerts to inform doctors for critical health situations;
- gives access to the central database to doctors and other health professionals;
- presents to the qualified users the health situation of the patients using different user-friendly interfaces.

All the monitoring system modules are able to run on a single computer without the need of dedicated high-end servers.

The final aim is to recognize those parameters that define an event. Several statistical tools based on a multifunctional analysis, such as principal component analysis (PCA) or independent component analysis (ICA), may be used for this purpose. In order to offer full mobility to the patient or the user, the acquired signals are transmitted wirelessly from the PPU to the remote monitoring system. The communication is based on TCP/IP that is the standard protocol for GPRS communication. For GPRS bandwidth limitation reason, the monitoring centre shall select the ECG lead to be transmitted (one at the time). All signals are sent in quasi real-time to the remote monitoring centre.

Off-line processing, depending on the application, is carried out at the monitoring centre. A preliminary list includes:

- tachogram;
- ST deviation;
- T wave area;
- spectral analysis of RR signal.

Combining these parameters and the information obtained by the other signals (movement, respiration, HR, etc.) the system generates automatic alerts. A set of rules for the determination of the alert criteria has been implemented in the alert module. New alerts are also possible to be included by authorised personnel, as well as modification of the alert criteria [3].

The user will be able to watch the health status of all patients connected to the central monitoring system (through the WEALTHY garments). The definition of the monitoring profiles will provide an easy to use monitoring of the patients' health status in real time and with different fully customisable views.

Simultaneously, the user will be able to review the generated alerts and using past medical data will determine the true and false alerts and correspondingly contact doctors through direct phone calls and online alerts. This central control module is not necessary in order for the monitoring system to work. It is an optional module ensuring the minimal generation of false alerts to the doctors and will be necessary for large scale hospitals dealing with hundreds of patients.

The WEALTHY platform will give the possibility to monitor and assist patients through a remote medical advice service. The use of intelligent systems provides to physicians the data to timely detect and manage health risks, diagnose early illness or injury, recommend treatment that would prevent further deterioration and, finally, to make confident professional decisions based on objective information all in a reasonably short time.

4. WEALTHY interface

Strain fabric sensors based on piezoresistive fabric or yarns, and fabric electrodes made with metal based yarns, enable the realization of wearable and wireless instrumented garments capable of recording physiological signals, to be used during the routinely activity, to be worn instead of a classical garment without discomfort for the user. Respiration, electrocardiogram, electromiogram, activity sensors, temperature, may be listed among the physiological variables that can be monitored through the proposed system.

Piezoresistive fabric sensors have been realized by using lycra® fabric coated with carbon loaded rubber, as well as by weaving a commercial electroconductive yarn (PAC 250 dtx x 1, by Europa NCT, Poland). These fabrics behave as strain gauge sensors and show piezoresistive properties in response to an external mechanical stimulus. The coated lycra® fabric has been used to detect respiration signal, due to the higher efficiency shown in term of quality of the signal, compared with the other fabric sensor. The Europa yarn has been used for the activity sensors and knitted in the multifunctional fabric. The behaviour of a knitted piezorestive sensor is different when stretched towards warp or weft direction. Preliminary tests have been done to select the more efficient technique of knitting and the direction of stretching. The fabric sensor have been integrated and oriented in a way to maximize the gauge factor according with the response shown during the preliminary tests.

Electrodes have been realized with a yarn where two stainless steel wires are twisted around a viscose textile yarn (Elitè by Lineapiù s.p.a., Italy). Electrodes were knitted by using the tubular intarsia technique [4] to get a double face, using the external – non conductive – part to isolate the electrode from the external environment. The basal yarn (not sensitive) was the same yarn used as core for the conductive electrode yarn. To improve the electrical signal quality in dynamic condition a hydro-gel membrane purchased by ST&D Ltd (Belfast-UK), has been used. The use of the membrane affects also the comfort as electrodes have a rough surface and a prolonged contact with the body can give rise to skin irritations. The contact between conductive fabric and skin can be improved by increasing the adherence of the garment with the use of a higher percentage of elastic component in the yarns. Another approach is the use of conducting rubber or silicon as coating layer for the electrodes; in our future work both the approaches will be investigated.

Connections have been realized by means of the tubular intarsia technique. A supplementary layer has been woven by using of vanise technique. The final connection is a multi layered structure where the conductive surface is sandwiched between two insulated standard textile surfaces. The same conductive yarn is used for the electrodes as well as for the realization of connections, a particular of the textile prototype is shown in Fig. 2.

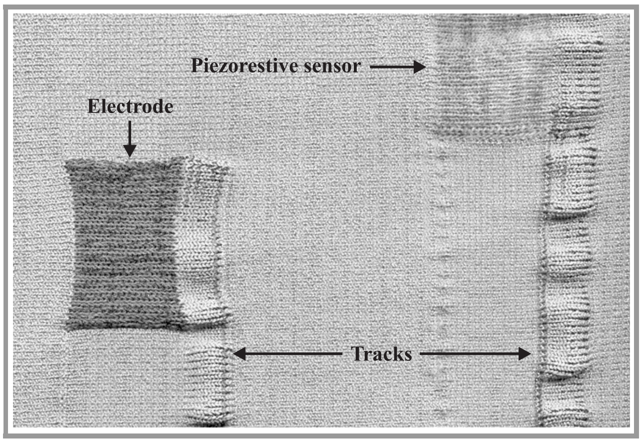


Fig. 2. Part of the WEALTHY interface.

The knitting fabric has been made with a flat-knitting machine (Vesta Vx 12 – Steiger, Switzerland). A draft position of sensors was implemented on the knitted fabric, and then by means of the use of models was possible to cut the fabric in a way to get the sensors in the desired configuration. The garment was finally sewed, which means that the final position of sensors and connections was achieved in the manufacturing phase.

The prototype model [5] is shown in Fig. 3 where the electrodes position is highlighted. In Fig. 3 the Einthoven

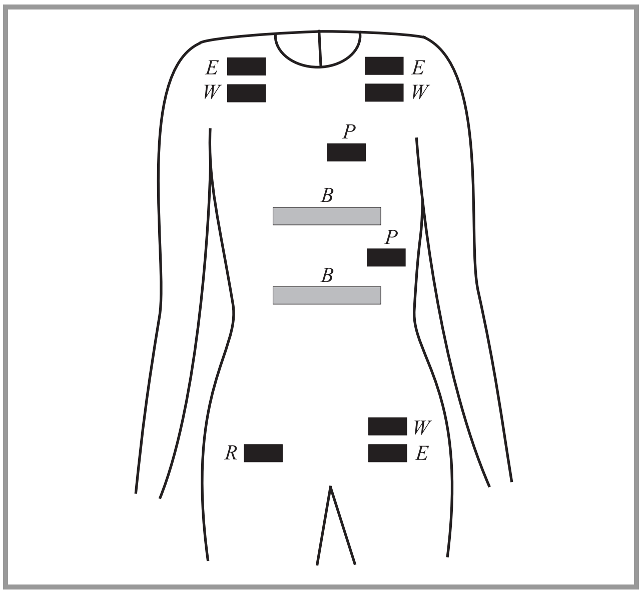


Fig. 3. Prototype model: *E* – Einthoven, *W* – Wilson, *R* – reference, *P* – precordial leads, *B* – breathing sensors.

and Wilson derivations (*E, W*), V2 and V5 as precordial leads (*P*) and the reference electrodes (*R*) are shown, while two breathing sensors (*B*) are positioned one on the thorax and the other on the abdomen. In Fig. 4 the position of the 6 movement sensors is shown.

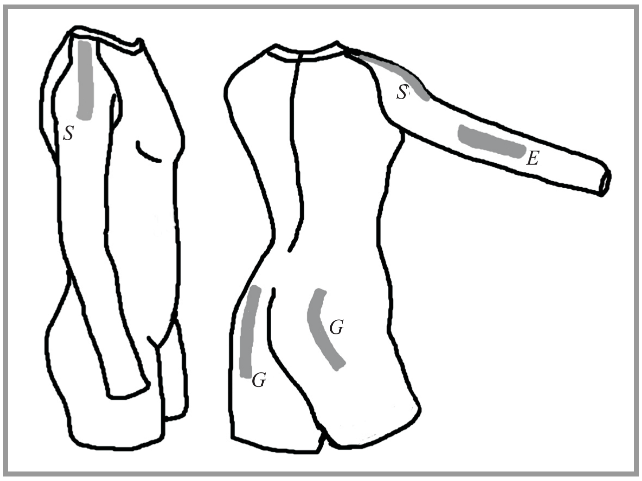


Fig. 4. Prototype model: *S* – shoulder movement, *E* – elbow movement, *G* – gluteus movement sensors.

The stainless steel threads have been selected for the realization of fabric electrodes for a series of reasons: first of all they are compatible with industrial textile processes, they are inert and stable in the presence of O_2, finally the cost of steel is very competitive compared with pure silver, or pure gold.

Naturally the fineness and flexibility of metal components have been chosen to get a final conductive yarn suitable for knitting, weaving and more in general for textile processing, which means that the metal threads used are washable, flexible and biocompatible. The same approach has been used for all the sensorial yarns and fabric developed in the project. It is also possible to work with silver coated threads that are occasionally employed for special fashion effects or for antibacterial purposes in textile world. Preliminary tests done with fabric containing polyester yarns coated with silver have shown that the use of stainless steel threads is more convenient: in fact during the experiments it has been observed that the conductivity of the silver electrode was lower than the stainless steel ones, when samples with the same dimension were compared. This is probably due to the small amount of metal components localized only in the coating layer of the threads. It is important to underline that the fabric cannot be realized only with metal yarns otherwise this region of the garment will be too rigid and not conformable, the amount of metal in the fabric is a compromise between the demand to increase the conductivity and the necessity to improve the touch sensation (the hand) of the cloth. Moreover the quality of silver adhesion was very poor, after several tests large metal coating regions looked removed; the electrodes need to be used with gel or conductive past and finally the electrodes have to be chlorinated.

Conductive and piezoresistive yarns are resistant to repeated washing in aqueous solutions, the physiological signals detected after washing have shown that the performances of the fabric sensors are not affected by the process.

5. Methods

The purpose of this publication is to evaluate the performance of textile sensors, electrodes and connections integrated in a garment (sensing part of the WEALTHY system), and to prove the possibility of the simultaneous acquisition of several biomedical signals during training session.

All the tests have been effected using the WEALTHY textile interface, adding two electrodes, not integrated but sewn, on the right leg, in order to monitor the EMG activity of the quadriceps muscle.

Piezoresistive signals have been conditioned by a voltage divider, followed by a Butterworth low pass filter (cut frequency at 10 Hz).

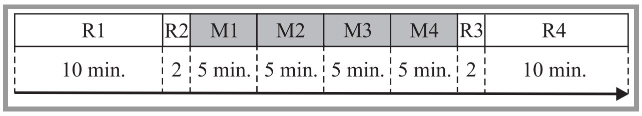
Signals from fabric electrodes have been conditioned by a GRASS-TELEFACTOR mod. 15LT device equipped with differential amplifiers mod. 15A54, with settable gain and band pass filter, notch filter at 50 Hz.

The ECG signals from fabric electrodes were conditioned setting gain 1000 and band pass filter with frequency range between 1 and 100 Hz. Surface EMG signals from fabric electrodes positioned on the right leg (quadriceps) were conditioned setting gain 2000 and band pass filter with frequency range between 10 and 500 Hz.

Every analogic signal has been acquired by an acquisition card (National Instruments PCI 6036) with sampling rate of 1000 Hz.

The experiments have been performed according the following experimental paradigm.

The baseline conditions were recorded when the subject was lying in supine position (R1) for a period of 10 minutes, followed by a control period of 2 minutes with the subject sitting on a cyclette (R2). This was followed by a period of progressively increasing physical exercise (cycling with increasing frequency and force) M1, M2, M3, M4, 5 minutes each. Then the period (R3), still in vertical position on the cyclette, for 2 minutes, as in R2.



R1	R2	M1	M2	M3	M4	R3	R4
10 min.	2	5 min.	5 min.	5 min.	5 min.	2	10 min.

Fig. 5. Experimental design.

Finally, the subject was asked to stand up and to lie in supine position for other 10 minutes (R4) as in R1. The experimental protocol is summarized in Fig. 5.

6. Results

The ECG leads used to evaluate the performances of the WEALTHY textile interface are:

- precordial V5;
- precordial V2;
- Einthoven D2.

These signals are acquired simultaneously with the respiratory activity (abdominal and thoracic) and the activity of the right quadriceps.

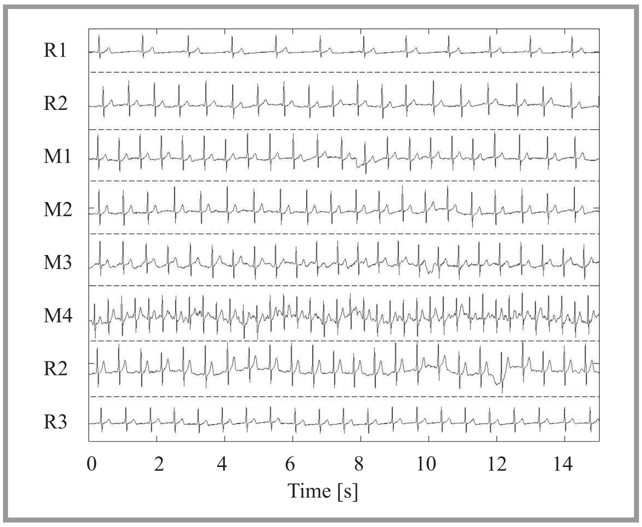


Fig. 6. V5 precordial lead signal in each experimental condition.

The results obtained in the whole sessions have been analyzed in order to demonstrate the robustness of the system.

In Fig. 6 the recordings from the V5 leads are shown, acquired according to the protocol previously described, for a period of 15 s each.

The analysis of the precordial leads shows that the quality of the signals is not affected by movement artefacts, in the frame of this trial.

In Fig. 7 it is possible to notice that the response of D2 lead is still satisfying for regular movement (M3). In fact only during M4 (very intense activity) the signal is noisy and could be very hard to get useful parameters (heart rate).

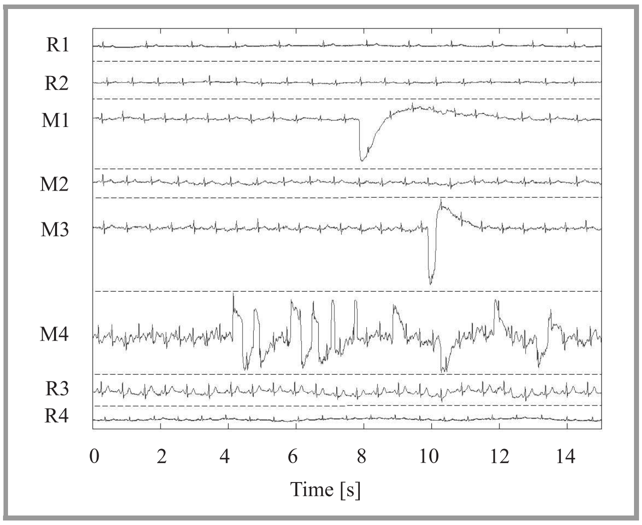


Fig. 7. D2 Einthoven lead signal in each experimental condition.

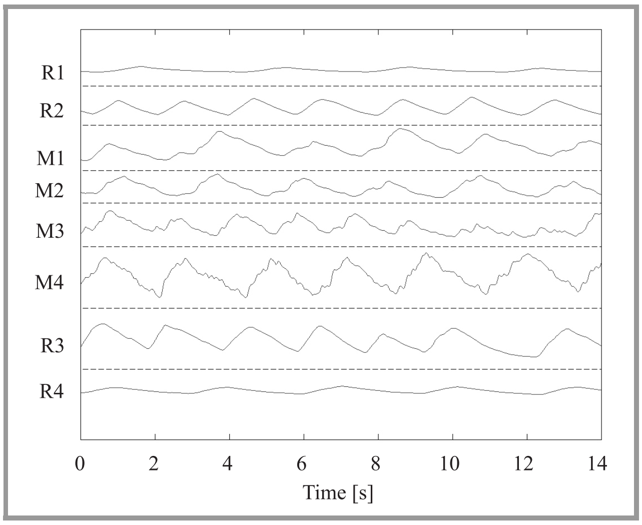


Fig. 8. Respiratory activity, reading plethysmography in thoracic position, in each experimental condition.

The signal obtained by the piezoresistive sensor placed on the thoracic position is shown in Fig. 8. It is affected by noise during the M-phase, but is still possible to obtain the respiratory rate and to have information about the plethysmography of thorax with an appropriate algorithm of analysis.

In Fig. 9 it is possible to notice the increasing of muscular activity by analyzing the results of surface EMG signal.

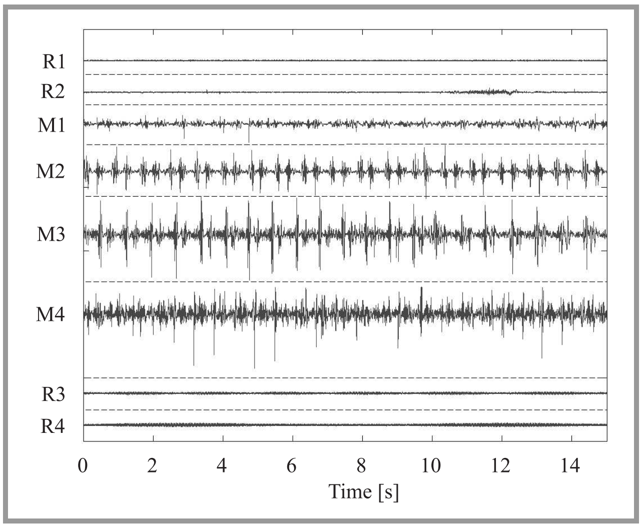


Fig. 9. Surface EMG signal from right quadriceps in each experimental condition.

The amplitude of signal and median frequency, defined as the frequency below which lies 50% of the total power of the PSD, increases with the speed of spinning movements and the effort required by the increasing resistance set on the cyclette, as shown in Table 1.

Table 1
Median frequency of PSD of the surface EMG during movements

Experimental condition	M1	M2	M3	M4
Median frequency [Hz]	17.58	24.41	25.39	28.32

7. Discussion

The achieved results show that fabric electrodes endowed in the sensing shirt allow a continuous and simultaneous monitoring of bioelectrical and biomechanical physiological signals in a behaving subject. In a previous work [6] has been shown that the signals recorded by fabric electrodes are comparable to those acquired with gold-standard electrodes commonly employed in research and clinical use. The electrical and mechanical properties of fabric electrodes have not been modified by their integration in the wearable shirt, as the characteristics of electrocardiographic (ECG), electromyographic (EMG) and respiratory (RESP*) signals are comparable to those obtained with standard electrodes in similar conditions. The response of the system during the different activity phases are clearly observable in Fig. 6, where the EGG data indicate that precordial leads exhibit a remarkable stability and are free from artefacts even during the maximal exercise intensity (M4), when also the background noise appears negligible. In the standard D2 derivation the signal is less stable and the amount of artefacts related to movement clearly increases (Fig. 7). This may be related to the lack of adherence of the garment to the upper chest when the subject had to grab the cyclette handle bar during exercise. Moreover, the strong engagement of the pectoral muscles in this type of exercise may be responsible for the higher background noise observed. The signal to noise ratio can be improved by trying to set the fabric electrode position on a rigid surface such as the clavicle and the sensing shirt will be modified accordingly in the future. The good quality of the ECG signal allows the computation of heart rate and its variability throughout the experimental cycle. As described in the literature during physical exercise there is a progressive increase of heart rate (Fig. 10), correlated to a parallel decrease of heart rate variability (Fig. 11).

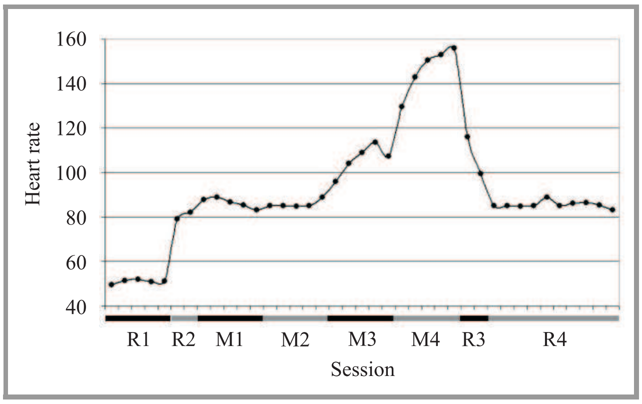


Fig. 10. Heart rate obtained analyzing V5 signal in each experimental condition.

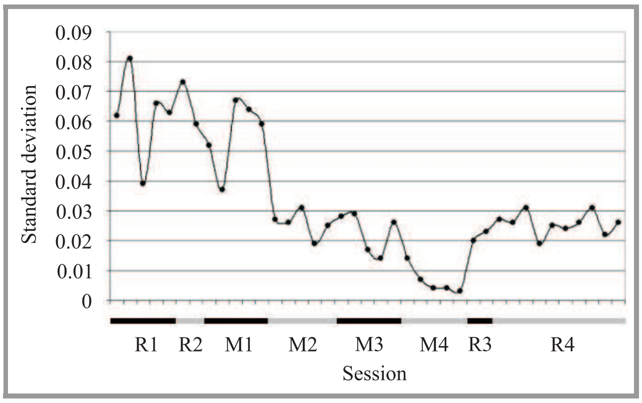


Fig. 11. Heart rate variability.

Due to the good quality of recorded signals, the ECG can be adequately employed to study non invasively and in behaving conditions more complex functional indexes related to the sympatho-vagal balance, such as low frequency and high frequency components derived by spectral analysis of RR interval variability [1], respiratory sinus arrhythmia and area under T wave of the ECG.

In Fig. 8, respiratory signals are shown detected through piezoresistive sensors. Also in this case is evident a remarkable stability and an excellent signal to noise ratio during the experimental session. Moreover the signal time course is adequate to reproduce the thoracic excursions without detectable phase shifts. Thus the respirogram yields accurate information about respiratory rate while the variations of signal amplitude can give only a qualitative estimation of the respiratory depth.

The surface activity of selected lower limb muscles such as the quadriceps can easily be recorded by fabric electrodes similar to those used for recording ECG. The EMG shown in Fig. 9 exhibits bursts of activity synchronous with the pedalling cycle which rise by increasing the frequency and the force required by the exercise.

As shown in Fig. 12, the sensing shirt makes possible a simultaneous and multi-parametric acquisition of several physiological variables in different behavioural conditions. This possibility represents a significant advantage when it is necessary to monitor the vital asset of workers in extreme environmental conditions as well as sportsmen during high physical performance or military personnel engaged in war sites.

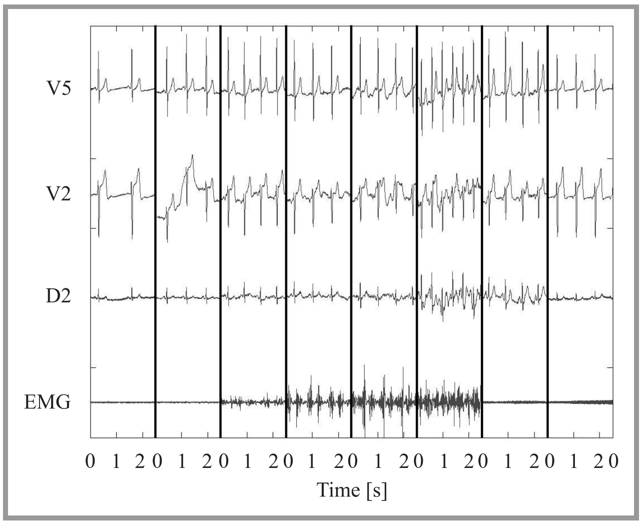


Fig. 12. Overview of changing in all electrical signal during experiment.

The most innovative feature of this system consists of the use of functionalized materials in form of fibres and yarns, which can be knitted or woven into a sensing fabric. Preliminary results [7] show that the basic sensing features on which vital sign recording is based can be implemented using integrated knitted sensors and electrodes. Previous authors works [8, 9] have shown that low frequency mechanical signals of cardiopulmonary origin (respiratory signals, ballistogram) or generated by body segments relative motion (kinaesthesia) could be recorded by textile strain gauges. Finally bioelectric potentials related to cardiac or skeletal muscle activity (ECG, EMC) have been faithfully recorded by metal based fabric electrodes. The integration of these different components with appropriate elastic electrical conductors and properly designed connectors to the wearable electronic unit, leads to a comfortable wearable cloth which has no counterpart in any existing monitoring system. These new integrated knitted systems enable applications extending even beyond the clinical area and open new possible applications in sport, ergonomics and monitoring operators exposed to harsh or risky conditions (fire fighters, soldiers, etc.). The possibility of simultaneously recording different physiological signals provides an integrated view of normal and abnormal pattern of activity which could be otherwise impossible to be detected by recording each signal in different time. Finally it must be outlined that the possibility of recording physiological variables in a more “natural” environment may help to identify the influence of the psycho-emotional state of the subject in the performance of a physical activity. This is not easily detectable when recording is done within a protected (medical) environment. A further innovation is the in-context data interpretation. While a simple telemonitoring system would just transmit or record real-time physiological signs, the WEALTHY system will be able to process physiological parameters in context, so that appropriate feedback can be given to the patient.

8. Conclusions

The innovative approach of this work is based on the use of standard textile industrial processes to realize the sensing elements. Transduction functions are implemented in the same knitted system, where movements and vital signs are converted into readable signals, which can be acquired and tele-transmitted. In our fabric sensors, electrodes and bus structure are all integrated in textile material, making possible to perform normal daily activity while the clinical status is monitored by a specialist, with a comfortable wearable cloth which has no counterpart in any existing monitoring system [10, 11]. WEALTHY system benefits from the performance of the textile sensing interface to guarantee a continuously remote monitoring of user vital signs, the signals are acquired and elaborated on body and a set of signals and parameters are teletransmitted and managed by a remote control system. The philosophy of this approach is focused on the realization of a friendly, human oriented textile based system, where the choose of sensing material is a compromise between comfort for the users and signal quality for the specialists.

Acknowledgments

The authors would like to thank the European Commission for funding (IST-2001-37778) and all partners involved in the project, University of Pisa, Italy; Centre Suisse d’Electronique et de Microtechnique, Switzerland; Atkosoft S.A., Greece; Institut National des Sciences Appliquees de Lyon, France; Istituto Scientifico H San Raffaele, Italy; Centre de Recherces du Service de Sante des Armees, France; Messe Frankfurt GmbH, Germany; Centre National de la Recherche Scientifique, France.

A special acknowledgement to Dr. Roberto Orselli for his useful contribution to the drawing up of the manuscript.

References

[1] Task Force of the European Society of Cardiology and the North America Society of Pacing and Electrophysiology, "Heart rate variability standards of measurement, physiological interpretation and clinical us", *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[2] R. Paradiso, A. Gemignani, E. P. Scilingo, and D. De Rossi, "Knitted bioclothes for cardiopulmonary monitoring", in *25th Ann. Int. Conf., IEEE-EMB, Engineering in Medicine and Biology Society*, Cancun, Mexico, 2003, pp. 3720–3723.

[3] R. Paradiso, J. Luprano, T. Vaovuras, and S. Coli, private communication, Jan. 2004.

[4] M. Bona, F. A. Isnardi, and S. L. Straneo, *Manuale di tecnologia Tessile*. Edizioni Scientifiche A. Cremonese-Roma, 1981.

[5] R. Paradiso, "Tessuto in maglia per il monitoraggio di segnali vitali", Italian Patent N. FI2003A000308, 2003.

[6] E. P. Scilingo, A. Gemignani, R. Paradiso, N. Taccini, B. Ghelarducci, and D. De Rossi, "Sensing fabrics for monitoring physiological and biomechanical variables", *IEEE Trans. Inform. Technol. Biomed.*, vol. 9, no. 3, pp. 337–344, 2005.

[7] M. Pacelli, R. Paradiso, G. Anerdi, S. Ceccarini, M. Ghignoli, F. Lorussi, P. Scilingo, D. De Rossi, A. Gemignani, and B. Ghelarducci, "Sensing threads and fabrics for monitoring body kinematic and vital signs", in *Conf. Fibr. Text. Fut.*, Tampere, Finland, 2001.

[8] D. De Rossi, A. Mazzoldi, F. Lorussi, and R. Paradiso, "From sensitive fabrics to distributed wearable sensors", in *Proc. SPIE's 8th Ann. Int. Symp. Smart Struct. Mater.*, Newport Beach, USA, 2001.

[9] D. De Rossi, F. Lorussi, A. Mazzoldi, P. Orsini, and E. P. Scilingo, "Monitoring body kinematics and gesture through sensing fabrics", in *1st Ann. Int. IEEE-EMBS Special Topic Conf. Microtechnol. Med. Biol.*, Lyon, France, 2000.

[10] M. A. Sackner and D. M. Inmann, "Systems and methods for ambulatory monitoring of physiological signs", Patent Application Publication US 2002/0032386.

[11] P. Sungmee and J. Sundaresan, "Full-fashioned garment in fabric having intelligence capability", International Publication number WO 02/100200 A2.

Rita Paradiso graduated in physics from the University of Genoa and received her Ph.D. in bioengineering in 1991. Molecular electronics, biosensors, biomaterials for biomedical applications have been her main research topics. In particular she worked on bio-functionalized surfaces and their characterization. She worked in London during the Ph.D. at the Physics Department of Queen Mary College. In 1993 she got a Post Doctor CE fellowship, in the frame of Human Capital and Mobility project at the Molecular Chemical Laboratory – CNE Saclay, France. In 1994 she was Post Doctor fellow from the Genoa University, at the Department of Material Engineering of the University of Trento. During 1998, she worked at the "IRST – Instituto Trentino di Cultura" on a project related to bio-activation of MEMS, FIBIA. From 1998–1999 she was Research Manager of Technobiochip s.r.l. – Marciana (LI) – Italy, working on two BRITE-EURAM II projects: BE97-4511 (PRO.MO.FILM) and BE97-5141 (BIO.M.I.ST.) and on a National Project PNR, Tema 10: Biosensors for the Environmental Control. She has over 20 scientific publications and conference presentation since 1989. She joined Smartex in 2000 as R&D Manager, and from Sept. 2001 is the coordinator of WEALTHY (IST-2001-37778), from January 2004, is working in MYHEART an Integrated Project (IST-2002-507816).

e-mail: rita@smartex.it
Milior s.p.a.
via Pistoiese 755/D
59100 Prato (PO), Italy

Giannicola Loriga graduated in electronic engineering in 2002, with a specialization in bioengineering with a dissertation on development and implementation of wearable systems for vital signals monitoring. In 2003 he got a fellowship for the WEALTHY project at the Interdepartmental Research Centre "E. Piaggio" of the University of Pisa. His research activity has been focused on the study of sensors for monitoring respiratory activity. He developed and patented a portable system for bioelectrical impedance measures. He joined Smartex in Apr., 2004, where he cares about software and hardware development and signal analysis.

e-mail: loriga@smartex.it
Smartex s.r.l.
via Giuntini 13 int. L
56023 Navacchio (PI), Italy

Nicola Taccini graduated in electronic engineering in 2002, with a specialization in bioengineering with a dissertation on development and implementation of systems for cinematic body variables detection. He joined Smartex in July, 2002, where he cares about software and hardware development and signal analysis.

e-mail: taccini@smartex.it
Smartex s.r.l.
via Giuntini 13 int. L
56023 Navacchio (PI), Italy

Angelo Gemignani graduated in medicine at the University of Pisa in 1991. He received the Ph.D. in basic neuroscience in 2000. He was a resident in psychiatry from 1991 to 1996 and since 2002, researcher at Department of Physiology and Biochemistry, Faculty of Medicine at University of Pisa, Italy. Since 1988 he has been involved in different scientific domains from sleep disorders to cognitive and emotional modulation of brain and autonomic activity. He spent three years (1995–1998) as a research fellow at the Laboratory of Neurobiophysics, "Unitè de Recherche en RMN Bioclinique – INSERM U438", Grenoble, where he worked on the post-processing of fMRI signal during actual movement and motor imagery. His main methodologies of research are neuropsychological tests, wake and sleep EEG, autonomic variables and fMRI. He has a good experience in the postprocessing of autonomic signal as well as of EEG and fMRI signal. His current research interests include cognitive neuroscience, experimental psychology and neurophysiology. Use of behavioural (EEG, EOG and EMG) and autonomic (ECG, respiration and skin resistance) parameters and functional magnetic resonance imaging (fMRI) in combination with cognitive modulation (i.e., hypnosis) to study brain functions and brain response to motor and cognitive/emotional activation in healthy subjects.
e-mail: gemignani@dfb.unipi.it
Università di Pisa
Dipartimento di Fisiologia e Biochimica
Via S.Zeno 31
56127, Pisa, Italy

Brunello Ghelarducci was born in 1942 and has obtained his medical degree at the University of Pisa in 1967. He specialized in neurology in 1970. In 1971 he begun his career in the Institute of Physiology in Pisa, directed by Prof. G. Moruzzi. Since 1986 he is full Professor of human physiology in the Medical Faculty of Pisa. His research activity has always been performed in the field of vestibulo-spinal sensorimotor integration and of cerebellar motor coordination. In 1973–1974 he has worked with Prof. Masao Ito in Tokyo University on cerebellar plasticity in the control of the vestibulo-oculomotor reflex. He has continued this line of research in Pisa studying the development and the characteristics of cerebellar control on the vestibulo-oculomotor reflex in the rabbit and in newborn humans. Since 1983, in collaboration with Prof. K. M. Spyer of the University College of London, he has begun a series of investigations on the role of the cerebellum in the control of autonomic reflex activity in the rabbit, in particular of the autonomic and behavioral responses to alerting stimuli. These investigations, funded by the Ministry of Scientific Research, by the National Research Council and by the European Training Programme have revealed the importance of the posterior cerebellar vermis in the control of the development of visceral responses to aversive stimulations in the newborn and in the coordination of complex autonomic and behavioral responses to the same stimuli in the adult. More recently he has been engaged in a research program aimed to study the autonomic and electroencephalographic responses evoked in humans by an aversive stimulation. In these studies a fear-like state has been induced by means of hypnosis. Professor Ghelarducci has published several papers in international journals and belongs to several scientific organizations such as IBRO, the Physiological Society of London, the European Neuroscience Association and the Japanese Physiological Society.
e-mail: ghelarducci@dfb.unipi.it
Università di Pisa
Dipartimento di Fisiologia e Biochimica
Via S.Zeno 31
56127, Pisa, Italy

Theoretical and practical aspects of military wireless sensor networks

Michael Winkler, Klaus-Dieter Tuchs, Kester Hughes, and Graeme Barclay

http://doi.org/10.26636/jtit.2008.2.870

Abstract—Wireless sensor networks can be used by the military for a number of purposes such as monitoring militant activity in remote areas and force protection. Being equipped with appropriate sensors these networks can enable detection of enemy movement, identification of enemy force and analysis of their movement and progress. The focus of this article is on the military requirements for flexible wireless sensor networks. Based on the main networking characteristics and military use-cases, insight into specific military requirements is given in order to facilitate the reader's understanding of the operation of these networks in the near to medium term (within the next three to eight years). The article structures the evolution of military sensor networking devices by identifying three generations of sensors along with their capabilities. Existing developer solutions are presented and an overview of some existing tailored products for the military environment is given. The article concludes with an analysis of outstanding engineering and scientific challenges in order to achieve fully flexible, security proved, ad hoc, self-organizing and scalable military sensor networks.

Keywords— wireless sensor networks, military sensor applications, joint intelligence surveillance reconnaissance (JISR), military sensors, energy efficient routing, WSN generations.

1. Introduction

There have been large amounts of research undertaken during the past decade in the areas of ad hoc networking and wireless sensor networks (WSNs) and significant progress has been achieved. Possible civilian use-cases for such networks include industrial plant monitoring and environmental monitoring. However, one area commonly cited as a primary use of sensor networks is for military benefit. Frequently, assumptions are stated regarding the requirements for military networks to motivate the work. The aim of this paper is to explore the military requirements of wireless sensor networks in the near to medium term (three to eight years) and to identify areas of research which would improve military usability.

2. The main characteristics of a sensor network

Wireless ad hoc sensor networks generally consist of a variable number of stationary sensors (also known as nodes) spread across a geographical area. The capabilities of these nodes typically comprise monitoring the environment and capturing specific information; the transmission of collected (and possibly preprocessed) data; as well as the forwarding of data obtained from neighbor nodes using wireless bearers[1]. A typical network structure is shown in Fig. 1.

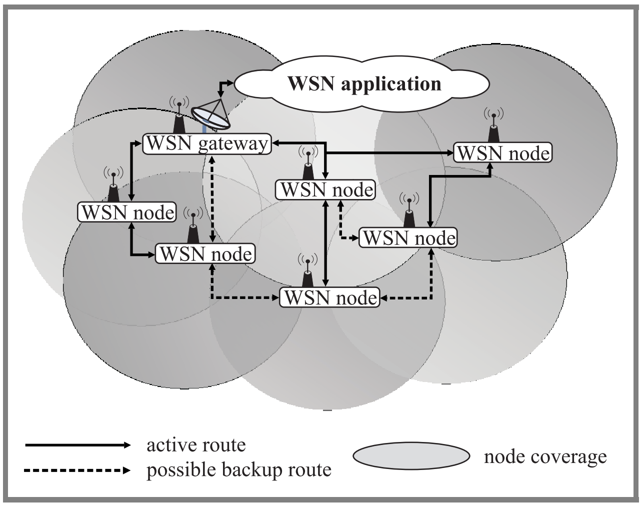


Fig. 1. Network set-up of a typical wireless ad hoc sensor network.

The information flow in a wireless sensor network will in general be from the sensor nodes to one or more wireless sensor network gateways. The network gateways can serve as data fusion points and provide reach-back capability. The reach-back capability can be based on different approaches such as:

- near real time connection, e.g., via longer range wireless transmissions (high frequency) or via a satellite link;
- asynchronous data transfer to passing unmanned aerial vehicles (UAVs).

Data processing can generally occur in three areas of the sensor network as shown in Fig. 2.

Processing can be carried out on the sensor node itself (such as the removal of unwanted signals from a target signal). Processing at the node reduces the amount of data to be passed over the network. This ensures that data loadings can be kept within the capacity capabilities of the radio system. In general, power consumption for the transmission of data is greater than the power consumption required to

[1] It is worth noting that by appropriately equipping sensor nodes with active capabilities, the network can operate actively as well as passively. The Defense Advanced Research Projects Agency (DARPA) Wolfpack concept of small, low-cost distributed jammers exemplifies an active network [1].

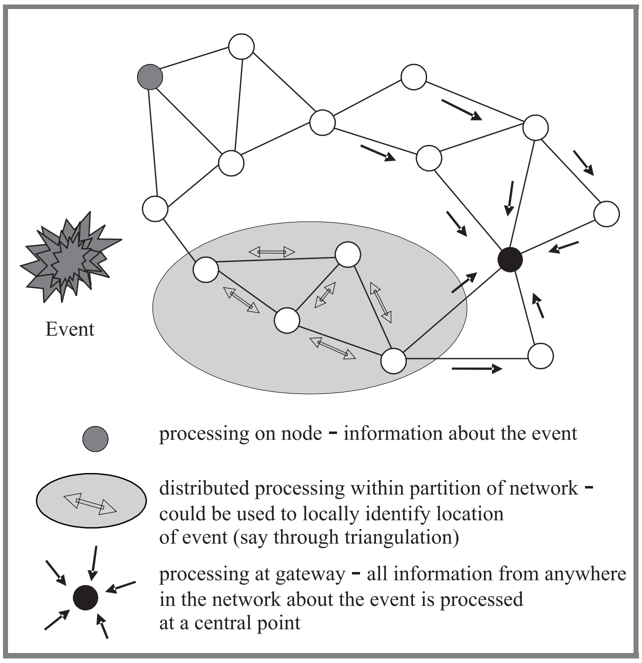

Event
processing on node – information about the event
distributed processing within partition of network – could be used to locally identify location of event (say through triangulation)
processing at gateway – all information from anywhere in the network about the event is processed at a central point

Fig. 2. Processing within a sensor network.

perform the same amount of processing data, thus there are power efficiency benefits in processing the data at source. Data processing can also be used to alleviate the amount of processing to be carried out at any gateways in the system. However, some data processing depends on data coming from multiple sources and therefore processing at source is not always possible.

Data processing can also be distributed within the network. This can be especially useful in large networks as it not only alleviates the amount of processing at the gateway but dramatically reduces the data loading which sensor nodes have to relay across the network. Hierarchical topologies lend themselves easily to perform distributed processing at "head" or "cluster" nodes (i.e., those nodes which logically "manage" other nodes in the hierarchy). However, there is an overhead associated with distributed processing. Either extra routing overhead is required to be able to pass data to be processed to specified nodes, or flooding techniques must be employed. Flooding techniques will forward user data to all or a limited subgroup of nodes thus negating the need for routing overhead traffic. These techniques allow unprocessed data to be exchanged between nodes adjacent to an "event" so that they can each do the processing required to locate the event. Then only the processed information is passed back to the gateway.

Finally, data can be processed at the gateway node(s). This allows the gateway to minimize the data it will send over the reach-back channel. Processing at the gateway thus will enable less power to be consumed in reach-back transmissions thus increasing the gateway's longevity and subsequently the lifetime of the whole network (as the gateway node is frequently the first node to fail due to depletion of its power source).

3. Military requirements

One of the main drivers for investigating wireless sensor networks is their use in military applications. The military use-cases for wireless sensor networks are diverse. They encompass applications such as:

- monitoring militant activity in remote areas of specific interest (e.g., key roads, villages);
- force protection (e.g., ensuring that buildings which have been cleared remain clear from infiltration by an adversary).

One prominent use-case which has received a great deal of interest from military personnel recently is base protection (or force protection in general). A possible set-up is depicted in Fig. 3.

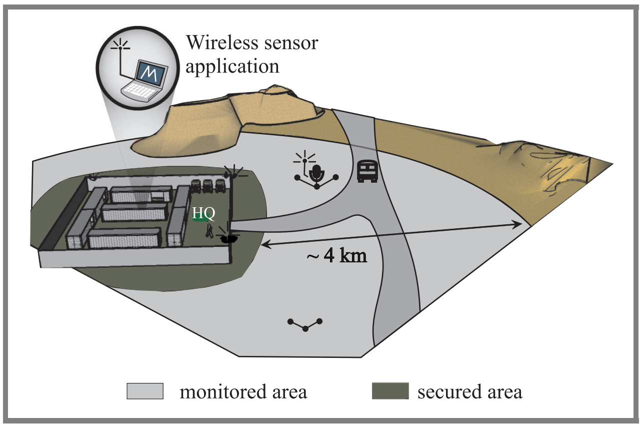


Fig. 3. Wireless sensors in support of base protection (e.g., making use of acoustic as well as electro-optical sensors).

Having deployed a headquarters in an area of active engagement it is essential to prevent the base from being attacked. The surrounding terrain may be undulating or mountainous and potentially could be obscured in trees and vegetation. Attack could come in the form of militant groups on foot or with motor vehicles.

In order to facilitate an early detection, the perimeter protection in Fig. 3 would cover a belt around the camp of up to 4 km, while in practice ranges of up to 10 km might be a requirement. Detection may be needed throughout the whole of this range whilst identification may only be required within a belt of around one to 1–2 km around the base.

3.1. Typical assumptions in the research community

Military applications are a primary use of wireless sensor networking and are best served by informed research that avoids making assumptions that are based on presumed military requirements. Many research papers propose algorithms for network sizes of thousands of sensor nodes and above. It is assumed that sensor nodes will be extremely small, lightweight and cheap. These are combined with the need for long battery lifetime. These assumptions have led to the following requirements:

- tailored routing and transport protocols are needed;
- short distances between nodes (often just a few metres) are taken for granted;
- special-purpose operating systems are required.

In practice these assumptions are more challenging than required in the near-term for current military needs while other aspects such as tamper-resistance are not sufficiently addressed. The following section gives an insight into current requirements for sensor networks in the military environment.

3.2. Realistic assumptions for military usage

In order to facilitate a meaningful operation of wireless sensor networks for military purposes in the near to medium term, there are a number of requirements which the military expect to be met.

Physical attributes of sensors. It is likely that the sensor nodes themselves could be hand deployed in advance of an operation. They could be transported to the area of deployment by vehicle. Thus the physical size and weight of the sensor need not be a major constraint. Sensor nodes the size of a matchbox, although desirable, are not currently expected and a sensor node (without including antenna) of order 20–30 cm in height would be acceptable. In occasional instances sensor nodes may be air dropped or deployed through a rocket launcher and would need to be suitably ruggedized.

Self-configuration after deployment. Sensor nodes must be able to rapidly identify neighbours within communications range and configure themselves into an ad hoc network. The network is likely to remain reasonably static as sensor nodes are unlikely to be moved during operation. The network should be able to cope with a node failing and reconfiguration of the network should occur without manual intervention.

Network size. For the majority of operations the area to be covered by the network may be between 5–20 km^2. Generally a communications range between nodes of around 250–500 m would be acceptable. This would amount to networks with less than 100 nodes being required. In occasional cases communication ranges of greater than 1 km would be desirable.

Information flows. Initially one-way communications can be seen as sufficient, i.e., from the sensor network to the WSN gateway and beyond. This is sufficient to achieve improved situational awareness for the warfighter as well as for the commander. In the medium term some degree of control within the network will be beneficial, e.g., the ability to orient cameras. This would however necessitate the need for communications in both directions. This need for two way communications should be reflected in the network security concept in order to avoid information leakage between a stub sensor network and the core military network to which it is attached.

Duration of usage. Some networks are only required to operate for periods of days, although generally periods of one to two months can be seen as a reasonable for military sensor networks. In the base protection example (Fig. 3) an exchange of batteries is practical and could extend the lifetime further. In some instances the network may not require to be functional throughout the whole day (perhaps only needed at night) or transmission of data from the WSN gateways may only be needed two or three times a day.

Physically and electronically inconspicuous operation. It would be beneficial if the nodes were covert in appearance with a small electromagnetic emission pattern so as to remain hidden from potential adversaries.

Data type. Even limited amounts of text ($<$ 30 bytes) can help to ensure information superiority by identifying an incident and providing location reports. This means data transmission rates do not need to be high. However, military commanders are likely to request imagery and video (both real-time and non-real time) in the future.

Data reliability. In many cases it is vital to ensure that data has been received by the end-user successfully, and techniques to guarantee delivery should be included. Also data should be received in a secure manner without the opportunity for interception and tampering by any eavesdropper.

Denial of service. Any network should be able to react against a denial of service attack by an adversary, at least by providing the means to report the incident of an attack such as jamming.

Tamper-proof. The data held on the node along with any crypto material must not be available to any third party even if the node itself is captured. The sensor nodes should have anti-tamper mechanisms in-built to address this.

Costs. As relevant information can be gained by the use of networks with just a few tens of sensors, and the retrieval of sensors after use might be desirable (e.g., for security reasons), the price for a single node is generally not as critical as in the "civil Bluetooth-focussed market".

4. Current technologies

4.1. Generations of sensor products

In a similar fashion to the evolution of mobile cellular technologies, it is possible to describe the evolution of military sensor devices in terms of generations.

First generation sensor networks (1GSN). Sensor networks consist of individual sensor devices. Deployment is via manual emplacement. The network is fully preconfigured. Access to information is via manual retrieval of the device itself, or long-range point-to-point communication links.

Second generation sensor networks (2GSN). Sensors work in collaboration to cover an area. The network is typically a hub and spoke formation with a small number of sensors (typically 3 or 4) communicating with a control node equipped with a reach-back link. They are typically manually deployed, relying heavily on preconfiguration.

Third generation sensor networks (3GSN). The latest generation of sensors encompasses self-organising, flexible and scalable networks. Sensors communicate with one another for two purposes, communications services (e.g., automatic relaying of messages to a network gateway) and in-network processing (data aggregation and data fusion). Sensor networks can contain many tens or even hundreds of nodes. Deployment can be hand-emplaced or remotely air-dropped. The sensors are able to establish and – if required – publish and make use of their own geographic location, e.g., based on global positioning system (GPS).

4.2. Fully integrated solutions

Companies such as SenTech, Textron and Lockheed Martin have systems with a variety of sensors (including seismic, acoustic, infrared) which transfer their data directly to a ground station over a number of long-range non-line of sight bearers (including satcom, very high frequency and high frequency bearers). These generally fit into the 1GSN category of networks where each node is equipped with its own backhaul system.

There is a number of 2GSN systems becoming available such as the Terrain Commander and Future Combat System from Textron Systems, or the Falcon Watch System from Harris which will provide processed information from a number of sensors (including acoustic, seismic, magnetic, electro-optical and passive infrared). However, in general, there are very few of the 2GSN systems on the market. Neither, the 1GSN or 2GSN systems are truly ad hoc multi-hop in nature requiring either a direct link back to a remote ground station or a direct link back to a gateway node. There are a few 3GSN ad hoc systems advertised although many of these appear to be immature and still at proof-of-concept stage.

The majority of the systems are aimed at military use (as well as industrial plant monitoring) and many of them cite perimeter protection as their main function (for both military assets and civilian assets such as airstrips).

4.3. Wireless sensor network components

Flexible ad hoc sensor networking needs to be supported by tailored network components such as the sensors themselves and special-purpose routing protocols. Significant scientific and engineering effort has been spent on some of these components which is reflected in the following.

Routing protocols should enable self-configuration after network deployment. They have influence on traffic latency (as some routing protocols will find routes at set-up whilst others require a route to be found prior to each transmission of user traffic), on networking overhead, on energy efficiency, on the speed of network recovery in case of failures, on traffic assurance. Three main classes of routing protocols for energy-efficient wireless sensor networks have been identified [2–4]:

- **Hierarchical/node-centric**. Most routing protocols follow this approach. These protocols aim at clustering the nodes so that “cluster heads” can perform some aggregation. This reduces the amount of data to be transmitted and saves energy. The scalability of these protocols is very good. However, their routing tables may take time to converge (i.e., choose the most appropriate route) if frequent network topology changes occur (which can happen if nodes can transition into suspend mode to conserve energy).
- **Location based/position-centric**. This routing class is based on the exact (GPS) or relative (triangulation, analysis of neighbor dependencies) position of the single nodes. The distance between sensor nodes can be used to estimate the required transmission power which facilitates energy efficient routing.
- **Data-centric**. In the data-centric approach the sensor network is seen from the application point of view as a pool of data. The interface to the network will forward a query and the network will return the data to satisfy the query condition. The routing is driven by the query of the application, not on the identity of the involved nodes or sensors. The underlying implementation of the routing protocol might still be hierarchical/node-centric, and it may only be the interface available to the user that is data-centric.

Other classification of routing protocols for wireless sensor networks can characterize the network by their ability to make use of multipath transmissions, to aggregate data and to eliminate redundant information:

- **Multipath**. The main reason for transmissions via several paths is to provide tolerance to faults in the network. The protocols address the fact that they take advantage of more than one route to the gateway. Mechanisms must be integrated to ensure that only limited (or ideally no) redundant information will be produced.
- **Data processing**. Data processing can be performed at different places in the network as discussed in the context of Fig. 2. Intelligent data aggregation allows the network to operate in an energy efficient manner as less data needs to flow over the network.
- **Negotiation based**. High level data descriptors can be used to eliminate redundant information through negotiation. The nodes will send negotiation messages to prevent or suppress the exchange of duplicated or unwanted information. It is important to ensure that the level of negotiation overhead is limited.

A selection of routing protocols for wireless sensor networks which are subdivided into classes and associated

Table 1
Routing protocols with associated characteristics

Routing protocol	Node-centric	Position-centric	Data-centric	Multipath	Data processing	Negotiation based
LEACH [5]	✓				✓	✓
PEGASIS [6]	✓				✓	
Tiny-AODV [7]	✓				✓	
MECN [8]		✓				
Geographic adaptive fidelity [9]		✓			✓	
GEAR [10]		✓				
SPIN [11]			✓	✓	✓	✓
Directed diffusion [12]			✓	✓	✓	✓
Rumor routing [13]			✓		✓	
Gradient-based routing [14]			✓		✓	
COUGAR [15]			✓		✓	

with the above-mentioned characteristics is shown in Table 1. The presented protocols are just a sample of the protocols discussed in literature. The usage of these protocols in available products is however still rare and generally non-specialized protocols such as optimized link state routing (OSLR) are used as these protocols are more mature.

In the future, disruption tolerant networking (DTN) techniques [16, 17] may receive further attention. These help to provide end-to-end communications in networks with large delays and/or frequent interruptions. Also the connection of the sensor network through the network gateways to the end application might profit from this approach – especially if this reach-back capability is not always present as in the case of the UAV relay.

Medium access control (MAC). The medium access control scheme defines how multiple radios will access the medium and is used to avoid collisions should two or more radios wish to transmit simultaneously. The MAC scheme has an influence on the efficiency of a distributed sensor network in three ways: throughput, delay and energy. Throughput can suffer due to collisions when two or more nodes transmit information at the same time. This wastes energy as well as introducing longer periods of idle listening. Within the range of specialized MAC protocols for wireless sensor networks, two generic types can be identified [18]:

- **Scheduled protocols**. These are time division multiple access (TDMA) based protocols mostly used in combination with hierarchical/node centric routing protocols as cluster heads are needed for synchronization purposes.
- **Contention protocols**. Carrier sense multiple access (CSMA) is an important part of the contention based protocols. Modifications of the MAC scheme of the Institute of Electrical and Electronic Engineers (IEEE) 802.11 family addressing frequency changes as well as protocol optimizations can be found.

Within existing wireless sensor products and developer kits the use of “Commercial off the Shelf” (COTS) protocols stemming from wireless local area network (WLAN), Bluetooth or Zigbee are common, and mature implementations of specialized MAC protocols are rare.

Transmission technologies. Based on an analysis of military use-cases it becomes apparent that low data rates of just a few kilobits per second can often be sufficient while transmission ranges of a few tens of metres or better a few hundreds of metres are desirable. Sufficient coverage can then be achieved based on multi-hopping (allowing intermediate nodes to relay data). This hopping concept has the additional positive effect, that the output power can be reduced facilitating a low probability of detection and interception.

In case of other signals being transmitted within the same frequency band – be it due to other users or to jamming – the transmission technology should provide some robustness against narrowband interference. Combined with the desire to achieve inconspicuous operation, the following transmission technologies can subsequently be seen as prominent for use in military wireless sensor networks:

- direct sequence spread spectrum (DS-SS),
- frequency hopping spread spectrum (FH-SS),
- pulsed ultra-wideband (UWB).

In many prototype networks, COTS chipsets are being used providing transmission based on:

- Bluetooth (FH-SS),
- ZigBee (IEEE 802.15.4/WPAN, DS-SS) or
- WLAN (IEEE 802.11b using DS-SS as well).

Table 2
Developer platforms and their operating systems (updated and expanded from [22])

Platform	MCU	RAM [KB]	Program memory [KB]	Nonvolatile data memory [KB]	Radio chip	Tiny OS	Tiny OS V2	Mantis OS	SOS
BTnode3	ATMega128	64	128	180	CC1000 ZV4002 Bluet.	✓	✓		
Cricket	ATMega128	4	128	512	CC1000	✓			
imote	ARM 7	64	512	0	ZV4002 Bluet.	✓			
imote2	Intel PXA271	256	$32 \cdot 10^8$	0	CC2420	✓	✓		
MANTIS nymph	ATMega 128	4	128	64	CC1000			✓	
mica	ATMega 128	4	128	512	TR1000	✓			
mica2	ATMega 128	4	128	512	CC1000	✓	✓	✓	✓
mica2Dot	ATMega 128	4	128	512	CC1000	✓	✓	✓	
micaz	ATMega 128	4	128	512	CC2420	✓	✓	✓	✓
rene2	ATMega 163	1	8	32	TR1000	✓			
TelosA	TI MSP430	2	60	512	CC2420	✓			
TelosB	TI MSP430	10	48	1000	CC2420	✓	✓	✓	
Tmote Sky	TI MSP430	10	48	1000	CC2420	✓			
tinynode	TI MSP430	10	48	512	XE1205	✓			✓
XYZ	ARM 7	32	256	256	CC2420				✓

However, while remaining within the legal power limits for the respective frequency bands, the transmission ranges are not generally sufficient with WLAN achieving only distances of around 200 metres in practice[2].

Sensor types. A wide range of different sensor types which are usable for wireless sensor applications are available on the market:

- acoustic sensors,
- seismic sensors,
- magnetic sensors,
- infrared sensors,
- electro-optical sensors (closed circuit TV, etc.),
- electromagnetic sensors.

Significant effort is necessary for proper integration into larger-scale sensor networks, and one of the greatest challenges is improving sensor accuracy to keep the false alarm rate to a minimum. The need for a reliable detection of critical incidents has led to the use of **multi-modal sensors**. The intelligent combination of sensors and their joint accuracy are essential for future robust sensor applications. Furthermore, multi-modal sensors can minimize the power consumption as well as the generated traffic, e.g., if a video camera is enabled by an acoustic sensor or an infrared sensor.

Security. There are a number of security solutions to the issues inherent in a wireless sensor network. A wireless sensor network is like any other data exchange network with generic vulnerabilities and associated solutions including:

- **Eavesdropping**. The potential for an enemy to intercept and decode messages passed between devices in the sensor network. Protection is possible using available civil crypto to prevent successful eavesdropping in a sensor network, particularly as the information is generally of only short-term utility.
- **Spoofing**. The potential for a (non-legitimate) node to pass itself off as a legitimate network node and thereby subvert network exchanges. Current cryptographic authentication mechanisms are available which would be appropriate for wireless sensor networks.
- **Message integrity**. The ability of messages to be passed between nodes unchanged or unmodified enroute. Cryptographic protection and strong integrity checks (e.g., secure hash) are available now and provide robust protection against message tampering and replay attacks.

[2] Dependent on terrain and other environmental factors and with an omni-directional antenna.

- **Denial of service**. Preventing nodes in the network from being able to access and use the radio network to pass messages. Low-cost transceivers currently do not have robust anti-jam capabilities making sensor networks susceptible to this type of attack.

- **Geolocation**. The ability to locate the geographical position of nodes in the sensor network by detecting and receiving emissions from the devices. Reducing transmissions to an absolute minimum both in duration and number reduces the chance that an adversary will detect or locate a sensor network. This optimization can be included in the protocol choice and design. However, there will always be the danger that an adversary will detect transmissions, particularly if they suspect that an area contains a sensor network.

- **Physical compromise**. The ability of an enemy to extract useful intelligence and information out of a sensor node that has been located and captured. It is likely physical compromise can be addressed through simple anti-tamper mechanisms, e.g., microswitches, and fill-purge mechanisms to purge system memory of sensitive data. These are tried and tested approaches to physical resilience.

4.4. Developer kits

A number of developer solutions ("developer kits") which include sensor platforms, operating systems and transmission technologies are currently available. These are useful for research purposes as well as to foster the development of versatile sensor network applications. Table 2 shows a summarized overview of existing platforms including some of their technical parameters.

5. Research opportunities

5.1. Engineering challenges

Current wireless sensor networks can make use of multiple years of research on ad hoc networking, energy-efficient routing and related areas. Consequently, the use-cases illustrated in this paper can be met to a greater or lesser extent by existing technologies. The key challenges to deploying military wireless sensor networks are more practical engineering problems than fundamental research issues as listed below:

- clear identification of several simultaneous events, and a reliable correlation of information from neighboring nodes;

- classification of objects and events in addition to their pure detection; an automatic identification and classification of objects and events would support a quick and appropriate reaction and would hence improve the use for military purposes;

- improved integration of different types of sensors (multi-modal sensors) for enhanced information reliability; as many events have a number of simultaneous effects such as creating not only noise but also emitting electromagnetic waves, combining sensors for a "joint detection" is expected to significantly improve the reliability especially in challenging environments;

- radio communications for use of wireless sensors in, specifically, urban warfare provides further challenges such as overcoming possibly strong interference from many sources, shadowing from buildings coupled with severe multipath transmission and at the same time achieving sufficient coverage and energy-efficiency with inconspicuous small-scale antennas and electromagnetic patterns;

- miniaturization of sensors allowing for a quick and automated network deployment and unsuspicious operation;

- robustness of sensors for deployment from planes or by rocket-launches;

- avoidance of data loops in large sensor networks;

- appropriate anti-tamper mechanisms;

- the optimization of sensor networks to provide the most efficient coverage of a geographic area; a number of trade-offs need to be considered including cost (minimizing the cost per unit area of coverage), communications range, sensor range, device size, weight, power, capability (e.g., detect and classify or just detect), and deployment mechanisms;

- agree on common formats and standards for sensor data and communications exchange.

5.2. Scientific challenges

Capabilities required for 3GSN sensor networks are far-reaching, and the step from existing 2GSN to future 3GSN truly ad hoc systems is huge. Research has still to address a number of challenges in order to increase the usability, flexibility and security, as well as to facilitate longer-term operations. These challenges include:

- security, particularly regarding the effectiveness of reputation approaches to protect against the injection of spoof messages or jamming;

- suitable power supplies and energy efficient protocols to meet the long-endurance applications where networks may be in place for several months; this includes power scavenging (e.g., EU IST VIBES project [19, 20]) and novel power sources [21];

– effective and efficient remote air delivery of sensors ensuring even density and coverage across area;

– robustness of data fusion and analysis, ensuring that data from multiple sensors can be appropriately processed to accurately detect and track moving objects even in the presence of measurement inaccuracies, distortions and communications delays.

6. Conclusions

Wireless sensor networks will have a role to play for a number of military purposes such as enemy movement detection and force tracking.

Comparing the actual military requirements with the current research and the available products, some misalignments become obvious. Much effort in current academic research is spent on optimization, e.g., routing protocols to work with tens of thousands of nodes, which are assumed to be small, lightweight and cheap. The paper has addressed the military requirements for actual costs per node, the current mode of deployment (mainly manual network set-up) and physical size. The limited existing products tend to address the current military requirements in that they are composed of larger sensor devices and consist only of small numbers of nodes (often even < 30 nodes).

The key challenges to deploying military wireless sensor networks are more practical engineering problems than fundamental research issues. However there are still outstanding scientific challenges as stated in this paper. Urban warfare scenarios are especially demanding and efforts such as the optimization of multi-modal sensors need to be addressed.

The use of common formats for sensor data such as textual information or images facilitates information exchange across network boundaries and promotes openness between sensor network vendors. This allows those requiring kit to purchase from multiple suppliers and keeps a competitive market place open (i.e., no one supplier can monopolize the supply base). The use of appropriate NATO STANAGs (standardization agreements) is encouraged.

References

[1] Wolfpack program, US Defense Advanced Research Projects Agency (DARPA), http://web-ext2.darpa.mil/sto/strategic/wolfpack.html

[2] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks", *Ad-hoc Netw.*, no. 3, pp. 325–249, 2005 (first published in Nov. 2003).

[3] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey", *IEEE Wirel. Commun.*, vol. 11, iss. 6, pp. 6–28, 2004.

[4] D. Niculescu, "Communication paradigms for sensor networks", *IEEE Commun. Mag.*, vol. 43, iss. 3, pp. 116–122, 2005.

[5] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocols for wireless microsensor networks", in *Proc. Int. Conf. Syst. Sci.*, Hawaii, USA, 2000.

[6] S. Lindsey and C. S. Raghavendra, "PEGASIS: power efficient gathering in sensor information systems", in *IEEE Aerosp. Conf. Proc.*, Montana, USA, 2002, vol. 3, pp. 3-1125–3-1130.

[7] TinyOS 2007, http://www.tinyos.net/

[8] V. Rodoplu and T. H. Meng, "Minimum energy mobile wireless networks", *IEEE JSAC*, vol. 17, no. 8, pp. 1333–1344, 1999.

[9] Y. Xu, J. Heidemann, and D. Estrin, "Geography informed energy conservation for ad hoc routing", in *Proc. 7th Ann. ACM/IEEE Int. Conf. Mob. Comp. Netw.*, Rome, Italy, 2001, pp. 70–84.

[10] Y. Yu, D. Estrin, and R. Govindan, "Geographical and energy-aware routing (GEAR): a recursive data dissemination protocol for wireless sensor networks", Techn. Rep., UCLA-CSD TR-010023, UCLA Comp. Sci. Dept., May 2001.

[11] J. Kulik, W. Rabiner, and H. Balakrishnan, "Adaptive protocols for information dissemination in wireless sensor networks", in *Proc. 5th ACM/IEEE MobiCom Conf.*, Seattle, USA, 1999.

[12] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion: a scalable and robust communication paradigm for sensor networks", in *Proc. ACM MobiCom 2000 Conf.*, Boston, USA, 2000, pp. 56–67.

[13] D. Braginsky and D. Estrin, "Rumor routing algorithm for sensor networks", in *Proc. 1st Worksh. Sens. Netw. Apps.*, Atlanta, USA, 2002.

[14] C. Schurgers and M. B. Srivastava, "Energy efficient routing in wireless sensor networks", in *Proc. MILCOM Conf.*, McLean, USA, 2001.

[15] Y. Yao and J. Gehrke, "The cougar approach to innetwork query processing in sensor networks", SIGMOD Record, Sept. 2002.

[16] Delay Tolerant Networking Research Group of the Internet Research Task Force (IRTF), http://www.dtnrg.org and http://www.irtf.org/

[17] V. Cerf, S. Burleigh, A. Hooke, L. Torgerson, R. Durst, K. Scott, K. Fall, and H. Weiss, "Delay-tolerant networking", RFC 4838, Apr. 2007.

[18] C. S. Raghavendra, K. M. Sivalingam, and T. Znati, *Wireless Sensor Networks*. New York: Springer, 2006.

[19] VIBES project, on vibration energy scavenging, EU IST, 2004–2007, http://www.vibes.ecs.soton.ac.uk/

[20] S. P. Beeby, M. J. Tudor, R. N. Torah, T. O'Donnell, and S. J. Roy, "Micro electromagnetic generator for vibration energy harvesting", *J. Micromech. Microeng.*, vol. 17, pp. 1257–1265, 2007.

[21] "Radio isotope micropower sources program", US Defense Advanced Research Projects Agency (DARPA), http://www.darpa.mil/sto/smallunitops/rims.html

[22] M. Healy, T. Newe, and E. Lewis, "A survey of operating systems for wireless sensor nodes", in *5th Worksh. Internet, Telecommun. Sig. Proces.*, Hobart, Australia, 2006.

Michael Winkler received his M.Sc. (1997) after studies at the University of Bristol and the ENST Bretagne, his M.Sc. (1998) and Ph.D. (2005) degrees of the University of Hannover, Germany. He worked as a scientist at the Institute for Communications of the University of Hannover and as technology consultant for an international media company. His main fields of research are wireless OFDM-based transmissions and high data rate communication networks. In 2005 he joined the NATO C3 Agency, where he is leading the team on ad hoc networking.

e-mail: Michael.Winkler@nc3a.nato.int
NATO C3 AGENCY (NC3A)
P.O. Box 174
2501 CD The Hague, Netherlands

Klaus-Dieter Tuchs received his M.Sc. on electrical engineering in 1996 and his Ph.D. in 2002 at University of Hanover, Germany. He worked as a scientist at the Institute for Communications of the University of Hannover until 2003. His main research area was on the optimization of fault management systems and the development of data mining algorithms. From 2003 to 2006 he worked as a team leader at a telecommunications planning and consulting company (DOK SYSTEME). In 2007 he joined the NATO C3 Agency (The Hague, Netherlands) as a Senior Scientist for network management and telecommunication protocols.

e-mail: Klaus-Dieter.Tuchs@nc3a.nato.int
NATO C3 AGENCY (NC3A)
P.O. Box 174
2501 CD The Hague, Netherlands

Kester Hughes is an experienced, senior team leader with QinetiQ's Communications Division with broad experience in defining, managing and delivering complex defence related research programmes. He has a detailed knowledge and understanding of the breadth UK military communications systems. He has broad experience of a wide variety of communications systems and technologies including military specific systems, communications technologies for sensor networks, IP and the Internet, cellular networks, ATM and wireless LAN technology. In over sixteen years of working, he has contributed and lead many applied research tasks and provided advice and consultancy to MOD's procurement teams and industry.

e-mail: knhughes@QinetiQ.com
QinetiQ Malvern
St Andrews Road, Malvern
Worcestershire, WR14 3PS, UK

Graeme Barclay has a background in mathematics and joined QinetiQ (formerly DERA) in 2000. He is a part of the networks group within the communications department and has worked primarily on research for the UK Ministry of Defence investigating the use of communications equipment and technologies within the military tactical environment. Much of his work has involved analysing wireless communications systems and the effects of mobility and quality of service issues on end-to-end delivery. Recently he has been involved in the use of MANET protocols and has led a team responsible for providing networking solutions within a sensor network.

e-mail: gbarclay@QinetiQ.com
QinetiQ Malvern
St Andrews Road, Malvern
Worcestershire, WR14 3PS, UK

A survey on mobility models for performance analysis in tactical mobile networks

Nils Aschenbruck, Elmar Gerhards-Padilla, and Peter Martini

http://doi.org/10.26636/jtit.2008.2.872

Abstract—In scenarios of military operations and catastrophes – even when there is no infrastructure available or left – there is a need for communication. Due to the specific context the communication systems used in these tactical scenarios need to be as reliable as possible. Thus, the performance of these systems has to be evaluated. Beside field-tests, computer simulations are an interesting alternative concerning costs, scalability, etc. Results of simulative performance evaluation strongly depend on the models used. Since tactical networks consist of, or, at least, contain mobile devices, the mobility model used has a decisive impact. However, in common performance evaluations mainly simple random-based models are used. In the paper we will provide classification and survey of existing mobility models. Furthermore, we will review these models concerning the requirements for tactical scenarios.

Keywords*— *mobility models, performance analysis, wireless networks, mobile networks, tactical networks.

1. Introduction

Military operations as well as catastrophes, be it natural ones (like hurricanes or tornados), man-made ones (like explosions or fires), or technical ones (like material-fatigue), cause an area of destruction. Buildings, bridges, as well as the infrastructure of the private and public systems for mobile communication might be destroyed. Hence, units working in these disaster areas need reliable communication which is independent of any infrastructure.

As the communication systems used in these tactical or disaster area scenarios need to be as reliable as possible, the performance of these systems has to be evaluated. Field-tests in manoeuvres may be the preferred evaluation method. However, they are expensive, as sufficient hardware is needed. Furthermore, the results concerning some characteristics (e.g., scalability) are limited – who can perform field-tests with several hundreds of devices? Thus, especially for the evaluation of algorithms and protocols, simulation is an alternative.

Naturally, the results of simulative performance evaluation strongly depend on the models used. Since tactical networks consist of, or, at least, contain mobile devices, the mobility model used has a decisive impact. However, in common performance evaluations mainly simple random-based models are used.

In the paper our aim will be to give a survey on mobility models used for performance evaluation in tactical mobile networks. As tactical networks may also be networks without infrastructure, the individual nodes and there movement characteristics need to be modeled. In this paper we will focus on models that realize the movement of individual nodes (microscopic models). In the literature there are already some surveys on mobility models [2, 4, 11]. However, these surveys are quite old or miss a lot of specific models. Furthermore, there is no review concerning the requirements for tactical scenarios. Thus, in this paper we will give a survey on existing mobility models and classify and review these models concerning the requirements of tactical communication systems.

The remaining part of this paper is structured as follows: Section 2 points out requirements for tactical communication. Next, we will introduce the way the existing models are classified (Section 3). After that, we will give a survey on existing models and review to which extent these models meet the requirements of tactical scenarios (Sections 4–8). Finally, we will conclude the paper (Section 9).

2. Requirements

The users of tactical communication systems are military or civil (e.g., civil protection) forces. These forces are strictly structured (e.g., platoons, groups, etc.) and their actions are strictly organized. The units do not walk around randomly. There is one leader or a group of leaders which tells everybody where and how to move or in which area to work. In general, the movements are driven by tactical reasons. Due to this, the units normally use the optimal path to a destination.

The destinations depend on the working site which is based on tactical issues. The tactics as well as the scene are usually hierarchically organized. Typically, the site is divided into different tactical areas. Each unit belongs to one of these areas. For example, in a disaster area scenario a firefighter belongs to an *incident site* and a paramedic will work at one place in the *casualties treatment area*. The units sent to a specific location once will typically stay close to this location. Some of them may have special tasks that make them move from one area to another (e.g., transport units). However, the major part of the units does not leave the area. Thus, the area in which a unit moves depends on tactical issues but is restricted to one specific area.

Furthermore, as tactical scenarios take place in areas of destruction, obstacles might be encountered. Smaller ones may be ignored, because they only have little impact on the movement. However, larger ones (walls, houses, etc.) will have a certain impact on movements.

In tactical networks, units and troops often move in tactical formation. Even if the detailed position may only have little impact, this fact implies group mobility. Moreover, there are units of different types. The units typically differ in their equipment. Some of them possess vehicles and use them resulting in faster movement. Others are pedestrians and move slower. Thus, there is heterogeneous velocity based on the type of node.

Finally, especially in tactical communication systems, it is quite common that units leave the scenario, while others join later on. In military scenarios there may be fatalities, and in civil protection scenarios there may be units that take patients to hospital. When some units leave the scenario, typically others are requisitioned.

As a conclusion, the analysis yields the following main requirements:

- heterogeneous velocity,
- tactical areas,
- optimal paths,
- obstacles,
- units join and leave the scenario,
- group movement.

The following sections present existing mobility models and examine which models meet these requirements.

3. Classification

In general, the mobility models can be classified according to the different kind of dependencies and restrictions that are considered.

- **Random based**. There are neither dependencies nor any other restriction modeled.
- **Temporal dependencies**. The actual movement of a node is influenced by the movement of the past.
- **Spatial dependencies**. The movement of a node is influenced by the nodes around it (e.g., group mobility).
- **Geographic restrictions**. The area in which the node is allowed to move is restricted.
- **Hybrid characteristics**. A combination of temporal dependencies, spatial dependencies, and geographic restrictions is realized.

4. Random based movement

The mobility model often used in the last years (especially in performance evaluation of ad hoc networks) is the *random-waypoint* model. The random-waypoint model is a simple stochastic model in which a node perpetually chooses destinations (waypoints) and moves towards them. In the original model [21] the nodes are distributed randomly over the simulation area. After waiting for a constant pause time, each node chooses a waypoint and moves towards it with a speed chosen from an interval $[\upsilon_{\min};\upsilon_{\max}]$. After arriving at the waypoint, the node again waits for a constant pause time and chooses the next waypoint. In [30] it is proposed to also choose the pause time from an interval $[p_{\min};p_{\max}]$. The different random variates are mostly chosen uniformly distributed.

In the last years, there were several studies that analyze the random-waypoint model with respect to implicit (unwanted) assumptions and characteristics. As the nodes are initially distributed randomly, it takes some time until the nodes reach a stationary distribution (cf. [28]). Thus, a long enough initial period should be discarded. In [36] it is shown that the average velocity is decreasing over simulation time if $v_{\min} = 0$. Thus, $v_{\min} > 0$ and $p_{\max} < \infty$ should be chosen. Furthermore, in several publications it was shown that the nodes cumulate in the middle of the simulation area (cf. [6, 7, 10]). For a square simulation area a density as shown in Fig. 1 results.

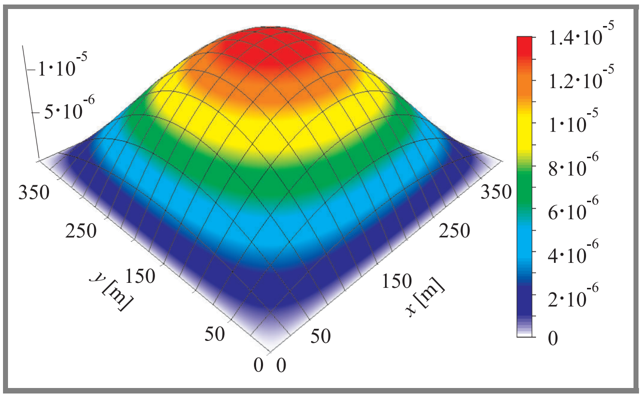


Fig. 1. Density for the random-waypoint model.

A distribution and movement of the nodes across the entire simulation area does not fit to the characteristics of most realistic movements. There are extensions (e.g., [7]) which add attraction points to this model in order to generate more realistic non-equally distributed mobility. The probability that a node selects an attraction point or a point in an attraction area as next waypoint is larger than the choice of other points. The nodes visit some points more frequently than others. Hence, they still move across the complete simulation area. The *clustered-mobility* model [24] is motivated by disaster areas and uses a similar approach. The difference is that the attraction of a point depends on the amount of nodes nearby. This implies that the areas of higher density variate concerning the intensity and position. Further approaches like the *random-direction* model [31], *random-border* model [7], and the *modified-random-direction* model [31] also result in fully random movement with different node density distributions.

All random-based models result in random movement across the complete simulation area. The models are quite simple to implement, but the only characteristics of an tactical scenario that is realized are the optimal paths. However, at least heterogeneous velocity may be integrated quite easily.

5. Temporal dependencies

Using one of the models of the previous section, the nodes suddenly may change speed or direction. This is quite unrealistic considering aspects like acceleration and deceleration. The models presented in this section realize such aspects by using temporal dependencies.

In the *Gauss-Markov* model [23] velocity and direction of the future (time interval $t+1$) depend on the current values (time interval t). Initially for each node position, velocity, and direction are chosen uniformly distributed. The movement of each node is variated after an interval δt. The new values are chosen based on a first-order autoregressive process. Further details can be found in [23].

The *smooth-random* model [4, 5] is a more detailed approach. The nodes are classified concerning their maximum velocity, preferred velocity, maximum acceleration, and deceleration. New velocities and directions are calculated based on these parameters and the current ones. Velocity and direction may also be chosen in correlation to each other. By doing so, more realistic movements like deceleration before a change of direction may be realized.

By using one of these models and realizing the temporal dependencies the movements of the nodes become smoother concerning direction and velocity. However, typical characteristics of tactical scenarios are not realized in this approach.

6. Spatial dependencies

Beside temporal dependencies there are also spatial ones. Nodes may move together in groups. Thus, the movement of one node may influence the movement of others around him.

One approach to realizing spatial dependence is the use of reference points. The *reference-point-group-mobility* model (RPGM) [15] models the movement of groups of nodes. The movement of the groups is modeled according to an arbitrary mobility model. The movement of the nodes inside a group is realized using a reference point for each node. The actual position of a node is a random movement vector added to the position of his reference point. The absolute positions of the reference points do change according to the arbitrary mobility model, but the relative positions of the reference points inside a group do not change. Hence, the spatial dependence is realized using the reference points.

In [9] a variance of the model called *structured-group-mobility* model is proposed. In this model there is no random movement vector. The nodes of a group move in a fixed non-changing formation. The formations are motivated by firefighter, police, and tanks. However, even if there is a formation of tanks, there may be some variances due to obstacles. In literature there are also found several other variances of the RPGM model, e.g., *column* model, *pursue* model, *nomadic-community* model (cf. [11, 34]).

Another approach to realize spatial dependence is to found on social networks. The *social-network-founded* mobility model [26] bases on interaction indicators for all pairs of nodes – the larger an interaction indicator, the larger the probability of a social relationship, the smaller the geographic distance. Initially the nodes are grouped in clouds according to their interaction indicator. The clouds as well as the nodes inside the clouds move according to a random-waypoint model, where the waypoints are chosen according to the interaction indicators as well. In [27] this approach is reinvented as *community-based* mobility model. Different more realistic algorithms are used for the classification of the nodes into groups and the movement inside the clouds. Furthermore, the interaction indicators are modified over time.

For realizing group mobility in tactical scenarios, the RPGM model seems to be the better approach, as with an appropriate choice of parameters relative positions of nodes inside the groups can be modeled explicitly. Using the RPGM model, beside the characteristic of group movement, other characteristics may be realized by using an appropriate model for the reference points.

7. Geographic restrictions

Beside considering temporal and spatial dependencies, for many scenarios it is unrealistic to assume that the nodes are allowed to move across the entire simulation area. There are very different approaches to restrict the nodes movement to certain parts of the simulation area. The following sections will describe several approaches realizing the different kind of geographic restrictions.

7.1. Graph-based approaches

A quite intuitive approach is to manage the allowed paths in a movement graph. The *graph-based* mobility model [35] realizes a graph whose vertices are the possible destinations and whose edges are the allowed paths. Based on this graph a random waypoint approach is used. The nodes initially start at a random position on the graph, choose a destination (vertex), move there at random velocity, and choose the next destination and velocity.

Another approach that is using graphs is the *weighted-waypoint* mobility model [16]. The vertices of the graph are specific areas (e.g., classroom, cafe, etc.). The nodes choose destinations inside these areas. The directed edges of the graph contain probabilities of choosing a destination in the directed area depending on the current area. Having chosen a waypoint, the nodes move there on the direct way similar to the random-waypoint model. Compared to the graph-based model, the movement is not restricted to distinct paths.

7.2. Voronoi-based approaches

One possibility of modeling simulation areas with obstacles is to determine the movement paths or areas using Voronoi-diagrams. This approach was first introduced with the *obstacle* mobility model [18, 19]. In this model, the edges of the buildings (e.g., of a campus) are used as an input to calculate a Voronoi-diagram. The movement graph consists of the Voronoi-diagram and additional vertices. These vertices are the intersection of the edges of the Voronoi-diagram and the edges of the obstacles. They model entrances to obstacles (e.g., buildings). The movement on the graph is realized similarly to the graph-based model. By using Voronoi-diagrams, the paths are modeled equidistant from all obstacles. Considering the requirements of tactical networks, these are not necessarily the optimal paths. Furthermore, even for a campus network it is a strong assumption that all streets are built equidistant from all buildings and all nodes move in the middle of the street. In [37] the approach is extended to realize buildings and streets more realistically. In the Voronoi mobility model movement, paths are refined to movement areas. The nodes choose their destinations inside these areas. The movement using this model is more realistic, as streets and buildings are realized more precisely. However, there is still no movement on optimal paths.

7.3. Division-based approaches

Another approach is to divide the simulation area in subareas and to use in them arbitrary mobility models.

The *area-graph-based* mobility model [8] tries to realize clusters (sub-areas) with higher node density and paths in between with lower node density. The clusters are regarded as vertices of the area graph while the paths are regarded as edges. A weight (probability) is assigned to each edge. A node moves inside the cluster for a randomly chosen time according to the random-waypoint model. After this time, he chooses one path according to probabilities at the edges. Next, the node moves on the path to the next area.

A similar approach is used in *CosMos* [14]. The simulation area is subdivided into non-overlapping zones. In each zone the nodes move according to an arbitrary mobility model. The transition between the zones is realized similarly to the area graph based mobility model using transition probabilities. If a node is chosen to change the zone, he moves to a handover area and switches to the other mobility model.

Considering tactical scenarios, both models contain interesting aspects as it is possible to realize tactical areas. However, neither of the model realizes all requirements of tactical scenarios.

7.4. Map-based approaches

A further approach to restrict the movement area geographically is to use information from road maps.

In the context of the UMTS standardization, the so-called *Manhattan-grid* model was specified [13]. The simulation area is divided into squared blocks. Nodes are modeled as pedestrians moving on the vertices of the squares (streets). Initially the nodes are randomly distributed on the streets. Each node chooses a direction and a velocity. If a node reaches a corner, the node changes direction with a certain probability. The velocity is changed over time.

The *random-waypoint-city* model [22] realizes vehicular traffic in urban environments. Therefore, road maps including speed informations and crossroads are retrieved. A node chooses a destination on the streets similar to the random-waypoint model and chooses a route after an arbitrary metric (e.g., smallest travel time). At the crossroads delays are modeled according to the amount of roads. Furthermore, an equal distribution of the nodes throughout the simulation area is realized.

In [25] two further models are described which realize mobility models (e.g., random-waypoint) on graphs based on road maps.

In respect to the requirements of tactical scenarios these models seem to be not applicable. On the one hand, the requirements are not realized, on the other, the streets on which the maps base may be destroyed.

8. Hybrid characteristics

In the previous sections several models were described that could quite clearly be assigned to one class of dependencies. However, there are also some models that realize hybrid dependencies and restrictions.

8.1. Complex vehicular traffic models

The *freeway* mobility model [3] realizes temporal and spatial dependencies as well as geographic restrictions. The nodes variate their velocity in dependence to their current velocity (temporal dependencies). Furthermore, the velocity is influenced by the velocity of a vehicle on the same line inside a certain radius (spatial dependence). The overall movement is restricted to a freeway (geographic restrictions).

The *street-random-waypoint* model (STRAW) [12] uses information from maps similar to the random-waypoint-city model. However, the actual movement of the vehicles is realized according to vehicular congestion and simplified traffic control mechanisms. The model realizes temporary dependencies (acceleration), spatial dependencies (to other vehicles) and geographic restrictions (streets).

Both models are specific for vehicular road-traffic and do not fit to a tactical scenario.

8.2. User-oriented meta-model

A general approach to modeling complex scenarios is described in [32] as *user-oriented mobility meta-model*. The model consists of three components:

1. Modeling the simulation area containing restrictions concerning the movements as well as attraction points.
2. Sequences of movement made by a user, e.g., a sequence of attraction points.
3. Temporal and spatial dependencies concerning the movements of a user.

Using this model, typical movements of node during a day may be modeled (cf. [33]). This abstract meta-model is generic and can be seen as general description of many other models. The requirements of tactical scenarios may be realized using this abstract meta-model. However, the concrete realization of the requirements is not specified in the meta-model.

8.3. Models for tactical scenarios

Apart from a lot of generic models, there are also some approaches to realize specific scenarios. In [20] three scenarios are considered. Beside a conference and a concert scenario there is also a *catastrophe scenario*. In the scenarios, obstacles, group movements, and tactical areas are considered. As one example for a military scenario in [17] a *hostage rescue scenario* was specified. The scenario is divided into periods (e.g., march, pull, fallback). The movement is modeled with regard to the specific phases. Another scenario [29] models the movement of a platoon in a city area. All these scenarios – the catastrophe, the hostage rescue as well as the platoon scenario – realize several requirements of tactical scenarios. However, they are only specific scenarios that are restricted concerning scalability, e.g., the amount of nodes and the size of the simulation area.

8.4. Disaster-area model

In [1] a model which realistically represents the movements in a disaster area scenario is provided. This model supports heterogeneous area-based movement on optimal paths avoiding obstacles with joining/leaving of nodes as well as group mobility.
To realize area-based movement, the simulation area is divided into polygonal tactical areas. The tactical areas are classified according to the civil-protection concept *separation of room* (cf. Fig. 2). Each node is assigned to one of these tactical areas. For some areas there are both stationary nodes, which stay in the distinct area moving according to a random based mobility model, as well as transport nodes that carry the patients to the next area following a movement cycle. Different areas and classes allow heterogeneous speeds. The area and the class (stationary or transport) the node belongs to define the movement of the node as well as the minimal and maximal speed distinguishing pedestrians from vehicles.

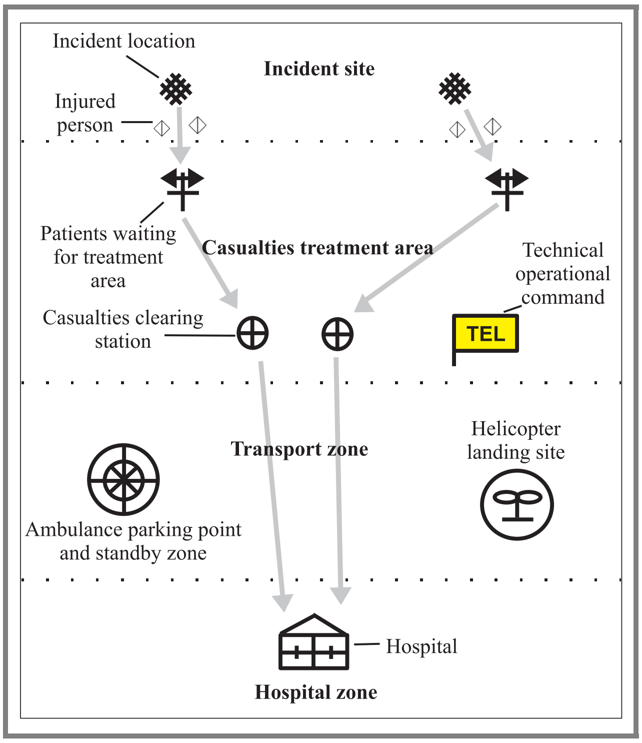

Incident location
Incident site
Injured person
Patients waiting for treatment area
Casualties treatment area
Technical operational command
Casualties clearing station
TEL
Transport zone
Helicopter landing site
Ambulance parking point and standby zone
Hospital
Hospital zone

Fig. 2. Separation of the room in civil protection.

The optimal path for the movement of the transport units between the different areas is determined by methods of robot motion planning. For finding the shortest paths and avoiding obstacles between the tactical areas, visibility graphs are used. A visibility graph is a graph where its vertices are the vertices of the polygons. There is an edge between two vertices, if the vertices can "see" each other – meaning the edge does not intersect the interior of any other obstacle. The shortest path between two points consists of an appropriate subset of the edges of the visibility graph. Thus, after having calculated the visibility graph containing all possible shortest paths between the areas avoiding obstacles, the direct path between two areas for each transport unit can be calculated.
Vehicular transport units (e.g., ambulances) typically leave the disaster area to carry patients to hospital. Thus, joining and leaving nodes are realized using specific entry and exit points (registration areas).
Group mobility is realized as an optional characteristic for disaster areas, as in civil protection there may only be one device for each group. Nevertheless, it is realized similar to RPGM [15] using reference points. The units of each area are grouped. The size of the group depends on the type of the area and the group. Similar to RPGM the nodes follow their reference point. The movement of each node in a group is calculated in relation to the movement of the reference point.

Table 1
Survey on an requirement analysis of existing mobility models

Model		Dependencies			Requirements for tactical scenarios					
		Temporary	Spatial	Geographical	Heterogeneous velocity	Tactical areas	Optimal paths	Obstacles	Units leave the scenario	Group movement
Random-waypoint	[21]				(√)	(√)	√			
Random-waypoint with attraction points	[7]				(√)	(√)	√			
Clustered-mobility	[24]		(√)		(√)	(√)	√			
Random-direction	[31]				(√)	(√)	√			
Random-border-model	[7]				(√)	(√)	√			
Modified random-direction	[31]				(√)	(√)				
Random-walk	[11]				(√)	(√)				
Gauss-Markov	[23]	√			(√)	(√)				
Smooth-random	[5]	√			√	(√)	√			
Reference-point-group	[15]	(√)	√	(√)	(√)	(√)	(√)	(√)	(√)	√
Structured-group	[9]		√		(√)	(√)				√
Social-network-founded	[26]		√		(√)	(√)				√
Community-based	[27]		√		(√)		√			√
Graph-based	[35]			√	(√)		√	(√)		
Weighted-waypoint	[16]			√	(√)		√			
Obstacle	[18]			√	(√)			√		
Voronoi	[37]			√	(√)			√		
Area-graph-based	[8]			√	(√)	√	(√)	(√)		
CosMos	[14]			√	(√)	√	(√)			
Manhattan-grid	[13]			√	(√)					
Ramdom-waypoint-city	[22]			√	(√)					
Graph-random-waypoint	[25]			√	(√)					
Graph-random-walk	[25]			√	(√)					
Freeway	[3]	√	√	√	(√)					
Street-random-waypoint	[12]	√	√	√	√					
User-oriented-meta-model	[32]	√	√	√	√		√	√		√
Catastrophe-scenario	[20]		√	√	√	√		√		
Hostage-rescue	[17]		√	√	√					
Platoon	[29]		√	√	√					√
Disaster-area-model	[1]		√	√	√	√	√	√	√	√

9. Conclusion

Finally, we want to discuss which requirements are realized and which approaches model tactical scenarios. Table 1 sums up the survey and requirements analysis that was provided in the paper. In the table for each model the dependencies considered as well as the requirements modeled are shown. A “√” means “explicitly modeled”, while a “(√)” means “not modeled but can be easily extended”. For example *heterogeneous velocity* is not considered in all models. However, it is quite easy to extend the models supporting heterogeneous velocities for different classes of nodes. *Tactical areas* are explicitly realized in some models. Others may be easily extended using an approach like the area-graph-based model. *Group movement* may be easily integrated in other models using the reference point approach. The other requirements *optimal paths*, *obstacles*, and *units join and leave the scenario* are considered in some specific models. However, beside the disaster area model there is no model that considers combinations of all of them.

The disaster-area model is a model that realizes mobility for one tactical scenario in detail, considering all the requirements. This scenario may also be used for the performance evaluation of communication systems for military usage. However, with respect to a military usage of a communication system, medical or humanitarian scenarios similar to civil protection are not the only ones to be considered. There may be totally different characteristics in other specific military scenarios that may have a certain impact on the performance of the communication systems. There are valuable first realizations of specific scenarios, e.g., the hostage rescue and the platoon scenario. However, in the future new scalable models for military scenarios should be invented. Furthermore, the characteristics of these, and, within this, the impact on existing performance evaluation results should be examined.

References

[1] N. Aschenbruck, E. Gerhards-Padilla, M. Gerharz, M. Frank, and P. Martini, "Modelling mobility in disaster area scenarios", in *Proc. 10th ACM IEEE Int. Symp. Model. Anal. Simul. Wirel. Mob. Syst. MSWIM*, Chania, Greece, 2007.

[2] F. Bai and A. Helmy, "Wireless ad hoc and sensor networks", Chapter 1: "A survey of mobility models", 2004, http://nile.usc.edu/ helmy/important/Modified-Chapter1-5-30-04.pdf

[3] F. Bai, N. Sadagopan, and A. Helmy, "IMPORTANT: a framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks", in *Proc. IEEE INFOCOM*, San Francisco, USA, 2003, pp. 825–835.

[4] C. Bettstetter, "Mobility modeling in wireless networks: categorization, smooth movement, and border effects", *ACM SIGMOBILE Mob. Comp. Commun. Rev.*, vol. 5, no. 3, pp. 55–66, 2001.

[5] C. Bettstetter, "Smooth is better than sharp: a random mobility model for simulation of wireless networks", in *Proc. 4th Int. Symp. Model. Anal. Simul. Wirel. Mob. Syst. MSWIM*, Rome, Italy, 2001, pp. 19–27.

[6] C. Bettstetter, G. Resta, and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks", *IEEE Trans. Mob. Comp.*, vol. 2, no. 3, pp. 257–269, 2003.

[7] C. Bettstetter and C. Wagner, "The spatial node distribution of the random waypoint mobility model", in *Proc. 1st German Worksh. Mob. Ad-Hoc Netw. WMAN'02*, Ulm, Germany, 2002, pp. 41–58.

[8] S. Bittner, W.-U. Raffel, and M. Scholz, "The area graph-based mobility model and its impact on data dissemination", in *Proc. IEEE PerCom*, Kuaai Island, Hawaii, USA, 2005, pp. 268–272.

[9] K. Blakely and B. Lowekamp, "A structured group mobility model for the simulation of mobile ad hoc networks", in *Int. Conf. Mob. Comp. Netw., Proc. 2nd Int. Worksh. Mob. Manag. Wirel. Acc. Protoc.*, Philadelphia, USA, 2004, pp. 111–118.

[10] D. M. Blough, G. Resta, and P. Santi, "A statistical analysis of the long-run node spatial distribution in mobile ad hoc networks", *Wirel. Netw.*, vol. 10, pp. 543–554, 2004.

[11] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research", *Wirel. Commun. Mob. Comp.*, vol. 2 no. 5, pp. 483–502, 2002.

[12] D. R. Choffnes and F. E. Bustamante, "An integrated mobility and traffic model for vehicular wireless networks", in *Int. Conf. Mob. Comp. Netw., Proc. 2nd ACM Int. Worksh. Veh. Ad Hoc Netw.*, Cologne, Germany, 2005, pp. 69–78.

[13] "Universal Mobile Telecommunicatios System (UMTS); Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0)", ETSI TR 101 112 V3.2.0 (1998-04).

[14] M. Günes and J. Siekermann, "CosMos – communication scenario and mobility scenario generator for mobile ad-hoc networks", in *Proc. 2nd Int. Worksh. MANETs Interoper. Iss. MANETII'05*, Las Vegas, USA, 2005.

[15] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang, "A group mobility model for ad hoc wireless networks", in *Proc. Int. Symp. Model. Simul. Wirel. Mob. Syst. MSWiM*, Seattle, USA, 1999, pp. 53–60.

[16] W.-J. Hsu, K. Merchant, H.-W. Shu, C.-H. Hsu, and A. Helmy, "Weighted waypoint mobility model and its impact on ad hoc networks", *ACM SIGMOBILE Mob. Comp. Commun. Rev.*, vol. 9, no. 1, pp. 59–63, 2005.

[17] M. Jahnke, J. Tölle, A. Finkenbrink, and A. Wenzel, "Dokumentation zum Forschungsvorhaben E/IB1S/6A661/2F005 – 1. Zwischenbericht", Tech. Rep., FGAN-FKIE im Auftrage des IT-AmtBw, 2006 (in German).

[18] A. Jardosh, E. M. Belding-Royer, K. C. Almeroth, and S. Suri, "Towards realistic mobility models for mobile ad hoc networks", in *Proc. IEEE MobiCom*, San Diego, USA, 2003, pp. 217–229.

[19] A. P. Jardosh, E. M. Belding-Royer, A. K. C., and S. Suri, "Real-world environment models for mobile network evaluation", *IEEE J. Selec. Areas Commun.*, vol. 23, no. 3, pp. 622–632, 2005.

[20] P. Johansson, T. Larsson, N. Hedman, B. Mielczarek, and M. Degermark, "Scenario-based performance analysis of routing protocols for mobile ad-hoc networks", in *Proc. IEEE MobiCom*, Seattle, USA, 1999, pp. 195–206.

[21] D. B. Johnson and D. A. Maltz, "Dynamic source routing in ad hoc wireless networks", in *Mobile Computing*, T. Imielinski and H. Korth, Eds. Norwell: Kluwer, 1996, vol. 353, pp. 153–181.

[22] J. Kraaier and U. Killat, "The random waypoint city model – user distribution in a street-based mobility model for wireless network simulations", in *Proc. 3rd ACM Int. Worksh. Wirel. Mob. Appl. Serv. WLAN Hotsp.*, Cologne, Germany, 2005, pp. 100–103.

[23] B. Liang and Z. J. Haas, "Predictive distance-based mobility management for multidimensional PCS networks", *IEEE/ACM Trans. Netw.*, vol. 11, no. 5, pp. 718–732, 2003.

[24] S. Lim, C. Yu, and C. R. Da, "Clustered mobility model for scale-free wireless networks", in *Proc. IEEE Conf. Loc. Comput. Netw. LCN 2006*, Tampa, USA, 2006, pp. 231–238.

[25] P. S. Mogre, M. Hollick, N. d'Heureuse, H. W. Heckel, T. Krop, and R. Steinmetz, "A graph-based simple mobility model", in *Proc. WMAN'07, Proc. Conf. KiVS'07*, Bern, Switzerland, 2007, pp. 421–432.

[26] M. Musolesi, S. Hailes, and C. Mascolo, "An ad hoc mobility model founded on social network theory", in *Proc. 7th ACM Int. Symp. Model. Anal. Simul. Wirel. Mob. Syst.*, Venice, Italy, 2004, pp. 20–24.

[27] M. Musolesi and C. Mascolo, "A community based mobility model for ad hoc network research", in *Proc. 2nd ACM/SIGMOBILE Int. Worksh. Multi-hop Ad Hoc Netw. Theory Real. REALMAN'06, Colocated with MobiHoc2006*, Florence, Italy, 2006, pp. 31–38.

[28] W. Navidi and T. Camp, "Stationary distributions for the random waypoint mobility model", *IEEE Trans. Mob. Comp.*, vol. 3, no. 1, pp. 99–108, 2004.

[29] S. Reidt and S. D. Wolthusen, "An evaluation of cluster head TA distribution mechanisms in tactical MANET environments", in *Proc. Conf. ACITA*, College Park, USA, 2007.

[30] G. Resta and P. Santi, "An analysis of the node spatial distribution of the random waypoint model for ad hoc networks", in *Proc. ACM Worksh. Princip. Mob. Comp. POMC*, Toulouse, France, 2002, pp. 44–50.

[31] E. M. Royer, P. M. Melliar-Smith, and L. E. Moser, "An analysis of the optimum node density for ad hoc mobile networks", in *Proc. IEEE Int. Conf. Commun.*, Helsinki, Finland, 2001, vol. 3, pp. 857–861.

[32] I. Stepanov, J. Hähner, C. Becker, J. Tian, and K. Rothermel, "A meta-model and framework for user mobility in mobile networks", in *Proc. 11th Int. Conf. Netw. ICON 2003*, Sydney, Australia, 2003, pp. 231–238.

[33] I. Stepanov, P. J. Marron, and K. Rothermel, "Mobility modeling of outdoor scenarios for manets", in *Proc. 38th Ann. Simul. Symp. ANSS'38*, San Diego, USA, 2005, pp. 312–322.

[34] M. Sánchez and P. Manzoni, "ANEJOS: a Java based simulator for ad hoc networks", *Fut. Gener. Comput. Syst.*, vol. 17, no. 5, pp. 573–583, 2001.

[35] J. Tian, J. Hähner, C. Becker, I. Stepanov, and K. Rothermel, "Graph-based mobility model for mobile ad hoc network simulation", in *Proc. 35th Ann. Simul. Symp.*, San Diego, USA, 2002, pp. 337–344.

[36] J. Yoon, M. Liu, and B. Noble, "Random waypoint considered harmful", in *Proc. IEEE INFOCOM*, San Francisco, USA, 2003, pp. 1312–1321.

[37] H.-M. Zimmermann and I. Gruber. "A Voronoi-based mobility model for urban environments", in *Eur. Wirel. 2005 Conf.*, Zypern, Greece, 2005.

Nils Aschenbruck received his Diploma in computer science (Dipl.-Inform.) from the University of Bonn, Germany, in 2003. Currently he is a Ph.D. candidate and research Assistant in the communication systems group at the University of Bonn. His research interests include mobile and wireless networks, especially mobility and traffic modeling as well as security.
e-mail: aschenbruck@cs.uni-bonn.de
Institute of Computer Science IV
University of Bonn
Roemerstr. 164
53117 Bonn, Germany

Elmar Gerhards-Padilla received his Diploma in computer science (Dipl.-Inform.) from the University of Bonn, Germany, in 2005. Currently he is a Ph.D. candidate and research Assistant in the communication systems group at the University of Bonn. His research interests include mobile and wireless networks, especially security in tactical mobile ad hoc networks and honeynets.
e-mail: padilla@cs.uni-bonn.de
Institute of Computer Science IV
University of Bonn
Roemerstr. 164
53117 Bonn, Germany

Peter Martini received his Diploma and Ph.D. degrees from the Aachen University of Technology, Germany, in 1986 and 1987, respectively. From 1986 to 1990 he was with the Institute of Computer Science IV at the Aachen University of Technology. From 1990 to 1996 he was Professor of operating systems and computer networks at the University of Paderborn, Germany. Since 1996, Professor Martini heads the Institute of Computer Science IV at the University of Bonn. His current research interests include IT security and security assistance systems, mobile communication systems and mobile devices, high speed networks, and performance engineering.
e-mail: martini@cs.uni-bonn.de
Institute of Computer Science IV
University of Bonn
Roemerstr. 164
53117 Bonn, Germany

Telecommunications network design and max-min optimization problem

Włodzimierz Ogryczak, Michał Pióro, and Artur Tomaszewski

http://doi.org/10.26636/jtit.2005.3.326

Abstract—Telecommunications networks are facing increasing demand for Internet services. Therefore, the problem of telecommunications network design with the objective to maximize service data flows and provide fair treatment of all services is very up-to-date. In this application, the so-called max-min fair (MMF) solution concept is widely used to formulate the resource allocation scheme. It assumes that the worst service performance is maximized and the solution is additionally regularized with the lexicographic maximization of the second worst performance, the third one, etc. In this paper we discuss solution algorithms for MMF problems related to telecommunications network design. Due to lexicographic maximization of ordered quantities, the MMF solution concept cannot be tackled by the standard optimization model (mathematical programme). However, one can formulate a sequential lexicographic optimization procedure. The basic procedure is applicable only for convex models, thus it allows to deal with basic design problems but fails if practical discrete restrictions commonly arriving in telecommunications network design are to be taken into account. Then, however, alternative sequential approaches allowing to solve non-convex MMF problems can be used.

Keywords*— *network design, resource allocation, fairness, lexicographic optimization, lexicographic max-min.

1. Introduction

Since the emergence of the Internet one has witnessed an unprecedented growth of traffic that is carried in the telecommunications networks. The pace at which the number of network users and the amount of traffic related to data-oriented applications are growing has been and still is much higher than several percent of growth that were typical for traditional voice-only networks; as a matter of fact data traffic almost doubles every year. It can also be observed that the distribution of traffic in data networks changes quickly, both – in the short and long time-scales, and is very difficult to predict. As a result, from the network operator's perspective the network extension process becomes very complicated – while it is not economically feasible to sufficiently over-dimension a network, it is also hard to decide when and where the network should be augmented. An inevitable effect of the situation that the capacity of a network does not match the traffic generated by network service users, is network overload – a phenomenon commonly encountered in current data-oriented networks.

Overloads influence the quality of service perceived by users – data transfer slows down because packet transfer delays increase and packet losses occur much more frequently. Overloads are one of the major concerns of network operators, because the guaranteed quality of service level is one of the basic elements of network operators' differentiation and a prerequisite of their success. In order to avoid overloads and provide the guaranteed quality of service level (instead of offering the so-called best-effort service) the network operator must control the amount of traffic that enters the network. The traffic admission control process is responsible for deciding how many users can be served and how much traffic each of these users can generate. What is important is that, in general, some users will be denied the service in order to reduce the overall stream of traffic that enters the network. Since the service denial probability is another important measure of the quality of service level, one of the primary objectives of the admission control process must be to guarantee that the users have fair access to network services. The most common "fairness-oriented" (as opposed to "revenue-oriented") approach is to admit equal amount of traffic from every stream – the amount being expressed in absolute or relative terms. Unfortunately, this approach can result in poor network capacity utilization, since for many streams much more traffic could still be admitted than this actual amount. Thus, one of the alternative approaches is to admit as much traffic as possible from every stream while making the smaller admitted amounts as large as possible.

The problem to determine how much traffic of every traffic stream should be admitted into the network, and how the admitted traffic should be routed through the network so as to satisfy the requirements of high network utilization and to guarantee fairness to the users, is one of the most challenging problems of current telecommunications networks design. In this paper we show how this problem is related to two well known OR problems – namely the max-min optimization problem and the lexicographic optimization problem. We study the general formulations of these problems and analyze how to use their notions to express the fairness of the traffic admission process. We go on to formulate basic network design problems and study the complexity of the obtained formulations. We analyze the methods of max-min and lexicographic optimization and examine how they can be applied to solve the presented network design problem.

The paper is organized as follows. In Section 2 we introduce the lexicographic max-min or the max-min fair (MMF) solution concept and summarize its major properties. In Section 3 we present details of three telecommunications problems leading to MMF formulations. Further in Section 4 we discuss solution algorithms for the lexicographic max-min optimization and analyze their applicability for telecommunications problems.

2. Max-min and the MMF concept

2.1. Max-min solution concepts

The problem we consider may be viewed in terms of resource allocation decisions as follows. Let us assume there is a set of m services. There is also a set Q of resource allocation patterns (allocation decisions). For each service j a function $f_j(\mathbf{x})$ of allocation pattern $\mathbf{x}$ has been defined. This function, called the individual objective function, measures the outcome (effect) $y_j = f_j(\mathbf{x})$ of the allocation pattern for service j. The outcomes can be measured (modeled) as service quality, service amount, service time, service costs as well as in a more subjective way the (client's) utility of the provided service. In typical formulations a greater value of the outcome means a better effect (higher service quality or client satisfaction); otherwise, the outcomes can be replaced with their complements to some large number. Therefore, without loss of generality, we can assume that each individual outcome y_j is to be maximized which results in a multiple criteria maximization model. The problem can be formulated as follows:

$$\max\ \{\mathbf{f}(\mathbf{x}) : \mathbf{x} \in Q\}, \tag{1}$$

where $Q \subseteq \Re^n$ is a feasible set and $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ is a vector of real-valued functions $f_j : Q \rightarrow \Re, j = 1, 2, \ldots, m$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is an n-vector. We refer to the elements of the criterion space as outcome vectors. An outcome vector $\mathbf{y}$ is attainable if it expresses outcomes of a feasible solution $\mathbf{x} \in Q$ (i.e., $\mathbf{y} = \mathbf{f}(\mathbf{x})$). The set of all the attainable outcome vectors is denoted by Y. Note that, in general, convex feasible set Q and concave function $\mathbf{f}$ do not guarantee convexity of the corresponding attainable set Y. Nevertheless, the multiple criteria maximization model (1) can be rewritten in the equivalent form

$$\max\ \{\mathbf{y} : y_j \leq f_j(\mathbf{x})\ \forall j,\ \mathbf{x} \in Q\}, \tag{2}$$

where the attainable set Y is convex whenever Q is convex and functions f_j are concave.

Model (1) only specifies that we are interested in maximization of all objective functions f_j for $j \in M = \{1, 2, \ldots, m\}$. Each attainable outcome vector $\mathbf{y} \in Y$ is called *nondominated* if one cannot improve any individual outcome without worsening another one. Each feasible solution $\mathbf{x} \in Q$ generating the nondominated outcome is called an *efficient* (Pareto-optimal) solution of the multiple criteria problem (1). In other words, a feasible solution for which one cannot improve any outcome without worsening another is efficient [33]. In order to make model (1) operational, one needs to assume some solution concept specifying what it means to maximize multiple objective functions. Simple solution concepts are defined by achievement functions $\theta : Y \rightarrow \Re$ to be maximized. Thus the multiple criteria problem (1) is replaced with the aggregation $\max\ \{\theta(\mathbf{f}(\mathbf{x})) :\ \mathbf{x} \in Q\}$.

The most commonly used achievement function is the mean (or simply the sum) of individual performances; this defines the so-called maxsum solution concept. This solution concept is primarily concerned with the overall system efficiency. As based on averaging, it often provides a solution where some services are discriminated in terms of performances. An alternative approach depends on the so-called max-min solution concept, where the worst performance is maximized:

$$\max\{\min_{j=1,\ldots,m} f_j(\mathbf{x})\ :\ \mathbf{x} \in Q\ \}. \tag{3}$$

The max-min solution concept has been widely studied in the multi-criteria optimization methodology [33, 35]. The optimal set of the max-min problem (3) always contains an efficient solution of the original multiple criteria problem (1). Thus, if unique, the optimal max-min solution is efficient. In the case of multiple optimal solutions, one of them is efficient but also some of them may not be efficient. It is a serious flaw since practical large problems usually have multiple optimal solutions and typical optimization solvers generate one of them (essentially at random). Therefore, some additional regularization is needed to overcome this flaw of the max-min scalarization.

The max-min solution concept is regarded as maintaining equity. Indeed, in the case of a simplified resource allocation problem, the max-min solution

$$\max\{\min_{j=1,\ldots,m} y_j\ :\ \sum_{j=1}^{m} y_j \leq b\ \} \tag{4}$$

takes the form $\bar{y}_j = b/m$ for all $j \in M$ thus meeting the perfect equity requirement $\bar{y}_1 = \bar{y}_2 = \ldots = \bar{y}_m$. In the general case, with possibly more complex feasible set structure, this property is not fulfilled [23]. Nevertheless, the following assertion is valid.

Theorem 1: If there exists a nondominated outcome vector $\bar{\mathbf{y}} \in Y$ satisfying the perfect equity requirement $\bar{y}_1 = \bar{y}_2 = \ldots = \bar{y}_m$, then $\bar{\mathbf{y}}$ is the unique optimal solution of the max-min problem

$$\max\{\min_{j=1,\ldots,m} y_j\ :\ \mathbf{y} \in Y\ \}. \tag{5}$$

Proof: Let $\bar{\mathbf{y}} \in Y$ be a nondominated outcome vector satisfying the perfect equity requirement. This means, there exists a number α such that $\bar{y}_j = \alpha$ for $j = 1, 2, \ldots, m$. Let $\mathbf{y} \in Y$ be an optimal solution of the max-min problem (5). Suppose, there exists some index j_0 such that $y_{j_0} \neq \bar{y}_{j_0}$. Due to the optimality of $\mathbf{y}$, we have:

$$y_j \geq \min_{1 \leq i \leq m} y_i \geq \min_{1 \leq i \leq m} \bar{y}_i = \alpha = \bar{y}_j \quad \forall\ j = 1, \ldots, m$$

which together with $y_{j_0} \neq \bar{y}_{j_0}$ contradicts the assumption that $\bar{\mathbf{y}}$ is nondominated. ■

According to Theorem 1, the perfectly equilibrated outcome vector is a unique optimal solution of the max-min problem if one cannot improve any of its individual outcome without worsening some others. Unfortunately, it is not a common case and, in general, the optimal set to the max-min aggregation (3) may contain numerous alternative solutions including dominated ones. While using standard algorithmic tools to identify the max-min solution, one of many solutions is then selected randomly.

Actually, the distribution of outcomes may make the max-min criterion partially passive when one specific outcome is relatively very small for all the solutions. For instance, while allocating clients to service facilities, such a situation may be caused by existence of an isolated client located at a considerable distance from all the location of facilities. Maximization of the worst service performances (equivalent to minimization of the maximum distance) is then reduced to maximization of the service performances for that single isolated client leaving other allocation decisions unoptimized. This is a clear case of inefficient solution where one may still improve other outcomes while maintaining fairness by leaving at its best possible value the worst outcome. The max-min solution may be then regularized according to the Rawlsian principle of justice. Rawls [30] considers the problem of ranking different “social states”which are different ways in which a society might be organized taking into account the welfare of each individual in each society, measured on a single numerical scale [30, p. 62]. Applying the Rawlsian approach, any two states should be ranked according to the accessibility levels of the least well–off individuals in those states; if the comparison yields a tie, the accessibility levels of the next–least well–off individuals should be considered, and so on. Formalization of this concept leads us to the lexicographic max-min concepts.

The lexicographic max-min solution is known in the game theory as the nucleolus of a matrix game. It originates from an idea, presented by Dresher [7], to select from the optimal (max-min) strategy set of a player a subset of optimal strategies which exploit mistakes of the opponent optimally. It has been later refined to the formal nucleolus definition [32] and generalized to an arbitrary number of objective functions [29]. The concept was early considered in the Tschebyscheff approximation [31] as a refinement taking into account the second largest deviation, the third one and further to be hierarchically minimized. Similar refinement of the fuzzy set operations has been recently analyzed [8]. Within the telecommunications or network applications the lexicographic max-min approach has appeared already in [3, 11] and now under the name max-min fair is treated as one of the standard fairness concepts. The approach has been used for general linear programming multiple criteria problems [1, 17], as well as for specialized problems related to (multiperiod) resource allocation [12, 16]. In discrete optimization it has been considered for various problems [4, 5] including the location-allocation ones [21].

2.2. Lexicographic optimization and MMF

Typical solution concepts for the multiple criteria problems are based on the use of aggregated achievement functions $\theta : Y \rightarrow \Re$ to be maximized, thus ranking the outcomes according to a complete preorder

$$\mathbf{y}' \succeq_\theta \mathbf{y}'' \quad \Leftrightarrow \quad \theta(\mathbf{y}') \geq \theta(\mathbf{y}''). \tag{6}$$

This allows one to replace the multiple criteria problem (1) with the maximization problem max $\{\theta(\mathbf{f}(\mathbf{x})) : \ \mathbf{x} \in Q\}$. However, there are well defined solution concepts which do not introduce directly any scalar measure, despite they rank the outcome vectors with a complete preorder. Especially, the lexicographic order is used for this purpose.

Let $\mathbf{a} = (a_1, a_2, \ldots, a_m)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_m)$ be two m-vectors. Vector $\mathbf{a}$ is lexicographically greater than vector $\mathbf{b}$, $\mathbf{a} >_{lex} \mathbf{b}$, if there exists index k, $0 \leq k < m$, such that $a_j = b_j$ for all $j \leq k$ and $a_{k+1} > b_{k+1}$. Consequently, $\mathbf{a}$ is lexicographically greater or equal $\mathbf{b}$, $\mathbf{a} \geq_{lex} \mathbf{b}$, if $\mathbf{a} >_{lex} \mathbf{b}$ or $\mathbf{a} = \mathbf{b}$. Contrary to the standard vector inequality $\mathbf{a} \geqq \mathbf{b} \Leftrightarrow a_j \geq b_j \forall \ j$, the lexicographic order is complete which means that for any two vectors $\mathbf{a}$ and $\mathbf{b}$ either $\mathbf{a} \geq_{lex} \mathbf{b}$ or $\mathbf{b} \geq_{lex} \mathbf{a}$. Moreover, for any two different vectors $\mathbf{a} \neq \mathbf{b}$ either $\mathbf{a} >_{lex} \mathbf{b}$ or $\mathbf{b} >_{lex} \mathbf{a}$. Vector inequality $\mathbf{a} \geqq \mathbf{b}$ implies $\mathbf{a} \geq_{lex} \mathbf{b}$ but the opposite implication is not valid. The lexicographic order is not continuous and it cannot be expressed in terms of any aggregation function. Nevertheless, it is a limiting case of the order (6) for the weighting aggregation functions $\theta(\mathbf{y}) = \sum_{j=1}^{m} \ w_j y_j$ defined by decreasing sequences of positive weights w_j with differences tending to the infinity.

The lexicographic order allows us to consider more complex solution concepts defined by several (say m) outcome functions $\theta_k : Y \rightarrow \Re$ to be maximized according to the lexicographic order. Thus one seeks a feasible solution $\mathbf{x}^0$ such that for all $\mathbf{x} \in Q$

$$(\theta_1(\mathbf{f}(\mathbf{x}^0)), \ldots, \theta_m(\mathbf{f}(\mathbf{x}^0))) \geq_{lex} (\theta_1(\mathbf{f}(\mathbf{x})), \ldots, \theta_m(\mathbf{f}(\mathbf{x}))).$$

In other words, the multiple criteria problem (1) is replaced with the lexicographic maximization problem

$$\text{lex}\max \ \{(\theta_1(\mathbf{f}(\mathbf{x})), \theta_2(\mathbf{f}(\mathbf{x})), \ldots, \theta_m(\mathbf{f}(\mathbf{x}))) : \mathbf{x} \in Q\}. \tag{7}$$

Problem (7) is not a standard mathematical programme. Nevertheless, the lexicographic inequality defines a linear order of vectors an therefore the lexicographic optimization is a well defined procedure where comparison of real numbers is replaced by lexicographic comparison of the corresponding vectors. In particular, the basic theory and algorithmic techniques for linear programming have been extended to the lexicographic case [10]. Certainly, the lexicographic optimization may also be treated as a sequential (hierarchical) optimization process where first $\theta_1(\mathbf{f}(\mathbf{x}))$ is maximized on the entire feasible set, next $\theta_2(\mathbf{f}(\mathbf{x}))$ is maximized on the optimal set, and so on. This may be implemented as in the following standard sequential algorithm.

Algorithm 1: Sequential lexicographic maximization

```
Step 0:  Put $k := 1$.
Step 1:  Solve programme $P_k$:
         $\max_{\mathbf{x} \in Q} \{\tau_k;\ \tau_k \leq \theta_k(\mathbf{f}(\mathbf{x})),\ \tau_j^0 \leq \theta_j(\mathbf{f}(\mathbf{x}))\ \forall j < k\}$
         and denote the optimal solution by $(\mathbf{x}^0, \tau_k^0)$.
Step 2:  If $k = m$, then stop ($\mathbf{x}^0$ is optimal solution).
         Otherwise, put $k := k+1$ and go to Step 1.
```

Note that directly from the properties of the lexicographic order it follows that for any achievement functions θ_k the lexicographic optimization problem always has unique values of those functions, as stated in the following assertion.

Theorem 2: For any two optimal solutions $\mathbf{x}^1, \mathbf{x}^2 \in Q$ of problem (7) the equalities $\theta_k(\mathbf{f}(\mathbf{x}^1)) = \theta_k(\mathbf{f}(\mathbf{x}^2))\ \forall\ k$ hold.

The most commonly used lexicographic models are based on simple functions $\theta_j(\mathbf{y}) = y_j$ thus introducing an hierarchy of original outcomes. In such a case, according to Theorem 2 the optimal solution is unique in the criterion space.

Theorem 3: In the case of problem (7) with $\theta_j(\mathbf{y}) = y_j \forall j \in M$, for any two optimal solutions $\mathbf{x}^1, \mathbf{x}^2 \in Q$ the equality $\mathbf{f}(\mathbf{x}^1) = \mathbf{f}(\mathbf{x}^2)$ holds and this unique outcome vector is nondominated.

Applying to achievement vectors $\Theta(\mathbf{y})$ a linear cumulative map one gets the cumulated achievements

$$\bar{\theta}_k(\mathbf{y}) = \sum_{j=1}^{k} \theta_j(\mathbf{y}) \quad \text{for } k = 1, 2, \ldots, m. \tag{8}$$

Note that for any two vectors $\mathbf{y}', \mathbf{y}'' \in Y$ one gets

$$\Theta(\mathbf{y}') \geq_{lex} \Theta(\mathbf{y}'') \quad \Leftrightarrow \quad \bar{\Theta}(\mathbf{y}') \geq_{lex} \bar{\Theta}(\mathbf{y}''). \tag{9}$$

Hence, the following assertion is valid.

Theorem 4: A feasible vector $\mathbf{x} \in Q$ is an optimal solution of problem (7), if and only if it is the optimal solution of the cumulated lexicographic problem

$$\operatorname{lex}\max\ \{(\bar{\theta}_1(\mathbf{f}(\mathbf{x})), \ldots, \bar{\theta}_m(\mathbf{f}(\mathbf{x}))) : \mathbf{x} \in Q\}. \tag{10}$$

The lexicographic order may also be used to construct refinements of various solution concepts [23]. We focus on application of the lexicographic optimization to refine the max-min solution concept according to the Rawlsian theory of justice. Let $\langle \mathbf{a} \rangle = (a_{\langle 1 \rangle}, a_{\langle 2 \rangle}, \ldots, a_{\langle m \rangle})$ denote the vector obtained from $\mathbf{a}$ by rearranging its components in the nondecreasing order. That means $a_{\langle 1 \rangle} \leq a_{\langle 2 \rangle} \leq \ldots \leq a_{\langle m \rangle}$ and there exists a permutation π of set M such that $a_{\langle i \rangle} = a_{\pi(i)}$ for $j = 1, \ldots, m$. Comparing lexicographically such ordered vectors $\langle \mathbf{y} \rangle$ one gets the so-called leximin order. The general problem considered in the balance of this paper depends on searching for the solutions that are maximal according to the leximin order. The problem called hereafter the max-min fair problem reads as follows:

P-MMF: Find $\mathbf{x}^0 \in Q$ such that $\langle \mathbf{f}(\mathbf{x}^0) \rangle \geq_{lex} \langle \mathbf{f}(\mathbf{x}) \rangle\ \forall\, \mathbf{x} \in Q$.

This problem may also be viewed as a standard lexicographic optimization (7) with the aggregation functions $\theta_j(\mathbf{y}) = y_{\langle j \rangle}$:

$$\operatorname{lex}\max\ \{(\theta_1(\mathbf{f}(\mathbf{x})), \ldots, \theta_m(\mathbf{f}(\mathbf{x}))) : \mathbf{x} \in Q\}. \tag{11}$$

Problem (11) represents the lexicographic max-min approach to the original multiple criteria problem (1). It is a refinement (regularization) of the standard max-min optimization, but this time, in addition to the smallest outcome, we also maximize the second smallest outcome (provided that the smallest one remains as large as possible), maximize the third smallest (provided that the two smallest remain as large as possible), and so on. Note that the lexicographic maximization is not applied to any specific order of the original criteria.

The lexicographic max-min is the only regularization approach of the max-min that satisfies the reduction (addition/deleting) principle [9]. Namely, if the individual outcome does not distinguish two solutions, then it does not affect the preference relation.

For the lexicographic max-min one may also take advantage of Theorem 4. Applying the cumulative map (8) to ordered outcomes $\theta_i(\mathbf{y}) = y_{\langle i \rangle}$ one gets $\bar{\theta}_k(\mathbf{y}) = \sum_{i=1}^{k} y_{\langle i \rangle}$ expressing, respectively: the worst (smallest) outcome, the total of the two worst outcomes, the total of the three worst outcomes, etc. Following Theorem 4, solution of the P-MMF is equivalent to the lexicographic problem

$$\begin{aligned} &\operatorname{lex}\max\ \{(\bar{\theta}_1(\mathbf{y}), \ldots, \bar{\theta}_m(\mathbf{y})) : \mathbf{y} \overset{\leq}{=} \mathbf{f}(\mathbf{x}),\ \mathbf{x} \in Q\}, \\ &\text{where } \bar{\theta}_k(\mathbf{y}) = \textstyle\sum_{j=1}^{k} y_{\langle j \rangle}. \end{aligned} \tag{12}$$

Note that

$$\bar{\theta}_k(\mathbf{y}) = \sum_{j=1}^{k} y_{\langle j \rangle} = \min_{\pi \in \Pi} \sum_{j=1}^{k} y_{\pi(j)},$$

where the minimum is taken over all permutations of the index set M. Hence, $\bar{\theta}_k(\mathbf{y})$ is a concave piecewise linear function of $\mathbf{y}$ which, due to (12) guarantees several important properties of the lexicographic max-min solution itself.

Recall, that every optimal solution of the lexicographic max-min model is an efficient solution of the original multiple criteria optimization problem. Note that every lexicographic max-min solution is also an optimal solution of the standard max-min problem. Hence, by virtue of Theorem 1, the lexicographic max-min model, generates efficient solutions satisfying the perfect equity of individual outcomes, whenever such an efficient solution exists. When there does not exist any efficient solution with perfectly equal individual outcomes, then the lexicographic max-min model generates another efficient solution but, due to concave functions $\bar{\theta}_k(\mathbf{y})$, still providing equitability of individual outcomes with respect to the Pigou-Dalton principle of transfers [14]. The principle of transfers states, in the context considered here, that a transfer of small amount from an individual outcome to any relatively worse-off individual outcome results in a more preferred outcome vector. Indeed, the following assertion is valid.

Theorem 5: For any outcome vector $\mathbf{y} \in Y$, $y_{j'} < y_{j''}$ implies

$$\langle \mathbf{y} + \varepsilon \mathbf{e}_{j'} - \varepsilon \mathbf{e}_{j''} \rangle >_{lex} \langle \mathbf{y} \rangle \quad \forall\ 0 < \varepsilon < y_{j''} - y_{j'}\,, \tag{13}$$

where $\mathbf{e}_j$ denotes the jth unit vector.

Proof: Let $\mathbf{y}^\varepsilon = \mathbf{y} + \varepsilon \mathbf{e}_{j'} - \varepsilon \mathbf{e}_{j''}$ for $\varepsilon < y_{j''} - y_{j'}$ and let $y_{\langle k' \rangle} = y_{j'}$, $y_{\langle k'' \rangle} = y_{j''}$. Then, $y_{j'} < y_{\langle k'' \rangle}$ and $\sum_{j=1}^{k} y^\varepsilon_{\langle j \rangle} \geq \sum_{j=1}^{k} y_{\langle j \rangle}$ for all $k = 1, 2, \ldots, m$ with at least one strict inequality for some $k' \leq k < k''$. Hence, $\langle \mathbf{y}^\varepsilon \rangle >_{lex} \langle \mathbf{y} \rangle$, due to (9). ■

Following Theorem 2, any two optimal solutions $\mathbf{x}^1, \mathbf{x}^2 \in Q$ of problem (11) result in the same ordered outcome vectors $\langle \mathbf{f}(\mathbf{x}^1) \rangle = \langle \mathbf{f}(\mathbf{x}^2) \rangle$. Hence, all the optimal solutions have the same distributions of outcomes. Nevertheless, they may generate different (differently ordered) outcome vectors themselves. The unique outcome vector is guaranteed, however, in the case of convex problems. It follows from the alternative convex formulation (12) of the MMF problem.

Theorem 6: In the case of convex feasible set Q and concave objective functions $f_j(\mathbf{x})$, for any two optimal solutions $\mathbf{x}^1, \mathbf{x}^2 \in Q$ of problem P-MMF the equality $\mathbf{f}(\mathbf{x}^1) = \mathbf{f}(\mathbf{x}^2)$ holds.

Proof: First of all, let us notice that problem P-MMF is equivalent (in the criterion space) to the following:

$$\operatorname{lex\,max}\ \{ \langle \mathbf{y} \rangle : y_j \leq f_j(\mathbf{x})\ \forall j,\ \mathbf{x} \in Q \} \tag{14}$$

and we need to prove that the problem has a unique optimal solution $\mathbf{y} \in Y$. Due to the convexity assumptions the attainable set Y is convex. Let, $\mathbf{y}^1 \neq \mathbf{y}^2 \in Y$ be optimal solutions of (14), thus $\langle \mathbf{y}^1 \rangle = \langle \mathbf{y}^2 \rangle$. Define $\mathbf{y}^\varepsilon = (1-\varepsilon)\mathbf{y}^1 + \varepsilon \mathbf{y}^2$ for some positive ε satisfying

$$0 < \varepsilon < \min_{y^1_{j'} \neq y^1_{j''}} |y^1_{j'} - y^1_{j''}| / \max_{y^1_{j'} \neq y^1_{j''}} |y^1_{j'} - y^1_{j''}|.$$

Due to the bound on ε, there exists a permutation π ordering both $\mathbf{y}^1$ and $\mathbf{y}^\varepsilon$, i.e., $y^1_{\pi(j)} \leq y^1_{\pi(j+1)}$ and $y^\varepsilon_{\pi(j)} \leq y^\varepsilon_{\pi(j+1)}$ for all $j = 1, \ldots, m-1$. Further, identifying the index j_o for which $y^1_{j_o}$ is the smallest value y^1_j such that $y^1_j \neq y^2_j$ one gets $y^\varepsilon_{\pi(j)} \geq y^1_{\pi(j)}$ for $j < j_o$ and $y^\varepsilon_{\pi(j_o)} > y^1_{\pi(j_o)}$ which contradicts optimality of $\mathbf{y}^1$. ■

The leximin order cannot be expressed in terms of any aggregation function. Nevertheless, it is a limiting case of the order (6) for the ordered weighted aggregation functions $\theta(\mathbf{y}) = \sum_{j=1}^{m} w_j y_{\langle j \rangle}$ defined by decreasing sequences of positive weights w_j with differences tending to the infinity [36, 38].

3. Telecommunications network design examples

Below we shall give three examples showing how the MMF concept can be used in formulations of multi-commodity network flow problems related to telecommunications applications.

3.1. Routing design for networks with elastic traffic

The first example is a problem of finding flows in a network with given link capacities so as to obtain the MMF distribution of flow sizes. This type of problem is applicable to networks carrying the so-called elastic traffic, which means that traffic streams can adapt their intensity to the available capacity of the network [28].

Problem 1: *Routing optimization for MMF distribution of demand volumes*

indices

$d = 1, 2, \ldots, D$	demands (pairs of nodes)
$p = 1, 2, \ldots, P_d$	allowable paths for demand d
$e = 1, 2, \ldots, E$	links

constants

δ_{edp}	equals 1 if link e belongs to path p of demand d; 0, otherwise
c_e	capacity of link e

variables

x_{dp}	flow (bandwidth) allocated to path p of demand d (non-negative continuous)
X_d	total flow (bandwidth) allocated to demand d (non-negative continuous), $X = (X_1, X_2, \ldots, X_D)$

objective

$$\operatorname{lex\,max}\ (X_{\langle 1 \rangle}, X_{\langle 2 \rangle}, \ldots, X_{\langle D \rangle}) \tag{15a}$$

constraints

$$\textstyle\sum_p x_{dp} = X_d \quad d = 1, 2, \ldots, D, \tag{15b}$$

$$\textstyle\sum_d \sum_p \delta_{edp} x_{dp} \leq c_e \quad e = 1, 2, \ldots, E, \tag{15c}$$

$$x_{dp} \geq 0 \qquad d = 1, 2, \ldots, D \quad p = 1, 2, \ldots, P_d. \tag{15d}$$

In the above formulation, Eq. (15b) defines the total flow, X_d, allocated to demand d, and constraint (15c) assures that the link load (left-hand side) does not exceed the link capacity. A solution of Problem 1 for an example network is discussed in Appendix A.

3.2. Restoration design for networks with elastic traffic

The second example corresponds to the problem of designing an optimal strategy of elastic traffic flows restoration in case of network failures ([27, Chapter 13]). It is assumed that a set of network failure situations have been identified. The adopted failure model is such that a failure may reduce the capacity of one or more network links. The design should determine optimal capacities of links and for each failure situation the optimal size and routing of every traffic flow so as to obtain the MMF distribution of revenue for all network failure situations. It is assumed that the revenue generated by a single traffic flow is proportional to the logarithm of this flow's size.

Problem 2: *Flow restoration optimization for MMF distribution of revenues*

indices

$d = 1, 2, \ldots, D$ demands
$p = 1, 2, \ldots, P_d$ allowable paths for demand d
$e = 1, 2, \ldots, E$ links
$s = 1, 2, \ldots, S$ states (including normal state)

constants

δ_{ed} equals 1 if link e belongs to the fixed path of demand d; 0, otherwise
r_{ds} revenue from demand d in situation s
ξ_e unit cost of link e
α_{es} fractional availability coefficient of link e in situation s ($0 \le \alpha_{es} \le 1$)
B assumed budget

variables

y_e capacity of link e (non-negative continuous)
x_{dps} flow allocated to path p of demand d in situation s (non-negative continuous)
X_{ds} total flow allocated to demand d in situation s (non-negative continuous)
R_s logarithmic revenue in situation s (continuous), $R = (R_1, R_2, \ldots, R_S)$

objective

$$\operatorname{lex\,max}\ (R_{\langle 1\rangle}, R_{\langle 2\rangle}, \ldots, R_{\langle S\rangle}) \tag{16a}$$

constraints

$$X_{ds} = \textstyle\sum_p x_{dps} \qquad d = 1, \ldots, D;\ s = 1, \ldots, S, \tag{16b}$$

$$R_s = \textstyle\sum_d r_{ds} \lg X_{ds} \qquad s = 1, \ldots, S, \tag{16c}$$

$$\textstyle\sum_d \sum_p \delta_{edp} x_{dps} \le \alpha_{es} y_e \qquad e = 1, \ldots, E, \tag{16d}$$

$$\textstyle\sum_e \xi_e y_e \le B, \tag{16e}$$

$$x_{dps} \ge 0 \tag{16f}$$

$$d = 1, \ldots, D;\ p = 1, \ldots, P_d;\ s = 1, \ldots, S.$$

3.3. Capacity protection design

The last example corresponds to the problem of designing the protection of network links' capacity [20]. It is assumed that the capacity of network links and the size and routing of all network flows are given. The design should determine how much capacity of each link should be freed and reserved so in case of any single-link failure the capacity of the failed link could be restored using the reserved protection capacity. In order to free the capacity of links the size of traffic flows should be reduced in such a way so as to obtain the MMF distribution of traffic flow sizes.

Problem 3: *Protection capacity optimization for MMF distribution of flow sizes*

indices

$d = 1, 2, ..., D$ demands
$p = 1, 2, ..., P_d$ allowable paths for demand d
$e, \ell = 1, 2, ..., E$ links
$q = 1, 2, ..., Q_\ell$ candidate restoration paths for link ℓ

constants

h_d "reference" volume of demand d
δ_{edp} equals 1 if link e belongs to path p realizing demand d; 0, otherwise
c_e total capacity of link e
$\beta_{\ell eq}$ equals 1 if link ℓ belongs to path q restoring link e; 0, otherwise

variables

y_e resulting normal capacity of link e
x_{dp} normal flow realizing demand d on path p
w_e protection capacity of link e
z_{eq} flow restoring capacity of link e on path q
X_d normalized realized demand volume for demand d, $X = (X_1, X_2, \ldots, X_D)$

objective

$$\operatorname{lex\,max}\ (X_{\langle 1\rangle}, X_{\langle 2\rangle}, \ldots, X_{\langle D\rangle}) \tag{17a}$$

constraints

$$X_d = \textstyle\sum_p x_{dp}/h_d \quad d = 1, \ldots, D, \tag{17b}$$

$$w_e + u_e \le c_e \quad e = 1, \ldots, E, \tag{17c}$$

$$\textstyle\sum_d \sum_p \delta_{edp} x_{dp} \le y_e, \quad e = 1, \ldots, E, \tag{17d}$$

$$y_e \le \textstyle\sum_q z_{eq} \quad e = 1, \ldots, E, \tag{17e}$$

$$\textstyle\sum_q \beta_{\ell eq} z_{eq} \le w_\ell \quad \ell, e = 1, \ldots, E;\ \ell \ne e, \tag{17f}$$

$$x_{dp} \ge 0 \qquad d = 1, \ldots, D \quad p = 1, \ldots, P_d. \tag{17g}$$

Note that the lexicographic max-min solution assures that all demand volumes will be in the worst case decreased by the same optimal proportion r^*, since in the optimal solution $\sum_p x^*_{dp} \ge r^* h_d$, $d = 1, 2, ..., D$, for some number r^*, such that $\sum_p x^*_{dp} = r^* h_d$ for some d.

3.4. Non-convex extensions of the example problems

All three problems presented in the previous subsections have convex sets of feasible solutions. As we will see in Section 4, this property allows for efficient solution algorithms of the introduced problems, but, unfortunately, it is not always present in telecommunications problems. For instance, we may require that the demand volumes are realized only on single paths and that the choice of these single paths is subject to optimization. This requirement usually leads to mixed-integer programme (MIP) formulations. In particular, Problem 1 in the single-path version requires additional multiple choice constraints to enforce nonbifurcated flows. Assuming existence of some constants U_d upper bounding the largest possible total flows X_d, this can be implemented with additional binary (flow assignement) variables u_{dp} used to limit the number of positive flows x_{dp} with constraints:

$$x_{dp} \leq U_d u_{dp} \qquad d = 1,\dots,D;\; p = 1,\dots,P_d, \tag{18a}$$

$$\sum_p u_{dp} = 1 \qquad d = 1,\dots,D, \tag{18b}$$

$$u_{dp} \in \{0,1\} \qquad d = 1,\dots,D;\; p = 1,\dots,P_d\ . \tag{18c}$$

In fact, as demonstrated in [13], such a modification makes Problem 1 *NP*-complete. The same requirement can be introduced to Problems 2 and 3 as well.

Another requirement leading to non-convex MIP problems is the modularity of the link capacity, which means that link capacities should be multiples of a given module C. Then, capacity variables become non-negative integers and respective constraints change. For example, for Problem 2 variables y_e are non-negative integers and constraints (16d) take the form

$$\sum_d \sum_p \delta_{edp} x_{dps} \leq \alpha_{es} C y_e, \qquad e = 1,\dots,E. \tag{19}$$

Certainly, the capacity variables in Problem 3 can also be made integral.

4. MMF solution algorithms

4.1. Sequential max-min algorithms for convex problems

The (point-wise) ordering of outcomes causes that the lexicographic max-min problem (11) is, in general, hard to implement. Note that the quantity $y_{\langle 1 \rangle}$ representing the worst outcome can be easily computed directly by the maximization:

$$y_{\langle 1 \rangle} = \max\ r_1 \quad \text{subject to} \quad r_1 \leq y_j \quad \text{for } j = 1,\dots,m.$$

Similar simple formula does not exist for the further ordered outcomes $y_{\langle k \rangle}$. Nevertheless, for convex problems it is possible to use iterative algorithms for finding the consecutive values of the (unknown) optimal unique vector $\mathbf{T}^0 = (T_1^0, T_2^0, \dots, T_m^0) = \langle \mathbf{f}(\mathbf{x}^0) \rangle$ by solving a sequence of properly defined max-min problems. Such algorithms are described below.

Suppose B is a subset of the index set M, $B \subseteq M$, and let $\mathbf{t}^B = (t_j : j \in B)$ be a $|B|$-vector. Also, let B' denote the set complementary to B: $B' = M \backslash B$. For given B and $\mathbf{t}^B$ we define the following convex mathematical programming problem in variables $\mathbf{x}$ and τ:

$\mathbf{P}(B,\mathbf{t}^B)$:

$$\begin{aligned} &\textbf{maximize} && \tau, && && (20a)\\ &\textbf{subject to} && f_j(\mathbf{x}) \geq \tau && j \in B', && (20b)\\ & && f_j(\mathbf{x}) \geq t_j^B && j \in B, && (20c)\\ & && \mathbf{x} \in \mathbf{X}. && && (20d) \end{aligned}$$

It is clear that the solution τ^0 of the convex problem $\mathbf{P}(\emptyset,\emptyset)$ (defined by (20) for empty set B and empty sequence $\mathbf{t}^B$) will yield the smallest value of $\mathbf{T}^0$, i.e., the value T_1^0 (and possibly some other consecutive entries of $\mathbf{T}^0$). This observation suggests the following algorithm for solving problem P-MMF specified by (11).

Algorithm 2: Straightforward algorithm for solving problem P-MMF

```
Step 0:  Put $B := \emptyset$ (empty set) and $\mathbf{t}^B := \emptyset$ (empty sequence).

Step 1:  If $B = M$ then stop ($\langle \mathbf{t}^B \rangle$ is the optimal solution
         of problem P-MMF, i.e., $\langle \mathbf{t}^B \rangle = \mathbf{T}^0$). Else, solve
         programme $\mathbf{P}(B,\mathbf{t}^B)$ and denote the resulting
         optimal solution by $(\mathbf{x}^0, \tau^0)$.

Step 2:  For each index $k \in B'$ such that $f_k(\mathbf{x}^0) = \tau^0$
         solve the following test problem $\mathbf{T}(B,\mathbf{t}^B,\tau^0,k)$:

             $\mathbf{max},\ f_k(\mathbf{x}),$                                  (21a)
             $\mathbf{s.t.}\ \ f_j(\mathbf{x}) \geq \tau^0 \quad j \in B' \backslash \{k\},$   (21b)
                   $f_j(\mathbf{x}) \geq t_j^B \quad j \in B,$                     (21c)
                   $\mathbf{x} \in \mathbf{X}.$                                (21d)

         If for optimal $\mathbf{x}^1$, while solving test
         $\mathbf{T}(B,\mathbf{t}^B,\tau^0,k)$ we have $f_k(\mathbf{x}^1) = \tau^0$, then
         we put $B := B \cup \{k\}$ and $t_k := \tau^0$.

Step 3:  Go to Step 1.
```

It can happen that as a result of solving the test in Step 2 for some index $k \in B'$, it will turn out that $f_l(\mathbf{x}^1) > \tau^0$ for some other, not yet tested, index $l \in B'$ $(l \neq k)$. In such an (advantageous) case, the objective function with index l does not have to be tested, as its value can be further increased without disturbing the maximal values $\mathbf{t}^B$. Observe that set B is the current set of blocking indices, i.e., the indices j for which the value $f_j(\mathbf{x}^0)$ is equal to t_j^B in every optimal solution of problem P-MMF. Note also, that although the tests in Step 2 are performed separately for individual indices $j \in B'$, the values of objective functions f_j for the indices $j \in B'$, where set B' is results from Step 2, can be simultaneously increased above the value of τ^0 in the next execution of Step 1. This follows from convexity of the set defined by constraints (21b–d): if $f_j(\mathbf{x}^j) = a^j > \tau^0$ and $\mathbf{x}^j$ satisfies (21b–d), then a convex combination of the points $\mathbf{x}^j$, $\mathbf{x} = \sum_{j \in B'} \alpha^j \mathbf{x}^j$ $(\sum_{j \in B'} \alpha^j = 1$, $\alpha^j > 0$, $j \in B')$ also satisfies (21b–d), and $f_j(\mathbf{x}) > \tau^0$ for all $j \in B'$.

Another version of Algorithm 2 may be more efficient, provided that the complexity of problems (20) and (21) is similar.

Algorithm 3: Algorithm for solving problem P-MMF

```
Step 0:  Put $B := \emptyset$ and $\mathbf{t}^B := \emptyset$.
Step 1:  If $B = M$ then stop ($\langle \mathbf{t}^B \rangle$ is the optimal solution of problem P-MMF, i.e., $\langle \mathbf{t}^B \rangle = \mathbf{T}^0$).
         Else, solve programme $\mathbf{P}(B, \mathbf{t}^B)$ and denote the resulting optimal solution by $(\mathbf{x}^0, \tau^0)$.
Step 2:  Start solving the test problem $\mathbf{T}(B, \mathbf{t}^B, \tau^0, k)$ for all indices $k \in B'$ such that $f_k(\mathbf{x}^0) = \tau^0$.
         When the first $k \in B'$ with $f_k(\mathbf{x}^1) = \tau^0$ is detected, then put $B := B \cup \{k\}$ and $t_k := \tau^0$, and go to Step 3.
Step 3:  Go to Step 1.
```

The idea behind the modification in Algorithm 3 is that in total it may involve solving less instances of problems $\mathbf{P}(B,\mathbf{t}^B)$ and $\mathbf{T}(B,\mathbf{t}^B,\tau^0,k)$ than Algorithm 2. If at optimum $\mathbf{x}^0$ all values $f_j(\mathbf{x}^0)$ are the same (equal to 0), then Algorithm 2 will require solving $m+1$ problems (problem $\mathbf{P}(\emptyset,\emptyset)$ and m tests $\mathbf{T}(\emptyset,\emptyset,\tau^0,k)$ for $k = 1,2,\ldots,m$), whilst Algorithm 3 will require solving $2m+1$ problems (problem $\mathbf{P}(\emptyset,\emptyset)$, m tests $\mathbf{T}(B,\mathbf{t}^B,\tau^0,k)$ and m problems $\mathbf{P}(B,\mathbf{t}^B)$). Hence, in this case, Algorithm 3 requires solving $O(m)$ more problems than Algorithm 2. Now let us consider a somewhat opposite case where all values $f_j(\mathbf{x}^0)$ are different. Additionally, assume that all optimal solutions $\mathbf{x}$ of the consecutively solved problems $\mathbf{P}(B,\mathbf{t}^B)$ and $\mathbf{T}(B,\mathbf{t}^B,\tau^0,k)$ yield the same values $f_j(\mathbf{x})$ for $j \in B'$. In this case Algorithm 3 will require solving $O(m^2/4)$ problems, while Algorithm 2 – $O(m^2/2)$ problems. This means that Algorithm 2 requires solving $O(m^2/4)$ more problems than Algorithm 3; this is a substantial difference.

Both algorithms presented above can be time consuming due to excessive number of problems $\mathbf{P}(B,\mathbf{t}^B)$ and $\mathbf{T}(B,\mathbf{t}^B,\tau^0,k)$ that may have to be solved in the iteration process. Therefore, below we give an alternative algorithm which is very fast provided that dual optimal variables problems $\mathbf{P}(B,\mathbf{t}^B)$ can be effectively computed (this is for instance the case for linear programmes and the simplex algorithm).

Suppose $\lambda = (\lambda_j)_{j \in B'}$ denotes the vector of dual variables (multipliers) associated with constraints (20b). It leads to the following Lagrangian function for problem $\mathbf{P}(B,\mathbf{t}^B)$:

$$\begin{aligned} \mathrm{L}(\mathbf{x};\tau;\lambda) &= -\tau + \textstyle\sum_{j \in B'} \lambda_j (\tau - f_j(\mathbf{x})) \\ &= (\textstyle\sum_{j \in B'} \lambda_j - 1)\tau - \textstyle\sum_{j \in B'} \lambda_j f_j(\mathbf{x}). \end{aligned} \tag{22}$$

The domain of Lagrangian (22) is defined by

$$\mathbf{x} \in \mathbf{Y}, \tag{23a}$$

$$-\infty < \tau < +\infty, \tag{23b}$$

$$\lambda \geq \mathbf{0}, \tag{23c}$$

where $\mathbf{Y}$ is determined by constraints (20c–d). Hence, the dual function is formally defined as

$$W(\lambda) = \min_{\tau, \mathbf{x} \in \mathbf{Y}} \mathrm{L}(\mathbf{x},\tau;\lambda) \quad \lambda \geq \mathbf{0} \tag{24}$$

and the dual problem reads:

$$\textbf{maximize}\; W(\lambda) \;\textbf{over}\; \lambda \geq \mathbf{0}. \tag{25}$$

The following theorem can be proved [27].

Theorem 7: Let λ^0 be the vector of optimal dual variables solving the dual problem (25). Then

1) $\sum_{j \in B'} \lambda_j^0 = 1,$ (26)

2) if $\lambda_j^0 > 0$ for some $j \in B'$, then $f_j(\mathbf{x})$ cannot be improved, i.e., $f_j(\mathbf{x}^0) = \tau^0$ for every optimal primal solution $(\mathbf{x}^0, \tau^0)$ of (20).

Note that in general the inverse of (2) in Theorem 7 does not hold: $\lambda_j^0 = 0$ does not necessarily imply that $f_j(\mathbf{x})$ can be improved (for an example see [27, 28]). In fact, it can be proved [27, Chapter 13] that the inverse implication holds if and only if set B is regular (set B is called regular if for any non-empty proper subset G of B, in the modified formulation $\mathbf{P}(B \backslash G, \mathbf{t}^{B \backslash G})$ the value of $f_k(\mathbf{x})$ can be improved for at least one of the indices $k \in B \backslash G$).

Whether or not the consecutive sets B are regular, the following algorithm solves problem P-MMF.

Algorithm 4: Algorithm for solving problem P-MMF based on dual variables

```
Step 0:  Put $B := \emptyset$ and $\mathbf{t}^B := \emptyset$.
Step 1:  If $B = M$ then stop ($\langle \mathbf{t}^B \rangle$ is the optimal solution of problem P-MMF, i.e., $\langle \mathbf{t}^B \rangle = \mathbf{T}^0$).
         Else, solve programme $\mathbf{P}(B, \mathbf{t}^B)$ and denote the resulting optimal solution by $(\mathbf{x}^0, \tau^0; \lambda^0)$.
Step 2:  Put $B := B \cup \{ j \in B' : \lambda_j^0 > 0 \}$.
Step 3:  Go to Step 1.
```

Observe that if for some $j \in B'$ with $\lambda_j^0 = 0$, $f_j(\mathbf{x})$ cannot be further improved, then in Step 1 the value of τ^0 will not be improved; still at least one such index j will be detected (due to property (5)) and included into set B in the next execution of Step 2. The regularity of set B simply ensures that in each iteration at least one $f_j(\mathbf{x})$ $(j \in B')$ will be improved.

In the case of LP problems, the dual quantities used in Algorithm 4 can be obtained directly from the simplex tableau. Indeed, it was a basis of early implementations of the lexicographic max-min solution for LP problems [1, 2, 12].

4.2. Conditional means

The sequential max-min algorithms can be applied only to convex problems, because, in general, it is likely that there does not exist a blocking index set B allowing for iterative processing. This can be illustrated with the following small example. Problem

$$\operatorname{lex}\max\ \{\langle(x_1+2x_2, 3x_1+x_2)\rangle : x_1+x_2=1,\ x_1,x_2\in\{0,1\}\}$$

has two feasible vectors $\mathbf{x}^1=(1,0)$, $\mathbf{x}^2=(0,1)$ and corresponding outcomes $\mathbf{y}^1=(1,3)$, $\mathbf{y}^2=(2,1)$. Obviously, $\mathbf{x}^1$ is the MMF optimal solution as $\langle(1,3)\rangle >_{lex} \langle(2,1)\rangle$. One can easily verify that both feasible solutions are optimal for max-min problem

$$\max\ \{\min\{x_1+2x_2, 3x_1+x_2\} : x_1+x_2=1,\ x_1,x_2\in\{0,1\}\}$$

but neither f_1 nor f_2 is a blocking outcome allowing to define the second level max-min optimization problem to maximize the second worst outcome. For the same reason, the sequential algorithm may fail for the single-path version of the routing optimization for the MMF distribution of demand volumes and other discrete models (refer to Subsection 3.4).

Following Yager [37], a direct, although requiring the use of integer variables, formula can be given for any $y_{\langle k\rangle}$. Namely, for any $k=1,2,\dots,m$ the following formula is valid:

$$\begin{array}{rll} y_{\langle k\rangle} = & \max\ r_k & \\ & \text{s.t.} & \\ & r_k - y_j \le C z_{kj},\ z_{kj}\in\{0,1\} \quad j=1,\dots,m & (27)\\ & \sum_{j=1}^{m} z_{kj} \le k-1, & \end{array}$$

where C is a sufficiently large constant (larger than any possible difference between various individual outcomes y_j) which allows us to enforce inequality $r_k \le y_j$ for $z_{kj}=0$ while ignoring it for $z_{kj}=1$. Note that for $k=1$ all binary variables z_{1j} are forced to 0 thus reducing the optimization in this case to the standard LP model. However, for any other $k>1$ all m binary variables z_{kj} are an important part of the model. Nevertheless, with the use of auxiliary integer variables, any MMF problem (either convex or non-convex) can be formulated as the standard lexicographic maximization with directly defined achievement functions:

$$\begin{array}{lr} \operatorname{lex}\max\ (r_1,r_2,\dots,r_m) & (28a)\\ \text{s.t.} & \\ \mathbf{x}\in Q & (28b)\\ r_k - f_j(\mathbf{x}) \le C z_{kj},\ z_{kj} \quad j,k=1,\dots,m & (28c)\\ \in\{0,1\} \quad j,k=1,\dots,m & (28d)\\ \sum_{j=1}^{m} z_{kj} \le k-1 \quad k=1,\dots,m. & (28e) \end{array}$$

Recall that one may take advantage of the formulation (12) with cumulated criteria $\bar{\theta}_k(\mathbf{y})=\sum_{i=1}^{k} y_{\langle i\rangle}$ expressing, respectively: the worst (smallest) outcome, the total of the two worst outcomes, the total of the three worst outcomes, etc. When normalized by k the quantities $\mu_k(\mathbf{y})=\bar{\theta}_k(\mathbf{y})/k$ can be interpreted as the worst conditional means [24]. The optimization formula (27) for $y_{\langle k\rangle}$ can easily be extended to define $\bar{\theta}_k(\mathbf{y})$. Namely, for any $k=1,2,\dots,m$ the following formula is valid:

$$\begin{array}{llr} \bar{\theta}_k(\mathbf{y}) = \max\ k r_k - \sum_{j=1}^{m} d_{kj} & & \\ \text{s.t.} & & \\ r_k - y_j \le d_{kj},\ d_{kj}\ge 0 & j=1,\dots,m & (29)\\ d_{kj} \le C z_{kj},\ z_{kj}\in\{0,1\} & j=1,\dots,m & \\ \sum_{j=1}^{m} z_{kj} \le k-1, & & \end{array}$$

where C is a sufficiently large constant. However, the optimization problem defining the cumulated ordered outcome can be dramatically simplified since all its binary variables (and the related constraints) turn out to be redundant. First let us notice that for any given vector $\mathbf{y}\in\Re^m$, the cumulated ordered value $\bar{\theta}_k(\mathbf{y})$ can be found as the optimal value of the following LP problem:

$$\begin{array}{rl} \bar{\theta}_k(\mathbf{y}) = & \min\ \sum_{j=1}^{m} y_j u_{kj}\\ & \text{s.t.}\\ & \sum_{j=1}^{m} u_{kj}=k,\ 0\le u_{kj}\le 1 \quad j=1,\dots,m. \end{array} \tag{30}$$

The above problem is an LP for a given outcome vector $\mathbf{y}$ while it becomes nonlinear for $\mathbf{y}$ being a variable. This difficulty can be overcome by taking advantage of the LP dual to (30) as shown in the following assertion.

Theorem 8: For any given vector $\mathbf{y}\in\Re^m$, the cumulated ordered coefficient $\bar{\theta}_k(\mathbf{y})$ can be found as the optimal value of the following LP problem:

$$\begin{array}{rl} \bar{\theta}_k(\mathbf{y}) = & \max\ k r_k - \sum_{j=1}^{m} d_{kj}\\ & \text{s.t.}\\ & r_k - y_j \le d_{kj},\ d_{kj}\ge 0 \quad j=1,\dots,m. \end{array} \tag{31}$$

Proof: In order to prove the theorem it is enough to notice that problem (31) is the LP dual of problem (30) with variable r_k corresponding to the equation $\sum_{j=1}^{m} u_{kj}=k$ and variables d_{kj} corresponding to upper bounds on u_{kj}. ■

It follows from Theorem 8 that

$$\begin{array}{rl} \bar{\theta}_k(\mathbf{f}(\mathbf{x})) = \max\{ & k r_k - \sum_{j=1}^{m} d_{kj}\ :\ \mathbf{x}\in Q;\\ & r_k - f_j(\mathbf{x}) \le d_{kj},\ d_{kj}\ge 0 \quad j\in M\} \end{array}$$

or in a more compact form $\bar{\theta}_k(\mathbf{f}(\mathbf{x})) = \max\ \{k r_k - \sum_{j=1}^{m} (f_j(\mathbf{x})-r_k)_+ : \mathbf{x}\in Q\ \}$ where $(.)_+$ denotes the nonnegative part of a number and r_k is an auxiliary (unbounded) variable. The latter, with the necessary adaptation to the minimized outcomes in location problems, is equivalent to the computational formulation of the k–centrum model introduced in [26]. Hence, Theorem 8 provides an alternative proof of that formulation.

Due to Theorem 4, the lexicographic max-min problem (11) is equivalent to the lexicographic maximization of conditional means

$$\operatorname{lex\,max}\ \{(\mu_1(\mathbf{f}(\mathbf{x})),\mu_2(\mathbf{f}(\mathbf{x})),\ldots,\mu_m(\mathbf{f}(\mathbf{x})))\ :\ \mathbf{x}\in Q\}.$$

Following Theorem 8, the above leads us to a standard lexicographic optimization problem with predefined linear criteria:

$$\begin{array}{ll}\operatorname{lex\,max} & (r_1-\sum_{j=1}^{m} d_{1j},\ldots,r_m-\frac{1}{m}\sum_{j=1}^{m} d_{mj})\\ & \text{s.t.}\\ & \mathbf{x}\in Q\\ & d_{kj}\ge r_k-f_j(\mathbf{x}) \quad j,k=1,\ldots,m\\ & d_{kj}\ge 0 \quad j,k=1,\ldots,m.\end{array} \tag{32}$$

Note that this direct lexicographic formulation remains valid for nonconvex (e.g., discrete) feasible sets Q, where the standard sequential approaches [16, 17] are not applicable [21].

Model (32) preserves the problem convexity when the original problem is defined with convex feasible set Q and concave objective functions f_j. In particular, for an LP original problem it remains within the LP class while introducing m^2+m auxiliary variables and m^2 constraints. Thus, for many problems with not too large number of criteria m, problem (32) can easily be solved directly. Although, in general, for convex problems such an approach seems to be less efficient than the sequential algorithms discussed in the previous subsection. The latter may require m iterative steps only in the worst case (only one blocking variable at each step), while typically there are more than two blocking variables identified at each step which reduces significantly the number of steps. The direct model (32) essentially requires the sequential lexicographic Algorithm 1 with m steps.

Further research on the increase of computational efficiency of model (32) seems to be very promising. Note that all lexicographic criteria of this problem express the conditional means which are monotonic with respect to increasing k. While solving the lexicographic problem with the standard sequential Algorithm 1, one needs to solve at Step 2 the following maximization problem:

$$\begin{array}{ll}\max\ \{\tau_k\ : & \tau_k\le r-w\sum_{j=1}^{m} d_j;\mu_l(\mathbf{f}(\mathbf{x}))\ge\tau_l^0\ \forall l<k;\\ & \mathbf{x}\in Q;\ r-f_j(\mathbf{x})\le d_j,\ d_j\ge 0\ \forall j\},\end{array}$$

where $w=1/k$. It may occur that the optimal solution of the above problem remains also optimal for smaller coefficients $w=1/\kappa$ thus defining conditional means for $\kappa>k$. In such a case, one may advance the iterative process to $\kappa+1$ instead of $k+1$. Hence, some parametric optimization techniques may allow us to reduce the number of iterations to the same level as in the sequential max-min algorithms.

Note that model (32) offers also a possibility to build some approximations to the strict MMF solution as it allows us to build lexicographic problems taking into account only a selected grid of indices k. In particular, the so-called augmented max-min solution concept, commonly used in the multiple criteria optimization [22, 35], is such an approximation, although very rough as based only on μ_1 and μ_k

$$\begin{array}{l}\operatorname{lex\,max}\{(r_1,\ \frac{1}{m}\sum_{j=1}^{m} f_j(\mathbf{x})):\ \ r_1\le f_j(\mathbf{x}) \quad j=1,\ldots,m,\\ \qquad\qquad\qquad\qquad\qquad\qquad \mathbf{x}\in Q\}.\end{array}$$

4.3. Distribution approach

For some specific classes of discrete, or rather combinatorial, optimization problems, one may take advantage of the finiteness of the set of all possible values of functions f_j on the finite set of feasible solutions. The ordered outcome vectors may be treated as describing a distribution of outcomes generated by a given decision $\mathbf{x}$. In the case when there exists a finite set of all possible outcomes of the individual objective functions, we can directly describe the distribution of outcomes with frequencies of outcomes. Let $V=\{v_1,v_2,\ldots,v_r\}$ (where $v_1<v_2<\cdots<v_r$) denote the set of all attainable outcomes (all possible values of the individual objective functions f_j for $\mathbf{x}\in Q$). We introduce integer functions $h_k(\mathbf{y})$ $(k=1,\ldots,r)$ expressing the number of values v_k in the outcome vector $\mathbf{y}$. Having defined functions h_k we can introduce cumulative distribution functions:

$$\bar{h}_k(\mathbf{y})=\sum_{l=1}^{k} h_l(\mathbf{y})\ , \quad k=1,\ldots,r. \tag{33}$$

Function $\bar{h}_k$ expresses the number of outcomes smaller or equal to v_k. Since we want to maximize all the outcomes, we are interested in the minimization of all functions $\bar{h}_k$. Indeed, the following assertion is valid [22]. For outcome vectors $\mathbf{y}',\mathbf{y}''\in V^m$, $\langle\mathbf{y}'\rangle\ge\langle\mathbf{y}''\rangle$ if and only if $\bar{h}_k(\mathbf{y}')\le\bar{h}_k(\mathbf{y}'')$ for all $k=1,\ldots,r$. This equivalence allows to express the lexicographic max-min solution concept for problem (1) in terms of the standard lexicographic minimization problem with objectives $\bar{\mathbf{h}}(\mathbf{f}(\mathbf{x}))$:

$$\operatorname{lex\,min}\ \{(\bar{h}_1(\mathbf{f}(\mathbf{x})),\ldots,\bar{h}_r(\mathbf{f}(\mathbf{x})))\ :\ \mathbf{x}\in Q\}. \tag{34}$$

Theorem 9: A feasible solution $\mathbf{x}\in Q$ is an optimal solution of the P-MMF problem, if and only if it is an optimal solution of the lexicographic problem (34).

The quantity $\bar{h}_k(\mathbf{y})$ can be computed directly by the minimization:

$$\begin{array}{ll}\bar{h}_k(\mathbf{y})= & \min\sum_{j=1}^{m} z_{kj}\\ & \text{s.t.}\\ & v_{k+1}-y_j\le Cz_{kj},\ z_{kj}\in\{0,1\} \quad j=1,\ldots,m,\end{array}$$

where C is a sufficiently large constant. Note that $\bar{h}_r(\mathbf{y})=m$ for any $\mathbf{y}$ which means that the rth criterion is always constant and therefore redundant in (34). Hence, the lexicographic problem (34) can be formulated as the following mixed integer problem:

$$
\begin{aligned}
&\operatorname{lex}\min \left[\sum_{j=1}^{m} z_{1j},\ \sum_{j=1}^{m} z_{2j}, \ldots,\ \sum_{j=1}^{m} z_{r-1,j} \right] \\
&\text{s.t.} \\
&v_{k+1} - f_j(\mathbf{x}) \le C z_{kj} \;\; j = 1, \ldots, m, \;\; k = 1, \ldots, r-1, \\
&z_{kj} \in \{0,1\} \quad j = 1, \ldots, m, \;\; k = 1, \ldots, r-1, \\
&\mathbf{x} \in Q.
\end{aligned}
\tag{35}
$$

Krarup and Pruzan [15] have shown that, in the case of discrete location problems, the use of the minisum solution concept with the outcomes raised to a sufficiently large power is equivalent to the use of the minimax solution concept. Formulation (34) allows us to extend such an approach to the lexicographic max-min solution concept. Note that the achievements functions in (34) can be rescaled with corresponding values $v_{k+1} - v_k$. When the differences among outcome values are large enough then the lexicographic minimization corresponds to the one-level optimization of the total of achievements which is equivalent to minimization of the sum of the original outcomes. In general, as shown by Burkard and Rendl [4], there is a possibility to replace then the lexicographic max-min objective function with an equivalent linear function on rescaled outcomes. Algorithms developed in [4, 5] take advantage of finiteness of the set of outcome values and they depend on making (explicitly or implicitly) differences among the outcomes larger (without changing their order) which does not affect the lexicographically maximal solutions of problem (11). When the differences are large enough the optimal solution of the maxisum problem is also the lexicographic max-min solution. In general, an unrealistically complicated scaling function may be necessary to generate large enough differences among different but very close outcomes. Therefore, the outcomes should be mapped first on the set of integer variables (numbered) to normalize the minimum difference, like in [4, 5] approaches. All these transformations are eligible in the case of finite outcome set. Nevertheless, while solving practical problems, large differences among coefficients may cause serious computational problems. Therefore, such approaches are less useful for large scale problems typically arriving in telecommunications network design.

Taking advantage of possible weighting and cumulating achievements in lexicographic optimization, one may eliminate auxiliary integer variables from the achievement functions. For this purpose we weight and cumulate vector $\bar{\mathbf{h}}(\mathbf{y})$ to get $\hat{h}_1(\mathbf{y}) = 0$ and:

$$
\hat{h}_k(\mathbf{y}) = \sum_{l=1}^{k-1} (v_{l+1} - v_l)\bar{h}_l(\mathbf{y}) \quad k = 2, \ldots, r. \tag{36}
$$

Due to Theorem 4 and positive differences $v_{l+1} - v_l > 0$, the lexicographic minimization problem (34) is equivalent to the lexicographic problem with objectives $\hat{\mathbf{h}}(\mathbf{f}(\mathbf{x}))$:

$$
\operatorname{lex}\min \; \{(\hat{h}_1(\mathbf{f}(\mathbf{x})), \ldots, \hat{h}_r(\mathbf{f}(\mathbf{x}))) \; : \; \mathbf{x} \in Q\} \tag{37}
$$

which leads us to the following assertion.

Theorem 10: A feasible solution $\mathbf{x} \in Q$ is an optimal solution of the P-MMF problem, if and only if it is an optimal solution of the lexicographic problem (37).

Actually, vector function $\hat{\mathbf{h}}(\mathbf{y})$ provides a unique description of the distribution of coefficients of vector $\mathbf{y}$, i.e., for any $\mathbf{y}', \mathbf{y}'' \in V^m$ one gets: $\hat{\mathbf{h}}(\mathbf{y}') = \hat{\mathbf{h}}(\underline{\mathbf{y}''}) \;\Leftrightarrow\; \langle \underline{\mathbf{y}'} \rangle = \langle \mathbf{y}'' \rangle$. Moreover, $\hat{\mathbf{h}}(\mathbf{y}') \le \hat{\mathbf{h}}(\mathbf{y}'')$ if and only if $\Theta(\mathbf{y}') \geq \Theta(\mathbf{y}'')$ [22].

Note that $\hat{h}_1(\mathbf{y}) = 0$ for any $\mathbf{y}$ which means that the first criterion is constant and redundant in problem (37). Moreover, putting (33) into (36) allows us to express all achievement functions $\hat{h}_k(\mathbf{y})$ as a piece wise linear functions of $\mathbf{y}$:

$$
\hat{h}_k(\mathbf{y}) = \sum_{j=1}^{m} (v_k - y_j)_+ = \sum_{j=1}^{m} \max\{v_k - y_j, 0\} \qquad k = 1, \ldots, r. \tag{38}
$$

Hence, the quantity $\hat{h}_k(\mathbf{y})$ can be computed directly by the following minimization:

$$
\begin{aligned}
\hat{h}_k(\mathbf{y}) = \quad &\min \sum_{j=1}^{m} t_{kj} \\
&\text{s.t.} \\
&v_k - y_j \le t_{kj}, \; t_{kj} \ge 0 \quad j = 1, \ldots, m.
\end{aligned}
\tag{39}
$$

Therefore, the entire lexicographic model (37) can be formulated as follows:

$$
\begin{aligned}
&\operatorname{lex}\min \left[\sum_{j=1}^{m} t_{2j},\ \sum_{j=1}^{m} t_{3j}, \ldots,\ \sum_{j=1}^{m} t_{rj} \right] \\
&\text{s.t.} \\
&v_k - f_j(\mathbf{x}) \le t_{kj}, \; t_{kj} \ge 0 \quad j = 1, \ldots, m, \;\; k = 2, \ldots, r \\
&\mathbf{x} \in Q.
\end{aligned}
\tag{40}
$$

Note that the above formulation, unlike the problem (35), does not use integer variables and can be considered as an LP modification of the original multiple criteria problem (1). Thus, this model preserves the problem's convexity when the original problem is defined with a convex feasible set Q and a concave objective functions f_j. The size of problem (40) depends on the number of different outcome values. Thus, for many problems with not too large number of outcome values, the problem can easily be solved directly and even for convex problems such an approach may be more efficient than the sequential algorithms discussed in the previous subsection. Note that in many problems of telecommunications network design, the objective functions express the quality of service and one can easily consider a limited finite scale (grid) of the corresponding outcome values. Similarly, in the capacity protection design (Subsection 3.3), one may focus on a finite grid of demand volumes. One may also notice that model (40) opens a way for the fuzzy representation of quality measures within the MMF problems.

5. Concluding remarks

Today, the major objective of telecommunications network design for Internet services is to maximize service data flows and provide fair treatment of all services. Fair treatment of services can be formalized through the MMF solution concept, which assumes that the worst service performance is maximized and the solution is additionally regularized with the lexicographic maximization of the second worst performance, the third one, etc. We have argued that the MMF solution concept is tightly related to the Rawlsian principle of justice and is equivalent the lexicographic max-min concept.

We have shown that with respect to telecommunications networks carrying the so-called elastic traffic, the problems of routing design, restoration design and protection capacity design are examples of important design problems that can be formulated with the use of the MMF notion to express design objectives. We have presented and evaluated several general efficient sequential algorithms that can be used to solve the basic variants of these problems as well as many other MMF problems. These algorithms are based on the idea to solve a sequence of properly defined max-min subproblems. The algorithms differ with respect to the strategy of choosing this sequence. We have shown that the efficiency of different strategies depends on the distribution of outcome values of the optimal solution to the original problem. Since the algorithms can still be time-consuming due to excessive number of subproblems that have to be solved in the iteration process, the values of subproblems' dual variables can be used to considerably reduce the number of solved subproblems. In the case of LP problem formulations the values of dual variables can be obtained directly from the simplex tableau.

Unfortunately, sequential algorithms are only applicable to convex problems. Hence if network design problems are augmented with the requirements that data flows are to be routed along single paths or that link capacity is modular, these algorithms cannot be applied any more. However, we have shown that the original problem of lexicographic maximization of the solution vector can be replaced with the lexicographic minimization of the vector that describes the distribution of outcome values, which, fortunately enough, is convex as long as an original problem is defined with a convex feasible set Q and a concave objective functions f_j. The complexity of the transformed problem is directly related to the number of different outcome values. As far as telecommunications network design is concerned, this number can be pretty small, for example if the objective functions express the quality of service. Therefore, further research on application of distribution approach to various classes of telecommunications MMF problems seems to be very promising.

Appendix A. Numerical example

In this appendix we present a numerical example of Problem 1 (Subsection 3.1). The structure of the considered network is shown in Fig. 1; c_e denotes the capacity of link e. We assume that the set of demands corresponds to the set of all pairs of nodes.

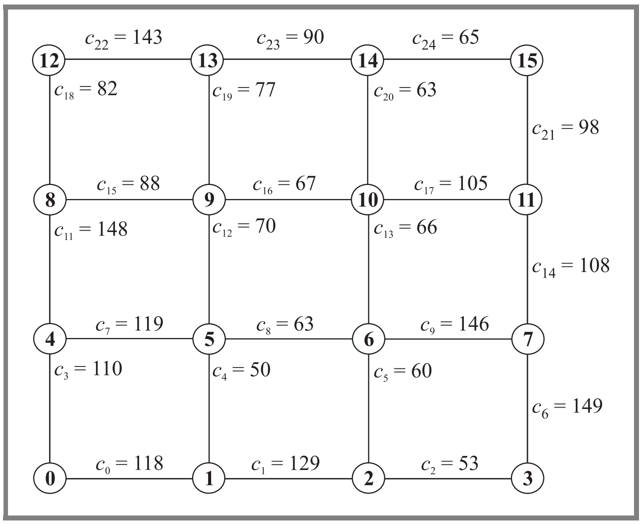


Fig. 1. 16-node square network.

The results of applying Algorithm 4 (Subsection 4.1) to Problem 1 are presented in Table 1. The table contains information pertaining to consecutive iterations of the algorithm. The information includes the number of demands blocked in an iteration and their flow size. To effectively solve the problem we applied a path (column) generation technique [27, Subsection 8.2.1] allowing for problem decomposition. The overall number of paths used in each iteration is presented in Fig. 2. The LP subproblems were

Table 1
Consecutive values of τ_n^0 and number of blocked demands in MMF allocation procedure

Iteration n	Value τ_n^0	Blocked demands
1	5.286	63
2	6.625	8
3	7.013	28
4	10.214	8
5	14.606	4
6	16.115	1
7	25.362	1
8	29.908	2
9	30.962	1
10	35.093	2
11	49.288	1
12	82.145	1

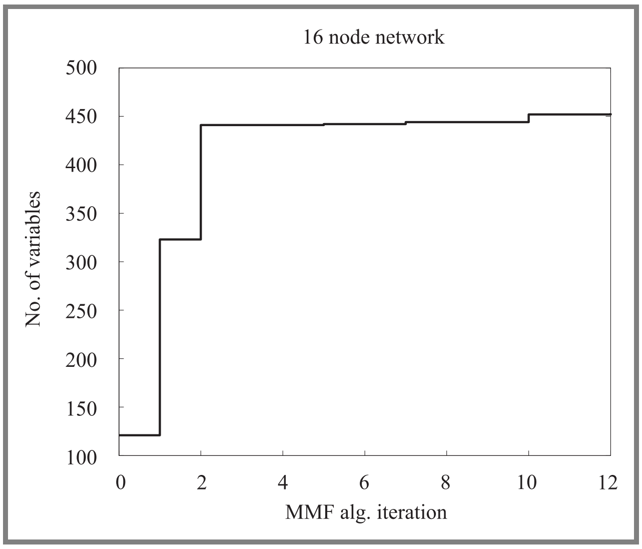


Fig. 2. Number of problem columns in function of MMF algorithm iterations.

solved with the use of the CPLEX 9.0 optimization package. Solving the problem on a PC-class computer equipped with a 2.4 GHz P4 HT processor required 0.2 s of the processor time, of which only 0.03 s in total was spent on solving the LP subproblems.

Acknowledgements

Włodzimierz Ogryczak is supported by The State Committee for Scientific Research under grant 3T11C 005 27 "Models and algorithms for efficient and fair resource allocation in complex systems".

Michał Pióro and Artur Tomaszewski are supported by The State Committee for Scientific Research under grant 3T11D 001 27 "Design methods for NGI core networks".

References

[1] F. A. Behringer, "A simplex based algorithm for the lexicographically extended linear maxmin problem", *Eur. J. Oper. Res.*, vol. 7, pp. 274–283, 1981.

[2] F. A. Behringer, "Linear multiobjective maxmin optimization and some Pareto and lexmaxmin extensions", *OR Spektrum*, vol. 8, pp. 25–32, 1986.

[3] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs: Prentice-Hall, 1987.

[4] R. E. Burkard and F. Rendl, "Lexicographic bottleneck problems", *Oper. Res. Lett.*, vol. 10, pp. 303–308, 1991.

[5] F. Della Croce, V. T. Paschos, and A. Tsoukias, "An improved general procedure for lexicographic bottleneck problem", *Oper. Res. Lett.*, vol. 24, pp. 187–194, 1999.

[6] R. Denda, A. Banchs, and W. Effelsberg, "The fairness challenge in computer networks", *LNCS*, vol. 1922, pp. 208–220, 2000.

[7] M. Dresher, *Games of Strategy*. Englewood Cliffs: Prentice-Hall, 1961.

[8] D. Dubois, Ph. Fortemps, M. Pirlot, and H. Prade, "Leximin optimality and fuzzy set-theoretic operations", *Eur. J. Oper. Res.*, vol. 130, pp. 20–28, 2001.

[9] M. Ehrgott, "Discrete decision problems, multiple criteria optimization classes and lexicographic max-ordering", in *Trends in Multicriteria Decision Making*, T. J. Stewart and R. C. van den Honert, Eds. Berlin: Springer, 1998, pp. 31–44.

[10] H. Isermann, "Linear lexicographic optimization", *OR Spektrum*, vol. 4, pp. 223–228, 1982.

[11] J. Jaffe, "Bottleneck flow control", *IEEE Trans. Commun.*, vol. 7, pp. 207–237, 1980.

[12] R. S. Klein, H. Luss, and D. R. Smith, "A lexicographic minimax algorithm for multiperiod resource allocation", *Math. Programm.*, vol. 55, pp. 213–234, 1992.

[13] J. Kleinberg, Y. Rabani, and E. Tardos, "Fairness in routing and load balancing", in *Proc. 40th Ann. IEEE Symp. Found. Comput. Sci.*, New York, USA, 1999.

[14] M. M. Kostreva and W. Ogryczak, "Linear optimization with multiple equitable criteria", *RAIRO Oper. Res.*, vol. 33, pp. 275–297, 1999.

[15] J. K. Krarup and P. M. Pruzan, "Reducibility of minimax to minisum 0–1 programming problems", *Eur. J. Oper. Res.*, vol. 5, pp. 125–132, 1981.

[16] H. Luss, "On equitable resource allocation problems: a lexicographic minimax approach", *Oper. Res.*, vol. 47, pp. 361–378, 1999.

[17] E. Marchi and J. A. Oviedo, "Lexicographic optimality in the multiple objective linear programming: the nucleolar solution", *Eur. J. Oper. Res.*, vol. 57, pp. 355–359, 1992.

[18] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press, 1979.

[19] P. Nilsson and M. Pióro, *Solving Dimensioning Problems for Proportionally Fair Networks Carrying Elastic Traffic*. Lund: Lund Institute of Technology, 2002.

[20] P. Nilsson, M. Pióro, and Z. Dziong, "Link protection within an existing backbone network", in *Proc. Int. Netw. Opt. Conf. INOC*, Paris-Evry, France, 2003.

[21] W. Ogryczak, "On the lexicographic minimax approach to location problems", *Eur. J. Oper. Res.*, vol. 100, pp. 566–585, 1997.

[22] W. Ogryczak, *Wielokryterialna optymalizacja liniowa i dyskretna. Modele preferencji i zastosowania do wspomagania decyzji*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego, 1997 (in Polish).

[23] W. Ogryczak, "Comments on properties of the minimax solutions in goal programming", *Eur. J. Oper. Res.*, vol. 132, pp. 17–21, 2001.

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

[25] W. Ogryczak and T. Śliwiński, "On solving linear programs with the ordered weighted averaging objective", *Eur. J. Oper. Res.*, vol. 148, pp. 80–91, 2002.

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

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

[28] M. Pióro, P. Nilsson, E. Kubilinskas, and G. Fodor, "On efficient max-min fair routing algorithms", in *Proc. 8th IEEE Int. Symp. Comput. Commun. ISCC'03*, Antalya, Turkey, 2003.

[29] J. A. M. Potters and S. H. Tijs, "The nucleolus of a matrix game and other nucleoli", *Math. Oper. Res.*, vol. 17, pp. 164–174, 1992.

[30] J. Rawls, *The Theory of Justice*. Cambridge: Harvard University Press, 1971.

[31] J. R. Rice, "Tschebyscheff approximation in a compact metric space", *Bull. Amer. Math. Soc.*, vol. 68, pp. 405–410, 1962.

[32] D. Schmeidler, "The nucleolus of a characteristic function game", *SIAM J. Appl. Math.*, vol. 17, pp. 1163–1170, 1969.

[33] R. E. Steuer, *Multiple Criteria Optimization: Theory, Computation & Applications*. New York: Wiley, 1986.

[34] A. Tomaszewski, "A polynomial algorithm for solving a general max-min fairness problem", in *Proc. 2nd Polish-German Teletraffic Symp. PGTS 2002*, Gdańsk, Poland, 2002, pp. 253–258.

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

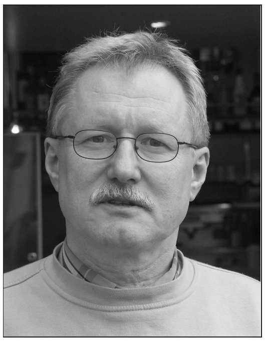
[36] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making", *IEEE Trans. Syst., Man Cyber.*, vol. 18, pp. 183–190, 1988.

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

[38] R. R. Yager, "On the analytic representation of the Leximin ordering and its application to flexible constraint propagation", *Eur. J. Oper. Res.*, vol. 102, pp. 176–192, 1997.

Włodzimierz Ogryczak is a Professor and the Head of Optimization and Decision Support Division in the Institute of Control and Computation Engineering (ICCE) at the Warsaw University of Technology, Poland. He received both his M.Sc. (1973) and Ph.D. (1983) in mathematics from Warsaw University, and D.Sc. (1997) in computer science from Polish Academy of Sciences. His research interests are focused on models, computer solutions and interdisciplinary applications in the area of optimization and decision making with the main stress on: multiple criteria optimization and decision support, decision making under risk, location and distribution problems. He has published three books and numerous research articles in international journals.

e-mail: W.Ogryczak@ia.pw.edu.pl
Institute of Control & Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Michał Pióro is a Professor and the Head of the Computer Networks and Switching Division in the Institute of Telecommunications at the Warsaw University of Technology, Poland, and a Full Professor at Lund University, Sweden. He received the Ph.D. degree in telecommunications in 1979 and the D.Sc. degree in 1990, both from the Warsaw University of Technology. In 2002 he received a Polish State Professorship. His research interests concentrate on modeling, design and performance evaluation of telecommunication systems. He is an author of four books and around 100 technical papers presented in the telecommunication journals and conference proceedings. He has lead many research projects for telecom industry in the field of network modeling, design, and performance analysis.

e-mail: mpp@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Artur Tomaszewski is an Assistant Professor of the Computer Networks and Switching Division in the Institute of Telecommunications at the Warsaw University of Technology, Poland. He received the M.Sc. and Ph.D. degrees in telecommunications in 1990 and 1993, respectively, both from the Warsaw University of Technology. His research interests focus on network architectures and network optimization methods. He is an author and co-author of about 30 papers on core and access network optimization, network planning processes and methods, and network planning support tools.

e-mail: artur@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Site selection for waste disposal through spatial multiple criteria decision analysis

Mohammed A. Sharifi and Vasilios Retsios

http://doi.org/10.26636/jtit.2004.3.253

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 tradeoff 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.

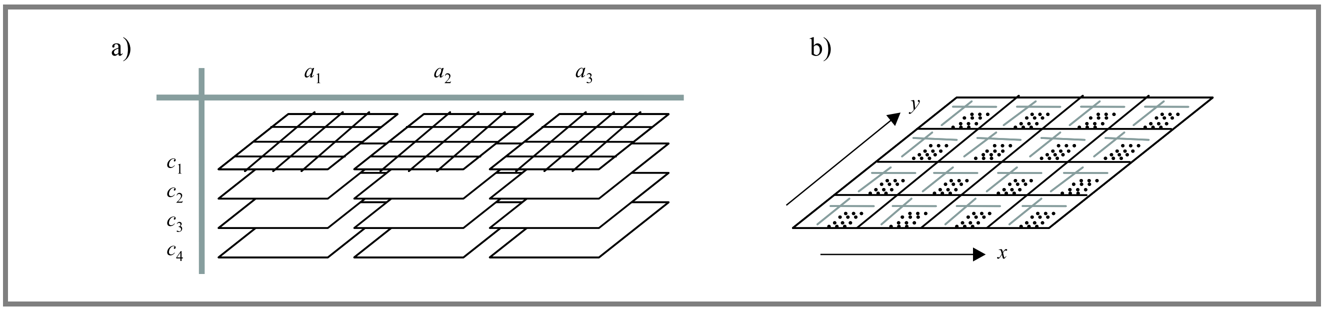


Fig. 1. Two interpretations of a 2-dimensional decision problem (a) table of maps; (b) map of tables.

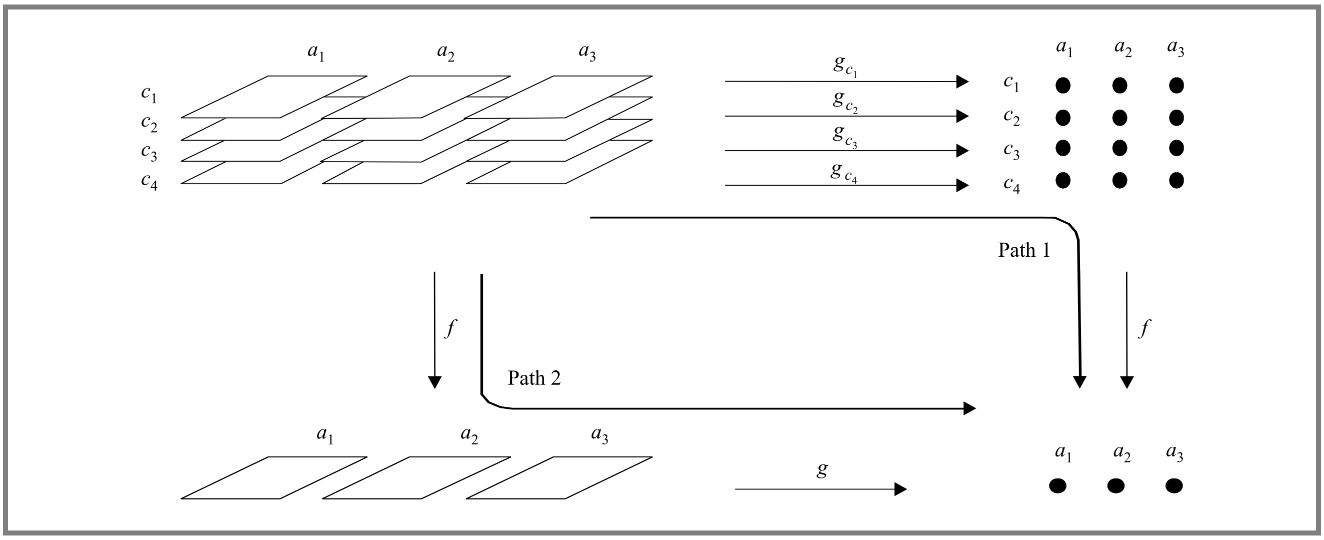


Fig. 2. Two paths of spatial multi-criteria evaluation.

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.

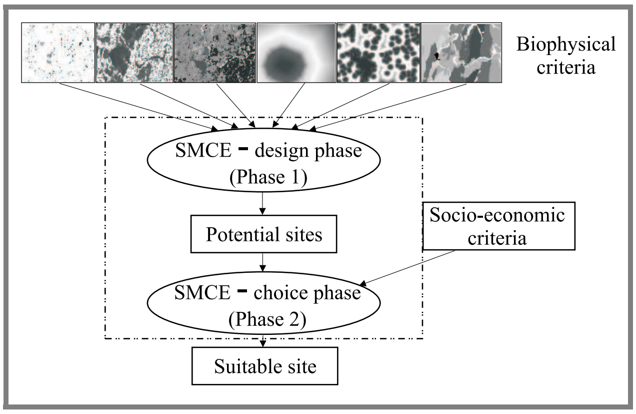


Fig. 3. Flowchart for the entire site selection process. Due to printing limitations, original color maps had to be converted to black and white, with loss of some detail.

[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):

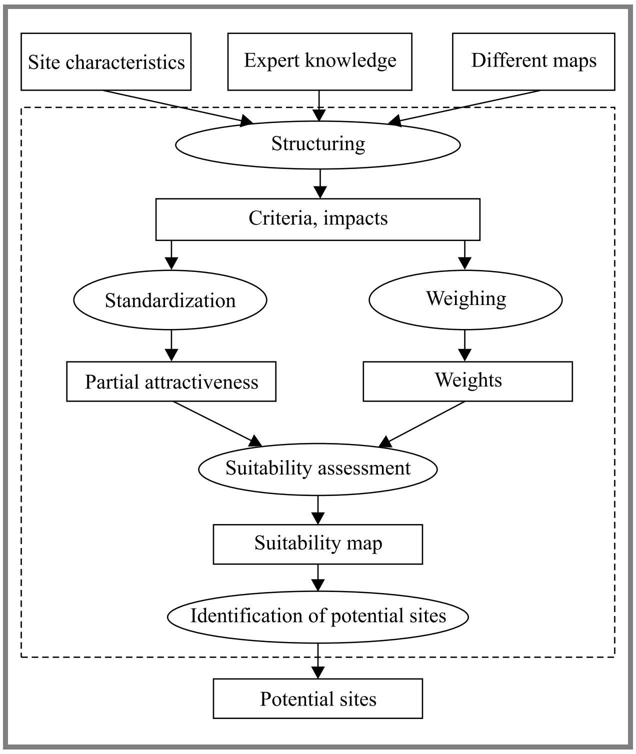


Fig. 4. Flowchart for the SMCE design phase, which results in selection of potential sites.

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

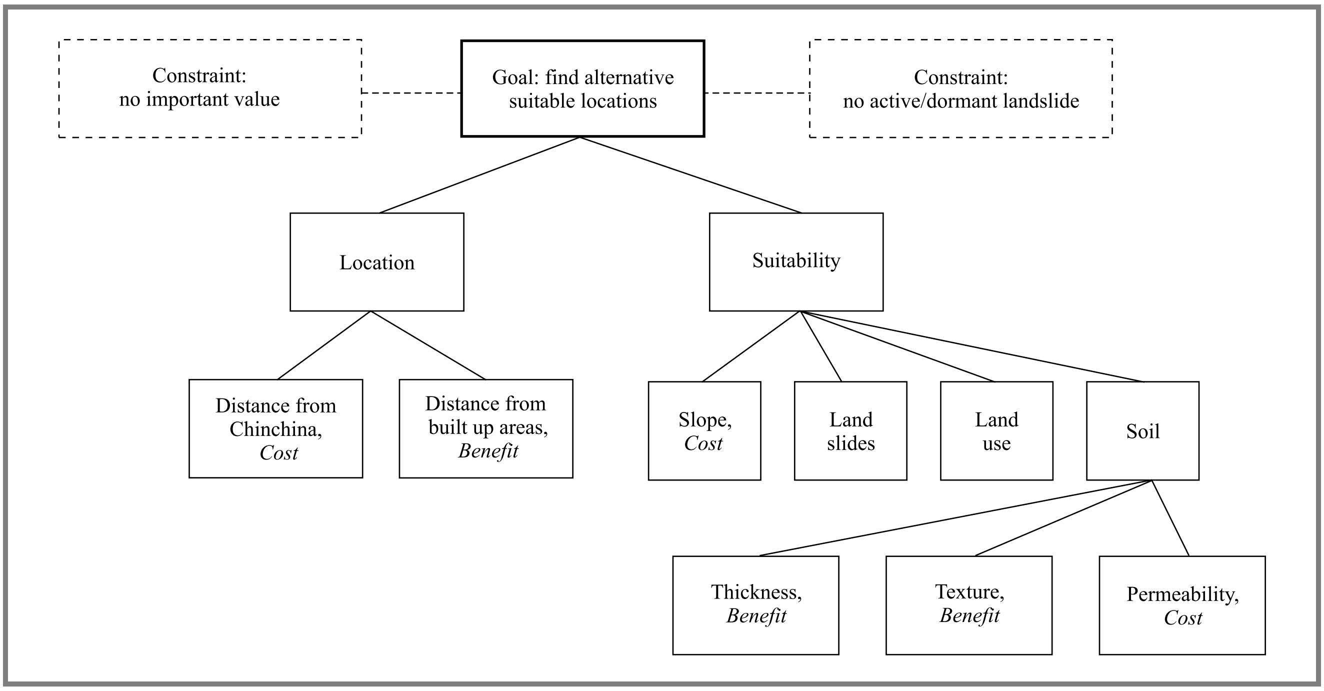


Fig. 5. Criteria tree for identifying the potential site for waste disposal.

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:

$\texttt{suitability\_map} = \texttt{w}_1\texttt{*Location} + \texttt{w}_2\texttt{*Suitability}$

Here $\texttt{w}_1$ and $\texttt{w}_2$ 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:

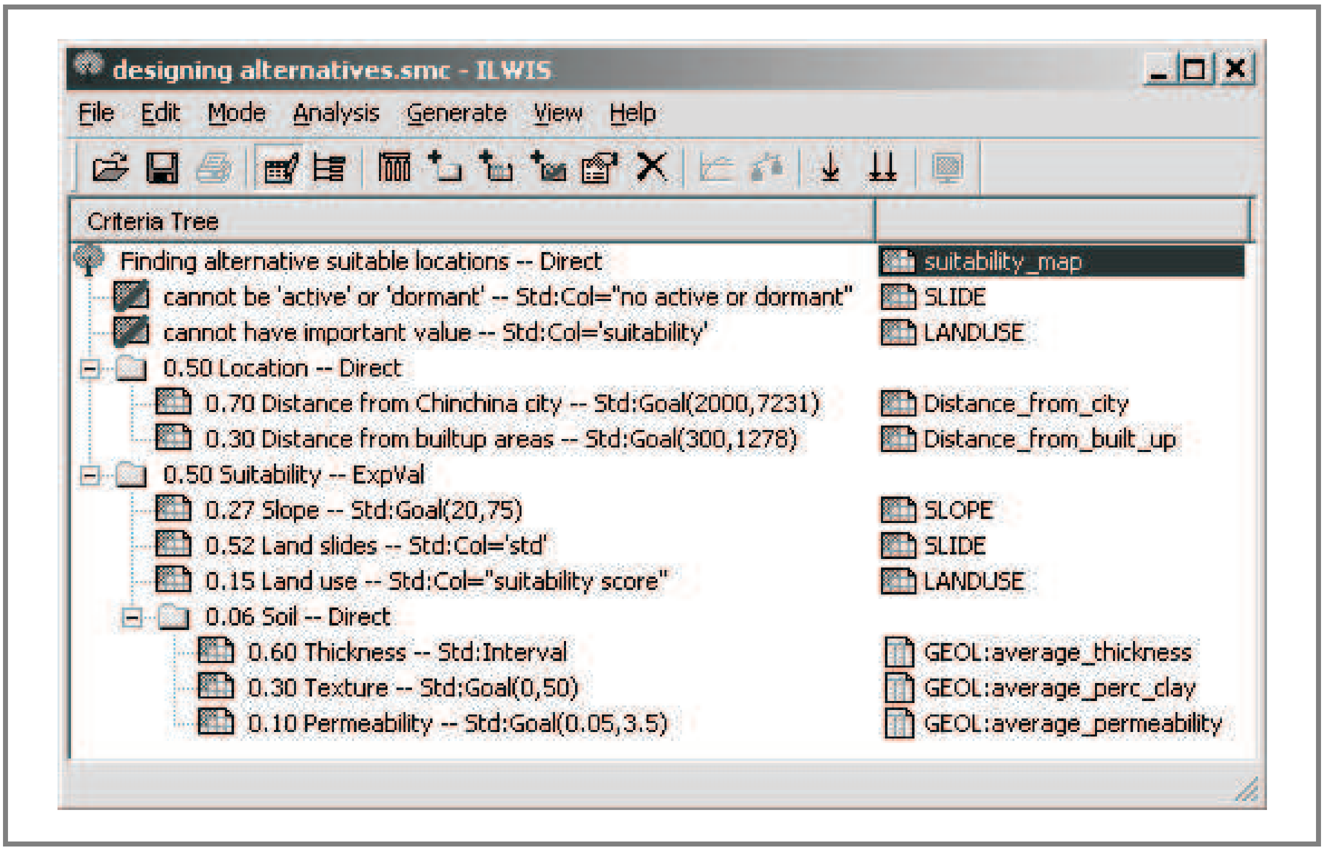
$\texttt{suitability\_map} =$
$\texttt{w}_1\texttt{*(w}_{11}\texttt{*Distance\_from\_Chinchina} + \texttt{w}_{12}\texttt{*Distance\_from\_builtup\_areas)} + \texttt{w}_2\texttt{*(w}_{21}\texttt{*Slope} + \texttt{w}_{22}\texttt{*Land\_slides} + \texttt{w}_{23}\texttt{*Land\_use} + \texttt{w}_{24}\texttt{*Soil)}$

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.



designing alternatives.smc - ILWIS

Criteria Tree

Finding alternative suitable locations -- Direct	suitability_map
cannot be 'active' or 'dormant' -- Std:Col="no active or dormant"	SLIDE
cannot have important value -- Std:Col='suitability'	LANDUSE
0.50 Location -- Direct	
0.70 Distance from Chinchina city -- Std:Goal(2000,7231)	Distance_from_city
0.30 Distance from builtup areas -- Std:Goal(300,1278)	Distance_from_built_up
0.50 Suitability -- ExpVal	
0.27 Slope -- Std:Goal(20,75)	SLOPE
0.52 Land slides -- Std:Col='std'	SLIDE
0.15 Land use -- Std:Col="suitability score"	LANDUSE
0.06 Soil -- Direct	
0.60 Thickness -- Std:Interval	GEOL:average_thickness
0.30 Texture -- Std:Goal(0,50)	GEOL:average_perc_clay
0.10 Permeability -- Std:Goal(0.05,3.5)	GEOL:average_permeability

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.

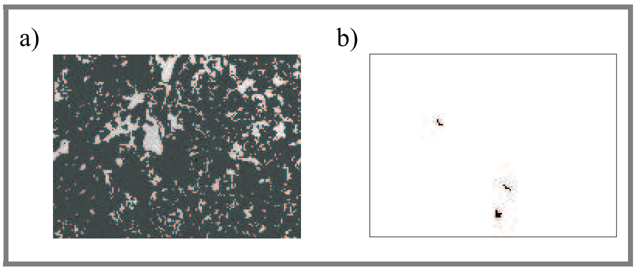


Fig. 7. The suitability map on the left (a) gives the three potential sites on the right (b). Due to printing limitations, the maps are printed as black and white, and the red–yellow–green gradation was converted to black–grey–white.

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 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):

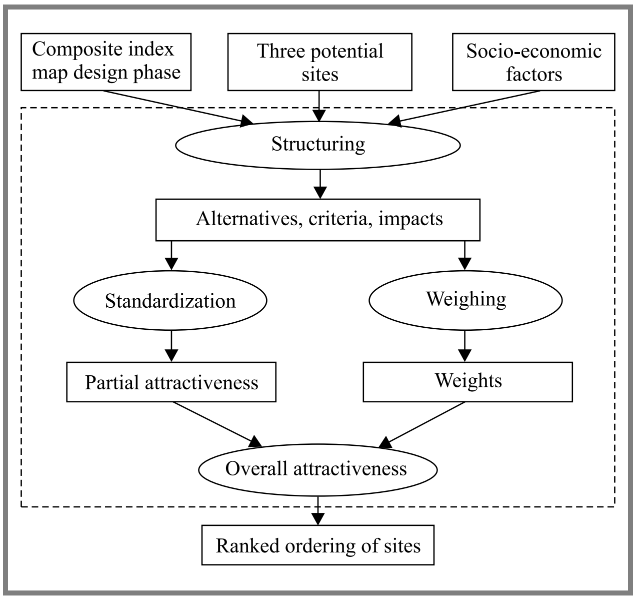


Fig. 8. Flowchart for the SMCE choice phase, which results in ranking of potential sites.

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).

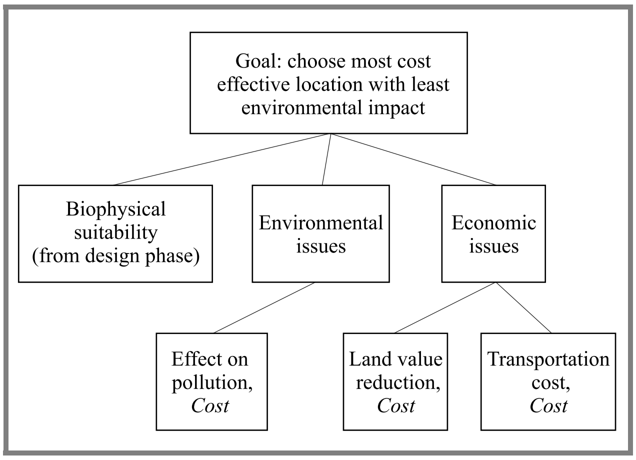
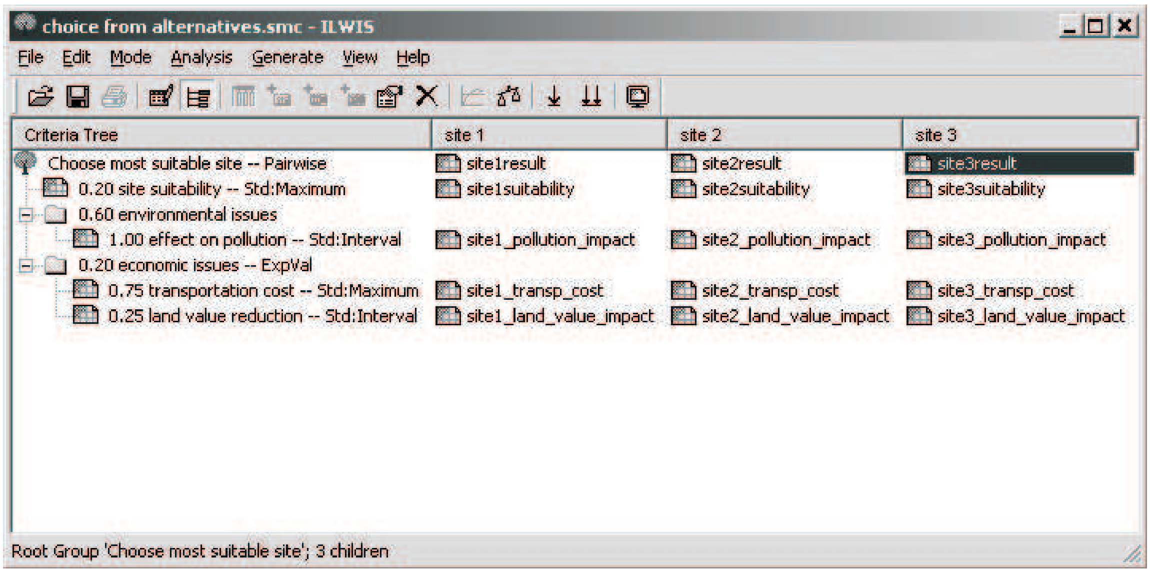

Goal: choose most cost effective location with least environmental impact
Biophysical suitability (from design phase)
Environmental issues
Economic issues
Effect on pollution, Cost
Land value reduction, Cost
Transportation cost, Cost

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



Criteria Tree	site 1	site 2	site 3
Choose most suitable site -- Pairwise	site1result	site2result	site3result
0.20 site suitability -- Std:Maximum	site1suitability	site2suitability	site3suitability
0.60 environmental issues			
1.00 effect on pollution -- Std:Interval	site1_pollution_impact	site2_pollution_impact	site3_pollution_impact
0.20 economic issues -- ExpVal			
0.75 transportation cost -- Std:Maximum	site1_transp_cost	site2_transp_cost	site3_transp_cost
0.25 land value reduction -- Std:Interval	site1_land_value_impact	site2_land_value_impact	site3_land_value_impact

Root Group 'Choose most suitable site'; 3 children

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).

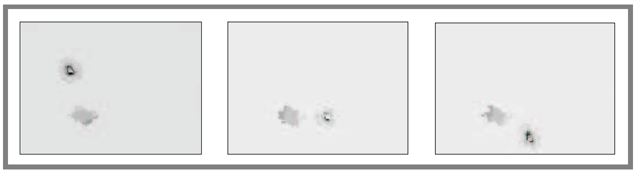


Fig. 11. The three composite index maps that correspond to the three sites. The maps are printed as black and white, and the red–yellow–green gradation was converted to black–grey–white.

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

Stochastic DEMATEL for structural modeling of a complex problematique for realizing safe, secure and reliable society

Hiroyuki Tamura and Katsuhiro Akazawa

http://doi.org/10.26636/jtit.2005.4.334

Abstract— In this paper we propose a revised Decision Making Trial and Evaluation Laboratory (DEMATEL), called stochastic DEMATEL, to extract structural model of a complex problematique and to represent the priority of each factor taking into account the uncertainty of structure. In the stochastic DEMATEL, the uncertainty of structure is expressed as a stochastic model. From numerical experiments and experimental analyses, the following results are obtained: when the structure is uncertain, stochastic DEMATEL could extract the features of structure by the degree of dispatching influences and the degree of central role; stochastic composite importance could express the uncertainty of priority and decide the priority taking into account the attitude of the decision maker; pessimistic, neutral or optimistic.

Keywords— *safe, secure and reliable society, structural modeling, stochastic DEMATEL, stochastic composite importance.*

1. Introduction

Decision Making Traial and Evaluation Laboratory (DEMATEL) has been widely used to extract a problem structure of a complex problematique [1–3]. By using DEMATEL we could quantitatively extract interrelationship among multiple factors contained in the problematique. In this case not only the direct influences but also the indirect influences among multiple factors are taken into account. Furthermore, we could find the dispatching factors that will rather affect the other factors, the receiving factors that will be rather affected by the other factors, the central factors that the intensity of sum of dispatching and receiving influences is big, and so forth.

It is important and useful to get the structural model of a problematique from which we could find the priority among multiple strategies to improve the structure. This is the main aim of DEMATEL. However, the conventional DEMATEL is insufficient to obtain significant implication of the priority of the strategies for decision making as follows:

1. Shortage of information on the importance of each factor

   Suppose we got three factors; “to get enough income”, “to get successor”, “to improve productivity”, in the problematique of agriculture. The decision maker is trying to find the order of priority among these three factors. Suppose the conventional DEMATEL found that “to improve productivity” is the most influential factor to improve the problem structure. However, if “to get successor” is the most important factor in the future agricultural problem, this factor should be the first priority for the strategic planning of agriculture. In the conventional DEMATEL it is hard to find the superiority of factors, since we could get only interrelationship of factors contained in the problematique. To overcome this difficulty we proposed a new criterion “composite importance (CI)” [4] combining the interrelationship of factors and the importance of each factor.

2. Shortage of flexibility to describe structural uncertainty

   Conventional DEMATEL describes the deterministic interrelationship among factors contained in the problematique. However, the strength of the interrelation among factors may be dependent on the various situations, and the fluctuation may depend on the factors taken into account. For example, in the agricultural problematique, “to improve productivity” may contribute “to get enough income”, but to what extent may be dependent on each farmhouse. “To get enough income” may contribute “to get successor” uniformly.

In this paper in the context of finding priority among multiple strategies to improve the structure of the problematique, we aim at three objectives as follows:

- We propose a stochastic DEMATEL to deal with flexible interrelationship among factors in the problematique.
- We show usefulness and future problem of stochastic DEMATEL through an empirical analysis of a simple numerical example where we deal with structural modeling of uneasy factors of university students and unmarried adults.
- We try to extract effective strategies to realize safe, secure and reliable society as the results of empirical analysis of uneasy factors of university students and unmarried adults.

2. DEMATEL and composite importance

2.1. Outline of DEMATEL

Suppose, in a complex problematique composed of n factors, binary relations and the strength of each relation are investigated. An example of binary relation is such that "How much would it contribute to resolve factor j by resolving factor i?" We would get $n \times n$ adjacent matrix X that is called the direct matrix. The (i, j) element x_{ij} of this matrix denotes the amount of direct influence from factor i to factor j. If the direct matrix X is normalizes as $X_r = \lambda X$, by using $\lambda = 1/$(the largest row sum of X), we would obtain

$$X^f = X_r + X_r^2 + \cdots = X_r(I - X_r)^{-1}. \tag{1}$$

Matrix X^f is called the direct/indirect matrix. The (i, j) element x_{ij}^f of the direct/indirect matrix denotes the amount of direct and indirect influence from factor i to factor j.

Suppose D_i denotes the row sum of ith row of matrix X^f. Then, D_i shows the sum of influence dispatching from factor i to the other factors both directly and indirectly. Suppose R_i denotes the column sum of ith column of matrix X^f. Then, R_i shows the sum of influence that factor i is receiving from the other factors. Furthermore, the sum of row sum and column sum $(D_i + R_i)$ shows the index representing the strength of influence both dispatching and receiving, that is, $(D_i + R_i)$ shows the degree of central role that the factor i plays in the problematique. If $(D_i - R_i)$ is positive, then the factor i is rather dispatching the influence to the other factors, and if negative, then the factor i is rather receiving the influence from the other factors. We call D_i, R_i, $(D_i + R_i)$ and $(D_i - R_i)$ the degree of dispatching influences, the degree of receiving influences, the degree of central role and the degree of cause, respectively.

There exist many case studies [5–10] of DEMATEL to get an appropriate structural model. Some of them are trying to get a structural model identifying the central factors and the causing factors based on the evaluation of the degree of central role and the degree of cause. The degree of cause denotes whether the factor is rather cause or effect. It does not reflect the amount of dispatching or receiving influence. Since the objective of this paper is to find the priority of the strategy to improve the overall structure, we turn our attention to the degree of dispatching influences.

2.2. Composite importance

Suppose based on the degree of dispatching influences we found a factor that may contribute to improve the overall structure. In this case to resolve this factor is not necessarily the best choice, since the factor that could contribute to resolve some important factors may be more efficient to resolve even if it may not contribute to improve overall structure. Since the original DEMATEL is not taking into account the importance of each factor itself, it is not possible to evaluate the priority among the factors. Similarly, it is not possible to evaluate the priority of each factor by just looking at the importance of each factor. We need to take into account both the strength of relationships among factors and the importance of each factor. To reflect both viewpoint we proposed the composite importance z as [4]

$$z = y_r + X^f y_r = (I + X^f)y_r, \tag{2}$$

where y_r denotes the normalized n-dimensional vector of y that denotes n-dimensional vector composed of the importance of each factor, where "normalized" means to divide each element of y by the largest element in y.

3. Stochastic DEMATEL

3.1. Stochastic direct matrix

In the ordinary DEMATEL the direct influence from factor i to factor j is written in the (i, j) element x_{ij} of the direct matrix X. Suppose the structure of the problematique is uncertain and x_{ij} is a random variable. Furthermore, suppose the stochastic parameter values of x_{ij} are different for different pair of i and j. When each element of the direct matrix is a random variable, each element of the direct/indirect matrix X^f is also a random variable. Furthermore, the composite importance z is also a random variable. Therefore, it is necessary to extend the ordinary DEMATEL to deal with uncertainty in the problem structure. We propose a stochastic DEMATEL in which we could take care of various uncertainties in the problem structure.

In stochastic DEMATEL let G be a set of stochastic direct matrices X^s generated by a Monte Carlo method from the direct matrix X^v with probabilistic information. The direct matrix with probabilistic information is an $n \times n$ matrix with (i, j) element x_{ij} and probability density function $g(x_{ij}|\theta_{ij})$, where θ_{ij} denotes the parameter including expectation and variance.

As described above in the ordinary DEMATEL we could get the element of direct matrix by asking such a question as "How much would it contribute to resolve factor j by resolving factor I?" On the other hand in the stochastic DEMATEL we need to collect the information on the variance as well as on the expectation of influence. Possible methods to collect information on variance are as follows:

- **Method 1.** We ask a respondent the best value and the worst value, by asking "How much would it contribute to resolve factor j at most by resolving factor i, and how much would it contribute to resolve factor j at least by resolving factor I?" From the best value and the worst value we could estimate the variance.

- **Method 2.** We ask multiple respondents on the value of direct matrix and compute the variance from these multiple direct matrices.

- **Method 3.** We combine Method 1 and Method 2. We ask each respondent the best value and the worst value of each element of the direct matrix. Then, we aggregate these data and estimate the variance of each element of the direct matrix.

3.2. Manipulation in stochastic DEMATEL

We normalize the stochastic direct matrix as

$$X_r^s = \lambda \cdot X^s, \tag{3}$$

where

$$\lambda = 1/(\text{the largest row sum of } X^s).$$

Then we obtain

$$X^{sf} = X_r^s + (X_r^s)^2 + \cdots = X_r^s(I - X_r^s)^{-1}, \tag{4}$$

where X^{sf} denotes a stochastic direct/indirect matrix that has the same property as the ordinary direct/indirect matrix. Stochastic composite importance is obtained as

$$z^s = y_r + X^{sf} y_r = (I + X^{sf}) y_r. \tag{5}$$

If we obtain stochastic direct/indirect matrices and stochastic composite importance for all the direct matrices contained in the set G, we could obtain the set G^f of the direct/indirect matrices and the set G^z of composite importance. Furthermore, we could obtain the set of the degree of dispatching, the set of the degree of receiving, the set of the degree of central role and the set of the degree of cause, respectively.

4. A simple numerical experiment

4.1. Structural modeling by stochastic DEMATEL

Suppose an overall structure is composed of three factors a, b and c, and the direct matrix is given by

$$X_e = \begin{bmatrix} 0 & 2 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{6}$$

In this structure factors a and b are mutually influenced, factor c is influenced by factor a, and factor b is influenced by factor c. Therefore, factor b is influenced by factor a both directly and indirectly. The intensity of direct influence is the largest from factor a to factor b.

As the degree of dispatching influences and the degree of central role, we obtained for factor a: 1.85 and 2.80, for factor b: 0.95 and 2.80 and for factor c: 0.65 and 1.30. As for the degree of dispatching influences, factor a is the largest and factor b is the next. Both factors a and b are the central factors, factor a is a cause factor and factor b is an effect factor. Suppose the structure of this simple numerical example is uncertain. Suppose besides the information on expectation given by the direct matrix, variance for each element is given by

$$Var_e = \begin{bmatrix} 0 & 0.04 & 0.04 \\ 0.04 & 0 & 0 \\ 0 & 0.04 & 0 \end{bmatrix}, \tag{7}$$

where the dispersion of the influence from factor a to factor b is assumed to be relatively small. It is assumed that cutting normal distribution between zero and infinity is assumed for probability density function.

We generated 1000 elements of a set G by using Monte Carlo method. Then, for each element of the set G, that is, for each stochastic direct matrix X_i^s ($i = 1, 2, \ldots, 1000$), we could obtain stochastic direct/indirect matrix and a set G^f.

Figure 1 shows the degree of dispatching influences and the degree of receiving influences obtained from the stochastic direct/indirect matrices. As seen in this figure the degree of dispatching influences of factor a is big and the degree of receiving influences of factor b is big. As the expectation (and the variance in the parenthesis) of the degree of dispatching influences and the degree of receiving influences we obtained for factor a: 1.8907 (0.0694), 1.0006 (0.1079), for factor b: 0.9936 (0.0966), 1.9064 (0.1167) and for factor c: 0.6805 (0.0418), 0.6577 (0.0175).

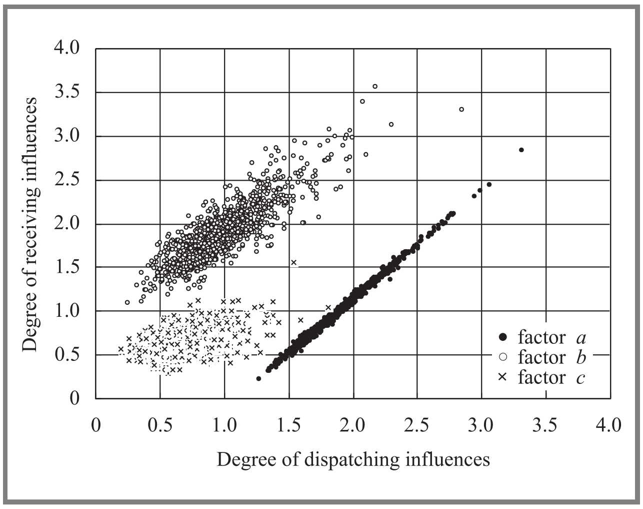


Fig. 1. Degree of dispatching influences and degree of receiving influences.

For factor a and factor b we found a big positive correlation between the degree of dispatching influences and the degree of receiving influences especially for factor a. The reason is that for both factors when they affect the other factor, the influence is fed back to themselves directly. On the other hand for factor c since the influence is fed back to itself indirectly, we did not find a big correlation (correlation coefficient = 0.51) between the degree of dispatching influences and the degree of receiving influences.

Table 1
Stochastic composite importance (numerical experiment)

Factors	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
Importance	0.4	0.4	0.4	0.3	0.1	0.7	0.1	0.7	0.4
Expected value	1.1563	0.7974	0.6722	0.8332	0.3860	0.8320	1.0670	1.0670	0.7637
2.5 percentile	0.9910	0.5936	0.5411	0.7049	0.2409	0.7622	0.9011	0.8791	0.5949
25 percentile	1.0822	0.7125	0.6166	0.7825	0.3254	0.7993	0.9930	0.9869	0.6973
Median	1.1414	0.7778	0.6621	0.8252	0.3733	0.8247	1.0549	1.0521	0.7544
75 percentile	1.2114	0.8658	0.7158	0.8748	0.4373	0.8550	1.1216	1.1313	0.8192
97.5 percentile	1.4006	1.0882	0.8679	1.0079	0.5971	0.9403	1.3114	1.3354	0.9915
CV	0.0911	0.1559	0.1216	0.0902	0.2283	0.0560	0.0992	0.1093	0.1290

In Fig. 1 we could draw many lines with gradient -1. The points on the same line have the same degree of central role, and the point located upper right side has a bigger degree of central role than the points on the line. These lines denote the indifference lines of the degree of central role. By using these indifference lines we could find that factors *a* and *b* are the central factors. As the expectations of the degree of central role we found for factors *a*, *b* and *c*: 2.8914, 2.9000 and 1.3382, respectively.

Next, we draw a line passing through the origin with gradient 1 in Fig. 1. Then, the points located lower right side of this line are the "cause" factors and the points located upper left side of this line are the "effect" factors. This fact implies that in every stochastic direct/indirect matrix it is found that factor *a* is a cause factor and factor *b* is an effect factor. Factor *c* is a cause factor or effect factor case by case.

If we compare the degree of dispatching influences, the degree of receiving influences and the degree of central role for ordinary DEMATEL and for stochastic DEMATEL, these values are almost identical. The values for stochastic DEMATEL are slightly larger than those for the ordinary DEMATEL. If we could find a precise probability distribution function and if we could generate infinitely many random numbers precisely, the expectation for both DEMATELs should agree each other in principle.

We found that we could get a proper structural model of a complex problematique under uncertainty by using the degree of dispatching influences and the degree of central role of the stochastic DEMATEL proposed in this paper.

4.2. Stochastic composite importance

If we assign the value of importance of each factor, we could evaluate the stochastic composite importance. Since we obtain 1000 values for each factor, we summarize the result in Table 1: percentiles (2.5%, 25%, median, 75%, 97.5%), expectation and coefficient of variation (*CV* = standard deviation/expectation).

In the ordinary DEMATEL we could decide the priority of each factor based on the value of composite importance itself. In the stochastic DEMATEL we use three stochastic decision principles as follows:

- **Expectation principle.** We decide the priority based on the expected value or median of composite importance.
- **Max-min principle.** We decide the priority of each factor by maximizing the worst value (either 2.5 percentile or 25 percentile) of composite importance. This principle reflects a pessimistic decision.
- **Max-max principle.** We decide the priority of each factor by maximizing the best value (either 75 percentile or 97.5 percentile) of composite importance. This principle reflects an optimistic decision.

As seen in Table 1 when the importance of each factor is 0.4, the composite importance of factor *a* is the largest under any of these three decision principles, therefore, the highest priority is given to factor *a*. When the importance of factors *a*, *b* and *c* is 0.3, 0.1 and 0.7, respectively, the priority of factor *a* is higher under the expectation principle and max-max principle, and the priority of factor *c* is slightly higher under the max-min principle. When the importance of factors *a*, *b* and *c* is 0.1, 0.7 and 0.4, respectively, the priority of factors *a* and *b* is higher under the expectation principle, the priority of factor *a* is higher under the max-min principle and the priority of factor *b* is higher under the max-max principle. In this case under the attitude of pessimistic decision, factor *a* is chosen to be resolved, and under the attitude of optimistic decision, factor *b* is chosen to be resolved. In this case the expectation for factors *a* and *b* is almost identical, *CV* for factor *a* is smaller than that for factor *b*, and factor *a* is chosen under the max-min principle and factor *b* is chosen under the max-max principle. This implies that the priority decided by max-min principle and max-max principle depends on the variance of composite importance of each factor.

As seen above the stochastic DEMATEL could describe the uncertainty of the structure of complex problematique, could describe the uncertainty of priority by the stochastic composite importance and could decide the priority of each factor reflecting the decision makers attitude whether he/she is pessimistic, neutral or optimistic.

5. Structural modeling of uneasy factors by stochastic DEMATEL

5.1. Data

We use the data previously obtained from university students and unmarried adults [4].

For university students 10 uneasy factors are chosen as follows:

1. Career to pursue (CAR)
2. Scholastic performance (SCH)
3. Home economy (HOE)
4. Health of myself (HEM)
5. Health of family (HEF)
6. Marriage (MAR)
7. Looks (LOO)
8. Ability/character (ABI)
9. Human relations (HUR)
10. Job and work (JAW)

For unmarried adults 9 uneasy factors are chosen as follows:

1. Home economy
2. Health of myself (HEM)
3. Health of family (HEF)
4. Unemployment (UNE)
5. Marriage (MAR)
6. Looks (LOO)
7. Ability/character (ABI)
8. Human relations (HUR)
9. Job and work (JAW)

Respondents to the questionnaire are 10 university students and 10 unmarried adults. The importance of each factor is asked to the respondents by 5-grade evaluation where the importance of each factor means the degree of feeling uneasy for each factor. Then, the strength of binary relation for each pair of factors is asked by 3-grade evaluation, We look at the binary relation such that "How much would it contribute to resolve factor b (the anxiety for SCH) by resolving factor a (the anxiety for CAR) ?"

The direct matrix is obtained by averaging the data of 10 people on the strength of binary relations. The data for the importance of each factor are first normalized between 0 and 1 and then averaged for 10 people.

Structural model for uneasy factors of university students is described as follows: the degree of central role for CAR (4.75) is high and CAR has the property of both cause factor and effect factor, but since the degree of cause for CAR (-0.35) is negative, CAR is rather an effect factor. Actually, CAR is greatly affected by ABI, SCH, HOE and JAW.

Besides CAR the degree of central role for HOE (3.63), ABI (3.63), JAW (3.54) and SCH (3.35) are high. Especially, the degree of cause for ABI (1.41) is high, this is a central factor with the property of cause factor.

Structural model for uneasy factors of unmarried adult is described as follows: the degree of central role for JAW (6.07) is high, and then ABI (5.79), HOE (5.40). JAW and ABI are mainly cause factor, however, they have the property of effect factor as well. On the other hand HOE is affected by UNE, JAW and others, and has the property of effect factor.

In Table 2 the degree of dispatching influences and composite importance of university students and unmarried adults are shown. Concerned with the degree of dispatching influences ABI, CAR are high for university students and ABI, JAW, HEM are high for unmarried adults with this order. Concerned with the composite importance CAR, ABI are high for university students and ABI, JAW, HEM are high for unmarried adults with this order. This implies that by resolving these factors overall uneasiness is resolved enormously.

We need to pay attention that for university students the order of factors for the degree of dispatching influences is different from the order of factors for the composite importance. The reason why is that in the composite importance the degree of dispatching influences as well as the importance of each factor and the importance of affecting factors are reflected. For example the degree of dispatching influences of CAR is not so high, but since the importance of CAR is high, the composite importance of CAR is high in consequence. Therefore, it is clarified that from the view point of importance CAR is to be resolved and from the view point of dispatching influences ABI is to be resolved.

5.2. Structural modeling by stochastic DEMATEL

Suppose the structure of the uneasy factors is uncertain. Expectation and variance of probability distribution is obtained by the dispersion of the data contained in multiple respondents reply in the direct matrix. Probability density function is assumed to be a cutting normal distribution defined on $[0, \infty)$. Based on these probabilistic information 1000 stochastic direct matrices are generated by using a Monte Carlo method.

Tables 3 and 4 show a structural model extracted by the stochastic DEMATEL from the uneasy factors of university students and unmarried adults, respectively. The ex-

Table 2
Composite importance

Factors		CAR	SCH	HOE	HEM	HEF	UNE	MAR	LOO	ABI	HUR	JAW
Students	D	2.200	1.643	1.825	1.052	0.607		0.866	0.753	2.521	1.510	1.826
	Importance	0.675	0.550	0.600	0.350	0.400		0.500	0.450	0.500	0.450	0.425
	CI	1.796	1.411	1.527	0.894	0.706		0.928	0.842	1.794	1.232	1.374
Adults	D			2.167	3.041	1.324	1.565	1.824	1.553	3.307	2.529	3.243
	Importance			0.475	0.550	0.550	0.400	0.425	0.550	0.600	0.425	0.475
	CI			1.528	2.028	1.182	1.163	1.316	1.299	2.189	1.659	2.048

Table 3
Structural extraction by stochastic DEMATEL (university students)

Values		CAR	SCH	HOE	HEM	HEF	MAR	LOO	ABI	HUR	JAW
D	Expected	2.248	1.818	2.217	1.613	0.948	1.422	1.063	2.523	1.859	2.296
	Median	2.140	1.679	2.156	1.472	0.861	1.290	0.947	2.456	1.730	2.254
	CV	0.359	0.422	0.347	0.427	0.466	0.439	0.511	0.263	0.404	0.342
$D+R$	Expected	4.934	3.799	4.286	3.044	2.112	2.706	3.209	3.939	3.762	4.223
	Median	4.689	3.570	4.031	2.809	1.926	2.505	2.953	3.753	3.454	4.023
	CV	0.339	0.366	0.341	0.368	0.398	0.381	0.383	0.294	0.365	0.337

Table 4
Structural extraction by stochastic DEMATEL (unmarried adults)

Values		HOE	HEM	HEF	UNE	MAR	LOO	ABI	HUR	JAW
D	Expected	2.181	2.630	1.661	1.734	1.970	1.813	2.772	2.268	2.703
	Median	1.992	2.551	1.490	1.571	1.767	1.634	2.665	2.089	2.601
	CV	0.453	0.345	0.486	0.484	0.476	0.498	0.341	0.447	0.362
$D+R$	Expected	4.820	4.283	3.347	3.987	4.495	3.696	5.074	4.575	5.182
	Median	4.453	4.015	3.035	3.656	4.136	3.372	4.779	4.305	4.815
	CV	0.405	0.368	0.433	0.422	0.416	0.427	0.368	0.405	0.385

pected value obtained in these tables did not agree with the results obtained by the ordinary DEMATEL. Since the expected value and median are little bit different, the assumption of cutting normal distribution may not be appropriate. However, since concerned with the order of the degree of dispatching influences (D) and the degree of central role $(D+R)$ the result obtained by the stochastic DEMATEL gave a good agreement with the result obtained by the ordinary DEMATEL, the stochastic DEMATEL is appropriate to extract the property of the structural model.

Since the value of coefficient of variation (CV) is around 0.4, the uncertainty of the structure is fairly big. If we look at CV of D and $D+R$ for each factor, we find that CV of D and $D+R$ for ABI is smaller than for the other factors and CV of D and $D+R$ for LOO is large for university students. This implies that by resolving ABI university students could expect a stable effect, but by resolving LOO university students should anticipate uncertain effect. For unmarried adults CV of D and $D+R$ for ABI and HOM is small and the variation of CV depending upon the different factors is relatively small.

5.3. Stochastic composite importance

In Tables 5 and 6 stochastic composite importance of each factor for university students and unmarried adults is shown, respectively.

For university students composite importance, that is, the priority of CAR, ABI and HOE are large with this order under the expectation principle and max-max principle, where the priority of ABI and HOE is reversed for 97.5 percentile. This is due to the fact that CV for HOE is larger than that for ABI. Under the expectation principle the priority obtained by the stochastic DEMATEL gave a good agreement with the priority obtained by the ordinary DEMATEL. This result implies that reliability of the stochastic com-

Table 5
Stochastic composite importance (university students)

Values	CAR	SCH	HOE	HEM	HEF	MAR	LOO	ABI	HUR	JAW
Expected	1.804	1.482	1.708	1.171	0.880	1.201	0.994	1.775	1.394	1.593
2.5 percentile	1.170	0.935	1.079	0.719	0.561	0.777	0.630	1.188	0.841	0.914
25 percentile	1.495	1.207	1.425	0.917	0.722	0.969	0.807	1.561	1.124	1.293
Median	1.752	1.411	1.671	1.093	0.834	1.134	0.938	1.741	1.328	1.566
75 percentile	2.053	1.676	1.935	1.365	0.995	1.370	1.112	1.956	1.591	1.846
97.5 percentile	2.726	2.392	2.574	2.001	1.475	1.976	1.725	2.497	2.348	2.454
CV	0.226	0.266	0.228	0.300	0.254	0.259	0.280	0.190	0.274	0.252

Table 6
Stochastic composite importance (unmarried adults)

Values	HOE	HEM	HEF	UNE	MAR	LOO	ABI	HUR	JAW
Expected	1.540	1.836	1.351	1.252	1.394	1.434	1.942	1.540	1.795
2.5 percentile	0.898	1.104	0.855	0.736	0.827	0.895	1.195	0.856	1.050
25 percentile	1.186	1.541	1.078	0.970	1.075	1.124	1.645	1.175	1.458
Median	1.449	1.801	1.266	1.177	1.292	1.345	1.889	1.454	1.739
75 percentile	1.795	2.080	1.522	1.442	1.593	1.635	2.196	1.808	2.074
97.5 percentile	2.636	2.740	2.294	2.231	2.554	2.573	2.923	2.664	2.779
CV	0.314	0.243	0.290	0.329	0.331	0.309	0.239	0.324	0.268

posite importance obtained by the stochastic DEMATEL is quite high.

Under the max-min principle the priority of ABI is the highest and then that of CAR and HOE. The reason why the priority of ABI is the highest is that the max-min principle reflects the pessimistic attitude of decision and that *CV* of *D* for ABI is small, and as the result *CV* of composite importance is also small. This will lead to the expectation of certain effect by resolving uneasiness of ABI.

For unmarried adults the priority of ABI, HEM and JAW are large with this order under all the three principles except that for 97.5 percentile the priority of JAW and HEM is reversed. This order of priority obtained by the stochastic composite importance is different from that obtained by the composite importance of ordinary DEMATEL: ABI, JAW and HEM. The reason why we get this result is that for unmarried adults the elements of the stochastic direct matrices are smaller than those of the direct matrix. As the result each element of the degree of dispatching influences (D) is reduced to be relatively small. As we know the composite importance reflects both D and importance of each factor. Since the value of D is reduced to be relatively small in the stochastic composite importance, weight for D is reduced to be smaller than that for importance of each factor. As the result, in the stochastic DEMATEL the priority of HEM became higher than that of JAW, because the importance of HEM is larger than that of JAW for unmarried adults.

The reason why the elements of the stochastic direct matrices are smaller than those of the direct matrix may be due to the error arisen from inappropriate assumption of probability density function. Although this error does not cause serious defects when we evaluate the degree of dispatching influences (D) and the degree of central role $(D+R)$, we may get some defects when we evaluate composite importance. Therefore, to overcome this difficulties we need to develop a method of identifying appropriate probability distribution function or to develop a non-parametric approach.

6. Concluding remarks

In this paper a stochastic DEMATEL is proposed for structural modeling of a complex problematique taking into account the uncertainty of structure. This method is obtained by extending the deterministic variables in the ordinary DEMATEL to random variables. To show the validity of the method a simple numerical example and a structural modeling of uneasy factors are included for the purpose of realizing safe, secure and reliable society.

New knowledge obtained in this study is as follows:

- Stochastic DEMATEL could extract the characteristics of the structure even when there exist uncertainty in the structure.
- Stochastic composite importance could describe the uncertainty of priority arising from the uncertainty of the structure, and could decide the priority taking into account the attitude of the decision maker towards risk; pessimistic, neutral or optimistic.

- In order to resolve uneasy factors of university students uneasiness of CAR and ABI is efficient to be resolved. CAR is to be resolved from the view point of the importance of the factor and ABI is to be resolved from the view point of the degree of dispatching influences. When the decision maker's attitude toward risk is pessimistic, it is desirable to resolve the uneasiness of ABI, since certain effect can be expected by doing so.
- To resolve the uneasiness of ABI is the most effective for unmarried adults.

It is demonstrated above that the stochastic DEMATEL and the information obtained by the stochastic composite importance are quite useful for structural modeling of complex problematique.

For further study we need to develop a method of identifying appropriate probability distribution function or we need to develop a non-parametric approach. We also need to develop a method of collecting information on variance. For these purposes we need to experience more empirical analysis of various case studies.

Acknowledgement

This research was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) under Grant-in-Aid for Creative Scientific Research (Project No. 13GS0018). The authors would like to thank Ms. Haruna Nagata who was an undergraduate student of Osaka University for her kind collaboration with this research project.

References

[1] E. Fontela and A. Gabus, "DEMATEL, innovative methods". Rep. No. 2, "Structural analysis of the world problematique (methods)", Battelle Geneva Research Institute, 1974.

[2] J. N. Warfield, *Societal Systems – Planning, Policy and Complexity*. New York: Wiley, 1976.

[3] *Large Scale Systems – Modeling, Control and Decision Making*, H. Tamura, Ed. Tokyo: Shokodo, 1986 (in Japanese).

[4] K. Akazawa, H. Nagata, and H. Tamura, "Structural modeling of uneasy factors for creating safe, secure and reliable society", *J. Pers. Finan. Econom.*, vol. 18, pp. 201–210, 2003 (in Japanese).

[5] T. Yamagishi, *From Safe and Secure Society to Reliable Society*. Tokyo: Chuko-shinsho, 1999 (in Japanese).

[6] A. Yuzawa, "A state and subjects of TMO conception for city core vitalization countermeasure – a case study of maebashi TMO conception", *Bull. Maebashi Institute of Technology*, vol. 5, pp. 61–67, 2002.

[7] I. Kimata, "Synthetic preliminary evaluation analysis on expectation of a sewerage improvement system in a rural community using the decision making and evaluation laboratory method-investigation of inhabitant consciousness of a sewerage improvement system (II)", *Trans. JSIDRE*, vol. 189, pp. 17–25, 1997 (in Japanese).

[8] I. Kimata, "Synthetic comparison analysis of inhabitant consciousness between before and after rural community sewerage improvement project in hilled rural area", *Trans. JSIDRE*, vol. 213, pp. 119–127, 2001 (in Japanese).

[9] M. Yamazaki, K. Ishibe, S. Yamashita, I. Miyamoto, M. Kurihara, and H. Shindo, "An analysis of obstructive factors to welfare service using DEMATEL method", Rep. Faculty of Engineering, Yamanashi University, vol. 48, pp. 25–30, 1997 (in Japanese).

[10] Y. Zhou, Y. Kawamoto, and Y. Honda, "A study on systematization of tourism development in China", *Mem. Fac. Eng. Fukui Univ.*, vol. 49, no. 2, pp. 177–184, 2001 (in Japanese).

Hiroyuki Tamura received the B.Sc., M.Sc. and Ph.D. degrees in engineering from Osaka University in 1962, 1964 and 1971, respectively. He was a research engineer with Mitsubishi Electric Corporation from 1964 to 1971. From 1971 to 1987 he was an Associate Professor, and from 1987 to 2003 he was a Professor in Osaka University. Since 2003 he has been a Professor in Kansai University, with the Department of Electrical Engineering, and Professor emeritus of Osaka University. His research interest lies in systems methodology for large-scale complex systems such as modeling, control and decision making, and its applications to societal systems and manufacturing systems. He has written more than 100 journal papers and more than 40 review papers in this field. He is a fellow of Operations Research Society of Japan, senior member of IEEE, member of INFORMS, SRA, etc.
e-mail: H.Tamura@kansai-u.ac.jp
Faculty of Engineering
Kansai University
Suita, Osaka 564-8680, Japan

Katsuhiro Akazawa received the B.Sc., M.Sc. and Ph.D. degrees in agriculture from Okayama University in 1993, 1995 and 1999, respectively. He was a Research Associate in Osaka University from 1998 to 2002. From 2002 to 2003 he was a lecturer in Shimane University. Since 2003 he has been an Associate Professor in Shimane University, with the Faculty of Life and Environmental Science. His research interest lies in modeling of consumer's preference for environmental and market goods. In particular, he deals with the improvement of the preference evaluation methods such as choice experiments and travel cost model.
e-mail: akazawa@life.shimane-u.ac.jp
Faculty of Life and Environmental Science
Shimane University
Matsue 690-8504, Japan

WannaCry Ransomware: Analysis of Infection, Persistence, Recovery Prevention and Propagation Mechanisms

Maxat Akbanov[1], Vassilios G. Vassilakis[2], and Michael D. Logothetis[3]

[1] *Department of Computer Science, University of York, York, United Kingdom*
[2] *University of York, York, United Kingdom*
[3] *WCL, Dept. of Electrical and Computer Engineering, University of Patras, Patras, Greece*

http://doi.org/10.26636/jtit.2019.130218

Abstract—In recent years, we have been experiencing fast proliferation of different types of ransomware targeting home users, companies and even critical telecommunications infrastructure elements. Modern day ransomware relies on sophisticated infection, persistence and recovery prevention mechanisms. Some recent examples that received significant attention include WannaCry, Petya and BadRabbit. To design and develop appropriate defense mechanisms, it is important to understand the characteristics and the behavior of different types of ransomware. Dynamic analysis techniques are typically used to achieve that purpose, where the malicious binaries are executed in a controlled environment and are then observed. In this work, the dynamic analysis results focusing on the infamous WannaCry ransomware are presented. In particular, WannaCry is examined, during its execution in a purpose-built virtual lab environment, in order to analyze its infection, persistence, recovery prevention and propagation mechanisms. The results obtained may be used for developing appropriate detection and defense solutions for WannaCry and other ransomware families that exhibit similar behaviors.

Keywords—dynamic malware analysis, ransomware, WannaCry.

1. Introduction

Ransomware threat is currently considered to be the main moneymaking scheme for cyber criminals and the key threat to Internet users [1], [2]. In recent years, the appearance of new types of ransomware has been observed, combining the use of worm-like spreading mechanisms and advanced recovery prevention schemes. Recent examples include WannaCry [3], [4] and Petya [5], [6], which exploit the weaknesses of Microsoft Windows, as well as BadRabbit [7], which spreads via insecure compromised websites.

From the defense perspective, the design of new countermeasures is considered, in addition to traditional security approaches, an important and trending task in this field. Such a design, however, requires a comprehensive analysis of ransomware functionality and behavior. This typically involves a wide range of malware analysis tools and techniques. Such techniques may be broadly classified as *static* and *dynamic*. Static analysis is performed without executing the malicious binary, while dynamic analysis involves executing the binary in an isolated environment.

In one of our previous works [8], we performed an initial static and dynamic analysis of WannaCry to identify its resources and functions, as well as its use of dynamic-link libraries (DLLs) and communication protocols. In this work, we have performed a comprehensive dynamic analysis, focusing on WannaCry's infection, persistence, recovery prevention and propagation mechanisms. The techniques presented are also applicable in the cases of other ransomware families whose characteristics are similar to that of WannaCry, such as worm-spreading mechanisms and public-key based encryption. In particular, the research presented examines WannaCry's behavior during its execution in a safe, purpose-built virtual lab environment at the University of York. The results obtained may form a basis for designing and developing effective ransomware defense solutions.

The rest of the paper is organized as follows. In Section 2, we present the relevant background information on ransomware in general and on WannaCry in particular. In Section 3, the main findings from the dynamic analysis of WannaCry we have performed, including its encryption process, recovery prevention and propagation mechanisms, are presented. Finally, Section 4 draws conclusions and discusses potential future directions.

2. Background

2.1. Ransomware

Ransomware is a type of malicious software (malware) that prevents users from accessing or limits their access to the system or files, either by locking the screen or by encrypting files, until a ransom is paid [9]. In most cases, ransomware leaves the user with very few options, such as only allowing the victim to communicate with the attacker and pay the ransom.

The most common types of ransomware use some form of encryption, including both symmetric and public-key based encryption schemes. Ransomware that relies on public-key encryption is particularly difficult to mitigate, since the encryption keys are stored in a remote command and control (C&C) server. There is usually a time limit for ransom to be paid, the users are provided with a special website to purchase cryptocurrency (e.g. Bitcoins) and step-by-step instructions on how to pay the ransom.

The lifecycle of modern day ransomware typically consists of the following steps [10]: distribution, infection, C&C communications, file search, file encryption and ransom demand.

2.2. WannaCry

WannaCry ransomware (also known as Wana Decrypt0r, WCry, WannaCry, WannaCrypt, and WanaCrypt0r) was observed during a massive attack across multiple countries on 12 May 2017 [11]. According to multiple reports from security vendors, the total of 300,000 systems in over 150 countries had been severely damaged. The attack affected a wide range of sectors, including healthcare, government, telecommunications and gas/oil production.

The difficulty in protecting against WannaCry stems from its ability to spread to other systems by using a worm component. This feature makes the attacks more effective and requires defense mechanisms that can react quickly and in real time. Furthermore, WannaCry has an encryption component that is based on public-key cryptography.

During the infection phase, WannaCry uses the *EternalBlue* and *DoublePulsar* exploits that were allegedly leaked in April 2017 by a group called The Shadow Brokers. EternalBlue exploits the server message block (SMB) vulnerability that was patched by Microsoft on March 14, 2017 and has been described in the security bulletin MS17-010 [12]. This vulnerability allows the adversaries to execute a remote code on the infected machines by sending specially crafted messages to an SMB v1 server, connecting to TCP ports 139 and 445 of unpatched Windows systems. In particular, this vulnerability affects all unpatched Windows versions starting from Windows XP to Windows 8.1, except for Windows 10.

DoublePulsar is a persistent backdoor that may be used to access and execute code on previously compromised systems, thus allowing the attackers to install additional malware on the system. During the distribution process, WannaCry's worm component uses EternalBlue for initial infection through the SMB vulnerability, by actively probing appropriate TCP ports and, if successful, tries to implant the DoublePulsar backdoor on the infected systems.

3. WannaCry Analysis

In this section, we present our findings based on the dynamic analysis of WannaCry we have performed. Samples of WannaCry were obtained from VirusShare [13]. Two executable files were analyzed: the worm component and the encryption component (Table 1).

Table 1
WannaCry components

	Worm component
MD5	db349b97c37d22f5ea1d1841e3c89eb4
SHA1	e889544aff85ffaf8b0d0da705105dee7c 97fe26
SHA256	24d004a104d4d54034dbcffc2a4b19a11 f39008a575aa614ea04703480b1022c
File type	PE32 executable (GUI) Intel 80386, for MS Windows
	Encryption component
MD5	84c82835a5d21bbcf75a61706d8ab549
SHA1	5ff465afaabcbf0150d1a3ab2c2e74f3a4 426467
SHA256	ed01ebfbc9eb5bbea545af4d01bf5f1071 661840480439c6e5babe8e080e41aa
File type	PE32 executable (GUI) Intel 80386, for MS Windows

3.1. Testbed

In order to analyze WannaCry, a virtual testbed shown in Fig. 1 was built. The characteristics of the host machine are as follows: Intel Core i7-4700MQ 2.40 GHz and 16 GB RAM. The host machine acts as a virtual switch and is running REMnux [14], which is a free Linux toolkit for reverse engineering and malware analysis. Two virtual machines (VMs), running Windows 7 SP1, were used. The first VM was infected with WannaCry, whereas the other VM was clean. A custom network VMnet 5 – 192.168.180.0/24 was created with the Virtual Network Editor option in VMWare hypervisor. This testbed allows observing domain name system (DNS) queries made by WannaCry during the infection and replication process across internal and external

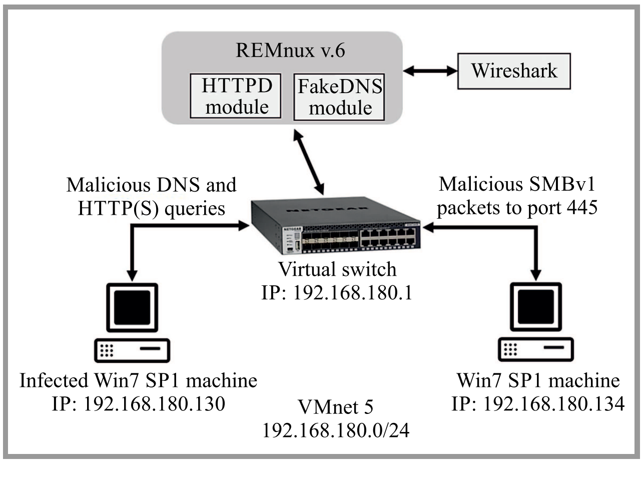


Fig. 1. Testbed for dynamic WannaCry analysis.

networks via port 445 of the SMB v1 protocol. The REMnux machine acts as a DNS and HTTP server, and is able to intercept all network communications using Wireshark. DNS and HTTP services in REMnux were enabled using FakeDNS and HTTP Daemon utilities, respectively.

The system level actions performed by WannaCry were observed on the infected Windows 7 SP1 machine with the 192.168.180.130 IP address. In order to observe and report the actions that WannaCry took while running on the system, the SysAnalyzer tool [15] was used. The main benefit of SysAnalyzer is that it is capable of taking system snapshots before and after malware execution, thus making it possible to inspect system attributes, such as running processes, open ports, DLLs loaded, registry key changes, run time file modifications, scheduled tasks, mutual exclusion objects (mutexes) and network connections. SysAnalyzer is also capable of taking memory dumps and scanning them for specific regular expressions. Before executing the WannaCry sample on the infected machine, the SysAnalyzer's configuration wizard was set to apply a 120 s delay between system snapshots, thus allowing to inspect all system attribute changes.

3.2. Libraries and Functions

Analysis performed with the Pestudio tool [16] revealed that the worm and the encryption components of WannaCry contain DLLs shown in Tables 2 and 3, respectively. During its execution, the worm component invokes *iphlpapi.dll* to retrieve network configuration settings for the infected host. *Kernel32.dll* and *msvcrt.dll* are the two libraries most frequently invoked by the encryption component. This may indicate that the main encryption functionality was implemented by these two malicious libraries. To confirm this, the imported functions of the libraries needed to be examined.

Table 2
DLLs of the worm component

Library	Imports	Description
ws2_32.dll	13	Windows Socket 2.0 32-bit DLL
iphlpapi.dll	2	IP Helper API
wininet.dll	3	Internet Extensions for Win32
kernel32.dll	32	Windows NT Base API Client DLL
advapi32.dll	11	Advanced Windows 32 Base API
msvcp60.dll	2	Windows NT C++ Runtime Library DLL
msvcrt.dll	28	Windows NT CRT DLL

Table 3
DLLs of the encryption component

Library	Imports	Description
kernel32.dll	54	Windows NT Base API Client DLL
advapi32.dll	10	Advanced Windows 32 Base API
user32.dll	1	Multi-User Windows User API Client DLL
msvcrt.dll	49	Windows NT CRT DLL

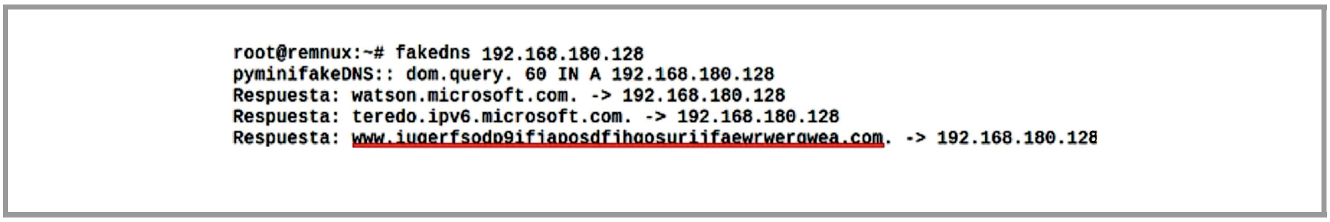
Table 4
Functions of the encryption component

Function	Location
GetCurrentThread	0xa53a
GetStartupInfoA	0xa97a
StartServiceCtrDispatcherA	0xa6f6
RegisterServiceCtrDispatcherA	0xa6d8
CreateServiceA	0xa688
StartServiceA	0xa662
CryptGenRandom	0xa650
CryptAcquireContextA	0xa638
OpenServiceA	0xa714
GetAdaptersInfo	0xa792
InternetOpenUrlA	0xa7c8

Table 5
Functions of the encryption component

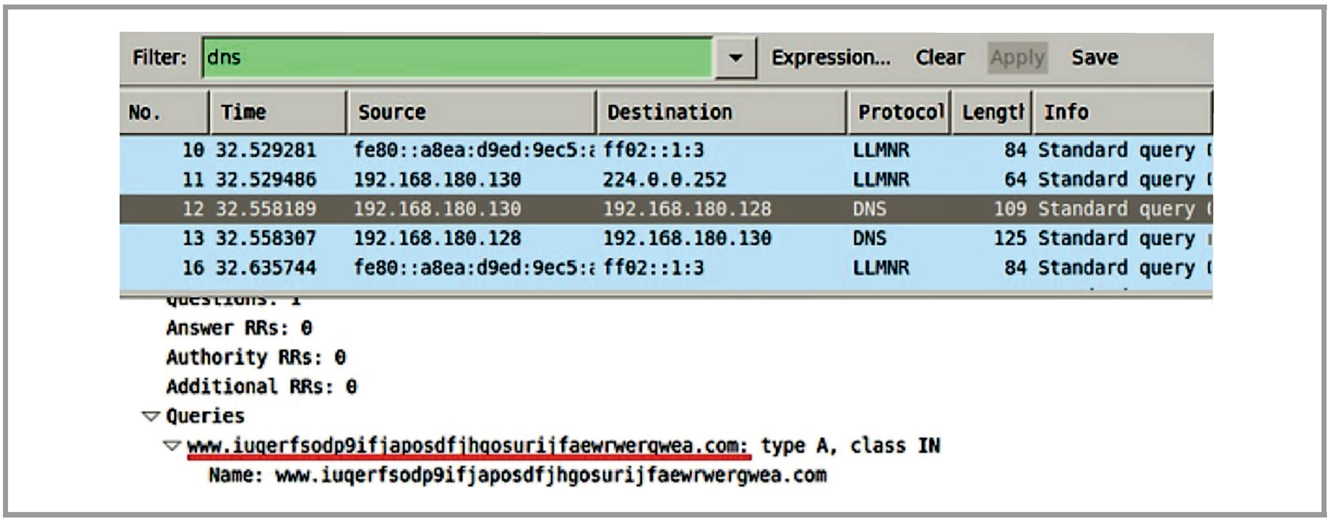
Function	Location
OpenMutexA	0xda84
GetComputerNameW	0xd8b2
CreateServiceA	0xdc2a
OpenServiceA	0xdc62
StartServiceA	0xdc52
CryptReleaseContext	0xdc14
RegCreateKeyW	0xdc04
fopen	0xdcd4
fread	0xdccc
fwrite	0xdcc2
fclose	0xdcb8
CreateFileA	0xd922
ReadFile	0xd964

The imported functions of the samples were observed by Pestudio. The most suspicious functions identified among them are shown in Tables 4 and 5. One may observe that in general, WannaCry uses Microsoft's crypto, file management and C runtime file APIs. The crypto API library is used to generate and manage random symmetric and asymmetric cryptographic keys.



```
root@remnux:~# fakedns 192.168.180.128
pyminifakeDNS:: dom.query. 60 IN A 192.168.180.128
Respuesta: watson.microsoft.com. -> 192.168.180.128
Respuesta: teredo.ipv6.microsoft.com. -> 192.168.180.128
Respuesta: www.iuqerfsodp9ifjaposdfjhgosurijfaewrwergwea.com. -> 192.168.180.128
```

Fig. 2. FakeDNS capture of the malicious DNS request.



```
No.  Time       Source                   Destination      Protocol Length Info
  10 32.529281  fe80::a8ea:d9ed:9ec5:ε  ff02::1:3        LLMNR       84 Standard query (
  11 32.529486  192.168.180.130          224.0.0.252      LLMNR       64 Standard query (
  12 32.558189  192.168.180.130          192.168.180.128  DNS        109 Standard query (
  13 32.558307  192.168.180.128          192.168.180.130  DNS        125 Standard query r
  16 32.635744  fe80::a8ea:d9ed:9ec5:ε  ff02::1:3        LLMNR       84 Standard query (
    Questions: 1
    Answer RRs: 0
    Authority RRs: 0
    Additional RRs: 0
  ▽ Queries
    ▽ www.iuqerfsodp9ifjaposdfjhgosurijfaewrwergwea.com: type A, class IN
        Name: www.iuqerfsodp9ifjaposdfjhgosurijfaewrwergwea.com
```

Fig. 3. Wireshark capture of the malicious DNS request.

3.3. Initial Interactions

The dynamic analysis conducted has revealed that, upon startup, the worm component tries to connect to the following domain, using the *InternetOpenUrl* function:

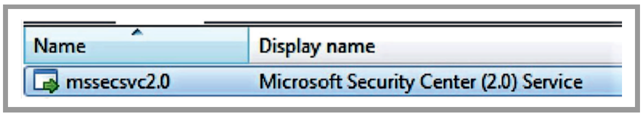
www.iuqerfsodp9ifjaposdfjhgosurijfaewrwergwea.com

The aforementioned domain is a kill-switch domain. This means that if the domain is active, the worm component stops running. On the other hand, if the worm component cannot establish a connection with this domain (e.g. if the domain is not active or if there is no connectivity), it continues to run and registers itself as a "Microsoft Security Center (2.0) Service" *mssecsvs2.0* process on the infected machine. Hence, this kill-switch domain may be used as part of a detection technique when developing a defense system.

The FakeDNS utility at REMnux captures the malicious DNS request on port 80 (Fig. 2), while Wireshark shows (Fig. 3) the DNS packet query field from the infected machine (IP 192.168.180.130) to the DNS server on REMnux (IP 192.168.180.128).

3.4. Persistence Mechanisms

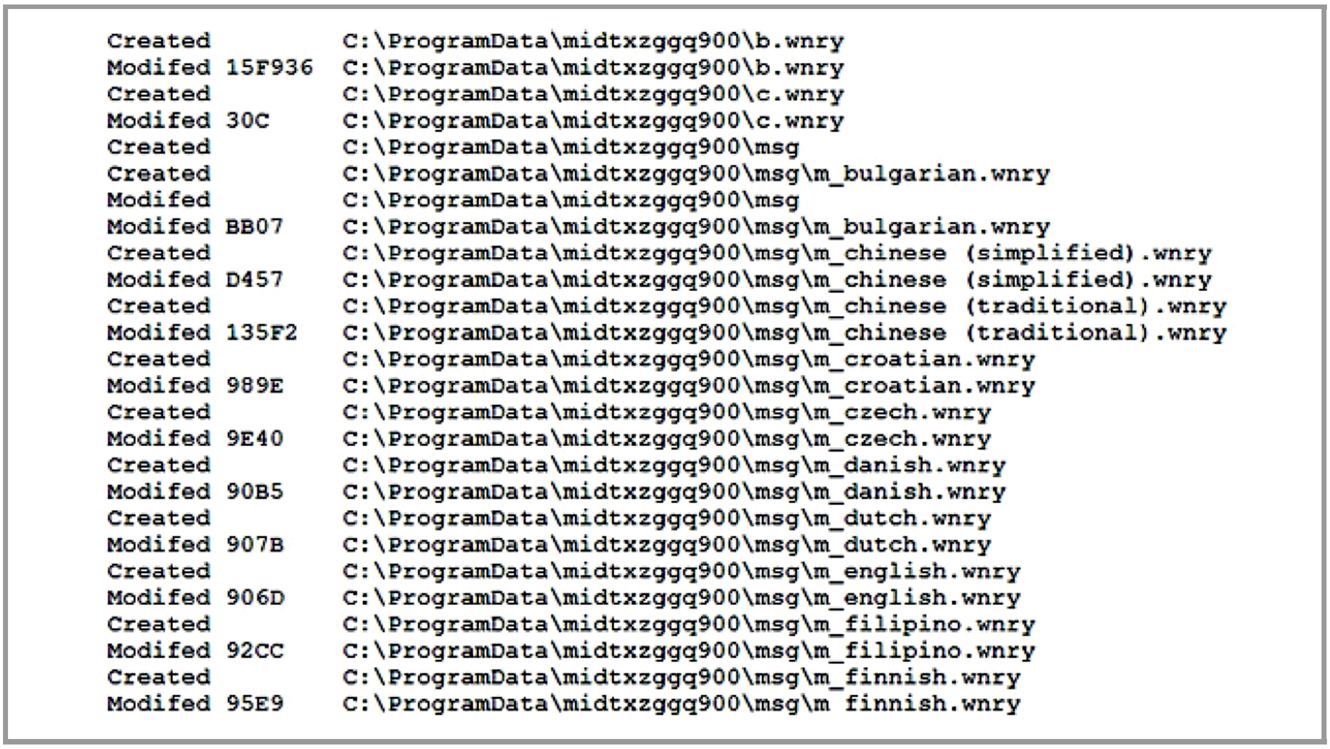
After connection failure with the kill-switch domain, the worm component attempts to create a *mssecsvs2.0* process with the DisplayName of "Microsoft Security Center (2.0) Service". This can be observed in the Process Hacker tool with 4016 PID, indicating that the service has been launched (Fig. 4). In addition to this, the worm component of WannaCry extracts the hardcoded *R resource* binary and then copies it to "C:\Windows\taskche.exe" directory path. The R resource represents the binary of the WannaCry encryption component. After that, the worm runs the executable with the following parameters in the command line: "C:\Windows\taskche.exe/i". Next, the worm tries to move the "C:\Windows\taskche.exe" file to "C:\Windows\qeriuwjhrf", to replace the original file if it exists. This is done to ensure multiple infections and avoid any issues with creating the tasksche.exe process.



Name	Display name
mssecsvc2.0	Microsoft Security Center (2.0) Service

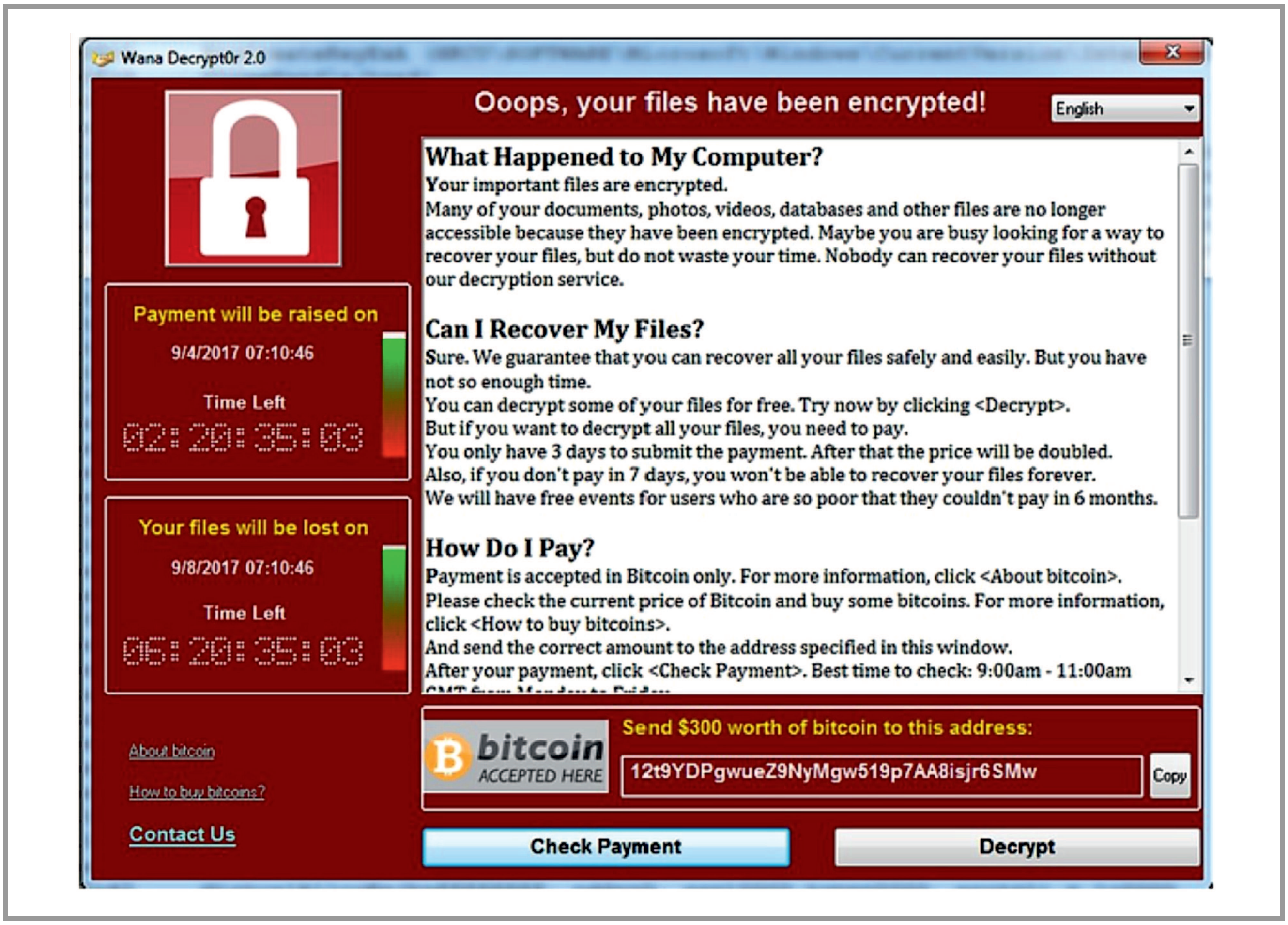
Fig. 4. Microsoft Security Center (2.0) Service.

Finally, WannaCry creates an entry in the Windows registry in order to ensure that it runs every time the computer is restarted. The new entry contains a string (e.g. "midtxzggq900"), which is a unique identifier randomly generated by using the computer name. Once the tasksche.exe component runs, it copies itself to a folder with a randomly generated name in the Common Appdata directory of the infected machine. Then, it attempts to establish memory persistence by adding itself to the AutoRun feature.



```
Created         C:\ProgramData\midtxzggq900\b.wnry
Modifed 15F936  C:\ProgramData\midtxzggq900\b.wnry
Created         C:\ProgramData\midtxzggq900\c.wnry
Modifed 30C     C:\ProgramData\midtxzggq900\c.wnry
Created         C:\ProgramData\midtxzggq900\msg
Created         C:\ProgramData\midtxzggq900\msg\m_bulgarian.wnry
Modifed         C:\ProgramData\midtxzggq900\msg
Modifed BB07    C:\ProgramData\midtxzggq900\msg\m_bulgarian.wnry
Created         C:\ProgramData\midtxzggq900\msg\m_chinese (simplified).wnry
Modifed D457    C:\ProgramData\midtxzggq900\msg\m_chinese (simplified).wnry
Created         C:\ProgramData\midtxzggq900\msg\m_chinese (traditional).wnry
Modifed 135F2   C:\ProgramData\midtxzggq900\msg\m_chinese (traditional).wnry
Created         C:\ProgramData\midtxzggq900\msg\m_croatian.wnry
Modifed 989E    C:\ProgramData\midtxzggq900\msg\m_croatian.wnry
Created         C:\ProgramData\midtxzggq900\msg\m_czech.wnry
Modifed 9E40    C:\ProgramData\midtxzggq900\msg\m_czech.wnry
Created         C:\ProgramData\midtxzggq900\msg\m_danish.wnry
Modifed 90B5    C:\ProgramData\midtxzggq900\msg\m_danish.wnry
Created         C:\ProgramData\midtxzggq900\msg\m_dutch.wnry
Modifed 907B    C:\ProgramData\midtxzggq900\msg\m_dutch.wnry
Created         C:\ProgramData\midtxzggq900\msg\m_english.wnry
Modifed 906D    C:\ProgramData\midtxzggq900\msg\m_english.wnry
Created         C:\ProgramData\midtxzggq900\msg\m_filipino.wnry
Modifed 92CC    C:\ProgramData\midtxzggq900\msg\m_filipino.wnry
Created         C:\ProgramData\midtxzggq900\msg\m_finnish.wnry
Modifed 95E9    C:\ProgramData\midtxzggq900\msg\m finnish.wnry
```

Fig. 5. WannaCry dropped files to the working directory.



Ooops, your files have been encrypted!

What Happened to My Computer?

Your important files are encrypted.

Many of your documents, photos, videos, databases and other files are no longer accessible because they have been encrypted. Maybe you are busy looking for a way to recover your files, but do not waste your time. Nobody can recover your files without our decryption service.

Can I Recover My Files?

Sure. We guarantee that you can recover all your files safely and easily. But you have not so enough time.

You can decrypt some of your files for free. Try now by clicking <Decrypt>.

But if you want to decrypt all your files, you need to pay.

You only have 3 days to submit the payment. After that the price will be doubled.

Also, if you don't pay in 7 days, you won't be able to recover your files forever.

We will have free events for users who are so poor that they couldn't pay in 6 months.

How Do I Pay?

Payment is accepted in Bitcoin only. For more information, click <About bitcoin>.

Please check the current price of Bitcoin and buy some bitcoins. For more information, click <How to buy bitcoins>.

And send the correct amount to the address specified in this window.

After your payment, click <Check Payment>. Best time to check: 9:00am - 11:00am GMT from Monday to Friday.

Send $300 worth of bitcoin to this address:

12t9YDPgwueZ9NyMgw519p7AA8isjr6SMw

Fig. 6. WannaCry extortion message.

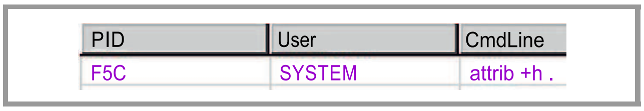
In summary, the dynamic analysis has revealed that, to achieve persistence on the infected machine, WannaCry performs the following actions:

- creates an entry in the Windows registry to ensure that it executes every time the machine is restarted,
- attempts to achieve memory persistence by adding itself to the AutoRun feature of Windows,
- uses Windows *icacls* command to grant itself a full access to all files on the machine,
- deletes all backup (shadow) copies and tries to prevent being booted in *safe mode* by executing several commands in the Windows command line,
- deletes all backup folders,
- by using the Windows command line, creates a VBScript program which generates a single shortcut of the *@WanaDecryptor@.exe* decrypter file,
- tries to kill SQL and MS Exchange database processes by executing several commands in the Windows command line.

3.5. Configuration Data Load

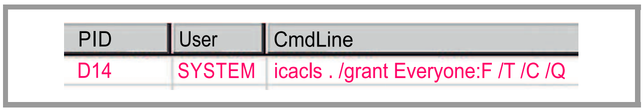
After the persistence phase, WannaCry loads the *XIA resource*, which corresponds to a password protected ZIP file. It decompresses the files and drops them to the working directory of the running process (Fig. 5), as observed in the DirWatch module of SysAnalyzer.

As one can see, WannaCry loads configuration data from the c.wnry file into memory. WannaCry randomly chooses one of the three available Bitcoin addresses and then writes this address back to the configuration data. This is done in order to display the payment address in the extortion message (Fig. 6). After that, WannaCry sets the hidden attribute (Fig. 7) for the working directory with the help of the CreateProcess function. Next, with the help of the Windows icacls command, WannaCry grants full access to all files on the target system (Fig. 8).



PID	User	CmdLine
F5C	SYSTEM	attrib +h .

Fig. 7. WannaCry sets the hidden attribute for the working directory.



PID	User	CmdLine
D14	SYSTEM	icacls . /grant Everyone:F /T /C /Q

Fig. 8. WannaCry grants full access on the target system.

The next step is to import one of the hardcoded public RSA keys as was identified at offset 0xec00 of the tasksche.exe process (Fig. 9). WannaCry then loads and executes, in memory, the contents of the t.wnry file (Fig. 10) which contains the default encrypted AES key required for decrypting the DLL responsible for the file encryption routine. The first 8 bytes of the file are checked to match the WANACRY! string. Then, the imported public RSA key hardcoded within binary is used to decrypt the AES key stored at the beginning of the t.wnry file. The AES key obtained is then used to decrypt and load the encryption DLL, which can be observed with the help of OllyDbg debugging tool [17] during WannaCry execution, as shown in Fig. 11. This DLL is responsible for file encryption on the infected machine and is summarized in Table 6.

Table 6
Encryption DLL

MD5	f351e1fcca0c4ea05fc44d15a17f8b36
SHA1	7d36a6aa8cb6b504ee9213c200c831e b8d4ef26b
Size	65536 bytes
File type	Dynamic-Link-Library
Internal name	kbdlv.dll
File description	Latvia keyboard layout
Timestamp	Mon, Jul 13 18:12:55 2009

3.6. Encryption Process

The encryption component of WannaCry is invoked with the TaskStart system thread. During its execution, the encryption component checks if one of the following mutexes exists:

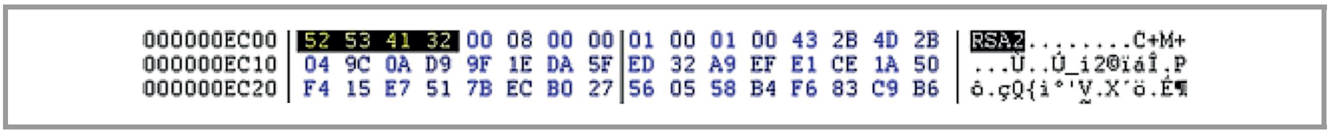
```
GlobalnMsWinZonesCacheCounterMutexA,
GlobalnMsWinZonesCacheCounterMutexW,
MsWinZonesCacheCounterMutexA.
```

If the mutex "MsWinZonesCacheCounterMutexA" is present, then the encryption component automatically stops without taking any further action. If the mutex is not present on the system, the encryption process starts. In particular, TaskStart creates a new mutex named "MsWinZonesCacheCounterMutexA" and reads the contents of the c.wnry file from the current directory. After that, WannaCry creates three configuration files shown in Table 7.

Table 7
WannaCry configuration files

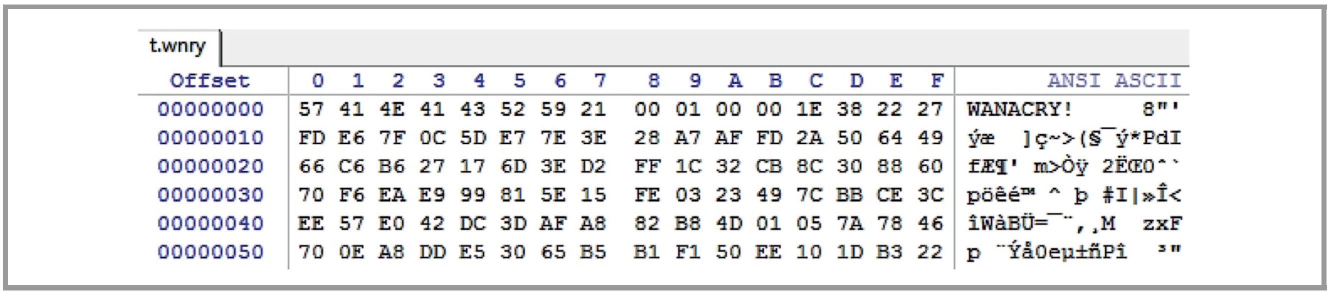
Filename	Description
00000000.res	TOR/C2 info
00000000.pky	Public RSA key
00000000.eky	Encrypted private RSA key

After the configuration files have been created, the encryption component is ready to start encrypting files on the system. To accomplish this, it spawns several threads. First,



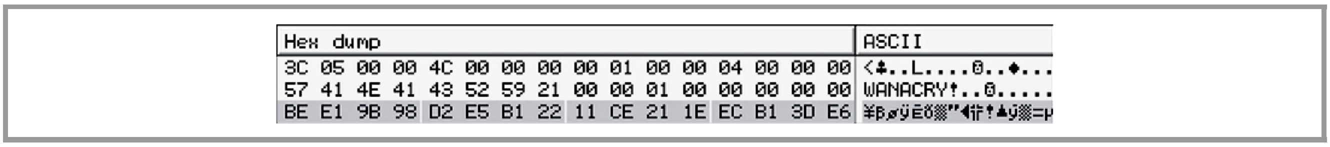
```
000000EC00 | 52 53 41 32 00 08 00 00|01 00 01 00 43 2B 4D 2B | RSA2........C+M+
000000EC10 | 04 9C 0A D9 9F 1E DA 5F|ED 32 A9 EF E1 CE 1A 50 | ...Ù..Ú_í2©ïáÎ.P
000000EC20 | F4 15 E7 51 7B EC B0 27|56 05 58 B4 F6 83 C9 B6 | ô.çQ{ì°'V.X´ö.É¶
```

Fig. 9. Imported RSA private key.



```
t.wnry
Offset     0  1  2  3  4  5  6  7   8  9  A  B  C  D  E  F      ANSI ASCII
00000000  57 41 4E 41 43 52 59 21  00 01 00 00 1E 38 22 27  WANACRY!     8"'
00000010  FD E6 7F 0C 5D E7 7E 3E  28 A7 AF FD 2A 50 64 49  ýæ  ]ç~>(§¯ý*PdI
00000020  66 C6 B6 27 17 6D 3E D2  FF 1C 32 CB 8C 30 88 60  fÆ¶' m>Òÿ 2ËŒ0ˆ`
00000030  70 F6 EA E9 99 81 5E 15  FE 03 23 49 7C BB CE 3C  pöêé™ ^ þ #I|»Î<
00000040  EE 57 E0 42 DC 3D AF A8  82 B8 4D 01 05 7A 78 46  îWàBÜ=¯¨‚¸M  zxF
00000050  70 0E A8 DD E5 30 65 B5  B1 F1 50 EE 10 1D B3 22  p ¨Ýå0eµ±ñPî  ³"
```

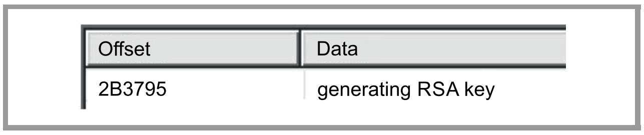
Fig. 10. Loaded and executed t.wnry file.



```
Hex dump                                        | ASCII
3C 05 00 00 4C 00 00 00 00 01 00 00 04 00 00 00 | <♣..L....0..♦...
57 41 4E 41 43 52 59 21 00 00 01 00 00 00 00 00 | WANACRY!..0.....
BE E1 9B 98 D2 E5 B1 22 11 CE 21 1E EC B1 3D E6 | ¥ßøÿÈõ▒"◄╬!▲ÿ▒=µ
```

Fig. 11. Decrypted AES key in a memory dump.

WannaCry attempts to load and check the existence of two keys in the 00000000.pky and 00000000.dky files. The 00000000.dky file presents a decryption RSA key which is received upon the payment has been verified. When the victim clicks the "Check Payment" button, WannaCry starts checking for the presence of the 00000000.dky file on the system. If the two aforementioned files do not exist, WannaCry generates a new unique RSA 2048-bit asymmetric key pair, which can be seen in the memory dump made with with SysAnalyzer tool at 0x2B3795 offset (Fig. 12).

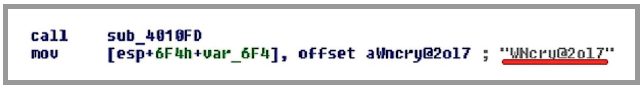


Offset	Data
2B3795	generating RSA key

Fig. 12. Generation of an RSA key pair.

Once the key pair has been generated, WannaCry exports the victim's public RSA key to a 00000000.pky file using Microsoft's *CryptExportKey* function. Next, WannaCry exports the victim's private RSA key and encrypts it with another hard-coded RSA public key. The encrypted private key is stored as a 00000000.eky file. After the key has been safely stored, WannaCry calls upon the *CryptDestroyKey* function to destroy the private key in memory, to limit any key recovery options.

Next, WannaCry starts enumerating, every 3 seconds, information about all logical drives attached to the system. If a new attached drive is not a CD ROM drive, then it begins the encryption process on the new drive. At this stage, WannaCry also starts iterating through all existing directories and searching for predefined file extensions of interest. To encrypt each file, it generates a 16-byte symmetric AES key using the *CryptGenRandom* function. Then, it encrypts every generated AES key with the public RSA key and stores it inside the file header starting with the WANACRY! string value. Encrypted files are renamed and appended with the .*WNCRY* file extension.



```
call    sub_4010FD
mov     [esp+6F4h+var_6F4], offset aWncry@2ol7 ; "WNcry@2ol7"
```

Fig. 13. Password for a ZIP archive in the encryption component.

The encryption component contains a password-protected ZIP archive. We managed to obtain the password, "WNcry@2ol7", by disassembling the encrypter with the IDA Pro tool [18] (see Fig. 13). The contents of the ZIP archive are summarized in Table 8 and described below:

- *msg* is a folder that contains a list of rich text format (RTF) files with the *wnry* extension. These files are the readme instructions used to show the extortion message to the victim in different languages, based on the information obtained from the system by malicious WannaCry functions;

- *b.wnry* is an image file used for displaying instructions for the decryption of user files. It starts with 42 4D strings, which indicates that this file is a bitmap image;

- *c.wnry* contains a list of Tor addresses with the *.onion* extension and a link to a zipped installation file of the Tor browser from Tor Project [19];

Table 8
Files in the password protected ZIP archive

Name	Size [bytes]	Modified
msg	1,329,657	2017-05-11
b.wnry	1,440,054	2017-05-11
c.wnry	780	2017-05-11
r.wnry	864	2017-05-10
s.wnry	3,038,286	2017-05-09
t.wnry	65,816	2017-05-11
taskdl.exe	20,480	2017-05-11
taskse.exe	20,480	2017-05-11
u.wnry	245,760	2017-05-11

- *r.wnry* is a text file in English with additional decryption instructions to be used by the decryption component (the *u.wnry* file mentioned below);
- *s.wnry* file is a ZIP archive (HEX signature 50 4B 03 04) which contains the Tor software executable. This executable has been obtained with the assistance of the WinHex tool [20] by saving raw binary data with the .zip extension;
- *t.wnry* is an encrypted file with the WANACRY! encryption format. The file header starts with the WANACRY! string;
- *taskdl.exe* is a supporting tool for the deletion of files with the .WNCRY extension. By observing the properties of the file, the following masquerade description can be found: "SQL Client Configuration Utility";
- *taskse.exe* is a supporting tool for malware execution on remote desktop protocol (RDP) sessions. The following file description was identified: "waitfor – wait/send a signal over a network";
- *u.wnry* is an executable file (HEX signature 4D 5A) with the name of "@WanaDecryptor@.exe", which represents the decryption component of WannaCry.

At the same time, another thread calls the taskse.exe process every 30 s, which tries to enumerate active RDP sessions on connected remote machines and to run the @WanaDecryptor@.exe binary file. This file is extracted from the u.wnry file and represents the decryption component of WannaCry. The persistence of RDP session injections is ensured by adding the value in the AutoRun registry key.

3.7. Recovery Prevention

After finishing the encryption process, WannaCry tries to prevent various common data recovery methods by executing several commands on the system. To prevent data recovery, WannaCry executes the following commands:

- vssadmin delete shadows/all/quiet. Deletes all the shadow volumes on the system without alerting the user. By default, these volumes contain backup data in the event of a system fault;
- wmic shadowcopy delete. Ensures deletion of any copies relevant to shadow volumes;
- bcdedit/set default bootstatuspolicy ignoreallfailures. Ensures that the machine is booted, even if errors are found;
- bcdedit/set default recoveryenabled no. Disables the Windows recovery feature, thus preventing the victims from the possibility to reverting their system to a previous build;
- wbadmin delete catalog $-q$. Ensures that victim can no longer use any backup files created by Windows Server.

3.8. Propagation

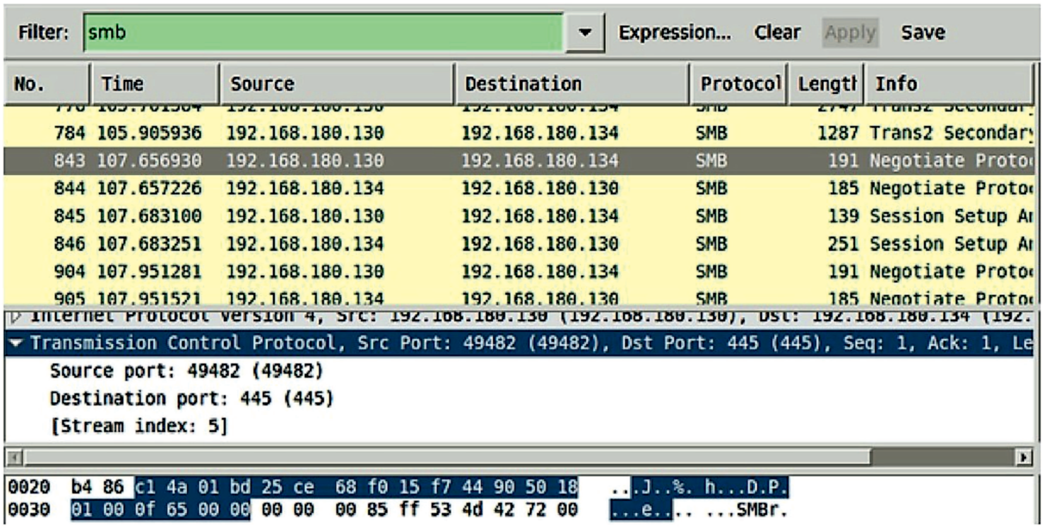
The worm component of WannaCry carries the main propagation and exploit functionality, which utilizes the EternalBlue exploit and the DoublePulsar backdoor to leverage the MS17-010 SMB vulnerability [12]. After performing the initial interactions and checking connectivity with the kill-switch domain, the worm functionality is established by initiating the *mssecsvs2.0* service, which WannaCry installs after being executed. This service tries to spread WannaCry payload through the SMB vulnerability on any vulnerable systems on both internal and external networks.

In order to perform this, WannaCry creates and spawns two separate threads that simultaneously replicate worm payload in all detected networks. In the internal network, before starting the propagation process, the component obtains the IP addresses of local network interfaces by invoking the *GetAdaptersInfo* function, and determines the subnets existing in the network.

After that, the worm component tries to connect to all possible IP addresses in any available local network on port 445, which is the default port for SMB over IP service. If successful, the worm attempts to exploit the service for the MS17-010 vulnerability. In our testbed, connection attempts were observed with Wireshark on a REMnux machine, when the infected machine (IP 192.168.180.130) sent SMB probe packets to the clean machine (IP 192.168.180.134), as shown in Fig. 14.

During the SMB probing, one of the unique features of the generated traffic is that it contains two hardcoded IP addresses: 192.168.56.20 and 172.16.99.5. They can be observed by extracting strings from the binary. In particular, WannaCry sends three NetBIOS session setup packets, where two of them contain the aforementioned hardcoded IP addresses.

At the same time, the worm component attempts to spread across the external networks by generating various IP addresses and by trying to connect to TCP port 445. This can be observed with Wireshark on REMnux, as shown

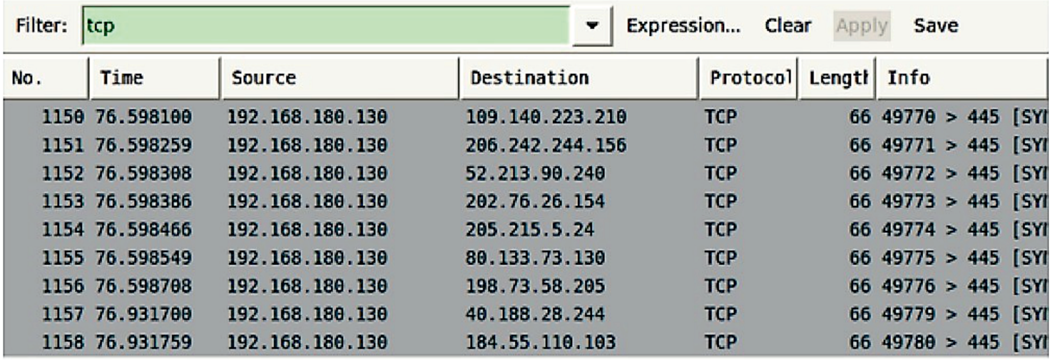


```
No.  Time        Source           Destination      Protocol Length Info
 784 105.905936  192.168.180.130  192.168.180.134  SMB      1287 Trans2 Secondary
 843 107.656930  192.168.180.130  192.168.180.134  SMB       191 Negotiate Proto
 844 107.657226  192.168.180.134  192.168.180.130  SMB       185 Negotiate Proto
 845 107.683100  192.168.180.130  192.168.180.134  SMB       139 Session Setup An
 846 107.683251  192.168.180.134  192.168.180.130  SMB       251 Session Setup An
 904 107.951281  192.168.180.130  192.168.180.134  SMB       191 Negotiate Proto
 905 107.951521  192.168.180.134  192.168.180.130  SMB       185 Negotiate Proto
```

```
▷ Internet Protocol Version 4, Src: 192.168.180.130 (192.168.180.130), Dst: 192.168.180.134 (192.
▼ Transmission Control Protocol, Src Port: 49482 (49482), Dst Port: 445 (445), Seq: 1, Ack: 1, Le
    Source port: 49482 (49482)
    Destination port: 445 (445)
    [Stream index: 5]
```

```
0020  b4 86 c1 4a 01 bd 25 ce  68 f0 15 f7 44 90 50 18   ...J..%. h...D.P.
0030  01 00 0f 65 00 00 00 00  00 85 ff 53 4d 42 72 00   ...e.... ...SMBr.
```

Fig. 14. WannaCry internal network traffic attempting the SMB exploit.



```
No.   Time       Source           Destination      Protocol Length Info
1150  76.598100  192.168.180.130  109.140.223.210  TCP      66 49770 > 445 [SYN
1151  76.598259  192.168.180.130  206.242.244.156  TCP      66 49771 > 445 [SYN
1152  76.598308  192.168.180.130  52.213.90.240    TCP      66 49772 > 445 [SYN
1153  76.598386  192.168.180.130  202.76.26.154    TCP      66 49773 > 445 [SYN
1154  76.598466  192.168.180.130  205.215.5.24     TCP      66 49774 > 445 [SYN
1155  76.598549  192.168.180.130  80.133.73.130    TCP      66 49775 > 445 [SYN
1156  76.598708  192.168.180.130  198.73.58.205    TCP      66 49776 > 445 [SYN
1157  76.931700  192.168.180.130  40.188.28.244    TCP      66 49779 > 445 [SYN
1158  76.931759  192.168.180.130  184.55.110.103   TCP      66 49780 > 445 [SYN
```

Fig. 15. WannaCry external network traffic attempting the SMB exploit.

in Fig. 15. As it can be seen, the worm attempts to probe external Internet IP addresses for the MS17-010 vulnerability. This explains the reason for the widespread infection seen during the massive outbreak on 12 May 2017. The full list of WannaCry generated IP addresses obtained during the analysis is presented in Table 9.

Table 9
External IP addresses generated by WannaCry

IP address : port
109.140.223.210 : 445
206.242.244.156 : 445
52.213.90.240 : 445
202.76.26.154 : 445
205.215.5.24 : 445
80.133.73.130 : 445
198.73.58.205 : 445
40.188.28.244 : 445
184.55.110.103 : 445

3.9. C&C Communication

During its execution, the software also tries to contact the C&C servers. To this end, WannaCry unpacked and dropped files from the s.wnry file, containing the Tor executable, into the installation directory as shown

```
Created         C:\ProgramData\midtxzggq900\s.wnry
Modifed 2E5C4E  C:\ProgramData\midtxzggq900\s.wnry
```

Fig. 16. Tor executable dropped into the installation directory.

in Fig. 16. Before unpacking, it starts listening on the localhost address 127.0.0.1:9050. This address, with the specified 9050 port, is typically used for configuring the

Tor browser application. If the contents of the s.wnry file are corrupted, then WannaCry tries to download the Tor executable from a hardcoded URL. After the successful extraction of the Tor executable, it copies "TaskData\Tor\tor.exe" to "TaskData\Tor\taskhsvc.exe" and executes it. Next, WannaCry parses the contents of the c.wnry file, which specifies the configuration data, including the following .onion addresses to connect and the zipped Tor browser installation file:

```
gx7ekbenv2riucmf.onion
57g7spgrzlojinas.onion
xxlvbrloxvriy2c5.onion
76jdd2ir2embyv47.onion
cwwnhwhlz52maqm7.onion
https://dist.torporject.org/torbrowser/6.5.1/tor
        -win32-0.2.9.10.zip
```

After that, WannaCry sends the first eight bytes of the 00000000.res file content to the C&C server. These bytes specify the host and user name of the infected machine. The 00000000.res file, which is dropped during encryption process, accumulates in total 88 bytes of configuration data, including internal flags, counters, and timestamps.

During its communication with Tor addresses, WannaCry establishes a secure HTTPS channel to port 443, and uses common Tor ports, 9001 and 9050, for network traffic and directory information.

4. Conclusions and Future Work

We have performed a comprehensive dynamic analysis of WannaCry ransomware in a purpose-built virtual testbed. We analyzed the WannaCry version which was observed during the massive attacks on 12 May 2017. The analysis has revealed that the given ransomware is composed of two distinctive components, which enable the worm-like self-propagating mechanism and combined encryption process. Both worm and encryption components of WannaCry have been examined.

The focus of this study was on WannaCry's initial interactions and the infection process, its persistence mechanism, encryption process, recovery prevention as well as its propagation mechanisms and communication with C&C servers. The analysis has revealed important characteristics and behaviors of WannaCry during its execution. In particular, we identified Tor addresses used for C&C, observed TCP and DNS connections, SMB probes, as well as actions related to WannaCry persistence and obfuscation.

The worm component of WannaCry weaponized by the functionality enabling it to exploit and propagate via Microsoft's MS17-010 on unpatched systems by sending SMB probing packets on port 445. In addition to the modular nature of WannaCry, it was also observed that it has embedded RSA keys used to decrypt the required malicious DLL representing the encryption component. It was identified that the worm component scans both internal and external networks for MS17-010 vulnerability, by generating a list of local and global IP addresses. The worm tries to probe the hosts from the generated list by sending packets to port 445. Before its execution, WannaCry also performs an initial check with the kill-switch domain.

At the same time, the analysis has identified two hardcoded IP addresses (192.168.56.20 and 172.16.99.5), which are sent during the SMB probing. Depending on the condition of the s.wnry file dropped during execution, WannaCry can also communicate with embedded .onion addresses via a secure channel on port 443 and via common Tor ports 900 and 9050 to download the Tor browser installation software from a specified URL.

The findings of this work could be used for designing effective mitigation mechanisms for WannaCry and other ransomware families that exhibit similar behavior. This is left as future work. In particular, we plan to investigate the use of software-defining networking (SDN) [21], [22] for ransomware detection and mitigation. SDN is an emerging paradigm of programmable networks that decouples the control and data planes. SDN controllers maintain a view of the entire network and implement policy decisions. On the other hand, each device at the data plane maintains one or more *flow tables*, where the packet handling rules are stored. This changes the way that networks are designed and managed, and enables new SDN-based security solutions [23]–[25], such as firewalls and intrusion detection systems for various types of malware, including ransomware mitigation [26], [27].

References

[1] D. O'Brien, "Ransomware 2017", Internet Security Threat Report, Symantec, July 2017 [Online]. Available: https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/istr-ransomware-2017-en.pdf

[2] K. Savage, P. Coogan, and H. Lau, "The evolution of ransomware", Security Response, Symantec, June 2015 [Online]. Available: http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/the-evolution-of-ransomware.pdf

[3] A. Zeichnick, "Self-propagating ransomware: What the WannaCry ransomworm means for you", May 2017 [Online]. Available: https://www.networkworld.com/article/3196993/security/self-propagating-ransomware-what-the-wannacry-ransomworm-means-for-you.html

[4] "Ransom.Wannacry", Symantec, May 2017 [Online]. Available: https://www.symantec.com/security-center/writeup/2017-051310-3522-99/

[5] "Petya – taking ransomware to the low level", Malwarebytes Labs, Jun. 2017 [Online]. Available: https://blog.malwarebytes.com/threat-analysis/2016/04/petya-ransomware/

[6] "Petya ransomware eats your hard drives", Kaspersky Labs, Jun. 2017 [Online]. Available: https://www.kaspersky.com/blog/petya-ransomware/11715

[7] "Bad Rabbit: A new ransomware epidemic is on the rise", Kaspersky Labs, Oct. 2017 [Online]. Available: https://www.kaspersky.com/blog/bad-rabbit-ransomware/19887/

[8] M. Akbanov, V. G. Vassilakis, I. D. Moscholios, and M. D. Logothetis, "Static and dynamic analysis of WannaCry ransmware", in *Proc. IEICE Inform. and Commun. Technol. Forum ICTF 2018*, Graz, Austria, 2018.

[9] C. Everett, "Ransomware: To pay or not to pay?", *Comp. Fraud & Secur.*, vol. 2016, no. 4, pp. 8–12, 2016 (doi: 10.1016/S1361-3723(16)30036-7).

[10] "Understanding ransomware and strategies to defeat it", McAfee Labs, White Paper, 2016 [Online]. Available: https://www.mcafee.com/enterprise/en-us/assets/white-papers/wp-understanding-ransomware-strategies-defeat.pdf

[11] "What you need to know about the WannaCry ransomware", Symantec, Threat Intelligence, Oct. 2017, [Online]. Available: https://www.symantec.com/blogs/threat-intelligence/wannacry-ransomware-attack

[12] Microsoft Security Bulletin MS17-010 – Critical, March 14, 2017 [Online]. Available: https://docs.microsoft.com/en-us/security-updates/securitybulletins/2017/ms17-010

[13] ViRus Share malware repository [Online]. Available: https://virusshare.com (accessed Nov. 30, 2018).

[14] "REMnux: A Linux toolkit for reverse-engineering and analyzing malware" [Online]. Available: https://remnux.org (accessed Nov. 30, 2018).

[15] SysAnalyzer – Automated malcode analysis system [Online]. Available: https://github.com/dzzie/SysAnalyzer (accessed Nov. 30, 2018).

[16] Pestudio, Malware Assessment Tool [Online]. Available: https://www.winitor.com (accessed Nov. 30, 2018).

[17] OllyDbg – A 32-bit assembler level debugger for Microsoft Windows [Online]. Available: http://www.ollydbg.de/ (accessed Nov. 30, 2018).

[18] IDA: Pro [Online]. Available: https://www.hex-rays.com/products/ida (accessed Nov. 30, 2018).

[19] Tor Project [Online]. Available: https://www.torproject.org (accessed Nov. 30, 2018).

[20] "WinHex: Computer forensics and data recovery software" [Online]. Available: https://www.x-ways.net/winhex (accessed Nov. 30, 2018).

[21] B. Nunes, M. Mendonca, X. N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, future of programmable networks", *IEEE Commun. Surveys & Tutor.*, vol. 16, no. 3, pp. 1617-1634, 2014 (doi: 10.1109/SURV.2014.012214.00180).

[22] V. G. Vassilakis, I. D. Moscholios, B. A. Alzahrani, and M. D. Logothetis, "A software-defined architecture for next-generation cellular networks", in *Proc. IEEE Int. Conf. on Commun. ICC 2016*, Kuala Lumpur, Malaysia, 2016 (doi: 10.1109/ICC.2016.7511018).

[23] C. Yoon, T. Park, S. Lee, H. Kang, S. Shin, and Z. Zhang, "Enabling security functions with SDN: A feasibility study", *Comp. Netw.*, vol. 85, pp. 19–35, 2015 (doi: 10.1016/j.comnet.2015.05.005).

[24] J. M. Ceron, C. B. Margi, and L. Z. Granville, "MARS: An SDN-based malware analysis solution", *Proc. IEEE Symp. on Comp. and Commun. ISCC 2016*, Messina, Italy, 2016 (doi: 10.1109/ISCC.2016.7543792).

[25] V. G. Vassilakis, I. D. Moscholios, B. A. Alzahrani, and M. D. Logothetis, "On the security of software-defined next-generation cellular networks", in *Proc. IEICE Inform. and Commun. Technol. Forum ICTF 2016*, Patras, Greece, 2016.

[26] K. Cabaj and W. Mazurczyk, "Using software-defined networking for ransomware mitigation: The case of CryptoWall", *IEEE Network*, vol. 30, no. 6, pp. 14–20, 2016 (doi: 10.1109/MNET.2016.1600110NM).

[27] K. Cabaj, M. Gregorczyk, and W. Mazurczyk, "Software-defined networking-based crypto ransomware detection using HTTP traffic characteristics", *Comp. & Elec. Engin.*, vol. 66, pp. 353–386, 2018 (doi: 10.1016/j.compeleceng.2017.10.012).

Maxat Akbanov received the B.Sc. degree in Information and Communications System Security from the National Technical University of Ukraine "Kyiv Polytechnic University", Kyiv, Ukraine, in 2011, and the M.Sc. degree in Cyber Security from the University of York, York, UK, in 2018. In 2008 and 2016, he received the prestigious Kazakhstan governmental "Bolashak" scholarship to fund his studies abroad. He holds merit and distinction awards for B.Sc. and M.Sc. degrees, respectively. He is currently working for the private sector in Kazakhstan and is involved in developing several startup projects for the government-sponsored "Digital Kazakhstan" and "Cyber Shield" strategies. His main research interests include network and malware forensics, software-defined networking, covert channels, cryptography, Internet of Things, machine learning and artificial intelligence.

E-mail: maxat.akbanov@gmail.com
Department of Computer Science
University of York
Deramore Lane
Heslington
York YO10 5GH, United Kingdom

Vassilios G. Vassilakis received his Ph.D. degree in Electrical and Computer Engineering from the University of Patras, Greece in 2011. He is currently a lecturer in Cyber Security at the University of York, UK. He's been involved in EU, UK, and industry funded R&D projects related to the design and analysis of future mobile networks and Internet technologies. His main research interests are in the areas of network security, Internet of Things, next-generation wireless and mobile networks, and software-defined networks. He has published over 90 papers in international journals/conferences. He has served as a Guest Editor in IEICE Transactions on Communications, IET Networks, and Elsevier Optical Switching & Networking, and in the TPC of IEEE ICC and IEEE Globecom.

E-mail: vv274@cl.cam.ac.uk
University of York
York YO10 5DD, United Kingdom

Michael D. Logothetis received his B.Eng. degree and Ph.D. in Electrical Engineering, both from the University of Patras, Patras, Greece, in 1981 and 1990, respectively. From 1991 to 1992 he was a Research Associate at NTT's Telecommunication Networks Laboratories, Tokyo, Japan. In 2009 he was elected (Full) Professor at the ECE Department of the University of Patras. His research interests include teletraffic theory, simulation and performance optimization of telecommunications networks. He has published over 200 conference/journal papers. He has become a Guest Editor in: Mediterranean Journal of Electronics and Communications, Mediterranean Journal of Computers and Networks, IET Circuits, Devices and Systems, IET Networks and Ubiquitous Computing and Communication Journal. He is a member of the IARIA (Fellow), IEEE (Senior), IEICE (Senior), FITCE and the Technical Chamber of Greece (TEE).

E-mail: mlogo@upatras.gr
Wire Communications Laboratory
Department of Electrical and Computer Engineering
University of Patras
265 04 Patras, Greece

Comparative Study between Several Direction of Arrival Estimation Methods

Youssef Khmou[1], Said Safi[1], and Miloud Frikel[2]

[1] Department of Mathematics and Informatics, Beni Mellal, Morocco
[2] Greyc UMR 6072 CNRS, ENSICAEN, Caen, France

http://doi.org/10.26636/jtit.2014.1.1004

Abstract—In this paper a comparative study, restricted to one-dimensional stationary case, between several Direction of Arrival (DOA) estimation algorithms of narrowband signals is presented. The informative signals are corrupted by an Additive White Gaussian Noise (AWGN), to show the performance of each method by applying directly the algorithms without pre-processing techniques such as forward-backward averaging or spatial smoothing.

Keywords—array processing, Direction of Arrival, geolocalization, propagation, smart antenna, spectral analysis.

1. Introduction

In array signal processing Direction of Arrival estimation (DOA) [1], [2] stands for estimating the angles of arrivals of received signals by an array of antennas. It is considered an important processing step in many sensors systems, i.e., radar, sonar, Measure Electronic Surveillance (MSE), submarine acoustics, geodesic location, optical interferometry, etc.

There are many types of DOA algorithms that have been proposed during the past four decades such as conventional spectral-based, subspace spectral-based and statistical methods. Beamforming techniques [3]–[7] are straightforward and require low computational power but these methods have low resolution [8]. That leads to introduction of subspace-based algorithms [9]–[11] that use the eigendecomposition of output data covariance matrix in order to obtain the so-called signal subspace or noise subspace. However these methods become limited in case of larger number of array sensors, many fast algorithms for DOA have been proposed in recent years such as the propagator method (PM) [12]–[14] without eigendecomposition with low computational load. Unfortunately, this method is only suitable to the presence of white Gaussian noise, and its performance will be degraded in spatial nonuniform colored noise. To overcome this problem, a modified PM algorithm has been proposed with different computation method for the propagation operator [15]. It is only obtained by the partially cross-correlation of array output data which makes it suitable for the case of spatially nonuniform colored noise due to using the off-diagonal elements of array covariance matrix.

This paper presents a comparative study that is restricted to one-dimensional stationary case (azimuth) between several DOA estimation algorithms of narrowband signals [16] that are corrupted by uniform Additive White Gaussian Noise (AWGN). The performance of each method is evaluated by applying directly the algorithms on Uniform Linear Array (ULA) without pre-processing techniques such as forward-backward averaging of the cross correlation of array output data R or spatial smoothing. The authors choose the key factor for this evaluation to be the Signal to Noise Ratio (SNR) of the environment surrounding the ULA and the radiating sources while the number of snapshots constant is maintained.

1.1. Problem Statement

Typical smart antenna architecture of base station can be divided into the following functional blocks as shown in Fig. 1 [16]. Radio signals arriving at the array antennas are conversed from analog to digital form by downconversion and sampling operations, next summation of the digitized signals over all array elements produces single stream output for further processing.

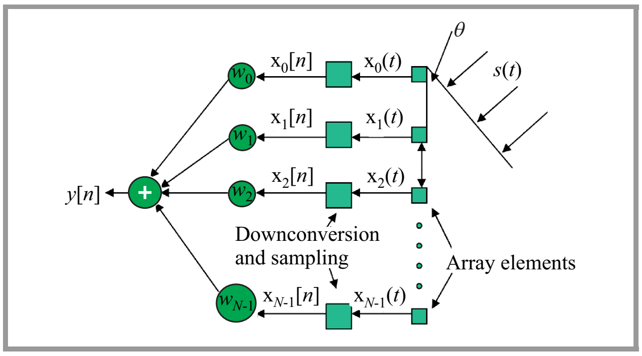


Fig. 1. Typical front-end architecture of base station receiver.

Let's consider an array of N elements receiving P signals such that each element of the array contains zero mean Gaussian noise, the output array is given by:

$$y[t] = \sum_{k=1}^{N} w_k x_k[t] \, , \tag{1}$$

where:

$$x(t) = A(\theta)s(t) + N(t) \tag{2}$$

$x[t] = [x_1(t), \ldots, x_N(t)]^T$, $A(\theta) = [a(\theta_1), \ldots, a(\theta_p)]$ are the received array data and the array manifold matrix respectively, $s(t) = [s_1(t), \ldots, s_p(t)]^T$ and $N(t) = [n_1(t), \ldots, n_N(t)]^T$ stand for the source waveform vector and sensor noise vector, respectively. In Eq. (2)

$$a(\theta_i) = \left[1,\, \mathrm{e}^{\frac{j2\pi d}{\lambda}\sin(\theta_i)}, \ldots, \mathrm{e}^{\frac{j2\pi d(N-1)}{\lambda}\sin(\theta_i)}\right]^T$$

is the steering vector, and d is the distance between elements of the Uniform Linear Array (ULA), λ is the wavelength of the propagating signals, θ_i is the angle of arrival of the i^{th} source and $(.)^T$ denotes the transposition of matrix.

The array signal waveform is considered as stationary process therefore the $N \times N$ correlation matrix can be defined as:

$$R = E\left[(X(t) - m_x(t)).(X(t) - m_x(t))^H\right], \tag{3}$$

where $(.)^H$ denotes the conjugate transposition of matrix.

In this study it is assumed that:

- the signals and the additive Gaussian noise are stationary and ergodic zero mean complex valued random processes,
- the signals sources are not correlated,
- the set of P steering vectors is linearly independent and the P signal sources are statistically independent of each other,
- the number of sources P is known and the number of sensors N satisfies the condition $N \geq 2P + 2$.

Under those assumptions the cross correlation matrix is given by:

$$\begin{aligned} R &= E\left[A(\theta)S(t)S^H(t)A^H(\theta)\right] + E\left[(N(t)).N^H(t)\right] \\ &= A(\theta)R_{ss}A^H(\theta) + \sigma^2 I_N\,, \end{aligned} \tag{4}$$

where $R_{ss} = E\left[S(t)S^H(t)\right]$ is $P \times P$ source signal covariance matrix, σ^2 is the noise variance and I_N stands for an $N \times N$ identity matrix.

In practice, the exact covariance matrix R is unavailable and must be estimated from the received data. The forward-only estimate of covariance matrix is given by:

$$\hat{R}_{xx} = \frac{1}{K}\sum_{k=1}^{K} XX^H\,. \tag{5}$$

In the Section 2 different algorithms for DOA estimation are presented.

2. DOA Algorithms

2.1. Beamforming Techniques

The beamforming techniques are based on scanning all possible angles in the range $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and measuring the output power of the array such that the power spectrum has peak when the given angle is the direction of arrival of one of the incoming signal. The output signal $y(t)$ is computed using a weight vector w with the received data x:

$$y(t) = w^H x(t)\,. \tag{6}$$

Given N spanshots, the total output power of an array is:

$$\begin{aligned} P(w) = \frac{1}{N}\sum_{n=1}^{N}|y(t_n)|^2 &= \frac{1}{N}\sum_{n=1}^{N} w^H x(t_n)x^H(t_n)w \\ &= w^H \hat{R}_{xx} w\,. \end{aligned} \tag{7}$$

Based on the Eq. (7) two main techniques have been developed.

2.2. Bartlett Method

Also known as method of averaged periodograms [3], Bartlett method computes the power spectrum as follows. Let $w = a(\theta)$ be the steering vector with arbitrary scanning angle:

$$a(\theta) = \left[1,\, \mathrm{e}^{j\mu}, \ldots, \mathrm{e}^{j(N-1)\mu}\right],$$

$$\mu = \frac{-2\pi f_c}{c} d\sin\theta\,,$$

where f_c is the carrier frequency of the incoming narrowband signals, c is the speed propagation of the wave signals and d stands for distance between array sensors.

The weight vector is normalized as the following:

$$w = \frac{a(\theta)}{\sqrt{a^H(\theta)a(\theta)}}, \tag{8}$$

and the spatial spectrum is then given by:

$$P(\theta) = P_{bart}(\theta) = \frac{a^H(\theta)\hat{R}_{xx}a(\theta)}{a^H(\theta)a(\theta)}\,. \tag{9}$$

The weight vector w can be considered as spatial filter, which has been matched to the incoming signal, the array weighting equalizes the delays experienced by the signal on various sensors to combine their respective contributions.

2.3. Capon Beamformer

Capon beamformer is an enhanced version of the Bartlett method, when the sources to be located are closer than the beamwidth, The Bartlett method fails in separating the sources, for this purpose Capon in [4] proposed the maximum likelihood method to solve the for Minimum Variance

Distortion Response (MVDR) of an array such that it maximizes the signal to interference ratio:

$$\min\big(P(w)\big) \qquad \text{subject to} \qquad w^H a(\theta) = 1\,.$$

The resulting weight vector is given by:

$$w = w_{Capon} = \frac{\hat{R}_{xx}^{-1} a(\theta)}{a^H(\theta)\hat{R}_{xx}^{-1} a(\theta)} \tag{10}$$

Replacing the weight vector w in the Eq. (7) yields to the power spectrum:

$$P(\theta) = P_{Capon}(\theta) = \frac{1}{a^H(\theta)\hat{R}_{xx}^{-1} a(\theta)}\,. \tag{11}$$

2.4. Linear Prediction

The linear prediction method [5] is widely used in spectral analysis and speech processing. It is based on the concept of minimizing the mean output signal power of the array elements subject to constraint that the weight on a selected element in ULA is unity. The array weight vector is given by:

$$w = \frac{\hat{R}_{xx}^{-1} u}{u^H \hat{R}_{xx}^{-1} u}\,,$$

where u is the m^{th} column vector of the identity matrix I_{NxN} such that the index m represents the m^{th} element of the ULA. No optimized criterion is proposed for the choice of this element.

The power spectrum can be computed as:

$$P(\theta) = P_{LP}(\theta) = \frac{u^H \hat{R}_{xx}^{-1} u}{|u^H \hat{R}_{xx}^{-1} a(\theta)|^2}\,. \tag{12}$$

The choice of the m^{th} element affects the resolution capability of this method which is dependent on the SNR, and the minimum angle separating the sources.

2.5. Maximum Entropy

Maximum entropy technique [9] is an improvement of the beamforming approach, based on extrapolation the covariance matrix. The extrapolation should be selected with maximized signal entropy where its maximum is achieved by searching for the coefficients of an auto-regressive (AR) model that minimize the expected prediction error:

$$a = \arg\min\left\{a^H \hat{R}_{xx}\right\},$$

subject to the constraint that the first AR coefficient satisfies $a^H e_1 = 1$ where $a = [a_1, a_2, \ldots, a_N]^T$ and e_1 is the first column of the identity matrix I_N. Applying the Lagrange multiplier technique yields to

$$a = \frac{\hat{R}_{xx}^{-1} e_1}{e_1^T \hat{R}_{xx}^{-1} e_1}\,.$$

Next the spatial spectrum can be computed as

$$P(\theta) = P_{ME}(\theta) = \frac{1}{|a(\theta)^H C_j|^2}\,, \tag{13}$$

where C_j represents the j^{th} column of the inverse cross correlation matrix $\hat{R}_{xx}^{-1}$.

The quality of the resolution of the maximum entropy method depends on the choice of column C_j.

2.6. Pisarenko Harmonic Decomposition

Pisarenko harmonic decomposition method [9] minimizes the Mean Square Error (MSE) of the array output under the constraint that the norm weight vector to be equal to unity. The eigenvector that minimizes the MSE corresponds to the smallest eigenvalue of the cross-correlation of array output data, the output power is given by:

$$P(\theta) = P_{PHD}(\theta) = \frac{1}{|a(\theta)^H \bar{e}_1|^2}\,, \tag{14}$$

where $\bar{e}_1$ is the eigenvector associated with the smallest eigenvalue σ_1.

2.7. Minimum Norm

The minimum norm technique [1], [9] is generally considered to be a high-resolution method which assumes a ULA structure.

The algorithm is described as the following. After estimating the cross correlation matrix $\hat{R}_{xx}$, a Singular Value Decomposition (SVD) is performed to extract the matrices U, S and V such that $\hat{R}_{xx} = USV'$. Next, a noise subspace is constructed by selecting the set of vectors $E_N = U(:, P+1:N)$ where P and N denotes the number of radiating sources and the number of elements in the ULA respectively. Constructing the spectrum is based on minimum norm vector lying in the noise subspace whose first element equals 1 and having minimum norm, this condition is satisfied by using the first column of the identity matrix $u = [1\ 0\ 0\ \ldots\ 0]^T$ to compute the following spatial spectrum:

$$P_{MN}(\theta) = \frac{1}{|a(\theta)E_N E_N^H u|^2}\,, \tag{15}$$

where $a(\theta)$ is the array steering vector and E_N is the noise subspace with columns representing the eigenvectors $[e_1, e_2, \ldots, , e_{N-P}]$.

2.8. MUSIC Algorithm

Multiple Signal Classification (MUSIC) method [10] is widely used in signal processing applications for estimating and tracking the frequency and emitter location.

This method is considered as a generalization of the Pisarenko's one [9]. It is based on spectral estimation which exploits the orthogonality of the noise subspace with the signal subspace.

Assume that $\hat{R}_{xx}$ is NxN matrix with rank P, therefore it has $N-P$ eigenvectors corresponding to the zeros/smallest eigenvalues in the absence/presence of noise. The eigendecomposition of $\hat{R}_{xx}$ is given by:

$$\hat{R}_{xx} = \sum_{i=1}^{N} \lambda_i q_i q_i^H = Q_s \Delta_s Q_s^H + Q_n \Delta_n Q_n^H \,, \tag{16}$$

where

$$\Delta_s = diag[\lambda_1, \lambda_2, \ldots, \lambda_P] \,,$$

$$\Delta_n = diag[\lambda_{P+1}, \lambda_{P+2}, \ldots, \lambda_N] \,,$$

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_P > \lambda_{P+1} = \lambda_{P+2} = \ldots = \sigma_N^2 \,,$$

$Q_s = [q_1, q_2, \ldots, q_P,]$ is the signal subspace corresponding to Δ_s and $Q_n = [q_{P+1}, q_{P+2}, \ldots, q_N]$ is the noise subspace corresponding to Δ_n.

The MUSIC spectrum is given by:

$$P_{MUSIC}(\theta) = \frac{1}{a^H(\theta) Q_n Q_n^H a(\theta)} \,. \tag{17}$$

When scanning the angles in range $[-\frac{\pi}{2}, \frac{\pi}{2}]$, if θ is DOA of one of signals, so $a(\theta) \perp Q_n$ the denominator is identically zero and the spectrum identifies the angle as a peak.

2.9. Propagator Method

Unlike the MUSIC algorithm, the propagator method [12]–[14] is computationally low complex because it does not need eigendecomposition of the covariance matrix, but it uses the whole of it, to obtain the propagation operator. Therefore, this algorithm is only suitable to the presence of white Gaussian noise and its performance will be degraded in spatial non-uniform colored noise. The propagator is constructed as the following. The covariance matrix can defined as:

$$\hat{R}_{xx} = \begin{bmatrix} R_1 & R_2 \end{bmatrix}^T ,$$

where R_1 and R_2 are PxN, $(N-P)xN$ matrices respectively.

In noiseless system:

$$R_2 = P^H R_1 \,. \tag{18}$$

In noisy environment the least mean squares technique (LMS) is used to estimate P that minimizes the Frobenius norm $||R_2 - P^H R_1||$:

$$P^H = R_2 (R_1^H R_1)^{-1} R_1^H \,. \tag{19}$$

Next, the matrix Q is constructed, such that:

$$Q^H = \begin{bmatrix} P^H & -I_{N-P} \end{bmatrix} . \tag{20}$$

The spectrum is given by:

$$P(\theta) = P_{propag}(\theta) = \frac{1}{||Q^H a(\theta)||^2} \,. \tag{21}$$

2.10. Partial Covariance Matrix

Partial covariance matrix technique [15] is an enhanced version of the propagator method, where no eigendecomposition is needed. The different approach for computing the propagation operator is based on using three submatrices of the estimated cross-correlation matrix $\hat{R}_{xx}$. The array manifold matrix can be portioned as:

$$A = \left[A_1^T, A_2^T, A_3^T\right] , \tag{22}$$

where A_i, $i=1,2,3$ is matrix with dimensions $P \times P$, $P \times P$ and $(N-2P) \times P$ respectively.

The following partial cross-correlation matrices of the array output are defined as :

$$R_{12} = E\left[X(1:P,:)X(P+1:2P,:)^H\right] = A_1 R_{ss} A_2^H \,, \tag{23}$$

$$R_{31} = E\left[X(2P+1:N,:)X(1:P,:)^H\right] = A_3 R_{ss} A_1^H \,, \tag{24}$$

$$R_{32} = E\left[X(2P+1:N,:)X(P+1:2P,:)^H\right] = A_3 R_{ss} A_2^H \,. \tag{25}$$

Based on these sub-matrices, the matrix Q is:

$$Q^H = [R_{32} R_{12}^{-1} \quad R_{31} R_{21}^{-1} \quad -2I_{N-2P}]$$

Multiplying Q with the steering matrix yields to:

$$Q^H A = 0, Q^H a(\theta_k) = 0 \quad (k = 1, 2 \ldots, p) \,. \tag{26}$$

The spectrum is then, similarly to the propagator method, given by:

$$P(\theta) = P_{partial}(\theta) = \frac{1}{||Q^H a(\theta)||^2} \,. \tag{27}$$

3. Simulation Results

A comparative study [17] has been made between 7 algorithms for DOA, using 4 elements and 2 sources with fixed SNR = 10 dB and the 2 sources were separated by $d = 80°$. This study focused on the performance of the algorithms based on the number of snapshots by simulating the first time with L1 = 10 then with L2 = 100 snapshots.

In this paper, real life scenario is simulated by studying the performance of each method based on the noise environment by testing with SNR1 = 1 dB (high noise level) and SNR2 = 20 dB (low noise level). To evaluate the Rayleigh angle resolution limit, for example the second and the third radiating sources were chosen to be separated by 6° while the number of snapshots was fixed.

The authors consider Uniform Linear Array (ULA) composed of $N = 10$ identical sensors with half wavelength inter-element spacing and $P = 4$ almost equally powered emitting sources with carrier frequency $fc = 1$ GHz. The distance between two sensors is $d = 15$ cm so the total distance of the array is 135 cm and $K = 200$ snapshots.

For simulation on evaluating each method the Monte-Carlo method was used such as each result is an average of $L = 100$ runs.

The sources are non-coherent as given by the normalized cross-correlation matrix R_{ss}:

$$R_{ss} = \begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.04 \\ 0.00 & 1.00 & 0.05 & -0.05 \\ 0.00 & 0.05 & 1.00 & 0.08 \\ -0.04 & -0.05 & 0.08 & 1.00 \end{pmatrix}.$$

In Table 1 the configuration of the described sources is presented.

Table 1
Sources characteristics

Sources	S1	S2	S3	S4
DOAS [°]	–24	15	21	70
Power [W]	1.20	1.30	1.44	1.50

Figure 2 shows the results of the Bartlett spectrum, apparently the maximum resolution for this method is more than 6°, which makes inappropriate for this case. In the previous studies [17], the authors show that ideal resolution of this algorithm is 20°.

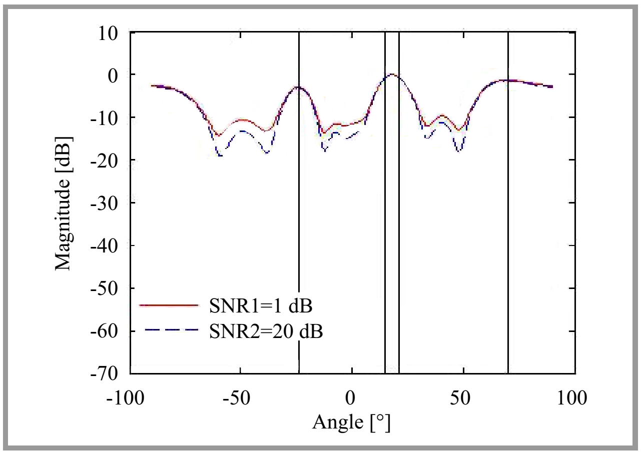


Fig. 2. Bartlett spectrum.

Figure 3 represents the Capon beamformer spectrum which is better performing than the Bartlett method, at SNR = 20 dB the algorithm detects well the sources, but in high-level noise it fails to separate the second and the third sources located at (15°, 21°). The numerical tests at SNR = 1 dB showed that the algorithm can separate the sources with minimal difference of 9°.

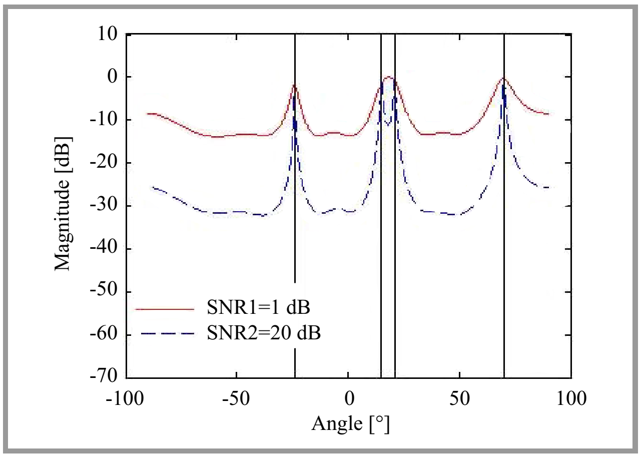


Fig. 3. Capon beamformer spectrum.

Figure 4 shows the result of Linear Prediction algorithm, by choosing the fifth element as the vector u, in the Eq. (12), from the identity matrix $I_{10\times 10}$

$$u = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$$

This algorithms performs better than the two previous techniques, it separates well the closed sources at low SNR.

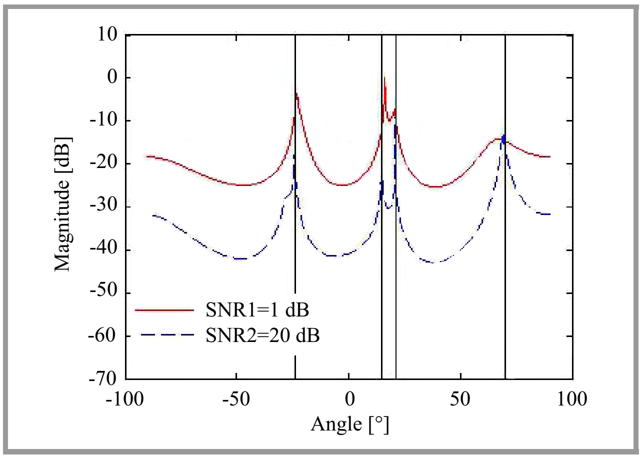


Fig. 4. Linear prediction spectrum.

Figure 5 shows the result of the maximum entropy DOA estimate, by choosing the vector $\bar{C}_j$ as the first column of the inverse cross-correlation matrix $\hat{R}_{xx}^{-1}$ in the Eq. (13).

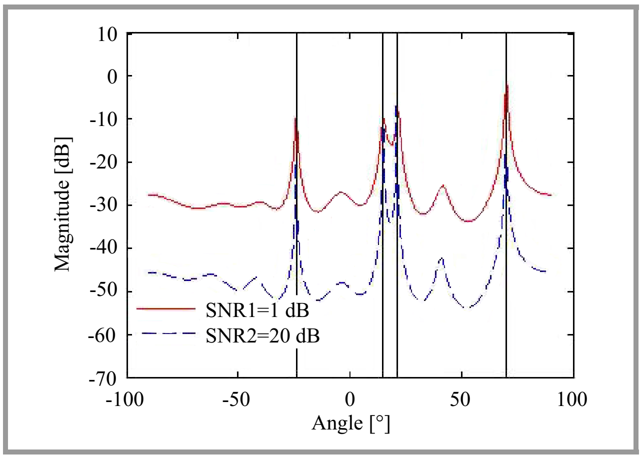


Fig. 5. Maximum entropy spectrum.

This technique performs well by separating the sources at both noise levels which makes it better than Bartlett, Capon and linear prediction methods, however the choice of the column $\bar{C}_j$ influences the performance. As in [17], the j^{th} column was chosen to be in the center of the cross correlation matrix, but in this study the first column was chosen which gives also good results.

In Fig. 6, the application of the Pisarenko harmonic decomposition, at SNR = 20 dB, gives almost the same spectrum of the maximum entropy method, while at SNR = 1 dB, the spectrum detected well the first source at $-20°$, could not separate the second and the third angles while the last source is detected at 67°, which makes this technique non convenient in low SNR condition.

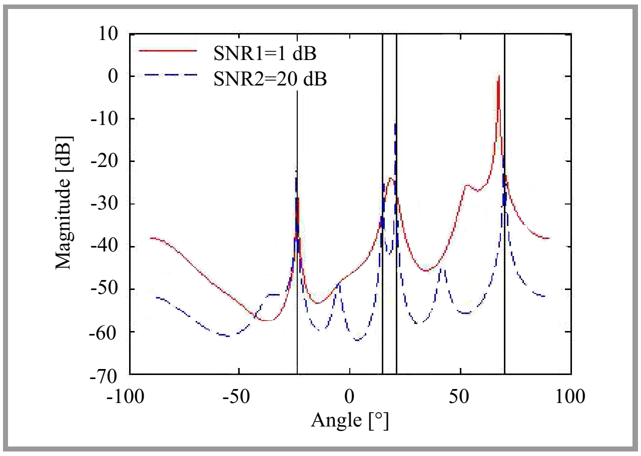


Fig. 6. Pisarenko harmonic decomposition spectrum.

Figure 7 illustrates the minimum norm spectrum which is almost identical with maximum entropy method but with higher number of floating point operations.

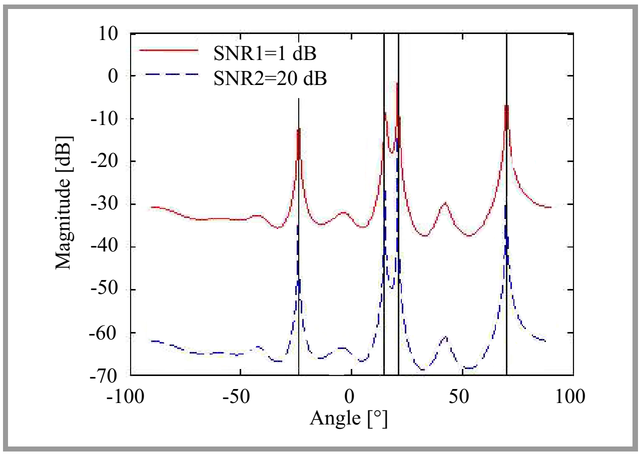


Fig. 7. Minimum norm spectrum.

It should be noted that all the methods are computed using MATLAB and the results are plotted in decibel using the formula

$$P(dB) = 10\log_{10}\left(\frac{spectrum}{Max[spectrum]}\right),$$

to produce a unique frame for comparison [18].

The MUSIC algorithm gives the best result compared to the previous algorithms, as illustrated in Fig. 8, because it detects well all the sources in any noise level and its spectrum does not contain side lobes unlike other techniques.

Note that in high level noise, the spectrum has minimum magnitude of –50 dB while the minimum norm presents a minimum at –60 dB.

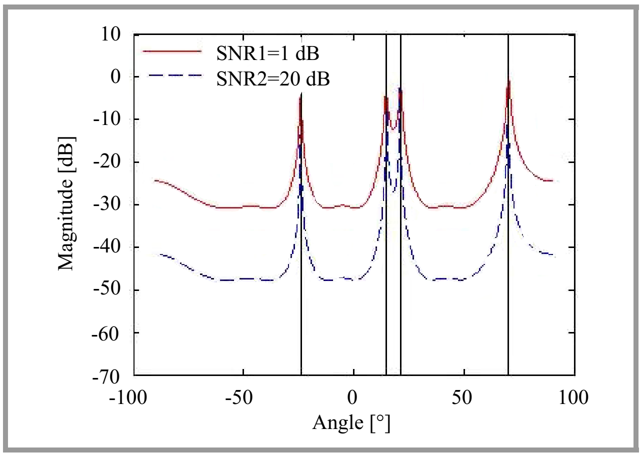


Fig. 8. MUSIC spectrum.

Although, the MUSIC algorithm may fail to resolve the high correlated sources which makes preprocessing techniques like the forward backward averaging or spatial smoothing mandatory to decorrelate the sources.

The propagator method, shown in Fig. 9, has identical performance in both noise levels with minimum apparition of side lobes.

The main advantage of the propagator method is that the constructed matrix Q in Eq. (20) does not need any eigendecomposition, hence the complexity is reduced to $NPK + O(P^3)$ [9].

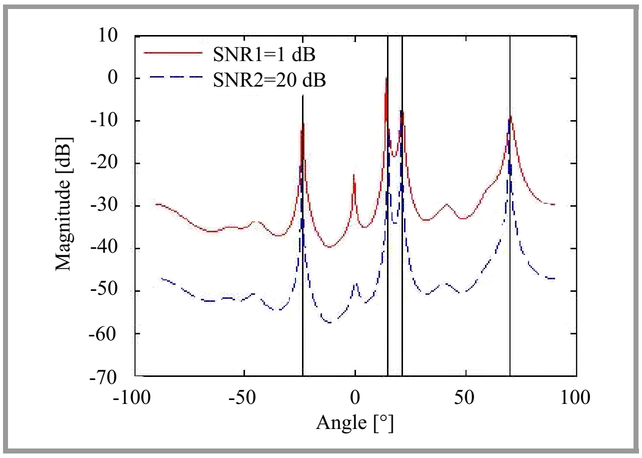


Fig. 9. Propagator spectrum.

Finally, the partial covariance matrix algorithm (without eigendecomposition) is shown in Fig. 10. The results are almost identical with the propagator method, except a noticeable increase in the two side lobes. What makes this technique better than that of the PM method is that the complexity [15] is reduced to $(N-P)PK + O(P^3)$ and takes only partial cross correlation matrices to compute the spectrum. Therefore it is effective in the case of nonuniform colored noise.

The second simulation is based on the average Root Mean Square Error (RMSE) over K = 100 runs between the true DOAs and the nine normalized spectrums, with chrono-

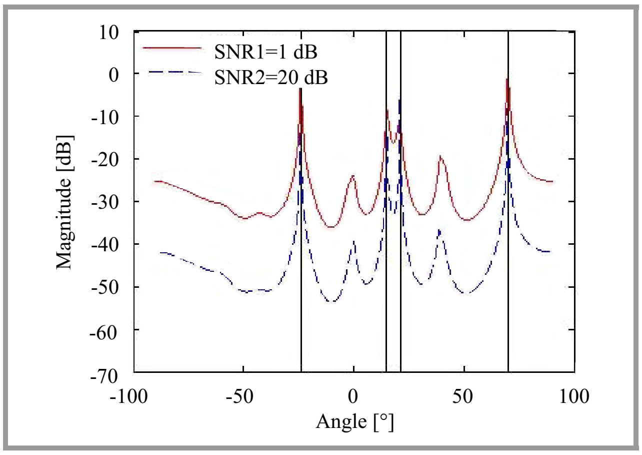


Fig. 10. Partial covariance spectrum.

logical order as described in this paper, computed for two values of SNR:

$$RMSE(\hat{P}(\theta), P(\theta)) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{P}(\theta_n) - P(\theta_n))^2}\,.$$

Figures 11–12 represent the RMSE between each method and the true spectrum for SNR = 1 dB and SNR = 20 dB respectively.

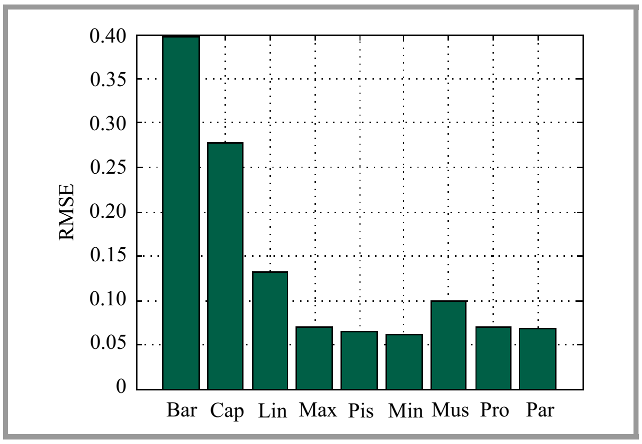


Fig. 11. RMSE, SNR = 1 dB.

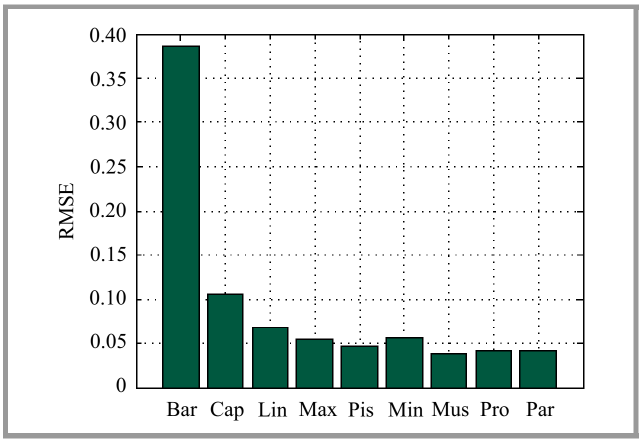


Fig. 12. RMSE SNR = 20 dB.

4. Conclusions

In this paper, some algorithms for one dimensional narrowband direction of arrival (DOA) estimation in stationary case for smart antennas, and for spatially uniform AWGN was compared, starting with the Bartlett method to the recent algorithm which is the partial covariance. In order to evaluate its performance four non-correlated and almost equally powered emitting sources was considered such that two of the sources are separated of 6°, the SNR of 1 dB and 20 dB was the key factor for evaluation. The results showed that in high-level noise, the minimum norm algorithm performs well while in the low-level noise the MUSIC, propagator and partial covariance matrix methods are almost the same and give good results.

In the perspective study, the authors will try to evaluate the partial covariance matrix algorithm in the case of two dimensional wideband sources.

References

[1] Z. Chen, G. Gokeda, and Y. Yu, *Introduction to Direction-of-Arrival Estimation*. Boston, USA: Artech House, 2010.

[2] H. Krim and M. Viberg, “Two decades of array signal processing research”, *IEEES Signal Proces. Mag.*, vol. 13, pp. 67–94, 1996.

[3] M. S. Bartlett, “Periodogram analysis and Continuous spectra”, Biometrika, vol. 37, no. 1–2, pp. 1–16, 1950.

[4] J. Capon, ”High resolution frequency-wave number spectral analysis”, *Proc. IEEE*, vol. 57, pp. 1408–1518, 1969.

[5] J. Makhoul, “Linear prediction: A tutorial review”, *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[6] W. Min and W. Shunjun, “A time domain beamforming method of UWB pulse array”, in *Proc. IEEE In. Radar Conf.*, Arlington, VA, USA, 2005, pp. 697–702.

[7] R. Sanudin *et al.*, ”Capon-like DOA estimation algorithm for directional antenna arrays”, in *Proc. Loughborough Anten. Propag. Conf. LAPC 2011*, Loughborough, UK, 2011, pp. 1–4.

[8] S. Ejaz and M. A. Shafiq, ”Comparison of spectral and subspace algorithms for Fm source estimation”, *Progr. Electromag. Res. C*, vol. 14, pp. 11–21, 2010.

[9] L. C. Godora, “Application of antenna arrays to mobile communications. beamforming and direction-of-arrival considerations”, *Proc. IEEE*, vol. 85, no. 8, pp. 1195–1245, 1997.

[10] R. O. Schimd, “Multiple emitter location and signal parameter estimation”, *IEEE Trans. Anten. Propag.*, vol. 34, no. 3, pp. 276–280, 1086.

[11] J. Xin and A. Sano, “Computationally efficient subspace based method for direction of arrival estimation without eigendecomposition”, *IEEE Trans. Sig. Proces.*, vol. 52, no. 4, pp. 876–893, 2004.

[12] J. Munier and G. Y. Delisle, “Spatial analysis using new properties of the cross-spectral matrix”, *IEEE Trans. Sig. Proces.*, vol. 39, no. 3, pp. 746–749, 1991.

[13] S. Marcos, A. Marsal, and M. Benidir, “The propagator method for source bearing estimation”, *Sig. Proces.*, vol. 42, no. 2, pp. 121–138, 1995.

[14] M. Frikel, “Localization of sources radiating on a large antenna”, in *Proc. 13th Eur. Sig. Proces. Conf. EUSIPCO 2005*, Antalya, Turkey, 2005.

[15] J. Chen, Y. Wu, H. Cao, and H. Wang, “Fast algorithm for DOA estimation with partial covariance matrix and without eigendecomposition”, *J. Sig. Inform. Proces.*, vol. 2, no. 4, pp. 266–259, 2011.

[16] J. Foutz, A. Spanias, and M. K. Banavar, *Narrowband Direction of Arrival Estimation for Antenna Arrays*. San Rafael, USA: Morgan and Claypool, 2008.

[17] Md. Bakhar, R. M. Vani, and P. V. Hunagund, "Comparative studies of direction of arrival algorithms for smart antenna systems", *World J. Sci. Technol.*, vol. 1, no. 8, pp. 20–25, 2011.

[18] X. WU and T. Guo, "Direction of arrival parametric estimation and simulation based on MATLAB", *J. Comput. Inform. Syst.*, vol. 6, no. 14, pp. 4723–4731, 2010.

[19] Q. Yuan, Q. Chen, and K. Sawaya, "Accurate DOA estimation using array antenna with arbitrary geometry", *IEEE Trans. Anten. Propag.*, vol. 53, no. 4, pp. 1352–1357, 2005.

[20] M. Frikel, B. Targui, S. Safi, and M. M'saad, "Bearing detection of noised wideband sources for geolocation", in *Proc. 18th Mediter. Conf. Control Autom. MED*, Marrakech, Morocco, 2010, pp. 1650–1653.

[21] X. Zhang, Y. Bai, and W. Zhang, "DOA estimation for wideband signals based on arbitrary group delay", in *Proc. World Congr. Engin. Comp. Sci. WCECS 2009*, San Francisco, USA, 2009, vol. 2, pp. 1298–1300.

Youssef Khmou obtained the B.Sc. degree in Physics and M.Sc. degree from poly disciplinary faculty, in 2010 and from Faculty of Science and Technics Beni Mellal, Morocco, in 2012, respectively. Now he is Ph.D. student and his research interests include statistical signal and array processing and statistical physics.

Email: khmou.y@gmail.com
Department of Mathematics and Informatics
Beni Mellal, Morocco

Said Safi received the B.Sc. degree in Physics (option Electronics) from Cadi Ayyad University, Marrakech, Morocco in 1995, M.Sc. degree from Chouaib Doukkali University and Cadi Ayyad Univerrsity, in 1997 and 2002, respectively. He has been a Professor of information theory and telecommunication systems at the National School for applied Sciences, Tangier, Morocco, from 2003 to 2005. Since 2006, he is a Professor of applied mathematics and programming at the Faculty of Science and Technics, Beni Mellal, Morocco. In 2008 he received the Ph.D. degree in Telecommunication and Informatics from the Cadi Ayyad University. His general interests span the areas of communications and signal processing, estimation, time-series analysis, and system identification – subjects on which he has published 14 journal papers and more than 60 conference papers. Current research topics focus on transmitter and receiver diversity techniques for single- and multi-user fading communication channels, and wide-band wireless communication systems.

E-mail: safi.said@gmail.com
Department of Mathematics and Informatics
Beni Mellal, Morocco

Miloud Frikel received his Ph.D. degree from the center of mathematics and scientific computation CNRS URA 2053, France, in array processing. Currently, he is with the GREYC laboratory (CNRS URA 6072) and the ENSICAEN as Assistant Professor. From 1998 to 2003, Dr. Frikel was with the Signal Processing Lab, Institute for Systems and Robotics, Institute Superior Tecnico, Lisbon, as a researcher in the field of wireless location and statistical array processing, after been a research engineer in a software company in Munich, Germany. He worked in the Institute for Circuit and Signal Processing of the Technical University of Munich. His research interests span several areas, including statistical signal and array processing, cellular geolocation (wireless location), space-time coding, direction finding and source localization, blind channel identification for wireless communication systems, and MC-CDMA systems.

E-mail: mfrikel@greyc.ensicaen.fr
GREYC UMR 6072 CNRS
Ecole Nationale Supérieure d'Ingénieurs
de Caen (ENSICAEN)
6, B. Maréchal Juin" 14050 Caen, France

MUPUS insertion device for the Rosetta mission

Jerzy Grygorczuk, Marek Banaszkiewicz, Karol Seweryn, and Tilman Spohn

http://doi.org/10.26636/jtit.2007.1.794

Abstract— An original mechanical device designed to insert a penetrator into a cometary nucleus in an almost gravity-free environment is described. The device comprises a hammer and a power supply system that stores electrical energy in a capacitor. The accumulated energy is discharged through a coil forming a part of electromagnetic circuit that accelerates the hammer. The efficiency of converting the electrical energy to kinetic energy of the hammer is not very high (amounts to about 25%), but the system is very reliable. Additionally, the hammer energy can be chosen from four power settings, hence adjustment of the stroke's strength to nucleus hardness is possible. The device passed many mechanical, functional, thermal and vibration tests and was improved from one model to another. The final, flight model was integrated with the lander Philae and started its space journey to comet Churyumov-Gerasimenko in March 2004.

Keywords— comets, penetrators, hammering device.

1. Introduction

The European Space Agency (ESA) cornerstone mission Rosetta to comet Churyumov-Gerasimienko comprises the main spacecraft that will become a comet companion for at least half a year and the Philae lander [1]. Philae weights about 100 km and includes eight instruments that will measure chemical composition and physical properties of the comet [2]. Space Research Centre participates in the experiment MUPUS (multi-purpose sensors for surface and sub-surface science) [3, 4] that is dedicated to obtain temperature profile of nucleus' subsurface layers to a depth of 40 cm and thermal conductivity of cometary material. The experiment MUPUS is developed by a multinational team led by Prof. T. Spohn from the Muenster University.

The main engineering problems that had to be solved in the design phase were:

- how to insert a 40 cm long penetrator equipped with thermal sensors into the nucleus composed of a porous ice-dust mixture;
- how to deploy the penetrator and its insertion device to a distance of about 1 m from Philae, in order to avoid thermal perturbations caused by the lander.

In this short paper we will address the first issue only.

The cometary environment is very unusual and poses severe requirements on the instrument. First of all, the nucleus is a small body (2–3 km in diameter), with almost negligible gravity (more than four orders of magnitude smaller that on the Earth) and with low pressure and temperature, 10^{-6} bar and 130 K, respectively. The gravity-free condition does not allow using the lander body as an inertial support; it will recoil easily after each stroke of the insertion device, in agreement with the conservation of momentum in an isolated system. In the vacuum and low temperature on the nucleus it is very important to avoid cold-welding effect between metal parts of the device and carefully design all moving and driving subsystems (e.g., motors). Additional constraints are due to sparse lander resources: an average power assigned to MUPUS is 1.5 W, and the instrument should not weight more than 2 kg.

2. Device concept and development

The first issue to be decided about was what type insertion device would be used:

- a single impulse engine (rocket), powered by gunpowder, chemical fuel or high-pressure gas, or
- a multi-stroke system that would slowly insert the penetrator into the nucleus.

For safety reasons and because of uncertainty in our knowledge about the material strength of the nucleus, the first option was abandoned and the hammering concept was chosen. The Philae power supply system delivers energy to its instruments in a form of low voltage electric current. In order to execute hammer strokes energetic enough to penetrate the surface this power must first be stored in the insertion device and then released with maximum efficiency possible. Three kinds of energy storage systems were considered:

- mechanical potential (spring);
- mechanical kinetic (reaction wheel);
- electric (capacitor).

The capacitor has an advantage that there is no need to convert the supplying current to any form of mechanical energy and was, therefore, accepted. Following this choice, an electromagnet was implemented as the hammer accelerating subsystem. The static hollow cylindrical carcass made of iron forms together with a hammer an electromagnetic circuit. The discharge of capacitor converts the current energy to magnetic field and forces the hammer to move inside the iron cylinder to close the circuit.

After having solved the hammer engine problem, one has to consider the basic issue of how to design the mechanical system that would be able to continuously insert the penetrator into the cometary ground without supporting gravity.

The concept applied in the MUPUS insertion device employs three masses: the penetrator, the hammer and the counter-mass. The counter-mass is connected to the other masses with weak springs. A single stroke can be described as a sequence of events:

a) acceleration of the hammer;

b) forward motion of the hammer and slower motion of the (heavy) counter-mass backwards (from momentum conservation);

c) the hammer hits the penetrator and recoils;

d) the penetrator moves forwards, into the ground, the hammer and the counter-mass together move backwards;

e) the penetrator stops or moves backwards for a short time;

f) the counter-mass and the hammer stop.

Since the counter-mass is much heavier than the hammer and the penetrator tube, therfore it moves slower and the average force acting from it through the spring on the penetrator in stages (d) and (e) is much smaller than the forward force pushing the penetrator into the ground just after the hammer hit. In Fig. 1 the insertion scenario during a single stroke is illustrated. The phase 1 in the figure corresponds to event (a) in the scenario. Events (b) and (c) are shown as phase 2, while events (d) and (e) are merged in the figure into phase 3. Finally, the last event (f) corresponds to phase 4.

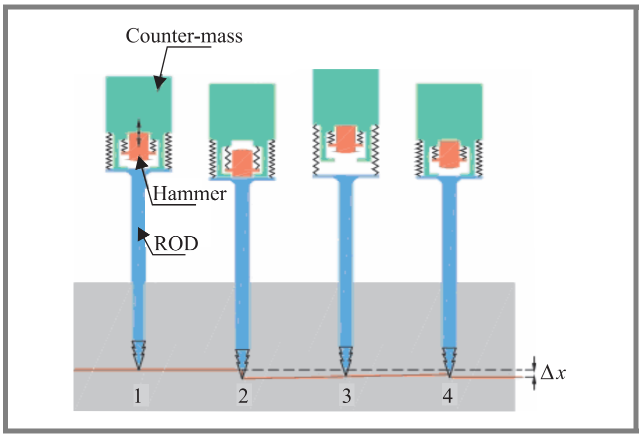


Fig. 1. Four phases of insertion during a single stroke: 1 – hammer acceleration; 2 – hammer hits the penetrator and recoils; 3 – the masses move backwards; 4 – the motion stops.

In Fig. 2 time dependence of the force acting on the penetrator is shown. The energetic phase of penetration, in which the force exerted on the penetrator is large, is followed by a much slower (due to the counter-mass inertia) recoil that tries to pull out the penetrator with a much weaker force. If this latter force is below the level of friction/anchoring force between the medium and the penetrator, then the penetrator will not be pulled out of the ground during recoil. An elementary analysis of the efficiency of momentum exchange in hammer-penetrator collisions as well as of push in and pulling out forces acting on the penetrator from the counter-mass shows that the mass distribution between the penetrator, the hammer and the counter-mass should be close to the relation 1:1:10. In practice, the masses are limited by functional constraints, choice of material, geometry, etc., therefore it is difficult to reach the ideal proportions. In the final, flight model of MUPUS the penetrator weights 60 g, the hammer 30 g, and the counter-mass 350 g.

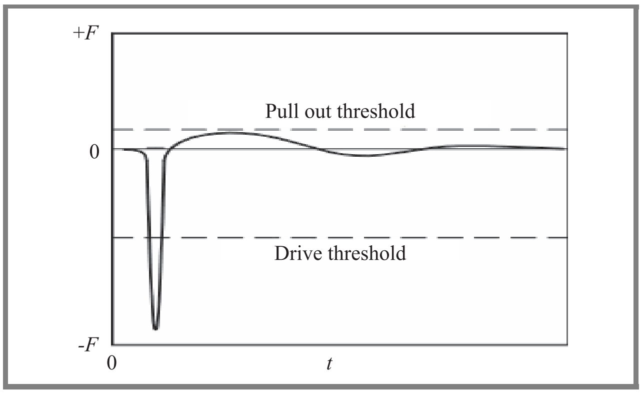


Fig. 2. Time dependence of the force exerted on the penetrator during a single stroke. The first part corresponds to the hitting of the penetrator by the hammer. The second part shows the pulling out force that acts during the recoil.

The last problem to consider is how to support the hammering device during the initial stage of insertion, when the penetrator is not yet stuck in the ground. Here, the deployment device composed of two expandable tubular booms comes to the rescue. It links the insertion device with the lander with a force of about 1 N that is strong enough to bring the penetrator tip back to the ground after the recoil following the stroke. The penetrator tip is equipped with a set of anchoring whiskers, which efficiently increase the resistance of the penetrator against the pulling out force.

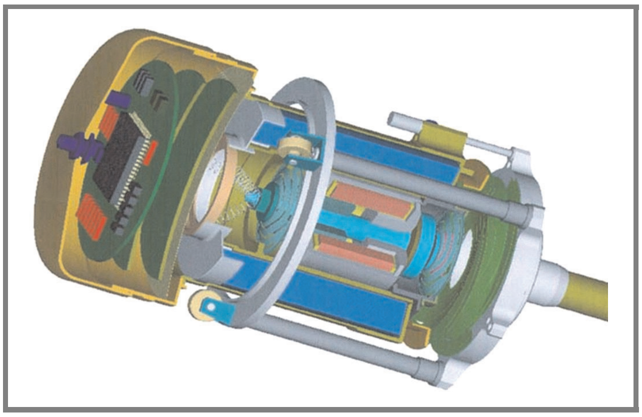


Fig. 3. Cross-section of the hammering device. Only the upper part of the penetrator rod is shown.

The cross-section of the hammering device is shown in Fig. 3. The housing contains the capacitor that surrounds the electromagnet with the hammer inside it. Above, there is an electronic compartment with the hammer controlling circuits and chips.

3. Performance and tests

The device was carefully tested, first at the level of subsystems (electromagnet, capacitor, mechanical part, etc.) then as a whole. The electromagnetic circuit was simulated by finite element method (FEM) and its parameters were optimized [5]. The most interesting were the functional tests of insertion into cometary like materials. Those were simulated by Ytong and solid foam blocks. To imitate the gravity free condition, the penetrator was suspended horizontally on a pair of strings (Fig. 4). The blocks were highly porous (up to 90%) that appropriately mimics cometary material but not at all weak; their compressive strength ranged from 0.79 MPa, through 1.75 MPa, to 5–7 MPa (for solid silica foam). The estimated strength of cometary nuclei vary from 25 kPa to 2.5 MPa. Since the test were passed successfully, one can assume that the hammering device would be able to insert the penetrator into the comet as well. The attempts to insert the penetrator into the ice were moderately successful. If its strength did not exceed 3 MPa, the penetrator worked fine. Above this limit, it could only be inserted to a certain depth and then got stuck in the ice.

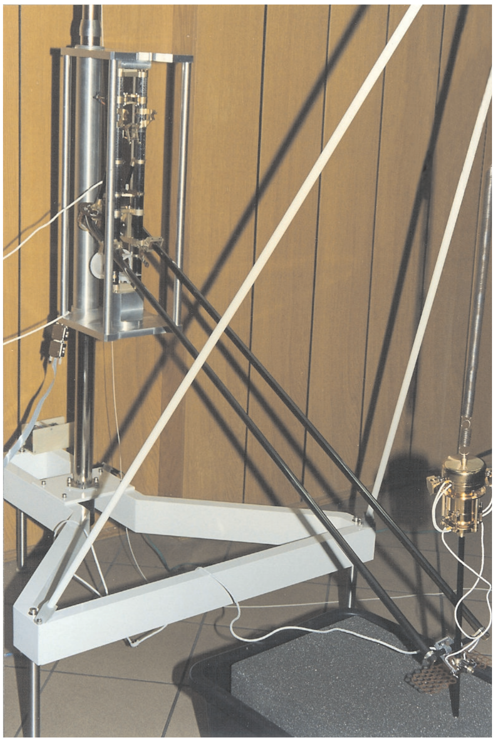


Fig. 4. The stand for functional test of the insertion device (left picture).

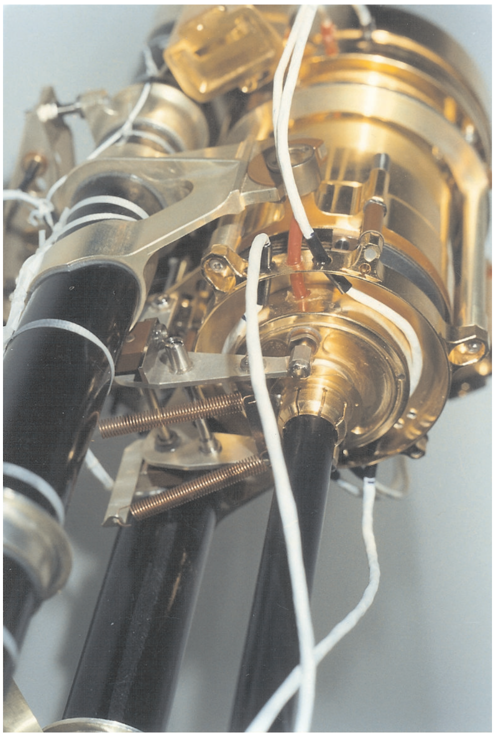


Fig. 5. The flight model of MUPUS insertion device in the stowed position on the lander balcony (right picture).

The functional tests of the insertion device were followed by several other tests of the whole MUPUS device: vacuum-thermal, electrical, vibrational, and electromagnetic compatibility (EMC). The most severe test were applied to the engineering model, so that the final flight model of MUPUS (Fig. 5) were subject to only moderately heavy loads and environmental tests.

4. Conclusions

The MUPUS is one of the most complicated space instruments developed in Space Research Centre. It is also the only so far designed and developed cometary penetrator. It took 6 years to develop the instrument, starting from the first conceptual study in 1996 till the delivery of the flight model to ESA in 2001. The insertion device described in this paper comprises only a pat of the whole experiment. It includes two other scientific elements: IR mapper that is placed on the lander balcony and accelerometer and temperature sensor located in the anchor that will be shot into the nucleus at the moment of landing. Those two units were developed by German and Austrian colleagues, respectively. What concerns Polish contribution, we developed the deployment device, thermal sensors for the penetrator, depth sensor to measure the insertion progress, part of the on-board and the whole hammer electronics, and flight software to control the experiment. The outcome of the tests are the data that are now used in computer simulations [6]. The involvement in MUPUS enabled the Polish team to gain the knowledge and expertise necessary to participate in most demanding space endeavors.

References

[1] ESA Rosetta mission, http://www.esa.int/esaMI/Rosetta/index.html

[2] DLR Resetta lander Philae, http://www.dlr.de/rs/forschung/roland/

[3] W. Marczewski *et al.*, "Prelounch performance evaluation of the cometary experiment MUPUS-TP", *J. Geophys. Res.*, vol. 109, pp. 1–17, 2004.

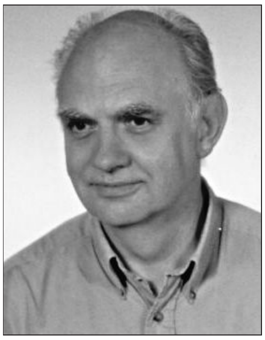
[4] T. Spohn *et al.*, "MUPUS – a thermal and mechanical properties probe for the Rosetta lander Philae" (submitted to *Space Sci. Rev.*).

[5] A. Demenko, L. Nowak, W. Szeląg, and J. Grygorczuk, "Symulacja dynamicznych stanów pracy elektromagnetycznego urządzenia wbijającego do penetratora gruntu komety", in *Symp. PPEE*, Ustronie, Poland, 1997, pp. 104–109 (in Polish).

[6] K. Seweryn, M. Banaszkiewicz, M. Grunwald, J. Grygorczuk, and T. Spohn, "Thermal model of MUPUS penetrator", *Int. J. Heat Mass Trans.*, vol. 48, pp. 3713–3721, 2005.

Jerzy Grygorczuk has been employed at SRC-PAS since 1977. Has participated in the development of mechanical subsystems of satellite instruments for over twenty space missions, including Cassini, Rosetta and Mars Express. His field of interest include: aerospace mechanisms and mechanical properties of the soil of extraterrestrial bodies.

e-mail: jurekgry@cbk.waw.pl
Space Research Centre
Polish Academy of Science
Bartycka st 18a
00-716 Warsaw, Poland



Marek Banaszkiewicz has graduated from Warsaw University, Poland, in solid state physics in 1968. He received a Ph.D. degree in theoretical physics (in 1982) and D.Sc. (habilitation in 2000) in astrophysics. He is a specialist in planetology and space physics and participated in several ESA missions: Cassini/Huygens, Ulysses, Rosetta, SMART1, Bepi Colombo.

e-mail: marekb@cbk.waw.pl
Space Research Centre
Polish Academy of Science
Bartycka st 18a
00-716 Warsaw, Poland

Karol Seweryn is currently a Ph.D. student in the Space Research Centre of the Polish Academy of Sciences, Poland. He received his bechelor's and master's of science degrees from Technical University of Kraków in 2003 in the automatics and robotics discipline. His main research interest are the dynamics and control of spacecraft and space robots. Other fields of activities are: spacecraft thermal control systems, finite elements analysis and mechanical engineering.

e-mail: kseweryn@cbk.waw.pl
Space Research Centre
Polish Academy of Science
Bartycka st 18a
00-716 Warsaw, Poland

Tilman Spohn
German Aerospace Center DLR
Rutherford st 2
12489 Berlin, Germany

Application of multiple criteria evolutionary algorithms to vector optimisation, decision support and reference point approaches

Marcin Szczepański and Andrzej P. Wierzbicki

http://doi.org/10.26636/jtit.2003.3.194

Abstract — Multiple criteria evolutionary algorithms, being essentially parallel in their character, are a natural instrument of finding a representation of entire Pareto set (set of solutions and outcomes non-dominated in criteria space) for vector optimisation problems. However, it is well known that Pareto sets for problems with more than two criteria might become complicated and their representation very time-consuming. Thus, the application of such algorithms is essentially limited to bi-criteria problems or to vector optimisation problems with more criteria but of simple structure. Even in such cases, there are problems related to various important aspects of vector optimisation, such as the uniformity of representation of Pareto set, stopping tests or the accuracy of representing Pareto set, that are not fully covered by the broad literature on evolutionary algorithms in vector optimisation. These problems and related computational tests and experience are discussed in the paper. In order to apply evolutionary algorithms for decision support, it would be helpful to use them in an interactive mode. However, evolutionary algorithms are in their essence global and of batch type. Nevertheless, it is possible to introduce interactive aspects to evolutionary algorithms by focusing them on a part of Pareto set. The results of experimental tests of such modifications of evolutionary algorithms for vector optimisation are presented in the paper. Another issue related to vector optimisation problems with more than two criteria is the computational difficulty of estimating nadir points of Pareto set. The paper describes the use of diverse variants of evolutionary algorithms to the estimation of nadir points, together with experimental evidence.

Keywords — evolutionary algorithms, vector optimisation, nadir point estimation, reference point techniques.

1. Evolutionary algorithms in vector optimisation: general comments

There are many excellent reviews of evolutionary algorithms used in vector optimisation [3–5, 10, 12]. Most of them, however, treat evolutionary or genetic algorithms as goals in themselves, as given tools that should be further developed and put into use. In this paper, we concentrate rather on the use of such algorithms for solving various tasks of vector optimisation or multiple criteria analysis for decision support.

First, let's recall the traditional distinction between genetic and evolutionary algorithms: genetic algorithms rely on binary representation of individuals, while evolutionary algorithms admit real-valued (computational) representations. For vector-valued representations, evolutionary algorithms are more appropriate. On the other hand, special methods developed for genetic algorithms can be also usefully translated into evolutionary algorithms.

Next, we observe that evolutionary algorithms are applied to vector optimisation in order to obtain accurate representation of the Pareto set (or any modified concept of a non-dominated set). Being inherently parallel, evolutionary algorithms are a natural approach to the problem of representing a complicated set. However, research on truly parallel or distributed implementations of evolutionary algorithms is scarce. Thus, the application of such algorithms is essentially limited to bi-criteria problems or very simple vector optimisation problems with more criteria. Accurate representation of more complicated Pareto sets using evolutionary algorithm still requires huge computation efforts.

On the other hand, practical applications of vector optimisation to decision support require interactive multiple criteria analysis [11], where instead of computing a single Pareto set, various characteristics of selected variants or parts of Pareto sets are needed for subsequent formulations of the problem being analysed. Such cases include utopia points, nadir points, neutral compromise points of Pareto sets and, finally and most importantly for interactive applications – representations of selected segments of Pareto sets. While evolutionary algorithms might be useful for obtaining such characteristics, little attention was given to such applications. Generally speaking, the same fact can be stated as follows: *since evolutionary algorithms are global and non-interactive in their nature, the challenge in their applications for multiple criteria analysis is to make them more local and interactive*. While this paper does not resolve all problems related to this challenge, it tries to move in this direction – by treating evolutionary algorithms not as main goal in itself, but as a way of addressing various tasks of multiple criteria analysis.

2. Modifications of evolutionary algorithms in vector optimisation

2.1. Representation of individual

By *individual* in genetic or evolutionary algorithms, we consider a current solution point together with additional parameters, typically characterising its mutation potential by specifying the dispersion σ. In vector optimisation or multiple criteria analysis, current solution is typically represented by a vector of decision variables $x \in R^n$ and vector of decision outcomes or criteria $q \in R^k$. Dispersion parameters are related to decision variables and can be represented by a vector of the same dimension. Thus, an individual is represented by:

$$ind = (x, \sigma, q) \in R^{2n+k}. \tag{1}$$

2.2. Constraints

Constraints on decision variables (either in equation or inequality form) define the permissible set of decisions:

$$X_0 = \left\{ x \in R^n : \begin{array}{c} g_i(x) \geq 0, i \in I \\ h_i(x) = 0, i \in E \end{array} \right\}. \tag{2}$$

In genetic algorithms, if x is not in X_0, the individual is simply discarded. This may lead, however, to quite long computations if the set X_0 has a complicated structure. Therefore, we shall use a method typically adopted in evolutionary algorithms to represent constraints – applying penalty functions. There are many types of penalty functions (internal, external, exact, shifted, etc. – see e.g. [11]). With evolutionary algorithms that do not need derivatives of optimised functions, it is best to use exact non-differentiable external penalty functions of the type $|h_i(x)|$ and $|g_i(x)_-| = \min\big(g_i(x), 0\big)|$ (with sufficiently large penalty coefficients), which are added to each criterion value – if it is minimised or subtracted – if maximised.

2.3. Cross-breeding

Cross-breeding is a typical evolutionary operation. In vector optimisation, cross-breeding applies to two parent individuals represented by decision variable vectors x_1 and x_2; their successor x' may be determined as follows:

$$x' = ax_1 + (1-a)x_2, \tag{3}$$

where the parameter a is a random variable from the interval $a \in [0,1]$. This is called *basic arithmetic cross-breeding*, while *extended arithmetic cross-breeding* applies to each component x'_i of the vector x' with separately generated random coefficients a_i. There are several other variants of cross-breeding, such as *heuristic cross-breeding*, not discussed here.

2.4. Mutation

In vector optimisation, mutation is applied to every component x_i of the decision variable vector x (usually, mutation is additionally applied to a successor of cross-breeding) by selecting a random variable with a normal distribution and modifying the component x_i by this variable with a corresponding dispersion coefficient:

$$\begin{aligned} \xi_i^x &\in N(0,1), \\ x'_i &= x_i + \sigma'_i \xi_i^x. \end{aligned} \tag{4}$$

Additionally, the dispersion coefficient is modified randomly, but usually slowly decreased after (or before) each mutation. This decreasing modification of dispersion parameters slows down mutations when approaching solutions. In vector optimisation, it results in coming closer to the Pareto set.

2.5. Selection

Selection is responsible for convergence of a genetic algorithm towards optimal solutions and applies to selection of parent individuals (selection in reproduction); there are numerous methods of such selection, not discussed here. In evolutionary algorithms, succession may substitute for selection. This means choosing the μ as best individuals from population $\mu + \lambda$ (so-called $\mu + \lambda$ succession strategy; μ denotes here a population from parent individuals, λ the corresponding population of successors) in some way. Another strategy consists of simply substituting parent population μ by successor population λ (the so-called μ, λ strategy). For vector optimisation purposes, succession is superior to selection.

2.6. Pareto ranking

Succession process includes multiple stages to uniformly approximate Pareto set by an evolutionary algorithm. First, we use *Pareto ranking of a population*, then apply special *niched methods* for preserving uniformity of representation, and finally use special *succession methods*. We will describe all of them below.

Pareto ranking consists of attaching a rank value (the lower the better) to each individual. Goldberg [2] has proposed to give rank 1 to each non-dominated individual in population. Next, we delete the non-dominated individuals and determine non-dominated individuals in remaining part of population, giving them rank 2. We continue the process with increasing rank values until each individual has a rank value. Then we can either select successor population of given number of individuals according to lowest rank values, or – as proposed by Goldberg – determine the probability of reproduction depending on rank value (which is actually a selection, not a succession mechanism).

Another ranking method proposed by Fonseca and Fleming [4] involves assigning each individual a rank value of 1 plus the number of other individuals dominating this individual. This method provides for more differentiation of a population than Goldberg method.

Having a rank value, it is easy to determine a fitness indicator $fit(x)$ – for example, by defining it as inverse of the rank value.

2.7. Niched methods

Having a fitness value or fitness function for Pareto ranking, it is easy to apply the basic principle of evolutionary algorithms – the *survival of the fittest* individuals.

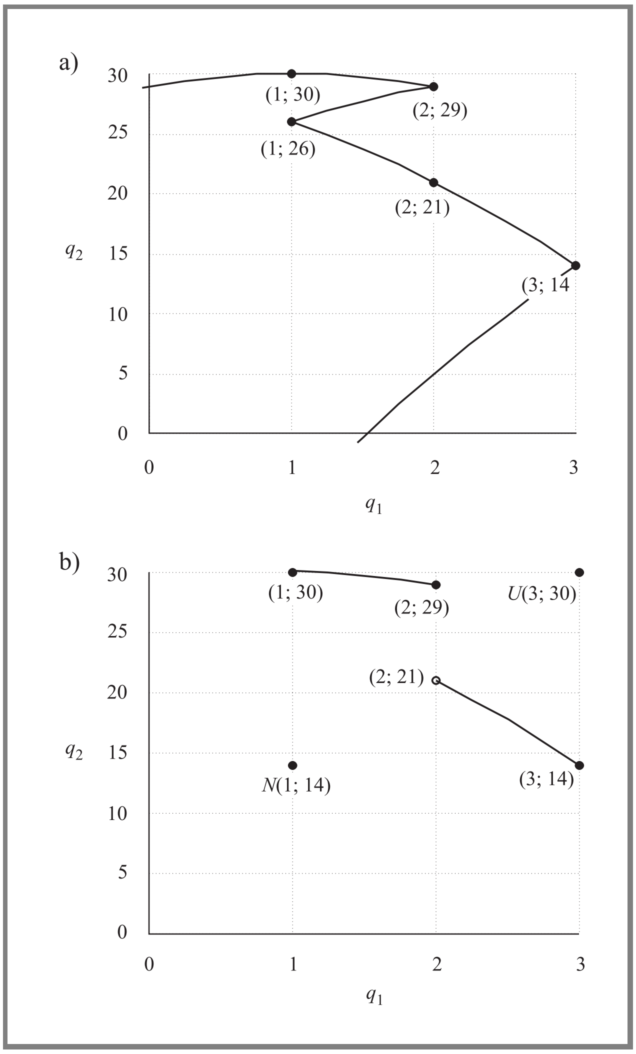


Fig. 1. The set of attainable criteria values (a) and the Pareto set (b) for the nonlinear example.

However, such a method does not result in a uniform representation of the Pareto set. The fittest individuals can form an elite close to each other, representing only an "easy" part of Pareto set. Such degeneration of the *survival of the fittest* principle can be illustrated by a relatively simple, but nonlinear example (Fig. 1). We maximize two criteria functions (with $-0.5 \le x \le 6$):

$$\begin{aligned}\max: q_1(x) &= \left\{\begin{array}{cc} x+2 & x\le 1\\ -x+4 & 1<x\le 3\\ x-2 & 3<x\le 4\\ -x+6 & x>4\end{array}\right\},\\ \max: q_2(x) &= -x^2+10x+5\,. \end{aligned} \tag{5}$$

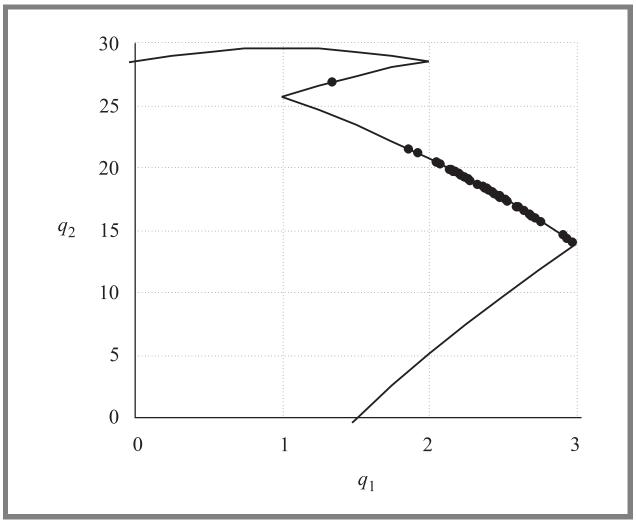


Fig. 2. Non-uniform representation of Pareto set with a simple *survival of the fittest* evolutionary algorithm (population size: 50, 200 generations).

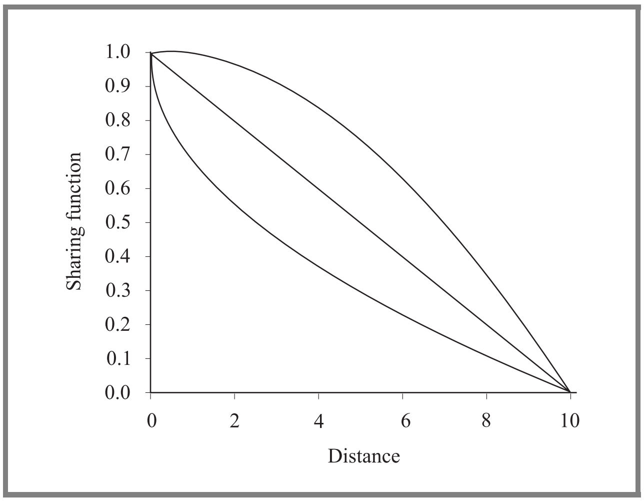


Fig. 3. Examples of sharing functions.

Application of a simple *survival of the fittest* algorithm here results in a degenerated representation of the Pareto set, concentrating on the "easy part" of the set (Fig. 2).

In order to overcome this difficulty, we must penalise the fitness function for individuals being too close to each other.

With this aim, we define a *sharing function* depending on a distance of two individuals, say x and x_0. This sharing function sh must have the following properties:

$$\begin{gathered} 0 \le sh(x - x_0) \le 1, \text{ for any distance } |x - x_0| \\ sh(0) = 1 \\ \lim sh(x - x_0) = 0, |x - x_0| \to \infty. \end{gathered} \tag{6}$$

Sharing functions shown in Fig. 3 belong to the family:

$$sh(x - x_0) = \begin{cases} 1 - \left(\frac{|x - x_0|}{D}\right)^p & |x - x_0| < D \\ 0 & |x - x_0| \ge D \end{cases}, \tag{7}$$

where D is a diameter of a *niche*.

The so-called *niched methods* consist of modifying fitness values $fit(x)$ for a given individual x, reciprocal to the sharing function:

$$fit'(x) = \frac{fit(x)}{1 + m(x)}, \tag{8}$$

where $m(x)$ is a sum of sharing functions over other non-dominated individuals y in given population:

$$m(x) = \sum_y sh\big(d(x, y)\big). \tag{9}$$

Figure 4 illustrates effectiveness of such niched methods in preventing degeneration through cross-breeding of too close individuals.

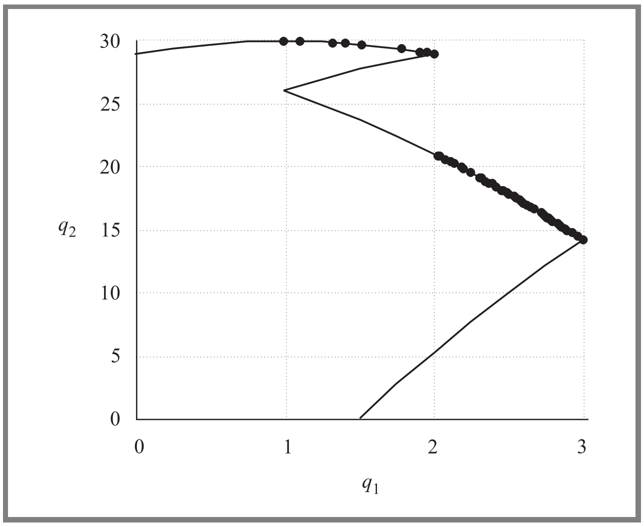


Fig. 4. Effectiveness of a *niched method* with $D = 0.1$ (population size: 50, 200 generations).

We see it is necessary to use niched methods in evolutionary algorithms of vector optimisation not only in order to obtain a uniform representation of Pareto set, but also to prevent degenerate populations resulting from naive direct application of the "survival of the fittest" principle.

2.8. Stopping tests

Before discussing succession methods, stopping tests for entire algorithm should be discussed. Stopping test for evolutionary and genetic optimisation algorithms are much less developed than for analytical optimisation methods. If the optimal value of an optimised function is known (which happens only in very special cases) then the distance from this optimal value can be used for a stopping test. Otherwise, one must limit the number of iterations in the algorithm (number of generations in a genetic or evolutionary algorithm) and hope for a good accuracy. Another stopping test is based on the speed of change of an approximation of the solution: work stops when changes fall below certain level.

For vector optimisation, the issue of stopping tests is more complicated. We can rely on a given number of iterations or generations, but cannot easily use the speed of change, because we approximate or represent an entire Pareto set and the uniformity of this representation is also a goal. A substitute for the speed of change might be a comparison of two subsequent generations and checking how many individuals in the next generation dominate some individuals in the former generation. Figure 5 shows example of such computation.

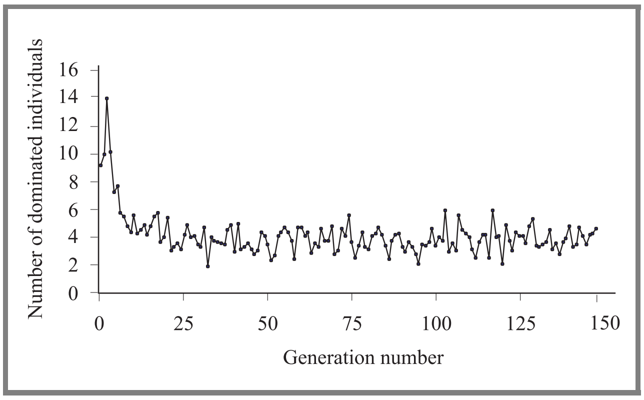


Fig. 5. Average numbers of dominated individuals between generations for a typical evolutionary algorithm.

We see such a stopping test cannot be very reliable. Other tests, however, might be related to special features of vector optimisation. One relates to the uniformity of Pareto set representation, which can be represented by average value of sharing function $m(x)$ as defined by Eq. (9). Another relates to the concept of utopia and nadir points for a Pareto set. For an approximation of Pareto set obtained in a subsequent generation numbered here by i, it is relatively easy (see also point **4**) to compute utopia points q_i^U ("lowest" points dominating entire Pareto set) as well as nadir points g_i^N ("highest" points dominated by the entire Pareto set). If the approximation of a Pareto set converges to the actual Pareto set, the distance between the approximations of utopia and nadir points:

$$un(i) = |q_i^U - q_i^N| \tag{10}$$

increases and converges to the value characterising the actual Pareto set. We illustrate both of these concepts on an example (we use here an example defined later by Eq. (22)) – see Figs. 6 and 7.

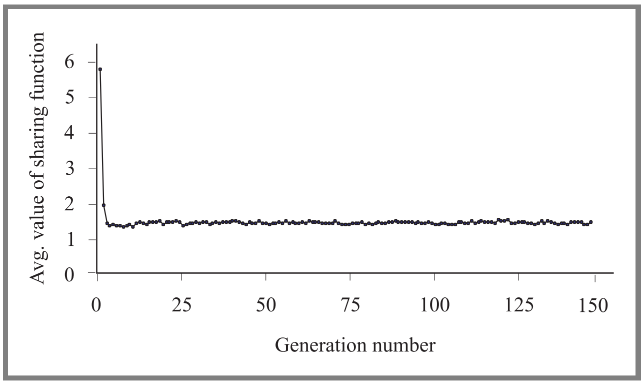


Fig. 6. Average values of sharing function in subsequent generations for Eq. (22).

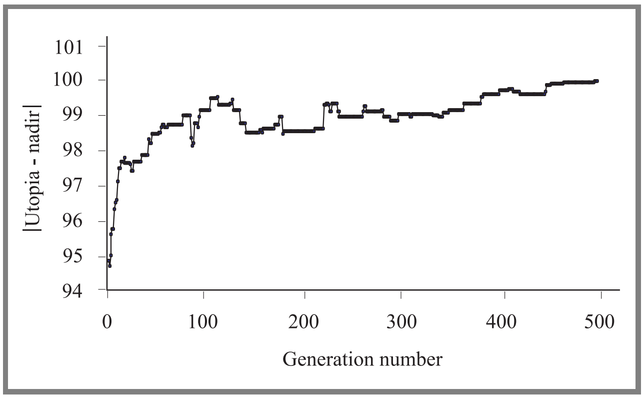
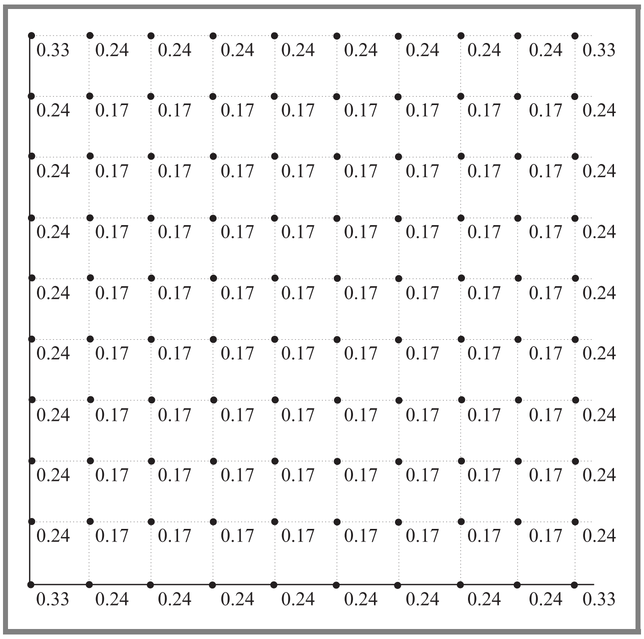


Fig. 7. Utopia-nadir distances in subsequent generations for Eq. (22).

We observe that after a small number of iterations most of the analysed measures oscillate around a constant value and thus are not particularly useful for stopping tests. An exception is the distance of utopia and nadir approximations, which converges to a constant value after a relatively large, but reasonable number of iterations. Thus, *the relative change of the distance of utopia and nadir approximations is the best stopping test for estimating a Pareto set by evolutionary algorithms.*

2.9. Succession methods

Application of niched methods results in decreasing fitness of an individual in densely represented parts of a Pareto set. However, this might lead to concentration on the boundaries of the Pareto set, demonstrated by the following example. Analysing how to choose successors in order to get a uniform representation of a Pareto set, we investigated a simple case: let the Pareto set in three-dimensional space belong to the plane $z = 0$ and be a square $x, y \in < 0; 9 >$. The simplest niched method with the niche diameter of 2 gives the following values of fitness function (100 points arranged in square table) shown in Fig. 8.



0.33	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.33
0.24	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.24
0.24	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.24
0.24	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.24
0.24	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.24
0.24	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.24
0.24	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.24
0.24	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.24
0.24	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.24
0.33	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.33

Fig. 8. Values of a fitness function for the simple case considered.

By applying the simplest succession method based on a simple ranking of the individuals to this case, we promote individuals located on the boundary of Pareto set (Fig. 9).

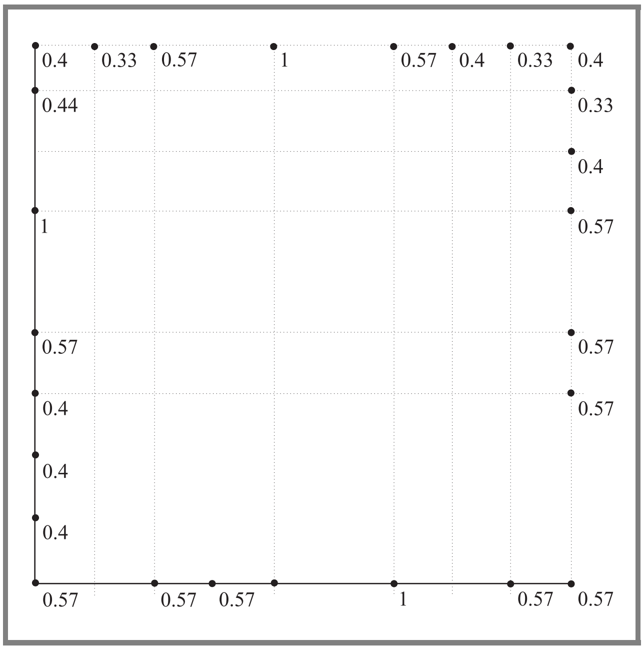


Fig. 9. Successors in the simple case with basic ranking succession rule ($\mu = 0.25$).

Table 1
Comparison of various succession methods

Parameter	Succession method				
	ranking	roulette	tournament	modified fitness	deterministic
	$\mu = 10$				
Computing time [ms]	36.59	4.84	7.40	28.56	1744
Average fitness	0.807	0.807	0.779	0.820	1
	$\mu = 25$				
Computing time [ms]	39.13	8.89	14.30	33.30	1665
Average fitness	0.520	0.579	0.548	0.601	1
	$\mu = 50$				
Computing time [ms]	46.92	19.14	29.70	43.63	1466
Average fitness	0.374	0.367	0.348	0.381	0.469

We can also imagine a deterministic (actually – non-evolutionary) succession rule in which we eliminate in a deterministic loop subsequent individuals, while increasing the fitness of its neighbours. The process is repeated until the population drops to a given number of individuals, as illustrated by Fig. 10.

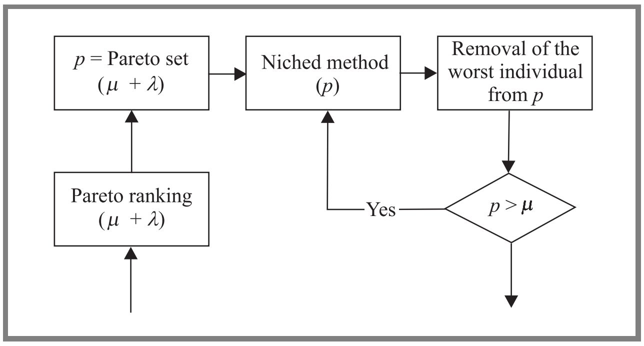

p = Pareto set ($\mu + \lambda$)
Niched method (p)
Removal of the worst individual from p
Pareto ranking ($\mu + \lambda$)
Yes
$p > \mu$

Fig. 10. Block-diagram of a deterministic succession rule.

Another succession rule is obtained by modifying definition of coefficient $m(x)$, needed to determine fitness. Instead of summing it up over all non-dominated individuals as in Eq. (9), it can be summed up only for individuals with lower index numbers on the list:

$$m(x) = \sum_{x=1}^{y-1} sh\big(d(x,y)\big)\,. \tag{11}$$

That way, the individuals considered first on the list obtain greater fitness indicators (Fig. 11).

Yet another methods of succession for evolutionary vector optimisation can be obtained by modifying roulette and tournament approaches to general evolutionary algorithms. Recall that a roulette approach determines successors (or selects individuals for cross-breeding) randomly, with probability increasing with the fitness indicator. Tournament approach determines successors by selecting randomly k individuals for a tournament and then selecting the tournament winner as the individual with highest fitness indicator (or randomly selects one of them, if there is a tie). Both approaches give similar results in our case (Fig. 12).

The above mentioned methods were compared in terms of their accuracy (defined by uniform coverage of the Pareto set, measured by average value of fitness indicator, that should be highest for a uniform coverage) and computational effort needed to solve this simple case. Table 1 gives results obtained by using a PC with 700 MHz Pentium III processor, after a large number of generations (10 000).

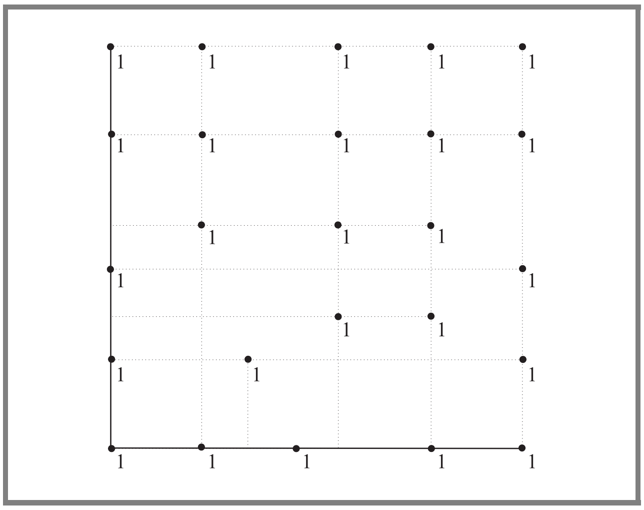


Fig. 11. Successors in the simple case with deterministic succession rule.

The most uniform representation of the Pareto set is obtained by deterministic method, though the required computing time is rather large. Among other methods, simple ranking method gives the least uniform representation – as can be expected since it favours individuals on the edge of Pareto set. For further experiments, either the roulette method (giving shortest computing time) or the determin-

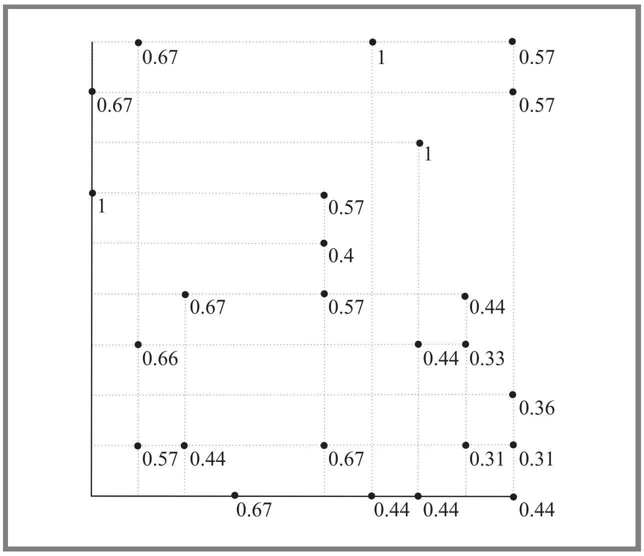


Fig. 12. Successors obtained by a roulette method in the simple case – the tournament method gives similar results.

istic method (ensuring uniform representation), were typically used. We will show later that performance of ranking succession method can be considerably improved if a more sophisticated ranking method is used.

2.10. Accuracy of representing Pareto set

When analysing more complicated Pareto sets than the simplest example presented before, it was observed that evolutionary algorithms do not converge precisely to the actual Pareto set. In a sense, this phenomenon is obvious: due to mutation necessary for evolutionary behaviour, only a few individuals come precisely to the Pareto set; most of them are oscillating just "below" the Pareto set. Even if obvious, this aspect was not sufficiently stressed and analysed in the literature. We give here results of investigating – in some cases for quite a long time with up to 30 000 generations – a simple example with known Pareto set, obtained by linear vector optimisation:

$$\max x_j,\ j = 0,\dots,i, \quad \sum_{j=0}^{i} x_j \le 1, \quad x_j \ge 0,\ j = 0,\dots,i. \tag{12}$$

We see that for the investigated example with $i = 2$, the average distance form the Pareto set oscillates about $4 \cdot 10^{-3}$ (actually, $3.76 \cdot 10^{-3}$) after only 200 generations (Fig. 13). Naturally, this value depends on the limit values for decreasing the dispersion parameter σ. This is because the oscillation of the distance from the Pareto set results from recombination and (predominantly) the mutation operation. Even if the original population were situated precisely on the Pareto set, mutation would put successors "below" this set, as illustrated by the following simple example (Fig. 14).

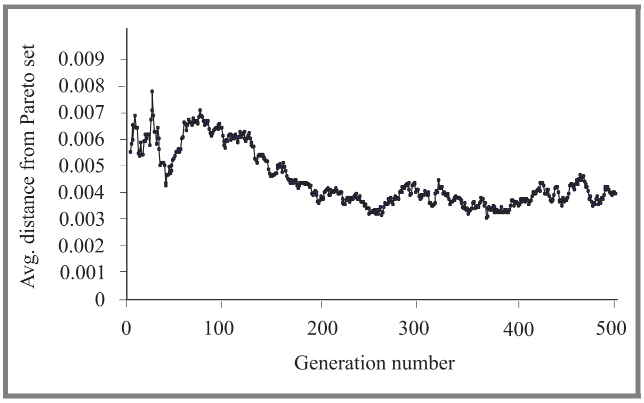


Fig. 13. Average distance from Pareto set for the example defined by Eq. (12) ($i = 2$, $\mu = 100$, $\lambda = 100$, $r = 0.01$).

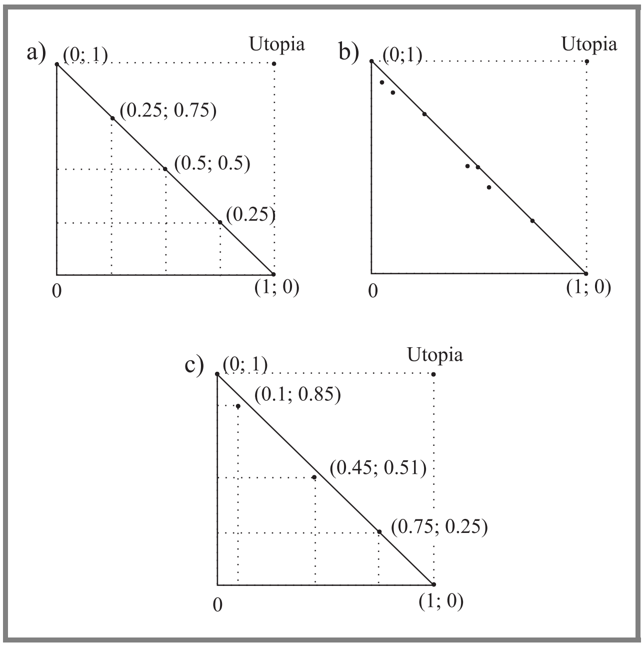


Fig. 14. Population on Pareto set (a), successor population after recombination and mutation (b) the same after succession (c).

We could of course force the algorithm to converge to the precise Pareto set, if we decided to decrease the mutation effect through decreasing dispersion parameter σ to zero. This would result, however, in losing exploratory powers of the evolutionary algorithm, considered a degeneration of the algorithm. Precise dependence of accuracy of approximating Pareto sets on the limit values of dispersion parameters requires further detailed study.

3. Use of reference points and achievement functions in evolutionary algorithms

A powerful and practical way of making vector optimisation algorithms interactive is to combine them with the concepts of reference points and to use order-preserving achievement functions [11]. We will investigate here, how to combine these concepts with evolutionary algorithms in order to either make them more interactive or to eliminate other deficiencies.

3.1. Segments of Pareto sets dominating a reference point

In interactive analysis of Pareto sets, it might be interesting to approximate a part of Pareto set "above" a given reservation point q^{res} – see the example shown in Fig. 15. We have to add constraints:

$$\begin{aligned} f(x) &\geq g_i^{res},\ i = 1, \ldots, k_1 \text{ (for maximised criteria)}, \\ f(x) &\leq g_i^{res},\ i = k_1 + 1, \ldots, k \text{ (for minimised ones)}. \end{aligned} \tag{13}$$

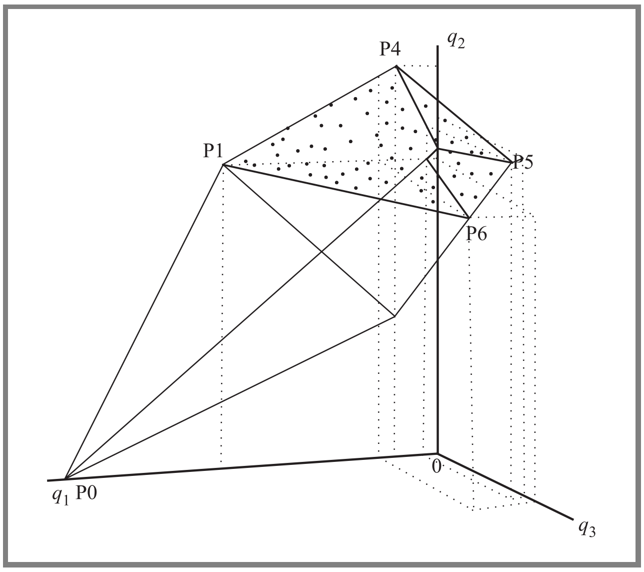


Fig. 15. A part of Pareto set above a given reservation point (0, 24, 0).

Provided that the resulting problem is feasible (the reservation point is not "above" Pareto set), specifying such additional requirement does not complicate the evolutionary algorithm. Additional constraints are simple and can be taken into account as selection conditions. We can also achieve a better approximation accuracy if the reservation point lies close to Pareto set. For the relatively simple examples of Pareto sets considered here, the necessary computational effort does not diminish, however: approximating a part of Pareto set is as expensive as approximating the entire set. On the other hand, the necessary computational effort is reasonable for simple examples. Interactive investigation by approximating first entire Pareto set, and approximating selected parts of it more precisely later is possible.

3.2. Using achievement functions for better ranking and for improving the accuracy of representing Pareto set

Ranking Pareto in evolutionary algorithms can be modified by using an order-consistent achievement function (see also [11]), e.g.:

$$\sigma(q, \overline{q}) = \min_{1 \leq i \leq m} \sigma_i(q_i, \overline{q}_i) + \varepsilon \sum_{i=1}^{m} \sigma_i(q_i, \overline{q}_i), \tag{14}$$

where $\overline{q}$ is a reference point in criteria space. The partial achievement functions can be defined for a simple case as follows:

$$\begin{aligned} \sigma_i(q_i, \overline{q}_i) &= \tfrac{q_i - \overline{q}_i}{q_i^U - q_i^N} \text{ (for maximised criteria)}, \\ \sigma_i(q_i, \overline{q}_i) &= \tfrac{\overline{q}_i - q_i}{q_i^N - q_i^U} \text{ (for minimised ones)}, \end{aligned} \tag{15}$$

where $\overline{q}^U$ and $\overline{q}^N$ are utopia and nadir point vectors or their approximations, respectively. Modification of Pareto ranking is based on the following property of the achievement function:

$$\overline{q} \in Q_0 \Rightarrow \left\{ \begin{array}{c} \max_{q \in Q_0} \sigma(q, \overline{q}) \geq 0 \\ \widehat{q} = \arg\max_{q \in Q_0} \sigma(q, \overline{q}) \geq \overline{q} \end{array} \right\}. \tag{16}$$

Thus, the value $\sigma(q, \overline{q})$ greater than 0 indicates (approximately), that point q dominates the reference point $\overline{q}$. The value 0 of the achievement functions indicates that point q is either equal or (approximately) equivalent to $\overline{q}$. Because of these properties, the Pareto rank of an individual can be determined by:

$$rank_j^{(t)} = 1 + \sum_{k=1}^{S_j} \sigma(q_k, q_j), \tag{17}$$

where q_k are individuals dominating q_j, thus $\sigma(q_k, q_J) \geq 0$, and S_j is the number of individuals dominating q_j. This way of ranking takes into account both distance of a given point from Pareto frontier and number of points dominating given point. The disadvantage is that estimation of utopia and nadir points must be available to construct the achievement function, hence this ranking method cannot be used when approximating Pareto set for the first time. It is applicable only to further, interactive analysis of selected parts of Pareto set.

Despite such drawback, the ranking method based on achievement function values has several advantages. It is more sensitive than the classical Golberg ranking method and the Fonseca and Fleming method, which can be illustrated by the simple example (Fig. 16).

Another, more practical advantage of Pareto ranking using achievement function values is that it might improve the accuracy of the entire evolutionary algorithm. We have seen

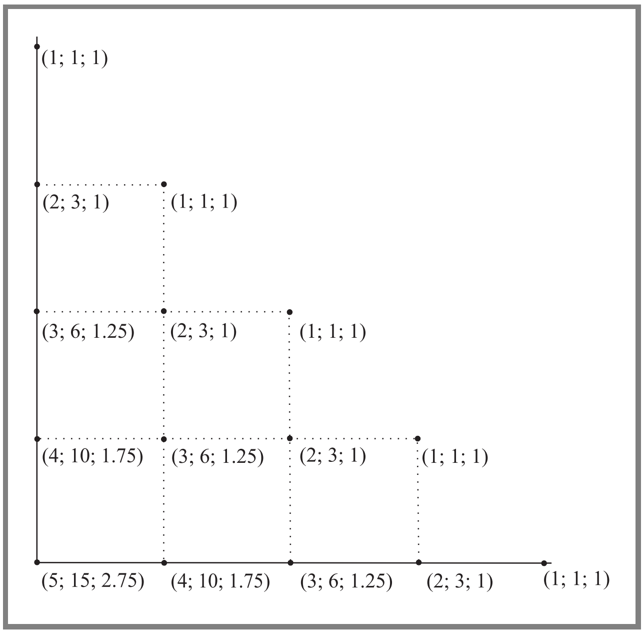


Fig. 16. Example of ranking values obtained by (1st value) Goldberg method; (2nd value) Fonseca and Flemming method; (3th value) by using achievement function values.

before that ranking methods did not behave well as succession mechanisms. A ranking method using achievement function values can perform much better: we can increase the accuracy of the entire evolutionary algorithm by increasing the value of the parameter ε, as suggested by the computation results shown in Table 2.

Table 2
Average distance from Pareto set after 30 000 generations depending on the parameter ε ($i = 2$, $\mu = 100$, $r = 0.01$)

ε	0	0.01	0.1	1	10	100
Average distance from Pareto set [$\cdot 10^{-3}$]	3.82	3.88	3.88	1.18	0.19	0.00012

We see that, using evolutionary algorithm interactively for more precise investigation of a part of Pareto set, we could actually obtain much better accuracy or use much shorter computation times for a ranking method based on achievement function values. On the other hand, very large values of ε (say, changing it from 10 to 100) mean only increasing the absolute value of achievement function, not its character that is dominated then by its linear part. This suggests that similar results would be obtained when using a slightly different form of the ranking formula:

$$rank_j^{(t)} = 1 + \beta \sum_{k=1}^{S_J} \sigma(q_k, q_j)\,, \tag{18}$$

while increasing the parameter β over its initial value 1.

Thus, *use of ranking values based on achievement functions not only increases flexibility of ranking, but also results in much better accuracy of approximating Pareto set.*

3.3. Neutral compromise points and their neighbourhoods

Given a reservation point q_{res} and an aspiration point q_{asp} in criteria value space, we can define a *relative neutral compromise point* as a point in Pareto set in criteria space being closest to the line joining points q_{res} and q_{asp} (Fig. 17).

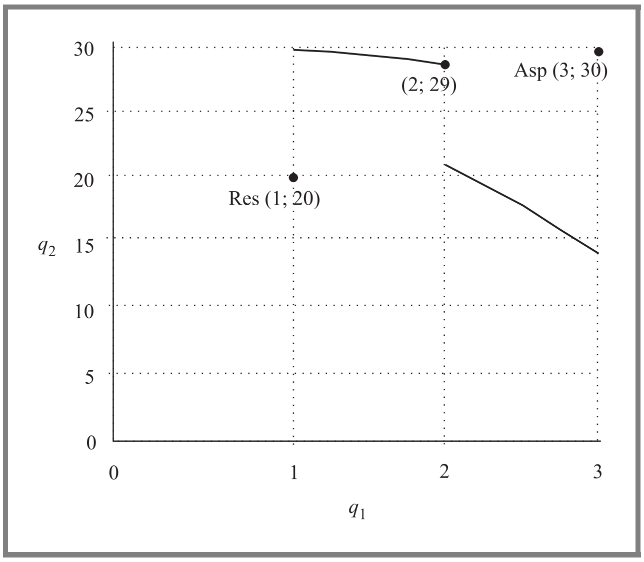


Fig. 17. Example of Pareto set with a reservation, aspiration and a relative neutral compromise points shown.

This point can be obtained by optimising the achievement function $\sigma(q, \overline{q})$ of the form (14) with partial achievement functions defined e.g. as follows:

$$\sigma_i(q_i, \overline{q}_i) = \frac{q_i - q_{asp.i}}{q_{asp.i} - q_{res.i}} \text{ (for maximised criteria)},$$

$$\sigma_i(q_i, \overline{q}_i) = \frac{q_{asp.i} - q_i}{q_{res.i} - q_{asp.i}} \text{ (for minimised ones)}. \tag{19}$$

For more sophisticated forms of partial achievement functions see e.g. [11]. In evolutionary algorithms, we can use the achievement function $\sigma(q, \overline{q})$ as a fitness measure and thus optimise it.

This results in interactive modification of evolutionary algorithms for vector optimisation: the user defines the aspiration and reservation points, the algorithms responds with the relative neutral compromise point or its approximation by a population of points (Table 3). This idea is illustrated by the following example. For the vector optimisation problem defined by Eq. (5), we define reservation and aspiration points as in Fig. 17. The line joining points q_{res} and q_{asp}

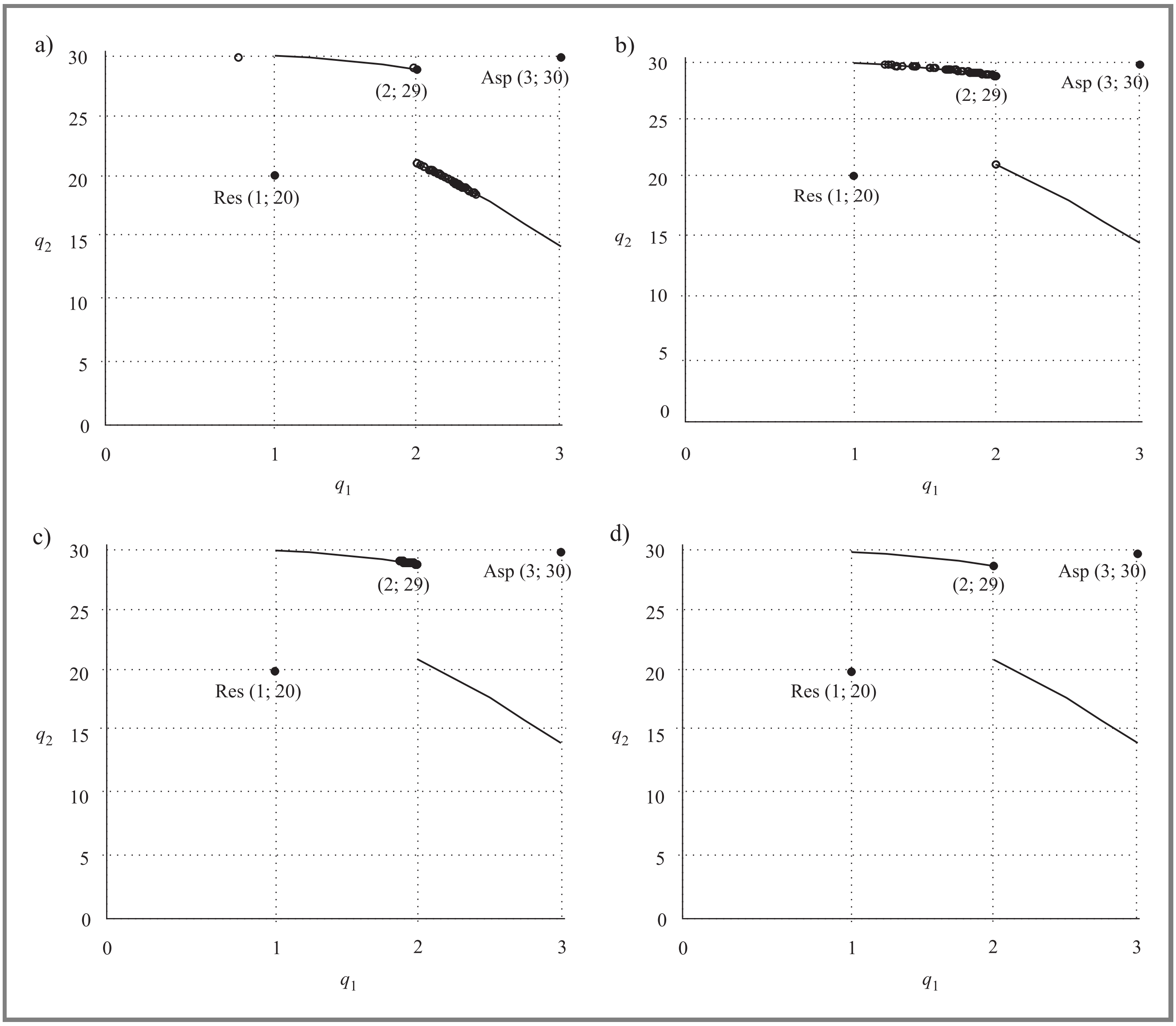


Fig. 18. The approximation of the relative neutral compromise point for the example from Fig. 17. Approximation cloud depending on generation number (a) $n = 5$; (b) $n = 10$; (c) $n = 15$; (d) $n = 30$ ($\mu = 50$, $\lambda = 50$).

Table 3
Diameter of approximation cloud depending on generation number $(\mu = 50, \lambda = 50)$

Generation number	5	10	20	40	60
$\Delta_{q_1}[\cdot 10^{-3}]$	1020	850	45.95	1.82	0.07
$\Delta_{q_2}[\cdot 10^{-3}]$	10640	9230	91.35	3.63	0.16

does not intersect Pareto set, but this makes the example more interesting. An evolutionary algorithm with achievement function used as a fitness measure produces a population approximating the relative neutral compromise point (2, 29) in the criteria space at first, and soon converges to this point (Fig. 18).

3.4. Parameterisation of representing Pareto set or its segment

The approach discussed above can be further parameterised combining a niched method with ranking based on achievement function. The niched method was originally used to provide a uniform representation of Pareto set in a global approach; here we use it to parameterise a local approach. Size of the niche can be related to e.g. the distance between aspiration and reservation points. Use of the niched method results in broadening the dispersion of a population around a neutral compromise point, as illustrated in Fig. 19.

We conclude that the evolutionary algorithms of vector optimisation, though traditionally understood as global and having non-interactive, batch character, can nevertheless be localised and used as local tools of interactive multiple criteria analysis.

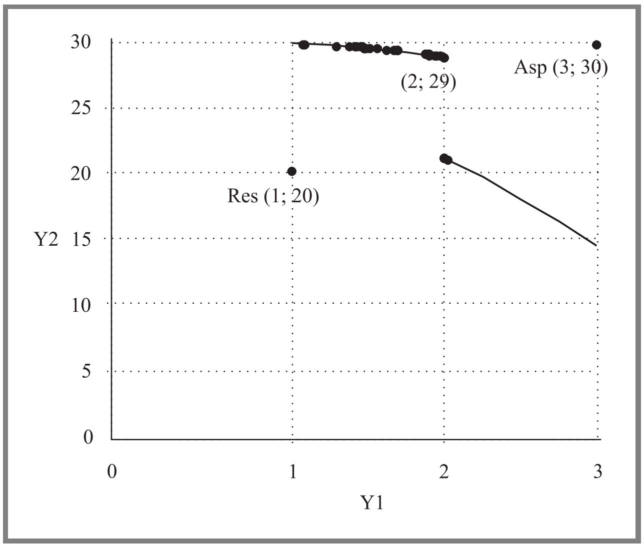


Fig. 19. Dispersion around the relative neutral compromise point for the example from Fig. 17, resulting from a niched approach with niche size equal to 10% of the range between aspiration and reservation points (population size 50, 200 generations).

4. Estimation of utopia and nadir points in evolutionary algorithms

4.1. Definitions and classical computations of utopia and nadir points

We recall that the *utopia point* q^U is defined as the "lowest" point dominating entire outcome set Q_0 (and thus entire Pareto set $\widehat{Q}_0$) in criteria value space. In other words, if some criteria are maximised and other minimised, we define:

$$\begin{aligned} q_i^U &= \max_{x\in X_0} f_i(x),\ i=1,\dots,k_1 \\ &\quad (\text{for maximised criteria}), \\ q_i^U &= \min_{x\in X_0} f_i(x),\ i=k_1+1,\dots,k \\ &\quad (\text{for minimised criteria}), \end{aligned} \tag{20}$$

where $f_i(x)$ are criteria functions, k is the number of them (while k_1 is the number of maximised criteria), X_0 is the set of admissible decisions and $Q_0 = f(X_0)$ is the outcome set of attainable criteria vectors.

The *nadir point* is defined as the "highest" point in criteria value space dominated by the entire Pareto set $\widehat{Q}_0$ – and not necessarily the entire outcome set Q_0. This difference explains the difficulty (see e.g. [9]) of precisely calculating the nadir point, since we must perform necessary computations not over entire X_0 or Q_0, but over their efficient subsets $\widehat{X}_0$ or $\widehat{Q}_0$. Thus, if some criteria are maximised and other minimised, we define:

$$\begin{aligned} q_i^N &= \min_{x\in \widehat{X}_0} f_i(x) = \min_{q\in \widehat{Q}_0} q_i,\ i=1,\dots,k_1 \\ &\quad (\text{for maximised criteria}), \\ q_i^N &= \max_{x\in \widehat{X}_0} f_i(x) = \max_{q\in \widehat{Q}_0} q_i,\ i=k_1+1,\dots,k \\ &\quad (\text{for minimised criteria}). \end{aligned} \tag{21}$$

We cannot replace $\widehat{Q}_0$ with Q_0 in the equation above, because this might lead to nadir estimation much lower than actual values. On the other hand, computation of precise value of the nadir point is very difficult when using classical methods. There are many methods that approximate nadir point components; the simplest of them is based on using only results of computations related to determining utopia components as in (20) and selecting the worst criteria values encountered during these computations:

$$\begin{aligned} q_i^U &= \max_{x\in X_0} f_i(x),\ \widehat{q}_i = \arg\max f_i(x),\ i=1,\dots,k_1 \\ &\quad (\text{for maximised criteria}), \\ q_i^U &= \min_{x\in X_0} f_i(x),\ \widehat{q}_i = \arg\min f_i(x),\ i=k_1+1,\dots,k \\ &\quad (\text{for minimised criteria}), \\ q_i^N &= \min_{1\le j\le k} \widehat{q}_i^j,\ i=1,\dots,k_1 (\text{for maximised criteria}), \\ q_i^N &= \max_{1\le j\le k} \widehat{q}_i^j,\ i=k_1+1,\dots,k\ (\text{for minimised criteria}), \end{aligned} \tag{22}$$

where q_j denotes the jth component of vector q. This method is accurate if $k=2$, for bi-criteria problems. However, in other cases it usually gives too optimistic estimations of the nadir value.

Matthias Ehrgott and Dagmar Tenfelde-Podehl [9] have proposed an algorithm computing the nadir point for three (or more) criteria by determining the Pareto sets for (each possible pair of) two criteria. For these bi-criteria Pareto sets, the values of the third missing criterion are attached, the resultant three-dimensional vectors are collected in one set, dominated results deleted, and the nadir values are directly computed from the resulting approximation of Pareto set.

4.2. Evolutionary algorithms and utopia and nadir points

Although the literature on evolutionary and genetic algorithms for vector optimisation is rather rich, it is focused more on the algorithms details than on their use for analysing Pareto set. Thus, an obvious fact was practically overlooked: since we approximate entire Pareto set by an evolutionary algorithm, the computations of utopia and nadir points should be much more easy than when using classical vector optimisation algorithms and should be actually by-products of the evolutionary algorithm applied. The questions that should be investigated are "only" how to provide for necessary accuracy of estimating these points –

especially the nadir point – while limiting the computational effort necessary for this estimation. We shall show on an example that these questions are by no means trivial.

We consider a slightly modified example from [3]:

$$\begin{aligned}
\text{maximise}: f_1(x) &= 100 - 7x_1 - 20x_2 - 9x_3 \\
\text{maximise}: f_2(x) &= 4x_1 + 5x_2 + 3x_3 \\
\text{maximise}: f_3(x) &= x_3 \\
1\tfrac{1}{2}x_1 + x_2 + 1\tfrac{3}{5}x_3 &\leq 9 \\
x_1 + 2x_2 + x_3 &\leq 10 \\
x_i \geq 0, i &= 1,2,3\,. \qquad (23)
\end{aligned}$$

The set of admissible decisions X_0 is illustrated by Fig. 20.

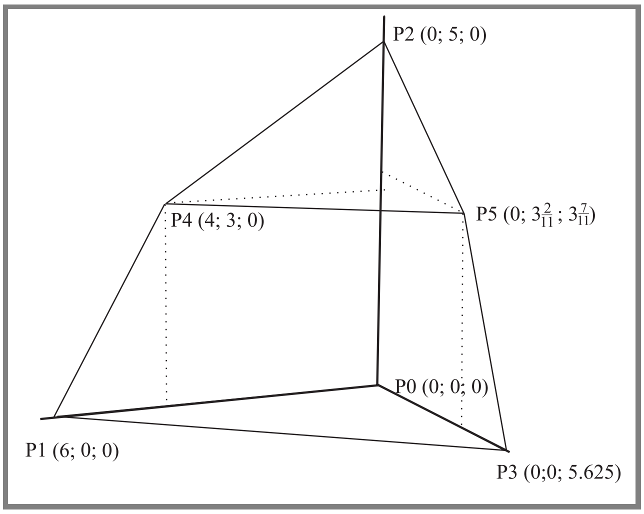


Fig. 20. Set of admissible decisions X_0 for the example defined by Eq. (23).

The set of admissible decisions X_0 is determined by its corner points:

$$\Big\{\mathrm{P0} = (0;0;0),\ \mathrm{P1} = (6;0;0),\ \mathrm{P2} = (0;5;0),\ \mathrm{P3} = \big(0;0;5\tfrac{5}{8}\big),\ \mathrm{P4} = (4;3;0),\ \mathrm{P5} = \big(0;3\tfrac{2}{11};3\tfrac{7}{11}\big)\Big\}\,.$$

Following the transformation $q = f(x)$ determined by Eq. (23), we can define also the corresponding corner points of the set of attainable criteria values Q_0 (Fig. 21). By direct examination, we can eliminate some of them as not belonging to Pareto set.

We can show in this way that the Pareto set is composed of surfaces determined by the following points in criteria space:

$$\Big\{\mathrm{P0} = (100;0;0),\ \mathrm{P1} = (58;24;0),\ \mathrm{P3} = \big(49\tfrac{3}{8};16\tfrac{7}{8};5\tfrac{5}{8}\big)\Big\} \quad \text{and}$$

$$\Big\{\mathrm{P1} = (58;24;0),\ \mathrm{P3} = \big(49\tfrac{3}{8};16\tfrac{7}{8};5\tfrac{5}{8}\big),\ \mathrm{P4} = (12;31;0),\ \mathrm{P5} = \big(3\tfrac{7}{11};26\tfrac{9}{11};3\tfrac{7}{11}\big)\Big\}\,.$$

By direct examination, we can find for these points the utopia point $q^U = \big(100;31;5\frac{5}{8}\big)$ and the nadir point $q^N = \big(3\frac{7}{11};0;0\big)$. Now we shall show the results of computing these points via three variants of evolutionary algorithms.

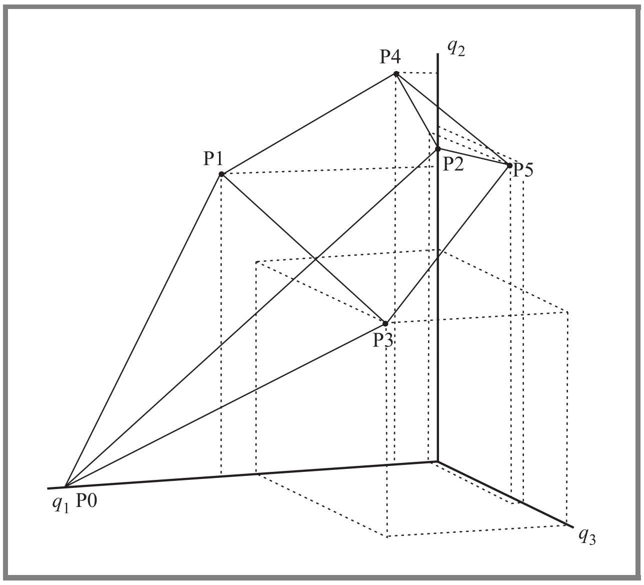


Fig. 21. The set of attainable criteria values Q_0 for the example defined by Eq. (23).

I. Evolutionary computations of utopia point with utopia based nadir approximations

The first variant uses direct determination – see Eq. (19) – of utopia point for an evolutionary approximation of a Pareto set and an indirect – see Eq. (21) approximation of the nadir point based on the data obtained in utopia point determination. An evolutionary algorithm with $(\mu,\lambda) = (200,100)$ and 200 generations gave the following results:

$$q_1^U = (100;\,0;\,0)$$

$$q_2^U = (11.9998;\,30.9999;\,0) \Rightarrow q^U = (100;\,30.9999;\,5.625)$$

$$q_3^U = (49.375;\,16.875;\,5.625)$$

with the corresponding quite inaccurate nadir approximation $q^N = (11.9998;\,0;\,0)$. By increasing the computing effort (measured below as the number of new computations of criteria values, because this, rather than organisation of the algorithm determines the computational effort) we can increase the accuracy of utopia approximations, but accuracy of nadir approximations remains inadequate, as shown in Table 4. Thus, we conclude that this method of nadir approximations is not worth using with evolutionary algorithms.

Table 4
Results of utopia and nadir point approximation by method I

The number of new computations of criteria values	Nadir point approximation	Utopia point approximation
30 000	(12.06; 0; 0)	(100.0; 30.79; 5.617)
60 000	(11.97; 0; 0)	(100.0; 30.95; 5.624)
120 000	(12.00; 0; 0)	(100.0; 31.00; 5.625)

II. Evolutionary computations of utopia point and nadir point

Since an evolutionary algorithm approximates entire Pareto set, we can also simply determine utopia and nadir points directly, according to their definitions, for the subsequent evolutionary approximations of Pareto set (Fig. 22). This simple method needs not, however, be the best, since a uniform approximation of Pareto set does not necessarily cover well the remote corners of this set, which are responsible for utopia and nadir points.

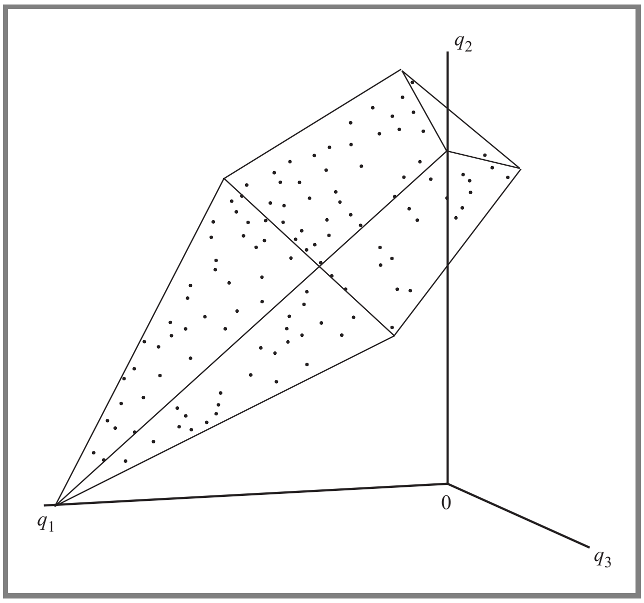


Fig. 22. Approximation of the Pareto set for the example defined by Eq. (23).

Thus, an evolutionary algorithm for vector optimisation must be modified in order to provide for a good approximation of utopia point and particularly the nadir point. It is necessary to increase fitness indicators for individuals with extreme values of criteria components.

Theoretically, such a method should give good approximations of Pareto set together with its utopia and nadir points. However, practical applications show that good approximations of the nadir point remain difficult to obtain. This is illustrated by results (Table 5) of an evolutionary algorithm with direct determination of nadir point for Pareto set approximations in subsequent iterations, with a modification of fitness indicators for individuals with extreme values of criteria vectors components. We observe that accuracy of the nadir point approximation, although much better than in method I, still remains inadequate even after very long computations.

Table 5
Results of nadir point approximation by method II

The number of new computations of criteria values	Nadir point approximation
Arbitrary starting population	
30 000	(6.78; 0; 0)
60 000	(5.90; 0; 0)
120 000	(5.91; 0; 0)
Starting population containing individuals responsible for utopia point	
30 000	(5.38; 0; 0)
60 000	(5.24; 0; 0)
120 000	(5.06; 0; 0)

III. Evolutionary approximations of Pareto sets for smaller number of criteria

The method proposed by Ehrgott and Tenfelde-Podehl [9] was not developed as an evolutionary algorithm, but can be easily combined with evolutionary approaches, involving the following steps:

– for each pair of criteria, Pareto sets are be approximated by using an evolutionary algorithm;

– for each individual in these approximations, the corresponding values of other criteria are computed;

– results obtained this way are combined and dominated points deleted, resulting in an approximation of Pareto set for the original problem;

– utopia and nadir points are computed according to their definitions for this approximation of Pareto set.

The advantage of this method over method II is that approximation of Pareto sets for bi-criteria problems in a natural way provides for more attention paid to extreme values of criteria components.

We illustrate the working of this method by showing the results of such approximations obtained by using an evo-

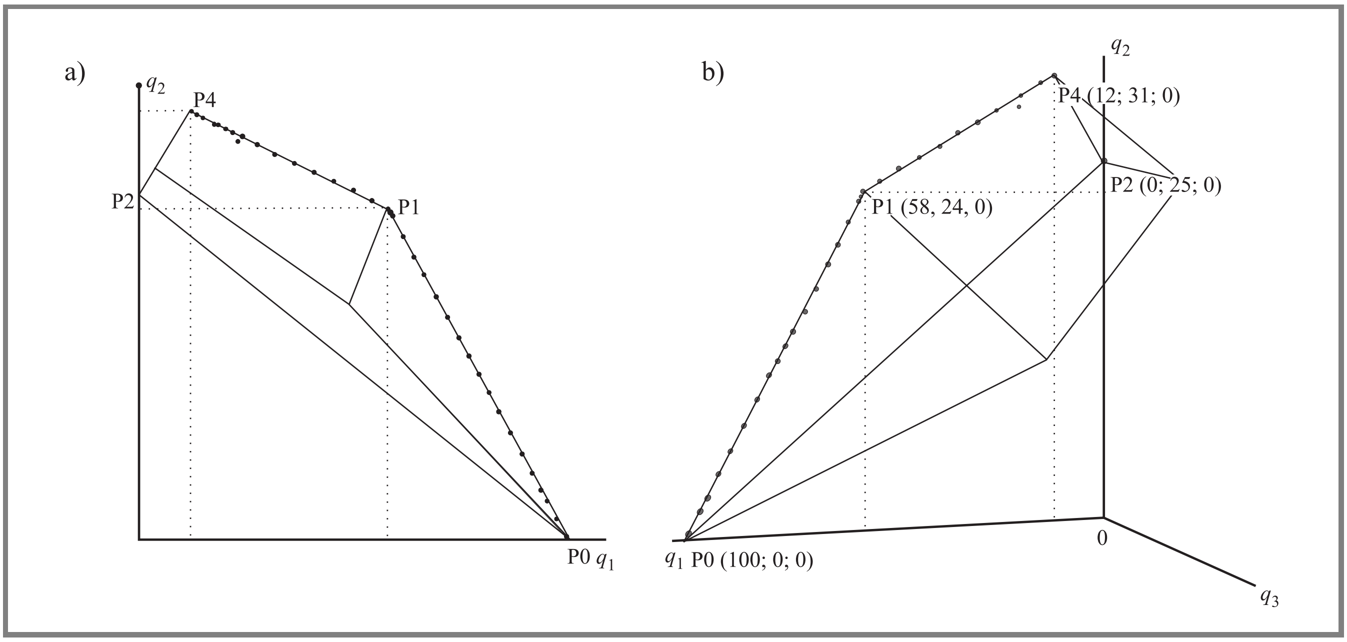


Fig. 23. Approximation of Pareto set for criteria 1 and 2 (a) with three-dimensional presentation (b).

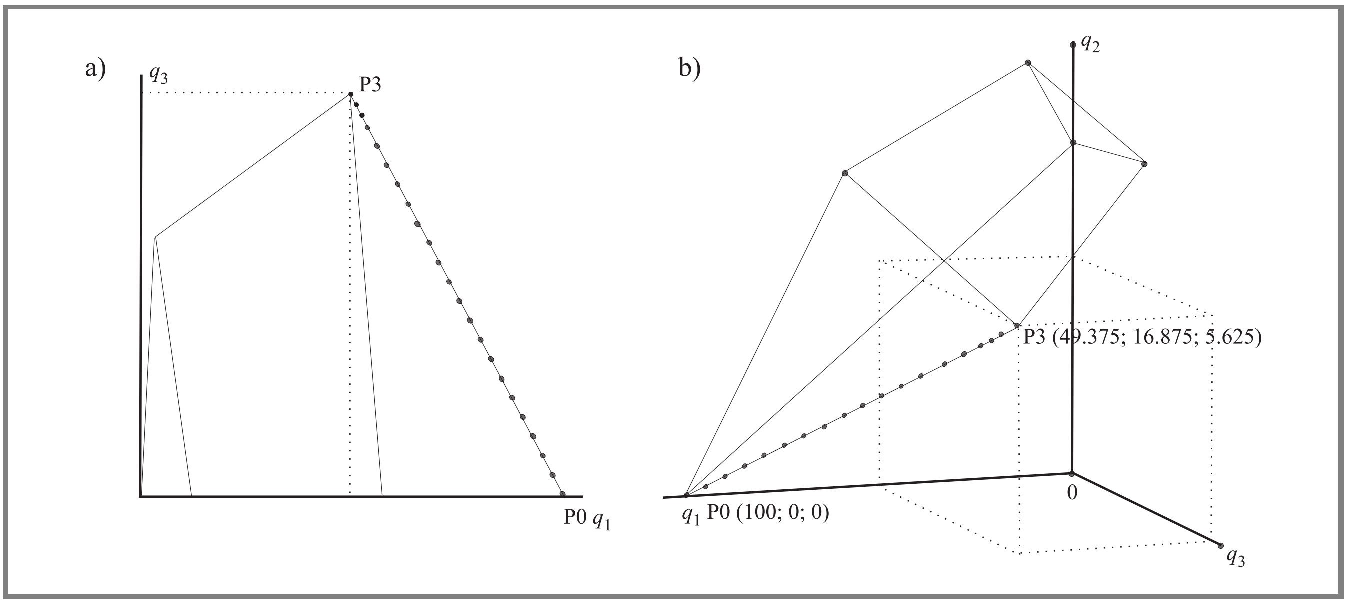


Fig. 24. Approximation of Pareto set for criteria 1 and 3 (a) with three-dimensional presentation (b).

lutionary algorithm with $(\mu,\lambda) = (200,100)$, 200 generations and alternative niche diameters 4,75; 1,55; 0,28) – see Figs. 23–25.

This way, after a large number (360 000) of computations of new criteria vectors, the following approximations were obtained: utopia point $q^U = (100; 30.999; 5.625)$ and nadir point $q^N = (4.36; 0; 0)$. We see that nadir point approximation, though much better than in other methods, still remains inadequate. Moreover, method III requires more computations (three times in this case) than methods II and I, and a fair way of comparing them is to compare nadir approximations after the same number of computations of new criteria vectors. Such a comparison is presented in Table 6.

When we compare the results of these three methods, we see that method III is most promising. The example defined by Eq. (23) might be especially difficult for nadir point approximation, hence we tried another variant

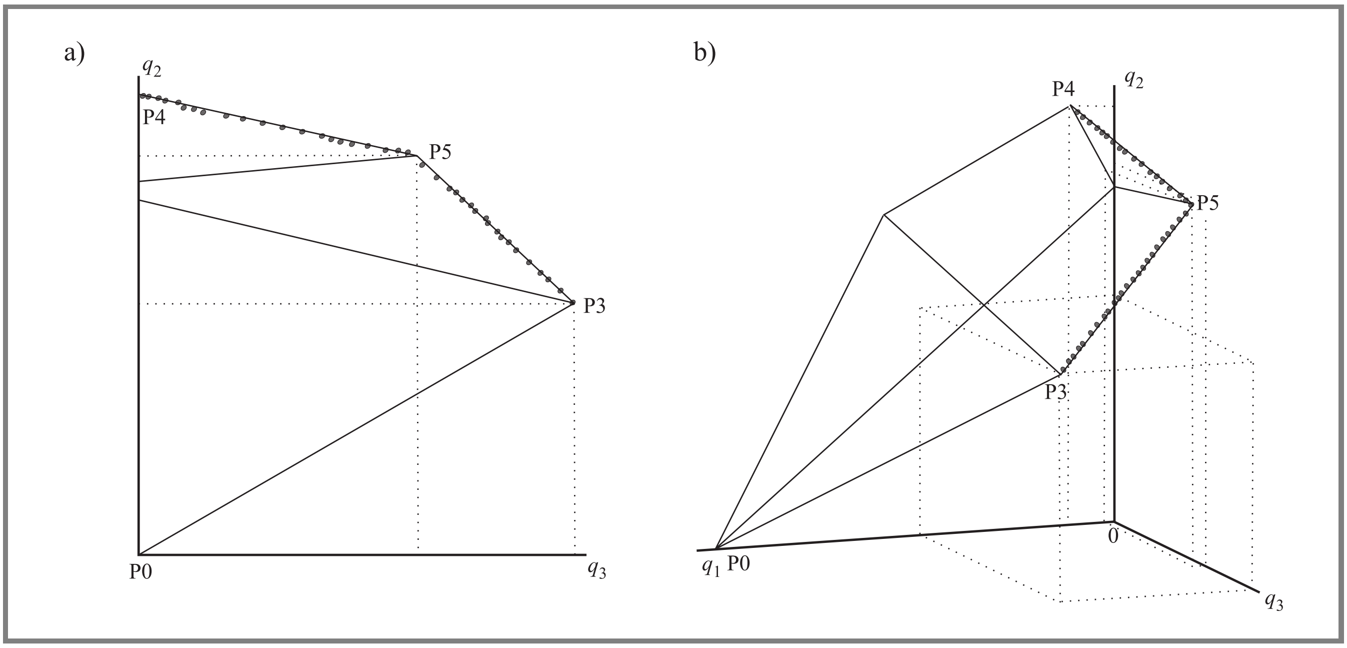


Fig. 25. Approximation of Pareto set for criteria 2 and 3 (a) with three-dimensional presentation (b).

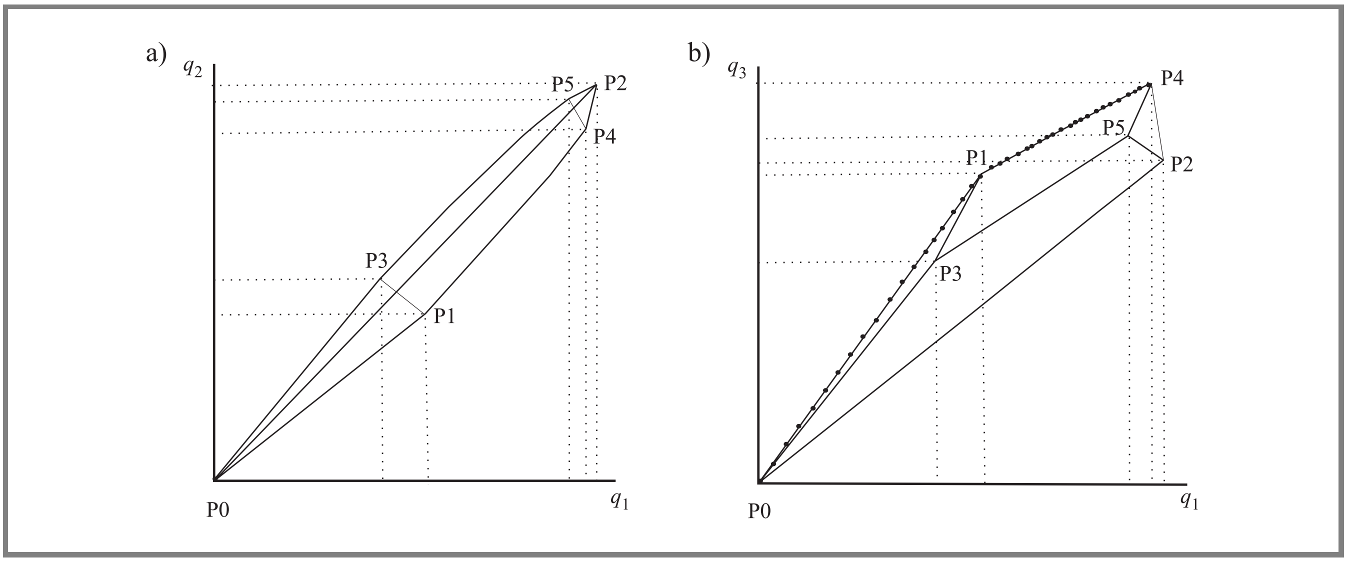


Fig. 26. Pareto sets for criteria 1 and 2 (a) or for criteria 1 and 3 (b), for example defined by Eq. (24).

Table 6
Results of nadir point approximation by method III

Number of new computations of criteria vectors	Nadir point approximation
30 000	(5.01; 0; 0)
60 000	(4.67; 0; 0)
120 000	(4.78; 0; 0)

of this example, at the same time testing the possibility of generalising method III for a larger number of criteria.

The original example from [3] is as follows:

$$\begin{aligned} \text{minimise}: f_1(x) &= 9x_1 + 19\tfrac{1}{2}x_2 + 7\tfrac{1}{2}x_3 \\ \text{minimise}: f_2(x) &= 7x_1 + 20x_2 + 9x_3 \\ \text{maximise}: f_3(x) &= 4x_1 + 5x_2 + 3x_3 \\ \text{maximise}: f_4(x) &= x_3 \\ 1\tfrac{1}{2}x_1 + x_2 + 1\tfrac{3}{5}x_3 &\le 9 \\ x_1 + 2x_2 + x_3 &\le 10 \\ x_i \ge 0, i &= 1, 2, 3. \end{aligned} \tag{24}$$

The set of admissible decisions X_0 is the same as in the example defined by Eq. (23) – see Fig. 20. However,

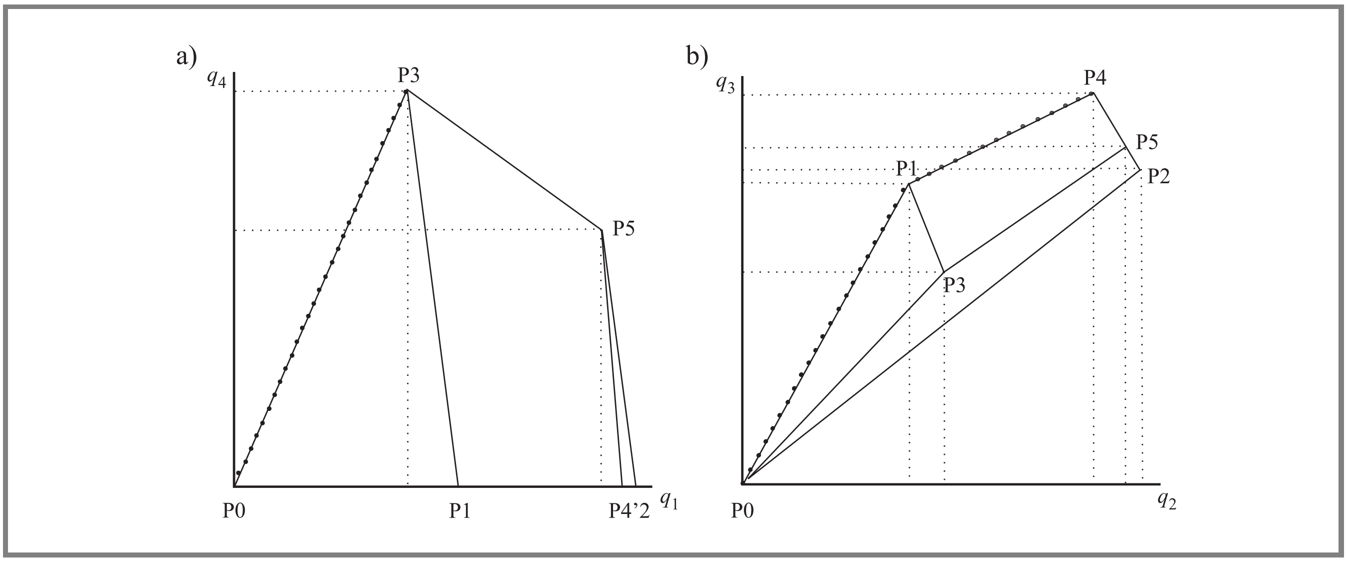


Fig. 27. Pareto sets for criteria 1 and 4 (a) or for criteria 2 and 3 (b), for example defined by Eq. (24).

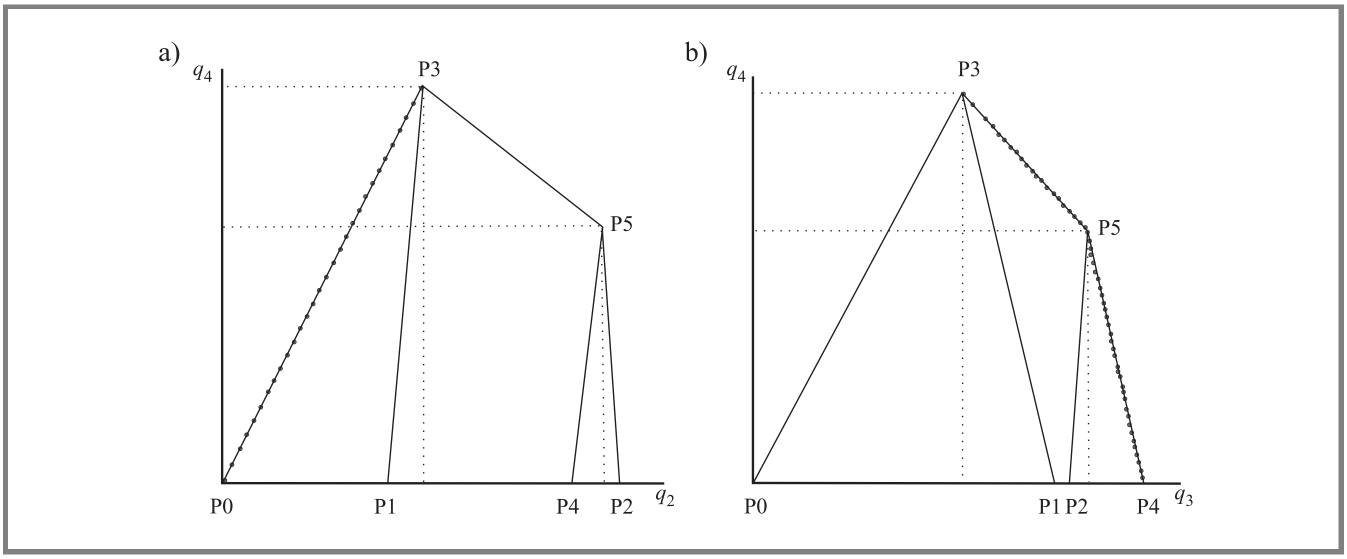


Fig. 28. Pareto sets for criteria 2 and 4 (a) or for criteria 3 and 4 (b), for example defined by Eq. (24).

utopia and particularly nadir points change with each change of criteria and they are here $q^U = \left(0; 0; 31; 5\frac{5}{8}\right)$ and $q^N = \left(94\frac{1}{2}; 96\frac{4}{11}; 0; 0\right)$. The Pareto sets for consecutive bi-criteria problems are shown in Figs. 26–28.

Utopia and nadir points obtained using evolutionary algorithm with $(\mu, \lambda) = (200, 100)$ and 200 generations, and a version of method III for four criteria: $q^U = (0; 0; 30.999; 5.625)$ and $q^N = (94.4998; 95.8747; 0; 0)$. Although the actual number of criteria value computations here increased 6 times (this is the drawback of using method III), we have obtained quite acceptable approximation of utopia and nadir points in this example.

5. Conclusions and future research

We shall point out only few conclusions, in particular those concerning future research:

- Although there is a very rich literature on evolutionary algorithms for vector optimisation, this literature focuses mostly on the tool – specific aspects of evolutionary algorithms, much less on the task – specific issues of vector optimisation, for which an evolutionary approach might be helpful.

- When concentrating on the task, evolutionary algorithms might be usefully extended – e.g. to obtain more precise approximations of selected parts of Pareto set, or better approximations of utopia and nadir points of Pareto set.

- In such extensions of evolutionary algorithms, an essential issue is to make them more interactive (e.g. first approximating entire Pareto set, then a selected part of it). For interactive extensions of evolutionary algorithms, combining them with reference point approaches and achievement function concepts might be useful.
- A particularly difficult issue (not only for evolutionary algorithms, but also in entire vector optimisation) is the determination of nadir points. Classical evolutionary approaches are not sufficient to solve this issue. Combinations of evolutionary algorithms with other approaches of vector optimisation are necessary.
- Many issues outlined in this paper should be treated as starting points only and require deeper future research. Starting from a different perspective, concentrating more on tasks than on tools, the paper serves only as identification of future research issues.

References

[1] K. Deb, "Non-linear goal programming using multi-objective genetic algorithms". Tech. Rep. no. CI-60/98, Department of Computer Science, University of Dortmund, 1998.

[2] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, Mass.: Addison-Wesley, 1989.

[3] C. M. Fonseca and P. J. Fleming, "An overview of evolutionary algorithms in multiobjective optimization", *Evol. Comput.*, vol. 3, no. 1, 1995.

[4] C. M. Fonseca and P. J Fleming, "Genetic algorithms for multiobjective optimization: formulation, discussion and generalization" in *Proc. Fifth Int. Conf. Genet. Algor.*, S. Forrest, Ed., San Mateo, USA, 1993, University of Illinois at Urbana – Champaign, Morgan Kauffman, pp. 416–423.

[5] J. Horn, *Multicriterion Decision Making*. Handbook of Evolutionary Computation, IOP & Oxford University Press, 1997.

[6] J. Horn and N. Nafpliotis, "Multiobjective optimization using the niched Pareto genetic algorithm" in *Proc. First IEEE Conf. Evol. Comput., IEEE World Congr. Comput. Intell.*, vol. 1, 1994.

[7] P. Korhonen, "Multiple objective programming support". IR-98-010, International Institute for Applied Systems Analysis, Laxenburg, 1998.

[8] Z. Kowalczuk, T. Białaszewski, and P. Suchomski, "Genetic poliop-timisation in Pareto sense with ranking and niched methods" in *Proc. III Nat. Conf. Evol. Algor. Glob. Optim.*, Warsaw, Poland, 1999.

[9] M. Ehrgott and D. Tenfelde-Podehl, "Nadir values: computation and use in compromise programming". Universitat Kaiserslautern Fachbereich Mathematik, 2000.

[10] D. A. Veldhuizen and G. B. Lamont, "Multiobjective evolutionary algorithm research: a history and analysis". Tech. Rep. TR-98-03, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, 1998.

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

[12] E. Zitzle, "Evolutionary algorithms for multiobjective optimization: methods and applications". Swiss Federal Institute of Technology (ETH), Zurich, 1999.

Marcin Szczepański was born in August 10, 1976. He has been graduated as Master of Engineering at Warsaw University of Technology at the Institute of Control and Computation Engineering at Optimization and Decision Support Division, in 2001. During his studies, he was especially interested in evolutionary algorithms and decision support.

e-mail: Marcin.Szczepanski@damovo.com
Damovo Poland
Jana Olbrachta st 94
01-102 Warsaw, Poland

Andrzej Piotr Wierzbicki born June 29, 1937 in Warsaw. Graduated as Master of Engineering at the Faculty of Electronics, Warsaw University of Technology (WUT), in 1960. Ph.D. degree at this University in 1964, for a thesis on nonlinear feedback systems; D.Sc. degree in 1968, for a thesis on optimisation of dynamic systems. In 1971–75 a Deputy Director of the Institute of Automatic Control, later a Deputy Dean of the Faculty of Electronics, WUT. In 1975–78 the Dean of the Faculty of Electronics WUT. Since 1978 worked with the International Institute for Applied Systems Analysis in Laxenburg n. Vienna, Austria; 1979–84 as the chairman of the theoretical branch, Systems and Decision Sciences Program, of this Institute. From 1985 back in the Institute of Automatic Control, WUT, as a Professor of optimisation and decision theory. In 1986–91 scientific secretary, currently member of presidium of the Committee of Future Studies "Poland 2000" (in 1990 renamed "Poland in XXI Century") of P.Ac.Sc. In 1991 elected a member of the State Committee for Scientific Research of Republic of Poland and the chairman of its Commission of Applied Research; contributed to basic reforms of Polish scientific system in 1991–94. Deputy chairman of the Council of Polish Foundation for Science in 1991–94, chairman of scientific councils of NASK (National Scientific and Academic Computer Network in Poland) and PIAP (the Industrial Institute of Measurements and Control). In 1991–96 the editor in chief of the quarterly "Archives of Control Sciences" of P.Ac.SC. In 1992 received (as first European researcher) the George Cantor Award of the International Society of Multiple Criteria Decision Making for his contributions to the theory of multiple criteria optimisation and decision support. Since 1996 the General Director of the National Institute of Telecommunications in Poland. In 2000 nominated as a member of the ISTAG (Information Society Technology Advisory Group) at European Commission. Since 2001 chairman of Advisory Group on Scientific International Cooparation of the State Committee for Scientific Research of Poland. Beside lecturing for over 40 years and promoting more than 80 master's theses at WUT (Warsaw University of Technology), he also lectured at the Department of Mathematics, Information Science and Mechanical Engineering of Warsaw University and in doctoral studies: at WUT, the Academy of Mining and Metallurgy, at the University of Minnesota, at the Illinois Technical University, Hagen University, and at the University of Kyoto. He also promoted 18 completed doctoral dissertations. Author of over 180 publications, including 11 books (4 monographs, 7 – editorship or co-authorship of international joint publications, over 50 articles in scientific journals (over 30 in international), 80 papers at conferences (68 at international, including over 48 published as chapters in books). He also authored 3 patents granted and applied industrially. Current interests include parallelisation of optimisation algorithms using multiple criteria approaches, diverse aspects of negotiation and decision support, including e.g. applications of fuzzy set theory for describing uncertainty in decision support models, multimedia software in computer networks, telematics in education, diverse issues of information society and civilisation. Languages: English, German, Russian (each fluent, beside native Polish). Member of IEEE, ISMCDM (International Society of Multiple Criteria Decision Making), SEP (Polish Society of Electrical Engineers), PTM (Polish Mathematical Society), PSKR (Polish Association for the Club of Rome).

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

Application of Recurrent Neural Networks for User Verification based on Keystroke Dynamics

Paweł Kobojek and Khalid Saeed

Faculty of Mathematics and Information Sciences, Warsaw University of Technology, Warsaw, Poland

http://doi.org/10.26636/jtit.2016.3.750

Abstract—Keystroke dynamics is one of the biometrics techniques that can be used for the verification of a human being. This work briefly introduces the history of biometrics and the state of the art in keystroke dynamics. Moreover, it presents an algorithm for human verification based on these data. In order to achieve that, authors' training and test sets were prepared and a reference dataset was used. The described algorithm is a classifier based on recurrent neural networks (LSTM and GRU). High accuracy without false positive errors as well as high scalability in terms of user count were chosen as goals. Some attempts were made to mitigate natural problems of the algorithm (e.g. generating artificial data). Experiments were performed with different network architectures. Authors assumed that keystroke dynamics data have sequence nature, which influenced their choice of classifier. They have achieved satisfying results, especially when it comes to false positive free setting.

Keywords—biometrics, GRU networks, keystroke dynamics, LSTM networks, recurrent neural networks, user verification.

1. Introduction

The problem of verification is most often solved by assigning some kind of a password, which should only be known to a given user and consists of finite sequence of characters. When the user provides this password, a party responsible for confirmation of an identity may tell whether the user is whom he claims, he is (based on an assumption that only the real user knows the password). However, such approach is not free from drawbacks. For example, there has to be some kind of a mechanism to handle a situation in which the user forgets his or her password. Moreover, traditional passwords can be broken with brute force method if only attacking person has enough time and computation power (and, of course, there are no other protections against it). Also, if the user stores the password somewhere else than in his or her own brain it has to be somehow secured as well. Alternative to this method is using a biometrics-based security.

Keystroke dynamics is a field within behavioral biometrics, which concerns humans typing patterns on a keyboard. It turns out that the way a user writes on a keyboard is one of his or her unique characteristics. Back in 1980s, the first work was done in order to develop an algorithm which could identify a user based on this trait [1]. Many experiments were performed which have shown it is a good indicator of identity [1]–[4].

In order to describe mathematically a typing pattern we first need to acquire specific data from the user. This data consists of a timestamp of the moment of pressing and/or leaving the button. Next, different measurements out of this can be computed, e.g. [5]:

- dwell time – time between moment of pressing and moment of leaving the button,
- flight time – time between pressing (or leaving) subsequent keys.

A user who types the text can make mistakes, which means that vectors representing different samples may differ in length.

In the next step, data is passed to some kind of a model, which task is to answer the question whether examined user is the one who he claims to be. This model may be anomaly detection system or classifier. Popular approach is to use algorithms based on database of samples. In this case, new sample is compared with those already in database in order to find similarity.

The algorithm consists of two parts: way of acquiring data along with features extraction and a model, which verifies/identifies the sample. Designing new solutions may affect both of these modules.

The accuracy may be influenced even by a way of acquiring data from a user as well as it nature. In the most basic approach, sample describing the user simply consists of timestamps mentioned earlier (from which dwell/flight time is computed). Besides this, it is sometimes useful to measure other values, e.g. eye motion. Humans often either follow their fingers with their eyes or look straight at the monitor. Taking this behavior into consideration may enhance classification accuracy. Mobile devices are supplied with additional sensors like gyroscope or accelerometer. Information from these sensors was proven useful [6]–[8]. [9] shows thoughts about authorization specific for mobile devices with focus on using biometrics techniques including keystroke dynamics. In addition to all these information, there is also meaningful signal in errors made by the user along with the way they correct them (e.g. by using *delete* vs. *backspace*).

For some applications, using only keystroke dynamics may not be accurate enough because of strict regulations. Even in such situation, it can be used as a valuable support for traditional data. Such approaches increase security and combined accuracy may be high enough to be used even in healthcare [10]. Such methods may be extended by even more biometrics techniques, e.g. face recognition [11].

As it was stated before, keystroke dynamics data may also find applications when it comes to user identification. In this paper this problem is reduced to of multiclass classification, i.e. each user is represented by a class. In this case, we usually have limited user count. This work focuses on verification because in a problem it tries to solve the user is already identified by his or her email address. Identification problem was broadly described in [12] along with proposed algorithm.

1.1. State of the Art Algorithms

Looking at the problem as an anomaly detection problem, statistical methods based on some kind of distance are often used. In standard approach, having some data set (let us treat every sample as a vector) we find its center, which is also a vector. This is a training phase. In testing phase on the other hand, the task is to tell that whether given vector (test sample) is an anomaly or not. In order to answer this question distance between center and test sample has to be computed. The distance may be classic Euclidean distance as well as something more sophisticated i.e. Manhattan distance. This simple algorithm can be further modified e.g. by applying distance norming. In *Filtered Manhattan* algorithm, after finding the center at first all samples, which are too far from it, are removed and then new center point is computed. Similar group of algorithms are those based on k-nearest neighbors idea. In this case, instead of designating a center point and comparing input with it, we find k (in particular, $k = 1$) closest, in terms of defined distance, samples. In this case, usually an *anomaly score* as distance from their center is computed. Another interesting approach is using fuzzy sets. In such sets each object belongs (to some degree) to ranges. The anomaly score is then computed as an average lack of belonging. The approach, which is most similar to the idea presented in this work, is probably one-class SVM. However, such a classifier is trained only on positive class (in opposition to this work's algorithm).

More thorough description of those algorithms (with references to exhaustive descriptions) can be found in [13]. Results of [13] are benchmark for results achieved by the algorithm described in this paper.

When it comes to multiclass classification with keystroke dynamics, the multiple classifiers were tested: HMM, SVM, k-nearest neighbors, and neural networks [14]. Presented algorithm does not solve multiclass classification problem. Nevertheless, with slight modification it could be trained for such problems as well. On the other hand, algorithms mentioned in this paragraph could be used as binary classifiers and replace the proposed one.

1.2. Algorithm Evaluation Methods

An important thing to consider is the evaluation of proposed algorithms. Let us introduce the following terms:

- True Positive Rate (TPR) or hit-rate $\frac{TP}{TP+FN}$,
- False Positive Rate (FPR) $\frac{FP}{FP+TN}$, informs about the probability of accepting an impostor,

where: TP – number of true positives, TN – number of true negatives, FP – number of false positives, FN – number of false negatives.

Besides standard accuracy or error measure, when it comes to keystroke dynamics (and also in other fields of biometrics) two more measures are often used to evaluate algorithms:

- Equal Error Rate (EER) – value for a threshold in which FPR and miss rate1 – TPR are equal,
- Zero-miss rate – FPR value for which TPR = 1 (no false positive errors).

Both these values can be easily read from ROC curve. Figure 1 shows sample ROC curve along with mentioned points marked on it. The values can be read from x axis of these points.

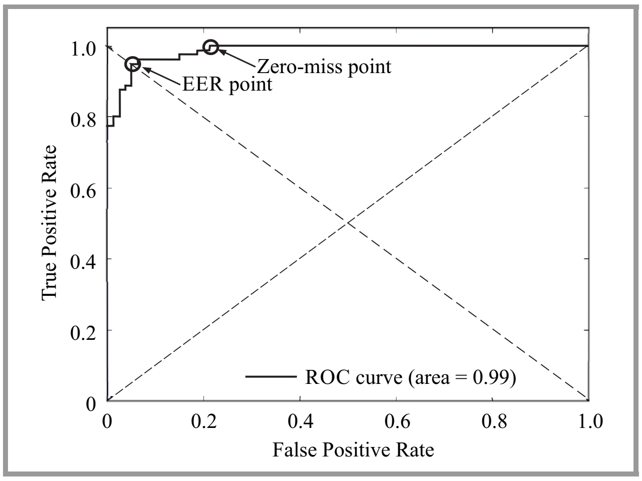


Fig. 1. Sample ROC curve (from the results of this research) with EER and zero-miss points marked.

1.3. Problems with Algorithms and Authors' Proposition

Some of the mentioned algorithms are based on assumption that we have some database of patterns for a user. In the moment when a new sample appears, we need to go through the whole database and find similarities (k-nearest neighbors is an example of this approach). Note that the keystroke dynamics is a behavioral feature, thus it changes with time more than physiological traits. When it comes to keystroke dynamics problem, maintaining a static database for a given user may end up with gradually decreasing accuracy. One of the solution, which comes to mind, is adding new samples. Unfortunately, the side effect of this approach is the growing need of memory of such a system. This disadvantage, combined with a big number of users may result in memory consumption as the main drawback. When it comes to multiclass classification, there is a need to add new class with each new user.

Another problem of these algorithms is the fact, that they treat input as a vector. Intuitively it seems that numbers representing the sample from a human are more like a sequence, i.e. there is some relation between them. As usual in machine learning problems – it is hidden and unknown. Because mentioned algorithms do not treat data as a sequence, some information natural to them must be encoded artificially. As an example, let us say that the user has mistaken and then corrected the errors. In case of sequence, such information is directly encoded in its length, because errors and corrections require more keystrokes.

Algorithms, which are described and compared in [13], reach relatively low accuracy when it comes to the situation where the threshold was set to avoid false positive errors (zero-miss). The best presented algorithm in this setting (k-nearest neighbors with Mahalanobis distance) achieved 0.468 zero-miss rate. In a problem of access control, this would mean the situation in which probability of rejecting a genuine user is close to 0.5.

A problem for which the presented algorithm could be useful is creating a centralized system serving authentication based on a way the user types his or her email address. Thus, the email along with the biological characteristic of a human being would be the only ID in the Internet and necessity of using multiple long passwords would disappear. Services of such a system could be used by external services which could supply it with sufficient information, i.e. keystroke dynamics data plus email address and in return get the information whether the user is verified or not.

Having all this in mind, the presented algorithm is a subject of more constraints. First, it should authenticate potentially everyone in the Internet. Given the enormous number of Internet users (almost 3 billion in 2014 [15]), infinite scalability in terms of user count have to be assumed, which cannot be constrained by the algorithm.

Such a system could potentially be used to grant the access to many services using the same one identification way. The most important feature of such a system is definitely securing resources from unauthorized people. Ignoring this problem would result not only in not solving the problem in which a user has one password to many accounts and someone has accessed it, but could even make it worse. It seems better to reject a genuine user from time to time than to accept the attacking one. The designed algorithm should thus focus on minimizing (ideally eliminating) false positive errors, which means accepting wrong user. Eliminating such errors should be a goal even at the cost of big drop in accuracy.

The proposed algorithm was designed with all that features in mind. Thus, the most important goals are scalability in terms of user count and high accuracy without false positive errors.

2. The Algorithm

2.1. General Idea and Motivation

The standard approach in a keystroke dynamics based verification is using anomaly detectors. Presented approach is different. It uses a binary classifier (recurrent neural networks). Data from the genuine user are positive and from the other people – negative. A big disadvantage, which may appear in readers mind, is the requirement of negative data for training phase. Some thoughts about it along with ways of mitigating this issue were described in latter sections.

In order to choose good classifier it is worth to consider the nature of a problem. First property of the examined data is they do not seem to be a vector describing some physical phenomenon or object (like images, where every element of a vector contains information about specific pixel). As it was stated before, it is assumed that data has a sequence nature. It is worth noting though, that there are no strict proofs of that. However, for some people it seems intuitive, because of (among other reasons) keyboard arrangement. This assumption has influenced the choice of a classifier.

2.2. Recurrent Neural Networks

Due to assumed sequence nature of input data authors have decided to use recurrent neural networks. These networks naturally operate on sequences. Plain recurrent neural networks are very simple (compared to other neural network architectures) models. They differ from feed forward networks in the way of processing input – here it is processed in a step-by-step manner. At step t the network receives x_t as input and having knowledge about state from last step h_{t-1} it computes its output according to the formulas (1) and (2). W_{hh}, W_{xh} and W_{hy} are matrices of network parameters.

$$h_t = tanh(W_{hh} \cdot h_{t-1} + W_{xh} \cdot x_t) \tag{1}$$

$$y = W_{hy} \cdot h_t \tag{2}$$

Unfortunately, in its simplest form, recurrent neural networks are very hard to train due to the problem known as exploding or vanishing gradient [16], [17]. However, there are modified architectures of recurrent neural networks, which solve this problem.

2.3. LSTM

Long Short-Term Memory (LSTM) networks along with training algorithm were proposed in 1997 in the paper [16] in order to solve mentioned problem of vanishing gradient. They are successfully used in many fields, especially when data is sequential, e.g. natural language processing, speech recognition, machine translation, image captioning [18] or even bioinformatics [19].

Core idea behind LSTM network is inclusion of a so-called cell state. It is a vector, which simply stores information, thus it is a kind of memory. This vector is passed through computation steps – modified or not. At each step the network can write or remove some information to/from the

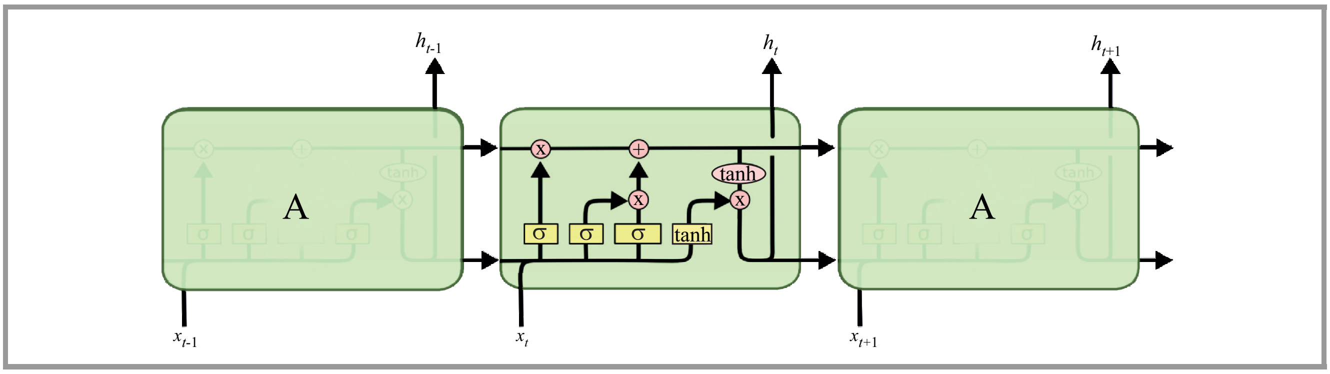


Fig. 2. LSTM computation in time.

memory. It is done using so-called gates. At each computation step the input and the cell state from previous step first go to the forget gate. The way it operates is very simple – it is a plain sigmoidal layer known from neural networks. To be more precise, its output is computed with formula:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

where:

- σ – sigmoid function $\left(\sigma(x) = \frac{1}{1+e^{-x}}\right)$,
- W_f, b_f - weight matrix and bias of forget gate,
- h_{t-1} - output from previous step,
- x_t - input in current step.

The vector resulting from this gate tells how much information should be forgotten and how much should be remembered. Degree of this "forgetting" is controlled by the value of sigmoid function which is in range $[0,1]$: 0 means forget everything, 1 means remember everything.

Next is the input gate. Input data along with the output from the previous step are used twice in this gate: in the sigmoid layer (similar to forget gate) and in another layer with hyperbolic tangent as activation. Results of these layers are going to be used in order to create a vector, which is then added to the cell state. This step is described by formulas:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)\,, \tag{4}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)\,. \tag{5}$$

After computing these 3 values they can be used to update the cell state. This computation is shown by equation:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t\,, \tag{6}$$

where: C_t is the cell state at the moment t. f_t is computed from Eq. (3), i_t from Eq. (4), $\tilde{C}_t$ from Eq. (5). Current memory value is first multiplied by an output from forget gate which potentially erases some information and then new information is added. The final LSTM result from the current step is computed not only from the input and the state but also from the cell state. This is described by following formulas:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)\,, \tag{7}$$

$$h_t = o_t \cdot tanh(C_t)\,. \tag{8}$$

Final result of this computation is some real value from range $[0,1]$ if the network is last layer of the model. If it is inner layer then it returns the whole sequence containing all values of h computed in "for" loop. If network is the last layer then its output is compared with the threshold, which determines the final class.

Figure 2 shows how the mentioned computations are performed in time [20].

2.4. GRU

Gated Recurrent Units (GRUs) were introduced in 2014 [21]. They are a similar to LSTM. What is different is that instead of two gates – forget and input gate, GRUs have only one – update gate. Another difference and simplification lies in fact, that GRUs do not have separate memory (cell state). The memory is associated with the state from previous step. Network computation is described by formulas:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])\,, \tag{9}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])\,, \tag{10}$$

$$\tilde{h}_t = tanh(W \cdot [r_t \cdot h_{t-1}, x_t])\,, \tag{11}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t\,. \tag{12}$$

Merged gates output is z_t vector. It is used for both forgetting and remembering.

As in LSTM, output is a real number from range $[0,1]$ if network is last layer of a model. Otherwise, it returns sequence consisting of every h values computed in "for" loop. The final model output is then compared with the threshold in order to determine the class.

Figure 3 presents diagram with GRU cell [20].

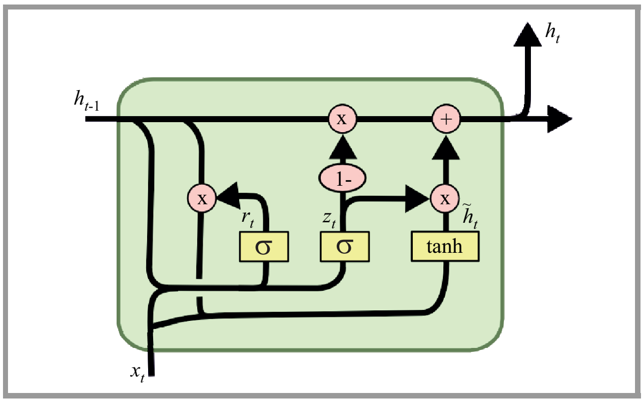


Fig. 3. GRU diagram.

2.5. Training

LSTM and GRU were trained by standard Back Propagation Through Time (BPTT) algorithm [22]. It is used to compute cost function derivative required for optimization algorithm, i.e. Adam optimizer in this case [23]. Networks were trained for 100 epochs, and cost is described by function:

$$C = -\frac{1}{n}\sum_x \left[y \ln a + (1-y)\ln(1-a) \right], \qquad (13)$$

where: n – samples count, x – single element, y – expected label for x, a – actual label for x.

Figure 4 shows an example loss over iterations graph.

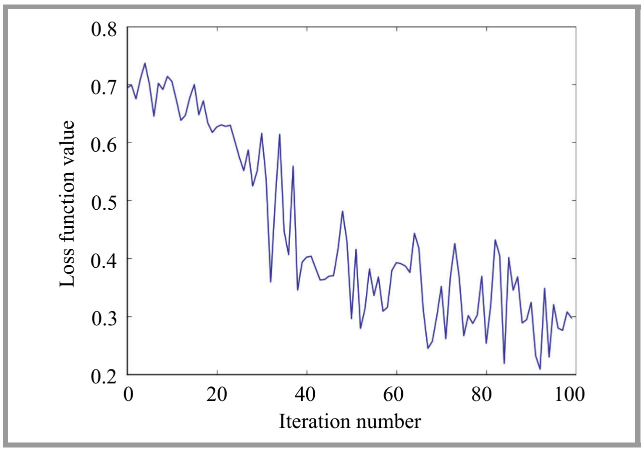


Fig. 4. An example graph showing loss value over iterations.

2.6. Small Training Set and Lack of Negative Data

As it was mentioned earlier, serious drawback of the presented algorithm is a need for negative data during the training phase. In real life applications, requiring user to type his address several times is already an inconvenience for him. Forcing other people to type, this address would be even harder. One of possible solutions could be using neural network as one-class classifier. In such a case, the model would be trained only on positive data, which is easier to acquire. Such methods are successfully applied for SVM classifiers [13] to solve verification problem. Unfortunately, the conducted experiments had not shown any good results with recurrent neural networks.

Another approach, which was tested, is generation of artificial negative data. From every positive sample authors got a negative by adding a random (with normal distribution, several values of standard deviation were tested) noise to it. The classifier was trained on positive and artificially created negative data and then evaluated only on real data.

Another problem is the size of a training set. Deep neural networks are models in which there are enormous number of parameters, which has to be adjusted during the training, thus they require big amount of data. Training for much iteration with small dataset tends to overfit. Unfortunately, in this case asking the user to type an address several hundred times would clearly be impractical. In this work, authors' dataset has only over a dozen samples for each user. Because of that, the authors had to apply regularization techniques in order to avoid overfitting.

It is worth mention that the challenges may not be a problem in some real applications. Large email services, e.g. Google Gmail, have (or might have) access to a large amount of data about address typing. Positive data could come from successful login or typing own email. Negative data on the other hand could be extracted from other people who type email of a given user in order to send him a message. In such a case, there would be no need to generate artificial data.

2.7. False Positive Errors Minimalization

As it was mentioned earlier, one of the challenges for the designed algorithm is the minimalization of false positive errors.

A standard approach is to select the acceptance thresholds. By increasing its value, number of samples classified as positive should decrease. Hopefully, first to drop will be samples classified as positive with low likelihood, which are potentially false positive errors. The idea to eliminate such errors is then to increase the threshold until every false positive error is gone on training set.

Unfortunately, networks tend to classify with a very high likelihood. Thus selecting the threshold, which eliminates unwanted errors will definitely decrease total accuracy, because it has to be pretty high, so many genuine samples are rejected.

The question arouses – why LSTM and GRU models tend to return high numbers even if they mistake? These models are very sophisticated and are based on strong type of neural networks (so-called deep neural networks) and are used for high dimensional problems like image recognition [24]. Compared to such problems, the presented task has much smaller dimensionality, which is probably a reason why network overfits.

There are different methods of regularization, which help to mitigate the problem of overfitting [25]. One of them, used in this work is dropout.

2.8. Dropout

Dropout was introduced by researchers from University of Toronto [26] as a regularization technique for deep neural networks. The idea behind it is to remove some random group of neurons along with connections during training phases. Since those random groups are different at each step, it prevents neurons from learning to copy other neurons, which in turn makes them better at approximating desired output. This is often compared to ensemble models, which is training several models and making them vote. Dropout is fully described in [26].

2.9. Tested Architectures

Several neural network architecture were trained and tested as classifiers. Architecture which was satisfying on chosen test, dataset turned out to be too weak for benchmark dataset (see results in Section 4) so it had to be adjusted. The tested architectures are shown in Figs. 5–8, where:

- LSTM – single LSTM cell,
- GRU – single GRU cell,
- Dropout – adding regularization using dropout,
- LR – sigmoid layer,
- Embedding – mapping value to vector space.

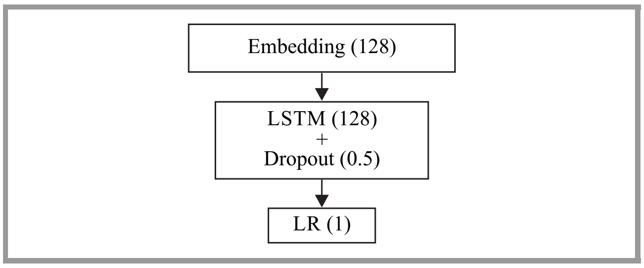


Fig. 5. Network embedding and one LSTM layer.

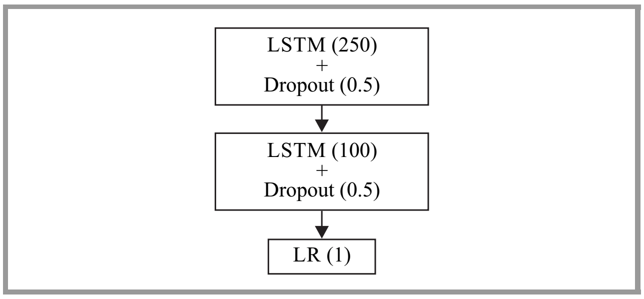


Fig. 6. Network structure with two LSTM layers.

Moreover, networks in Figs. 5 and 6 were trained with and without dropout, which turned out to have major influence on results. Process of architecture selection was empirical, which means that many architectures have been tested and hyperparameters based on results were adjusted.

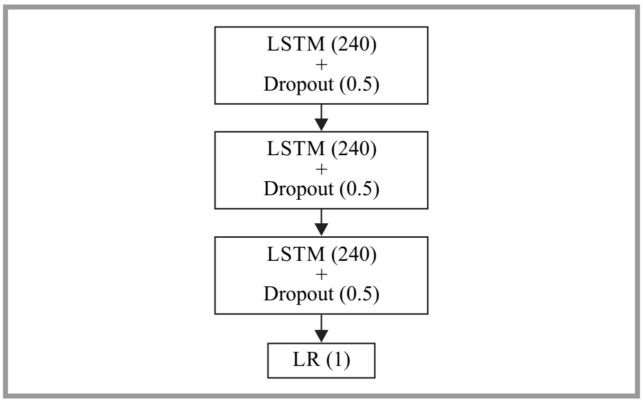


Fig. 7. Network with three LSTM layers (it was tested only on benchmark set).

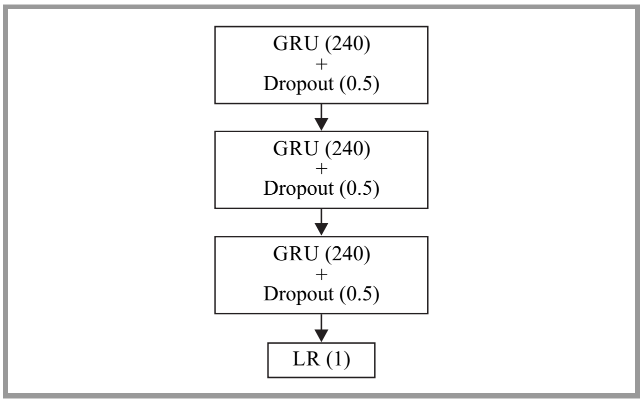


Fig. 8. Network with three GRU layers (it was tested only on benchmark set).

3. Datasets

3.1. Authors' Dataset

In order to conduct experiments we prepared own dataset. To achieve this task a website gathering keystroke data was used. The recording software was implemented as a student project by Albert Wolant. Every user was asked to type his address 5 times and type other addresses once. Nine students have taken part in this experiments, but due to low quality of some data (e.g. copy-paste method), samples batch from 3 people was rejected. Final dataset consisted of data from 6 people ranging from 12 to 20 samples each. It is worth mentioning that samples include information about mistakes made by typists.

3.2. Benchmark Dataset

Because the dataset described in previous section has only few samples, in addition a benchmark dataset available on the website was used [27]. This site provides exhaustive description of this set and acquiring method. This dataset was used by its authors to compare anomaly detectors [13]. It consists of data gathered from 51 typists, each has typed the same phrase 400 times. Even though it was created for anomaly detectors, it turned out to be very valuable for presented algorithm. Because users typed the same phrase (to be more precise – same password), it is possible to create a dataset for each user containing this user's samples as positive data and all other samples as negative. The problem here is that in such case there are 50 times as many negative as positive data. Classifier trained on such data will most likely tend to classify new samples as negative. It would also give false impression of high accuracy [28]. Because only small part of test data would be positive, so just classifying everything as negative gives high accuracy. With this problem in mind, authors decided to balance positive and negative data. For every user, only 400 random samples from other users were chosen as negative. Eventually, 51 sets were obtained (one for every user) containing 800 samples each: 400 hundred positive, 400 negative.

One drawback of this set is that it is cleared from mistakenly typed samples. Thus, there is no information about frequency of errors done by a user.

It should be clearly stated, that since this set was not used in its direct shape and that evaluation methods were different from those used by its authors results of this research cannot be directly compared to original results achieved by authors in [13].

4. Experiments and Results

For both (own and benchmark) datasets the different but not disjoint sets of models have been tested. In addition, a scenario in which negative data is artificial was also included in test process. In this case was tested and compared with *k*-nearest neighbors' classifier.

In order to achieve repeatability of experiments, they were all performed with the same random number generator seed.

4.1. Authors' Dataset

Due to small size of this set, a non-standard evaluation has been employed (by "standard" we mean dividing set for training, validation and test sets). This is why leave-n-out-cross-validation with $n = 1$ [29] is used. In this validation with ($n = 1$) the one model for every dataset element was trained. This selected element acts as one element test set. The model is trained on the rest of the set. This means, having n elements in a set, n models should be trained. Then the model computation on this selected element is performed. The total accuracy is an average computed from all those results.

4.1.1. Model with one LSTM Cell and Embedding

For this model, the total accuracy for all users reached only 58%. Table 1 shows results for all users. Note that because of validation type, single sample is included multiple times here. FP-free thresholds cell shows score when acceptance thresholds were chosen to eliminate false positive errors on training set. Because total accuracy was low, we got rid of embedding layer in favor of another LSTM cells.

Table 1
Results for all users

	Accepted	Rejected
Genuine user	359	138
Impostor	164	72
	Threshold 0.5	FP-free thresholds
Accuracy	0.58	0.55

Table 2
Results for all users

	Accepted	Rejected
Genuine user	456	41
Impostor	75	161
	Threshold 0.5	FP-free threshold
Accuracy	0.85 (0.88)	0.52 (0.8)

4.1.2. Model with two LSTM Cells

This model was tested in two versions – with and without dropout. Table 2 shows its accuracy. The numbers in parentheses relate to models with dropout.

The accuracy of LSTM model with 2 cells is satisfying. The dropout's influence on results is clear, especially if false positive free thresholds are used. By only adding dropout, the accuracy raised from 0.52 to 0.8. Unfortunately, the size of this dataset is small and the evaluation method had negative impact on results. Hence, it is hard to judge the algorithm quality by this data only. Despite this problem, those results hold some value, because in real life applications of verification based on keystroke dynamics usually only have small datasets are available.

4.2. Benchmark Set – Limited Data

Benchmark dataset, as it was described earlier, contains more data (in terms of both user count and samples per user). It is thus more reliable when it comes to the algorithm evaluation. Samples in this dataset contain more than just dwell-time. However, because only dwell-time was recorded in author's dataset, first study was performed only including this measure.

4.2.1. Model with two LSTM Cells

This is the same model, which turned out to be good enough for our custom dataset. This time, only version with dropout was tested as it achieved better results. Results are presented in Tables 3 and 4.

Unfortunately, model with two LSTM cells, even though it performed well on small dataset, does not give satisfying

Table 3
Results on benchmark dataset for two LSTM cells

	Threshold 0.5	FP-free
Average accuracy	0.759	0.59
Maximum	0.975	0.994
Minimum	0.4875	0.5
Standard deviation	0.101	0.1344

Table 4
Results on benchmark dataset for two LSTM cells

EER	0.227 (0.094)
Zero-miss rate	0.764 (0.221)

results on benchmark dataset. We have then made it more complex by adding one more LSTM cell as well as increasing the total number of neurons. This model has achieved results presented in Tables 5 and 6.

Table 5
Accuracy on benchmark dataset for model
with three LSTM cells

	Threshold 0.5	FP-free
Average accuracy	0.764	0.61
Maximum	0.9875	0.9875
Minimum	0.5187	0.5
Standard deviation	0.114	0.1399

Table 6
Benchmark dataset results for three model
with three LSTM layers

EER	0.219 (0.106)
Zero-miss rate	0.747 (0.221)

4.2.2. Model with Three LSTM Cells

In this case results are only slightly better than previous. It seems like simply increasing complexity of this model is not enough. Therefore, we decided to try again, swapping LSTM cells with GRU equivalents.

4.2.3. Model with Three GRU Cells

Results of experiments with this model (Tables 7 and 8) are comparable with those achieved on own dataset. However, if we compare it with results achieved by author's dataset, proposed algorithm would be placed 8th in terms of EER and 6th when it comes to zero-miss rate. Especially zero-miss rate is high which we would like to minimize.

Table 7
Accuracy on the benchmark dataset for model
with three GRU cells

	Threshold 0.5	FP-free
Average accuracy	0.83	0.68
Maximum	0.9875	0.9875
Minimum	0.5	0.5
Standard deviation	0.099	0.1397

Table 8
Benchmark dataset results for three model
with three GRU layers

EER	0.150 (0.087)
Zero-miss rate	0.613 (0.260)

4.3. Benchmark Dataset – All Data

As it was mentioned earlier, original dataset contains more than just dwell-time. Authors decided to try testing proposed algorithm using all data provided by the dataset. Achieved results are presented in Table 9. Only measures, which are easily comparable with algorithms presented in [13] are shown.

Table 9
Models results on benchmark dataset with all data

Model	EER	Zero-miss rate
LSTM 2 cells	**0.136 (0.176)**	0.379 (0.314)
LSTM 3 cells	0.165 (0.191)	**0.333 (0.282)**
GRU 3 cells	0.224 (0.319)	0.389 (0.325)

The results are significantly better than those which only included dwell time. In addition, the best model here is the one built with 3 LSTM cells.

4.4. Artificially Generated Data

One of the most important disadvantages of the algorithm is the need of negative data during training. In this work the method of artificial generation of negative data based on positive samples have been tested. Proposed algorithm was to *k*-nearest neighbor classifier. In total 417 different combinations of distance definition, number of neighbors and standard deviation of the normal distribution used for negative data generation was tested. It is worth noting that the algorithm is used as a classifier and not as an anomaly detector. The algorithm presented in this work has achieved results shown in Tables 10 and 11.

Table 10
Accuracy for data with artificial negative samples

Model	Accuracy threshold 0.5	Accuracy FP-free
LSTM 2 cells	0.622 (0.094)	0.562 (0.094)
LSTM 3 cells	0.633 (0.106)	0.561 (0.100)
GRU 3 cells	0.629 (0.093)	0.707 (0.101)

Table 11
EER and zero-miss rate for data with artificial negative samples

Model	EER	Zero-miss rate
LSTM 2 cells	0.441 (0.336)	0.598 (0.303)
LSTM 3 cells	0.768 (0.219)	0.592 (0.291)
GRU 3 cells	0.527 (0.402)	0.597 (0.334)

For a comparison, best result of k-nearest neighbor was for $k = 1$ and dice distance and it was 58%.

A similar experiment was also conducted on author's dataset. In this case, the best k-nearest neighbor algorithm accuracy was 87% while neural networks achieved 100% accuracy. However, this cannot be used as an argument for high accuracy of the model, because experiments on the larger dataset have not confirmed such high accuracy. The question is, however, why artificial data has actually increased accuracy (using only real data 80% accuracy was achieved). The reason is probably that without artificial data, the dataset was imbalanced in terms of negative to positive samples ratio, which made the network to be more eager to answer with class, which was overrepresented in the training set. Since we generated one artificial sample for each positive, this gave us the perfectly balanced set.

Unfortunately, presented method of generating artificial data turned out to be not very effective. The accuracy is significantly lower compared to training with only real (positive and negative) data. However, as was expected, recurrent neural networks performed generally better than k-nearest neighbor classifier.

5. Conclusions and Algorithm Evaluation

Compared to results from authors of the benchmark dataset, achieved best result (EER 0.136) would be on 7th place in terms of EER for total 14 places. It is equal to the one achieved by filtered Manhattan algorithm, yet its standard deviation is better: 0.083 compared to 0.176. The presented algorithm performed better than other neural networks tested by authors.

However, zero-miss rate is more interesting. The best result achieved by authors of [13] is 0.468. In this research the best result is 0.333, a lot better, however, those results cannot be directly compared, because of different nature of algorithms – this work shows binary classifier, authors of the mentioned paper tested anomaly detectors, different training method and different evaluation method. Despite that, presented algorithm performed well in terms of zero-miss rate and lets us recommended it as valuable when it comes to such a case. It is worth reminding, that high accuracy without false positive errors was one of the main objectives of the designed algorithm. It is worth noting, that the big leap in accuracy was caused by including additional data (both flight and dwell time). As it turned out, this had more influence than classifier architecture.

Another important feature, which was required from the algorithm, was scalability in terms of user count. Because for each user we train separate classifier, there is no problem with too many similar classes – each model is binary classifier trained for a given individual. Because neural networks are based on parametric models, they require the access to samples database only in the training phase. Thus, the increasing sample count for the user will not increase the size of the model when it comes to memory usage. Each model requires about 3.5 MB. This seems reasonable size (1 million users would require 3.5 TB of disk space). Therefore, the objective of unconstrained scalability is achieved.

We have stated the hypothesis that input data are sequences, and not just vectors and that a valuable signal comes from this information. Because recurrent neural networks are the natural choice for sequences processing, it could have direct impact on the accuracy. However, we cannot say with strong belief that this statement is more than just hypothesis. If our results were significantly better then others, it would be the strong evidence for it.

Unfortunately, the algorithm is not free from flaws. Most of them, however, were known at the beginning of the work. We have tried to mitigate the problem by generating artificial negative data. Results were admittedly better than k-nearest neighbors, yet they are noticeably worse than those achieved by the same model with access to real negative data. Perhaps, there is a method of generating better data, but further studies are needed here.

Certain drawback of the algorithm is how much time it requires to be fully trained. Neural networks are complicated models with a large number of parameters, so it requires time to adjust them. On a typical desktop 2.9 GHz Intel Core i5 CPU training and evaluating most sophisticated models took about 8 hours, which means about 10 minutes per user (there were 51 typists in benchmark dataset). Even if this seems quick, it is very long time compared to many anomaly detectors, which often only require one pass through the database. 10 minutes is a big issue for so-called continuous verification, i.e. constant monitoring of keyboard usage in order to detect impostors. However, training time directly depends on the dataset size. In this case, for each user we had 640 samples. Acquiring this number of samples (with assumption that exactly half of them are negative) requires time and has to be finished before training. Having said that, 10 minutes becomes less significant. Nevertheless, full training and evaluation requires 8 hours, which makes the hyperparameters adjustment a tougher task.

Paper [30] presents LSTM networks used as anomaly detectors. By incorporating this idea, we could use the same evaluation method as it was used in [13], which introduced benchmark dataset. This would allow us direct comparison. Moreover, it would solve the problem of negative data requirement. Results achieved in the mentioned work give hope for increase of usability if those models for keystroke dynamics in the future.

References

[1] R. Gaines, W. Lisowski, S. J. Press, and N. Shapiro, "Authentication by keystroke timing: some preliminary results", R-2526-NSF RAND Report, RAND Corporation, Santa Monica, CA, USA, May 1980.

[2] F. Monrose and A. D. Rubin, "Keystroke dynamics as a biometric for authentication", *Future Gener. Comp. Syst.*, vol. 16, pp. 351–359, 2000.

[3] R. Joyce and G. Gupta, "Identity authorization based on keystroke latencies", *Commun. of the ACM*, vol. 33, no. 2, pp. 168–176, 1990.

[4] D. Mahar, R. Napier, M. Wagner, W. Laverty, R. Henderson, and M. Hiron, "Optimizing digraph-latency based biometric typist verification systems: Inter and intra typists differences in digraph latency distributions", *Int. J. Human-Comp. Stud.*, vol. 43, no. 4, pp. 579–592, 1995.

[5] P. R. Dholi and K. P. Chaudhari, "Typing Pattern Recognition Using Keystroke Dynamics", in *Mobile Communication and Power Engineering*, Vi. V. Das and Y. Chaba, Eds. *Communications in Computer and Information Science*, vol. 296, pp. 275–280. Springer, 2012.

[6] G. Ho, "TapDynamics: Strengthening User Authentication on Mobile Phones with Keystroke Dynamics", Tech. Rep., Stanford University, San Francisco, CA, USA, 2014.

[7] Y. Deng and Y. Zhong, "Keystroke dynamics advances for mobile devices using deep neural network", in *Recent Advances in User Authentication Using Keystroke Dynamics Biometrics*, Y. Zhong and Y. Deng, Eds. Science Gate Publishing, 2015, vol. 2, pp. 59–70.

[8] C. Giuffrida, K. Majdanik, M. Conti, and H. Bos, "I sensed it was you: Authenticating mobile users with sensor-enhanced keystroke dynamics", in *Detection of Intrusions and Malware, and Vulnerability Assessment*, S. Dietrich, Ed. *LNCS*, vol. 8550, pp. 92–111. Springer, 2014 (doi: 10.07/978/3-319-08509-8_6).

[9] M. Rogowski, K. Saeed, M. Rybnik, M. Tabędzki, and M. Adamski, "User authentication for mobile devices", in *Computer Information Systems and Industrial Management*, K. Saeed, R. Chaki, A. Cortesi, and S. Wierzchoń, Eds. *LNCS*, vol. 8104, pp. 47–58. Springer, 2013.

[10] T. Bhattasali and K. Saeed, "Two Factor Remote Authentication in Healthcare", in *Proc. 3rd Int. Conf. Advan. in Comput., Commun. & Inform. ICACCI 2014*, Delhi, India, 2014, pp. 380–386.

[11] T. Bhattasali, K. Saeed, N. Chaki, and R. Chaki, "Bio-authentication for layered remote health monitor framework", *J. Medical Inform. & Technol.*, vol. 23, no. 1, pp. 131–139, 2014.

[12] M. Rybnik, P. Panasiuk, K. Saeed, and M. Rogowski, "Advances in the keystroke dynamics: the practical impact of database quality", in *Computer Information Systems and Industrial Management*, A. Cortesi, N. Chaki, K. Saeed, and S. Wierzchoń, Eds. *LNCS*, vol. 7564, pp. 203–214. Springer, 2012 (doi: 101007/978-3-642-33260-9_17).

[13] K. S. Killourhy and R. A. Maxion, "Comparing anomaly detectors for keystroke dynamics", in *Proc. 39th Annual IEEE/IFIP Int. Conf. Dependable Syst. & Netw. DSN 2009*, Lisbon, Portugal, 2009, pp. 125–134.

[14] Y. Deng and Y. Zhong, "Keystroke Dynamics User Authentication Based on Gaussian Mixture Model and Deep Belief Nets", *ISRN Sig. Process.*, vol. 2013, article ID 565183, 2013 (doi: 10.1155/2013/565183).

[15] "ICT Facts and Figures" 2005, 2010, 2014, Telecommunication Development Bureau, International Telecommunication Union (ITU), 24 May 2015.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computat.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult", *IEEE Trans. on Neural Netw.*, vol. 5, no. 2, pp. 157–166, 1994.

[18] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent Neural Network Regularization", in *Proc. Int. Conf. on Learn. Representat. ICLR 2015* San Diego, CA, USA, 2015 [Online]. Available: https://arxiv.org/pdf/1409.2329.pdf

[19] S. K. Sønderby, C. K. Sønderby, H. Nielsen, and O. Winther, "Convolutional LSTM Networks for subcellular localization of proteins", in *Algorithms for Computational Biology*, A.-H. Dediu, F. Hernández-Quiroz, C. Martín-Vide, and D. A. Rosenblueth, Eds. *LNCS*, vol. 9199, pp. 68–80. Springer, 2015.

[20] C. Olah, "Understanding LSTM Networks", Aug. 27, 2015 [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed on Aug. 9, 2016)

[21] K. Cho, F. Bougares, H. Schwenk, D. Bahdanau, and Y. Bengio, "Learning phrase representations using RNN Encoder-decoder for statistical machine translation", in *Proc. of the Conf. on Empir. Methods in Natural Language Process. EMNLP 2014*, Doha, Qatar, 2014, pp. 1724–1734.

[22] P. J. Werbos, "Backpropagation through time: What it does and how to do it", *Proc. of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1998.

[23] D. P. Kingma and J. L. Ba, "Adam: A Method for stochastic optimization", in *Proc. Int. Conf. on Learn. Representat. ICLR 2015* San Diego, CA, USA, 2015 [Online]. Available: https://arxiv.org/pdf/1412.6980v8.pdf

[24] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition", *IEEE Trans. Pattern Anal. & Mach. Intelig.*, vol. 31, no. 5, pp. 855–868, 2009.

[25] A. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance", in *Proc. 21st Int. Conf. on Machine Learn. ICML'04*, Banff, Canada, 2004.

[26] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting", *J. of Mach. Learn. Res.*, vol. 15, no. 1, 1929–1958, 2014.

[27] K. Killourhy and R. Maxion, "Keystroke Dynamics – Benchmark Data Set", Carnegie Mellon University, Pittsburgh, PA, USA [Online]. Available: http://www.cs.cmu.edu/∼keystroke/ (accessed on May 27, 2016).

[28] H. He and E. A. Garcia, "Learning from imbalanced data", *IEEE Trans. on Knowl. & Data Engin.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[29] J. Schneider, "Cross Validation", Carnegie Mellon University, Pittsburgh, PA, USA [Online]. Available: https://www.cs.cmu.edu/∼schneide/tut5/node42.html (accessed on May 27, 2016).

[30] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series", in *Proc. of European Symp. on Artif. Neural Netw., Computat. Intellig. and Machine Learning ESANN 2015*, Bruges, Belgium, 2015.

Paweł Kobojek received the M.Sc.degree in the field of Computer Science from Warsaw University of Technology in 2016. His scientific interests are mainly focused on Bioinformatics and Artificial Intelligence and its applications in biometrics and language processing.

E-mail: PawelKobojek@gmail.com
Faculty of Mathematics and Information Sciences
Warsaw University of Technology
Koszykowa st 75
00-662 Warsaw, Poland

Khalid Saeed is a B.Sc., M.Sc., Ph.D. and D.Sc. degrees holder. He is a Computer Science full professor (Biometrics, Image Analysis and Processing) at Białystok University of Technology. He also works with the Faculty of Mathematics and Information Sciences, Warsaw University of Technology. He was with AGH Krakow in 2008–2014. He has published more than 200 publications and edited 27 books, journals and conference proceedings, 11 text and reference books. He received 19 academic awards. Mr. Khalid Saeed is a member of more than 15 editorial boards of international journals and conferences. He is an IEEE Senior Member and has been selected as IEEE Distinguished Speaker for 2011–2017. He is the editor-in-chief of International Journal of Biometrics with Inderscience Publishers.

E-mail: k.saeed@mini.pw.edu.pl
Faculty of Mathematics and Information Sciences
Warsaw University of Technology
Koszykowa st 75
00-662 Warsaw, Poland

The Kummer confluent hypergeometric function and some of its applications in the theory of azimuthally magnetized circular ferrite waveguides

Georgi Nikolov Georgiev and Mariana Nikolova Georgieva-Grosse

http://doi.org/10.26636/jtit.2005.3.317

Abstract— Examples of the application of the confluent hypergeometric functions in miscellaneous areas of the theoretical physics are presented. It is suggested these functions to be utilized as a universal means for solution of a large number of problems, leading to: cylindrical, incomplete gamma, Coulomb wave, Airy, Kelvin, Bateman, Weber's parabolic cylinder, logarithmic-integral and exponential integral functions, generalized Laguerre, Poisson-Charlier and Hermit polynomials, integral sine and cosine, Fresnel and probability integrals, etc. (whose complete list is given), which are their special cases. The employment of such an approach would permit to develop general methods for integration of these tasks, to generalize results of different directions of physics and to find the common features of various phenomena, governed by equations, pertaining to the same family. Emphasis is placed here on the use of the Kummer function in the field of microwaves: the cases of normal and slow rotationally symmetric *TE* modes propagation in the azimuthally magnetized circular ferrite waveguide are considered. Lemmas on the properties of the argument, real and imaginary parts, and positive purely imaginary (real) zeros of the function mentioned in the complex (real) domain, of importance in the solution of boundary-value problem stated for normal (slow) waves, are substantiated analytically or numerically. A theorem for the identity of positive purely imaginary and real zeros of the complex respectively real Kummer function for certain parameters, is proved numerically. Tables and graphs support the results established. The terms for wave transmission are obtained as four bilaterally open intervals of variation of the quantities, specifying the fields. It turns out that the normal (slow) modes may exist in one (two) region(s). The theoretically predicted phase curves for the first waves of the two *TE* sets examined show that the structure explored is suitable for ferrite control components design.

Keywords*— *microwave propagation in anisotropic media, microwave guides and components, ferrite phase shifters, switches and isolators, eigenvalue problems, function-theoretic and computational methods in electromagnetic theory, theoretical and numerical analysis of special functions.

1. Introduction

The Kummer confluent hypergeometric function (CHF) belongs to an important class of special functions of the mathematical physics [1–19] with a large number of applications in different branches of the quantum (wave) mechanics [2, 5–7, 9, 10, 12, 17, 20, 21], atomic physics [2, 5, 22, 23], quantum theory [23], nuclear physics [23], quantum electronics [24, 25], elasticity theory [2, 5, 7, 9, 26], acoustics [5, 10, 27, 28], theory of oscillating strings [2, 5, 29], hydrodynamics [5, 10, 30], random walk theory [2, 7], optics [31], wave theory [2, 7], fiber optics [32–34], electromagnetic field theory [5, 7, 35, 36], plasma physics [37–39], the theory of probability and the mathematical statistics [5, 7, 10, 13, 40], the pure [5, 41–43] and applied mathematics [44]. In the microwave physics and in particular in the theory of waveguides, such examples are the problems for rotationally symmetric wave propagation in closed and opened circular guiding structures, containing: radially inhomogeneous isotropic dielectric [45–48] or azimuthally magnetized radially stratified anisotropic media (e.g., ferrite or semiconductor) [48–74]. The possibility to obtain signal phase shifting at microwaves makes the geometries of the second type of filling attractive for the development of nonreciprocal devices for this frequency band and is the reason for their extensive study [48–88].

In this paper some properties of the complex and real Kummer CHF and its positive purely imaginary, respectively real zeros are investigated, which are employed in the analysis of normal and slow rotationally symmetric TE modes in the simplest canonical structure of the aforesaid family of anisotropic transmission lines: the circular waveguide, entirely filled with ferrite. Obtained are the propagation conditions and phase characteristics in both cases, too. It is found that there is one (there are two) area(s) of normal (slow) wave transmission, available for both signs (only for the negative sign) of magnetization. The potentialities of the configuration as phaser, switch or isolator are discussed. Symbols with (without) hats "^" stand for quantities, relevant to the slow (normal) TE modes, respectively to the real (complex) Kummer functions.

Besides, the idea is also expressed to replace in the applications the special cases of the CHFs (that are enumerated) by the functions themselves (to replace the multitudinous schemes, utilized at present by a more universal technique) as much as possible. In this way lots of the common traits of different processes which usually remain hidden, owing to the usage of a rather diverse mathematics, would come into sight.

2. Confluent hypergeometric functions

2.1. Basic concepts

Confluent hypergeometric are called four functions: the Kummer and the connected with it Tricomi function $\Phi(a, c; x)$ and $\Psi(a, c; x)$, respectively, and the Whittaker first, and second ones $M_{\kappa,\mu}(x)$ and $W_{\kappa,\mu}(x)$ [10]. The functions $\Phi(a, c; x)$ and $\Psi(a, c; x)$ are solutions of the confluent hypergeometric equation (CHE), written in the standard form of Kummer [1–14, 16–19, 44, 54, 55, 57–59, 61, 69, 72], whereas $M_{\kappa,\mu}(x)$ and $W_{\kappa,\mu}(x)$ – of the same equation, presented in its modified form, suggested by Whittaker [3, 5, 8, 10, 11, 13, 15–17, 19, 44, 55]. The quantities a and c (κ and μ) are called parameters and x – variable [3]. The CHFs except the Kummer one are multiple-valued for which the zero is a branch point. Their main branch is taken in the complex x – plane with a cut along the negative real axis. Both $\Phi(a, c; x)$ and $M_{\kappa,\mu}(x)$ are regular at zero, whereas $\Psi(a, c; x)$ and $W_{\kappa,\mu}(x)$ tend to infinity for $x \to 0$ [1–19, 55, 57–59, 61, 69, 72]. The greater symmetry with respect to the parameters observed in the formulae, involving Whittaker functions [5, 15], as well as the symmetry in the functions themselves (in their values) [55], is the reason for discussing them in parallel with the Kummer and Tricomi ones. In our opinion however, though not symmetrical, the couple $\Phi(a, c; x)$ – $\Psi(a, c; x)$ is to be preferred in the applications in view of the simpler character of power series, determining them. In addition to above definition, due to L. J. Slater [10], worth mentioning also is the one, given by Tricomi who ascertains that CHF is called any solution of CHE, considered in whichever of its forms [3]. Accordingly, such are for example the $\Phi^{*}(a, c; x)$, $\mathscr{M}_{\kappa,\mu}(x)$ and $N_{\kappa,\mu}(x)$ functions, too, introduced by Tricomi [2, 3, 7, 9, 61], Buchholz [5] and Erdélyi [10], respectively. Beside the notations, accepted here following F. G. Tricomi [2–4, 7, 9] and our previous works [54, 55, 57–61, 63, 64, 66–74], the symbols $M(a, b, x)$, ${}_1F_1[a; b; x]$, $\overset{\infty}{u}(a, b, x)$, and $F(\alpha, \beta, x)$ are employed also in literature instead of $\Phi(a, c; x)$, the symbols $U(a, b, x)$, $\overset{\infty}{v}(a, b, x)$ and $G(a, b, x)$ – instead of $\Psi(a, c; x)$, and the ones $\sqrt{2x/\pi}\, m_{\kappa}^{(2\rho)}(x)$ and $\sqrt{2x/\pi}\, w_{\kappa}^{(2\rho)}(x)$ – instead of $M_{\kappa,\mu}(x)$ and $W_{\kappa,\mu}(x)$, respectively [1, 5, 10, 12, 13]. The term “confluent” in the name of the functions is used, since the Kummer one might be deduced from the Gauss hypergeometric function ${}_2F_1(a, b; c; x)$ through a limiting process, leading to a confluence of two of its three regular singularities (1 and ∞) into an irregular one (the point ∞) [3, 5, 10]. (The regular singularity 0 remains unchanged.) The word “hypergeometric” is applied, as the expressions for the functions can be obtained by adding factors to the terms of the infinite geometric progression [10].

2.2. Special cases

A lot of special functions can be regarded as special cases of CHFs, or combinations of them:

- the ordinary and modified cylindrical and spherical Bessel functions: $J_\nu(x)$, $I_\nu(x)$, $\sqrt{\pi/(2x)}J_{n+1/2}(x)$ or $\sqrt{\pi/(2x)}J_{-n-1/2}(x)$ and $\sqrt{\pi/(2x)}I_{n+1/2}(x)$, respectively [1–3, 7, 9, 10, 12, 13, 15, 16];
- the Hankel functions $H_\nu^{(1)}(x)$ and $H_\nu^{(2)}(x)$ [1, 2, 7, 12, 13, 16];
- the Neumann function $N_\nu(x)$ [3, 7];
- the cylindrical and spherical McDonald functions $K_\nu(x)$ and $\sqrt{\pi/(2x)}K_{n+1/2}(x)$ [7, 13, 15, 16];
- the Coulomb wave functions: the two pairs $P_L(a, x)$ and $Q_L(a, x)$, and $U_L(a, x)$ and $V_L(a, x)$, considered by Curtis [17], the couples $G_L(\sigma)$ and $H_L(\sigma)$, defined by Hartree [17], and $U(\alpha, \gamma, Z)$ and $V(\alpha, \gamma, Z)$, introduced by Jeffreys and Jeffreys [17] and the most preferable in the applications standard pair $F_L(\eta, \rho)$ and $G_L(\eta, \rho)$, discussed by Abramowitz and Stegun [5, 10, 13, 17];
- the function $H(m,n,x)$, named Coulomb wave function and function of the paraboloid of revolution by Tricomi [2, 7, 9] or confluent hypergeometric function by Miller [13];
- the Laguerre functions $L_\nu^{(\mu)}(x)$ and $U_\nu^{(\mu)}(x)$ [3, 5, 10, 16], denoted also as $S_\nu^{\mu}(x)$ and $V_\nu^{\mu}(x)$ by Mirimanov [5, 10, 35];
- the Airy functions $Ai(x)$ [13, 16, 44] and $Bi(x)$ [13, 16];
- the incomplete $\gamma(a, x)$, the complementary $\Gamma(a, x)$, the modified $\gamma^{*}(a, x)$ and the fourth incomplete $\gamma_1(a, x)$ gamma functions [1–3, 7, 9, 10, 13, 15, 16, 44], as well as the derivative of them $g(a,x)$, $g_1(a,x)$, $G(a,x)$ and $k(a,x)$ ones, treated by Tricomi [2, 7, 9];
- the Kelvin (Thomson) functions $bei_n(x)$, $ber_n(x)$, $kei_n(x)$, $ker_n(x)$ [13, 16], $hei_n(x)$ and $her_n(x)$ [3, 11], met also like $bei_n x$, $ber_n x$, etc. [13];
- the Bateman function $k_\nu(x)$ [3, 5, 7, 10, 13, 16];
- the Weber’s parabolic cylinder functions $D_\nu(x)$ in the Whittaker’s notation [2, 3, 5, 7, 9, 10, 13, 15, 16, 44], $E_\nu^{(0)}(x)$ and $E_\nu^{(1)}(x)$ in the Buchholz’s one [5, 7, 10, 13, 16], $D_\nu^{+}(x)$ and $D_\nu^{-}(x)$, proposed by Tricomi [7], $U(a,x)$, $V(a,x)$ and $W(a,x)$ in the Miller form [13], or $\delta(\xi, \nu)$ and $\rho(\xi, \nu)$ in the symbols by Magnus [5, 10] and $\varphi_n(x)$ and $\Psi_n(x)$, suggested by Janke, Emde and Lösch [11];
- the Cunningham function $\omega_{m,n}(x)$ [5, 10, 13], known as Pearson-Cunningham function, too [15];
- the Heatly Toronto function $T(m, n, r)$ [3, 5, 10, 13, 16];
- the Meixner’s functions $F_1(\alpha, \beta, x)$ [3, 5, 10, 16] and $F_2(\alpha, \beta, x)$ [10];

- the MacRobert's function $E(\alpha, \beta :: x)$ [3, 5, 16];
- the Erdélyi function ${}_2F_0(\alpha,\beta;x)$ [3, 10, 16];
- the Poiseuille functions $pe(r, w)$ and $qe(r, w)$ [10];
- the Krupp functions ${}_1R(v, l; x)$ and ${}_2R(v, l; x)$ [5, 10];
- the Schlömilch function $S(v, x)$ [5, 15];
- the Chappell function $C(x, k)$ [5, 10];
- the logarithmic-integral function $li(x)$ or lix [1, 3, 7, 9, 10, 12, 13, 15, 16];
- the exponential integral functions Eix or $Ei(x)$ and $E_1(x)$ [3, 7, 9, 10, 13, 16], the generalized exponential integral function $E_n(x)$ [13, 16], marked also as $\mathscr{E}_n(x)$ [16] and the modified exponential integral one $Ein(x)$, used by Tricomi [2, 7, 13];
- the error $erfx$ or $Erf(x)$ and $Erfi(x)$, and complementary error $erfcx$ or $Erfc(x)$ functions (the error and probability integrals) [2, 3, 7, 9, 10, 13, 15, 44], as well as the ones $\Phi(x)$ and $F(x)$ [1, 5, 11–13], $\phi(x)$ and $L(x)$ [10], $\Theta(x)$, $H(x)$ and $\alpha(x)$ [3, 11], connected with them, the multiple probability integral $i^n\ erfcx$ [13] and the Hh – probability function $Hh_n(x)$ [13];
- the normal (Gauss) $P(x)$ and $Z(x)$ [13], and the χ^2-distribution $P(\chi^2|v)$ and $Q(\chi^2|v)$ functions [40], and the F-distribution $P(F|v_1, v_2)$ one [13];
- the Lagrange-Abel function $\phi_m(x)$ [15];
- some elementary (exponential e^x, power x^n, circular $\sin x$ and hyperbolic shx) functions [1, 3, 7, 13, 16];
- the reduced to $n+1$th degree exponential series $e_n(x)$ [7, 9];
- the Laguerre and generalized Laguerre polynomials $L_n(x)$ and $L_n^{(\alpha)}(x)$ [1–3, 7, 9, 10, 12, 13, 16, 44];
- the Sonine polynomials $T_\mu^{(n)}(x)$ [5, 15];
- the Poisson-Charlier polynomials $\rho_n(v, x)$ [10, 13] or $p_n(x)$ in the Tricomi's notation [3];
- the Hermit and modified Hermit polynomials $He_n(x)$ and $H_n(x)$ [1–3, 7, 10, 12, 13, 16, 44];
- some polynomials (in general incomplete) in $1/x$ of nth degree [7];
- the integral sine $Si(x)$ and cosine $Ci(x)$ [1–3, 5, 10, 12, 13, 15, 16]; and the modified cosine $Cin(x)$, employed by Tricomi [2, 7, 13];
- the Fresnel integrals $C(x)$ and $S(x)$ [3, 7, 10, 12, 13, 16], the related to them $C^*(x)$ and $S^*(x)$, and the generalized Fresnel ones $C(\alpha, x)$ and $S(\alpha, x)$ [2, 7].

2.3. Examples of application

The CHFs play an exceptional role in many branches of physics and mathematics. Several examples of their applications are:

- the solution of Schrödinger equation for charged particle motion (e.g., electron motion) in Coulombian field in the quantum mechanics, atomic physics and quantum theory [2, 5–7, 9, 10, 12, 17, 20–23];
- the energy spectrum specification of the isotropic (spherically symmetric) harmonic oscillator in nuclear physics and other related areas [5, 12, 23];
- the quantum mechanical treatment of the operation of the masers and lasers [24, 25];
- the elasticity problem for the flexion of circular or annular plates of lenticular form (resembling to a concave or convex lens), resting on, or rabbeted along its contour, subjected to a normal load whose value at certain point depends on its radial elongation from the center of the plate [2, 5, 7, 9, 26];
- the theory of the reflection of sound waves by a paraboloid [5, 10, 27];
- the consideration of sound waves propagation in parabolic horn, excited by a point source in its focus, and in the space between two co-focal paraboloids of revolution and the construction of the three-dimensional Green function for the homogeneous boundary-value problem of the first kind (Dirichlet problem) and of the second one (Neumann problem) for the wave equation in both cases [5, 28];
- the inquiry of the natural oscillations of a tight stretched string whose mass is distributed symmetrically with respect to its middle, following a parabolic law [5, 29];
- the investigation of a heat generation in a laminary Poiseuille flow through (in a viscous incompressible liquid, flowing through) a thin cylindrical capillary tube of circular cross-section [5, 10, 30];
- the determination of the length of the resultant of a large number of accidentally directed vectors (a special case, connected with the problems of random walk) [2, 7];
- the task for cylindrical-parabolic mirrors [31];
- the description of sea waves motion against a sheer coast [2, 7];
- the analysis of guided modes along a cladded optical fiber of parabolic-index core and homogeneous cladding [32–34];
- the portrayal of electromagnetic waves transmission in parabolic pipes [5];
- the study of the reflection of electromagnetic waves by a parabolic cylinder [2, 5, 7];
- the solution of the diffraction problem for a plane and a spherical electromagnetic wave in a paraboloid of revolution of infinite dinemsions [5, 35];

- the exploration of radiation electromagnetic field in a hollow paraboloid of revolution, launched by an axially oriented electric or magnetic dipole, placed at or before its focus, and between two co-focal paraboloids [5];
- the electrodynamic characterization of the field in an excited by a loop cavity resonator, consisting of two co-focal caps of the form of paraboloids of revolution [5];
- the finding of the normal (Gauss), the χ^2- and the F-distribution for arbitrary quantities in the theory of probability and mathematical statistics [13, 40];
- the development of a mathematical model of the electrical oscillations in a free ending wire [5];
- the assessment of the noise voltages transfer over a linear rectifier [5];
- the explanation of radiation of magnetized dipole in a stratified medium of spherical symmetry (in a globular layered atmosphere) [5];
- the case of electromagnetic waves in plasma with electron density changing linearly along one of the co-ordinate axes, if an infinitely large constant magnetic field is applied along the latter [37];
- the problem for electromagnetic waves in an inhomogeneous plasma whose collision frequency is a constant and the electron density varies in one direction only as a second-degree polynomial of the last-mentioned (or following a parabolic profile) [38];
- the examination of the radiation field from a uniform magnetic ring current around a cylindrical body of infinite length covered by a plasma sheath in the presence of a uniform azimuthal static magnetic field which is of practical application to improve radio communications during the blackout period in the re-entry of a conical space vehicle in the earth's atmosphere at hypersonic speed [39];
- the Tricomi euristic approximate evaluation of the distribution of the positive integers which can be presented as sums of two kth powers of possible value in the theory of probability [7];
- the finding of the normal (Gauss) and χ^2 – and F – distribution for arbitrary quantities in the theory of probability and mathematical statistics [13, 40];
- the series expansion of an arbitrary function in terms of eigenfunctions, of significance in the theory of hydrogen atom to describe the point (discrete) and continuous energy spectrum [5, 41];
- some continued fractions expressions of analytic functions in the complex plane, employable in the computational methods [13, 42];
- the realization of irreducible (simple) representations of a group of third order triangular matrixes, in which integral operators whose kernels are written through Whittaker functions, correspond to certain of its elements [43];
- the inspection of TE_{0n} and TM_{0n} modes, sustained in radially inhomogeneous circular dielectric waveguides (plasma columns or optical fibers) whose permittivity alters in radial direction following certain profiles [45–48];
- the theory of normal and slow surface TM_{0n} waves in the azimuthally magnetized millimeter-wave semiconductor (solid-plasma) coaxial waveguides, using n-type InSb and GaAs cooled to 77K as a plasma material [49, 53, 56, 62, 65];
- the problems for normal and slow TE_{0n} modes in the azimuthally-magnetized ferrite and ferrite-dielectric circular and co-axial waveguides and for slow waves, propagating along cylindrical helices, closely wound around (or surrounded by) an azimuthally magnetized ferrite rod (toroid) [49–52, 54, 55, 57–61, 63, 64, 66–74];
- the study of microwave radiation from a magnetic dipole in an azimuthally magnetized ferrite cylinder [89] which may also be explored by means of the functions considered.

3. The confluent hypergeometric functions – a universal means for solution of problems of mathematical physics

The above analysis shows that: a lot of tasks from different areas of mathematical physics lead to various representations of CHFs and a large number of functions are special cases of the latter and can be expressed in terms of them. In view of this one might expect to meet the CHFs throughout the literature. In fact, as Lauwerier wrote, "they are only sparingly used" [30]. Even one of the problems from the class examined was categorized as "not a particularly fortunate one" in the words of Suhl and Walker [49]. An attempt to substantiate these inferences is the following assertion (standing nowadays in plenty of fields): "The reason may be that these functions are still too little known, and are therefore evaded as much as possible." [30].

Indeed, the CHFs are more complicated than many other special functions, since they possess two parameters and an independent variable. The lack of numerical tables, or the insufficient tabulation of the functional values and their zeros were a grave obstacle in their applications [30, 49, 75, 80]. Serious computational predicaments arise, if the parameters and variable get large and especially, provided they are complex. The relations between these three quantities also influence the speed of convergence of power series, determining the functions. Due to this, coming upon them, some authors gave only formal analytical results [2, 5, 7, 9, 12, 21, 23, 24, 27, 30, 36–39, 49, 51], whereas others tried to avoid them through:

- reducing the CHFs to their special cases (if possible) [5, 10, 12];
- defining new functions which replace them [75, 79, 80] or harnessing such ones [83, 89];
- elaborating various numerical methods [48, 76, 82, 86, 87].

In our opinion the usage of so many very diverse artificially devised approaches hampers tracing out the connections among the different phenomena explored (which obviously exist, since the latter could be described by the same mathematical language), and impedes the establishment of their common characteristics. It is our conviction that in spite of the drawbacks pointed out, or the difficulties, appearing as a result of their complexity, the CHFs have indisputable advantages: generality and well developed theory together with valuable properties, such as for example symmetry in case of Whittaker functions. Therefore, a way out of this complicated situation, is to find means to overcome the computational challenges, instead of inventing contrivances to obviate the CHFs.

In essence the employment of the special cases, debated in Subsection 2.2, has a similar effect on the process of investigation of the phenomena and their properties, as the just discussed one, when the CHFs have been excluded from the solutions. Utilizing such a great number of functions entails as well a fragmentation of the analysis methods of corresponding tasks. However, unlike before, this state of affairs has sprung up in a natural way, when different problems have been attacked by different schemes.

As a set-off to that, it is suggested to replace the functions in question (the special cases) everywhere, where they attend by the having more universal character CHFs. To this end, the following statement is formulated:

Statement for universality**: The confluent hypergeometric functions, considered in any of their forms, could be used as a universal means instead of any of the functions, being their special cases and the related to them, such as: the cylindrical, incomplete gamma, Coulomb wave, Weber's parabolic cylinder functions, etc. (whose complete list is given above), in the tasks in which they are met.**

Corrolary: Moving from a fragmentation to a generalization would permit:

- to solve enormous number of problems by the same universal mathematical technique;
- to develop general methods for their solution;
- to generalize results of different branches of physics;
- to find common features in different phenomena, governed by equations from the same family.

An undoubted benefit could be derived even from the partial realization of the programme proposed (when the computational hardships are surmountable).

4. Kummer confluent hypergeometric function

4.1. Definition

The Kummer CHF is defined by the absolutely convergent infinite power series [1–14, 16–19, 54, 55, 57–59, 61, 69, 72]:

$$\Phi(a,c;x)=\sum_{0}^{\infty}\frac{(a)_\nu}{(c)_\nu}\frac{x^\nu}{\nu!}\,. \tag{1}$$

It is analytic, regular at zero entire single-valued transcendental function of all a, c, x, (real or complex) except $c = 0, -1, -2, -3, \ldots$, for which it has simple poles. $\Phi(a,c;x)$ is a notation, introduced by Humbert, $(\lambda)_\nu = \lambda(\lambda+1)(\lambda+2)\ldots(\lambda+\nu-1) = \Gamma(\lambda+\nu)/\Gamma(\lambda)$, $(\lambda)_0 = 1$, $(1)_\nu = \nu!$, where λ stands for any number (real or complex) and ν for any positive integer or zero, is the Pochhammer's symbol and $\Gamma(\lambda)$ is the Euler gamma function. The series (1) is a solution of the Kummer CHE that is a second order ordinary differential equation [1–14, 16–19, 54, 55, 57–59, 61, 69, 72]:

$$x\frac{d^2y}{dx^2}+(c-x)\frac{dy}{dx}-ay=0\,, \tag{2}$$

having regular and irregular singularities at 0 and at ∞, respectively.

4.2. Asymptotic expansion

The asymptotic expansion of $\Phi(a,c;x)$ for large values of variable $x = |x|e^{j\varphi}$, $0 < \varphi < \pi$, is [6, 54, 57–59]:

$$\Phi(a,c;x)\approx\frac{\Gamma(c)}{\Gamma(a)}|x|^{a-c}e^{j(a-c)\varphi}e^{|x|e^{j\varphi}}+\frac{\Gamma(c)}{\Gamma(c-a)}|x|^{-a}e^{ja(\pi-\varphi)}. \tag{3}$$

If $x = jz$ ($z = |x|$ – real, positive), i.e., $\varphi = \pi/2$, both terms in the expression are approximately equally large and should be taken into account. Provided x is real, positive ($\varphi = 0$), the first term in formula (3) is considered only, since the second one becomes less than the unavoidable error, inherent to the asymptotic expansions. When x is real, negative ($\varphi = \pi$), the second term is used solely for the same reason [6, 54, 57–59].

5. Some properties of the complex Kummer function

5.1. Properties due to the analytical study

The case a – complex ($a = \mathrm{Re}a + \mathrm{jIm}a$), $c = 2\mathrm{Re}a$ – positive integer, $\mathrm{Im}a = -k$, k – real $[a = c/2 - \mathrm{j}k,\ k = \mathrm{j}(a - c/2)]$, $x = \mathrm{j}z$ – positive purely imaginary ($x = \mathrm{Re}x + \mathrm{jIm}x$, $\mathrm{Re}x = 0$, $\mathrm{Im}x = z$, $|x| = z$, z – real, positive, $\varphi = \arg x = \mathrm{Im}x/\mathrm{Re}x$, $\varphi = \pi/2$), is discussed. Under these assumptions an application of the first Kummer theorem [1–3, 5–7, 9–13, 16] facilitates to prove the statement [57–59]:

Lemma 1: If $c = 2\mathrm{Re}a$, $\mathrm{Re}x = 0$ ($x = \mathrm{j}z$ – purely imaginary), then

$$\arg\Phi(a, 2\mathrm{Re}a; \mathrm{j}z) = z/2, \qquad (4)$$

where $\arg\Phi$ stands for the argument of the Kummer function.

In addition, a new modulus-argument representation of the asymptotic expansion (3) is obtained [57, 58]:

$$\Phi(a, 2\mathrm{Re}a; \mathrm{j}z) \approx 2F(\cos v)\mathrm{e}^{\mathrm{j}(z/2)} = 2F|\cos v|\mathrm{e}^{\mathrm{j}(z/2+n\pi)}, \qquad (5)$$

where $F = [\Gamma(2\mathrm{Re}a)/|\Gamma(a)|]\mathrm{e}^{-(\pi/2)\mathrm{Im}a}z^{-\mathrm{Re}a}$, $v = (z/2) + \mathrm{Im}a\ln z - \arg\Gamma(a) - \mathrm{Re}a(\pi/2)$ and $n = 1, 2, 3\ldots$ denotes the number of corresponding zero of cosine, $\arg\Gamma(a)$ is the argument of gamma function. An inspection of expression (5) permits to formulate further to Lemma 1.

Lemma 2: If $c = 2\mathrm{Re}a$, $\mathrm{Re}x = 0$ ($x = \mathrm{j}z$ – positive purely imaginary), the function $\Phi(a, 2\mathrm{Re}a; \mathrm{j}z)$ has an infinite number of simple zeros $\zeta_{k,n}^{(c)}$ in z both for $k > 0$ and $k < 0$ ($k = -\mathrm{Im}a$, $n = 1, 2, 3\ldots$), at which $\mathrm{Re}\Phi = \mathrm{Im}\Phi = |\Phi| = 0$ [57, 58, 69].

Lemma 3: If $c = 2\mathrm{Re}a$, $\mathrm{Re}x = 0$ ($x = \mathrm{j}z$ – positive purely imaginary) and z exceeds the nth zero $\zeta_{k,n}^{(c)}$ of Kummer CHF $\Phi(a, 2\mathrm{Re}a; \mathrm{j}z)$ in z ($k = -\mathrm{Im}a$, $n = 1, 2, 3\ldots$) then its argument

$$\arg\Phi(a, 2\mathrm{Re}a; \mathrm{j}z) = (z/2) + n\pi \qquad (6)$$

is a linear function of z with finite increase by π at each consecutive zero of the function [57, 58, 69].

Lemma 4: If $c = 2\mathrm{Re}a$, $\mathrm{Re}x = 0$, ($x = \mathrm{j}z$ – positive purely imaginary), then for the real and imaginary parts of Kummer CHF it holds $\mathrm{Re}\Phi(a, 2\mathrm{Re}a; \mathrm{j}z) = 0$ for $z = (2m+1)\pi$, whereas $\mathrm{Im}\Phi(a, 2\mathrm{Re}a; \mathrm{j}z) = 0$ for $z = 2m\pi$, $m = 0, 1, 2, 3, \ldots$, irrespective of the value of $\mathrm{Im}a$, (k) [57, 58].

Corollary: An infinite decreasing (if $\mathrm{Re}a > 0$) or increasing (if $\mathrm{Re}a < 0$ and $\mathrm{Re}a \neq t/2$, $t = 0, -1, -2, -3, \ldots$) sequence of maxima of $|\Phi(a, 2\mathrm{Re}a; \mathrm{j}z)|$ and a sequence of its zeros alternate with each other when z grows in case $c = 2\mathrm{Re}a$, $\mathrm{Re}x = 0$ ($x = \mathrm{j}z$ – positive purely imaginary) [57].

5.2. Properties due to the numerical study

The statements of Lemmas 1–4 are confirmed by the numerical evaluation of the function $\Phi(1.5 - \mathrm{j}k, 3; \mathrm{j}z)$ made, using series (1). Figure 1 is a plot of the loci curves of Φ in the complex plane for $k = +0.5$, 0 and -0.5 (solid, dotted and dashed lines, respectively), Fig. 2 visualises the variation of $\mathrm{Re}\Phi$ (solid lines) and $\mathrm{Im}\Phi$ (dashed lines) versus z for $k = 0, \pm0.5, \pm1.0$ and Fig. 3 gives the dependence of modulus and argument of Φ on z for $k = +0.5$, 0 and -0.5 (solid, dotted and dashed lines, re-

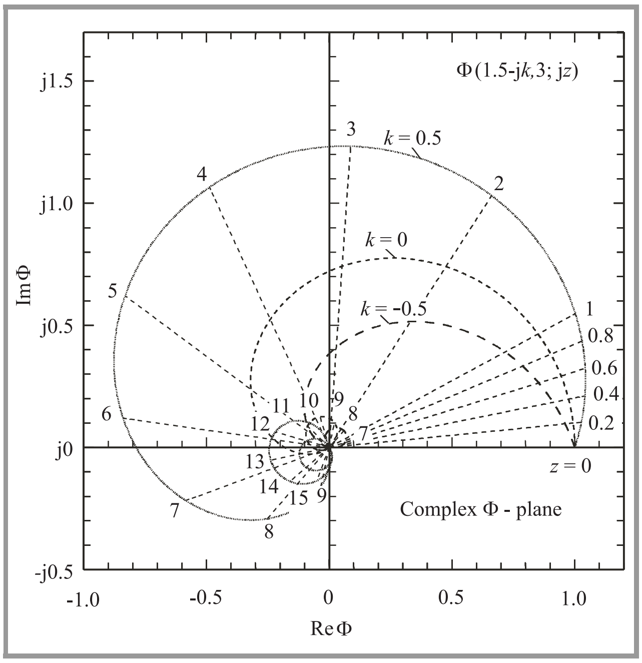


Fig. 1. Loci curves of $\Phi(1.5 - \mathrm{j}k, 3; \mathrm{j}z)$ in the complex plane for $k = +0.5$, 0 and -0.5.

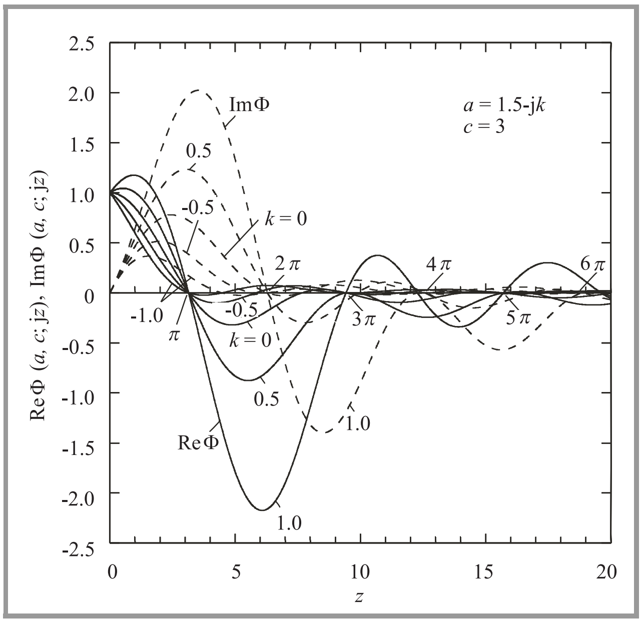


Fig. 2. Real and imaginary parts of Kummer function $\Phi(1.5 - \mathrm{j}k, 3; \mathrm{j}z)$ against z for $k = 0$, ±0.5 and ±1.0.

Table 1
First six positive purely imaginary zeros $\zeta_{k,n}^{(3)}$ of $\Phi(1.5-\mathrm{j}k,3;\mathrm{j}z)$ for $k=-1.0\,(0.2)+1.0$

k	$\zeta_{k,1}^{(3)}$	$\zeta_{k,2}^{(3)}$	$\zeta_{k,3}^{(3)}$	$\zeta_{k,4}^{(3)}$	$\zeta_{k,5}^{(3)}$	$\zeta_{k,6}^{(3)}$
−1.0	4.4750 5671	9.5777 9569	15.0744 6601	20.7758 5770	26.6000 3381	32.5053 0790
−0.8	4.9618 8564	10.3259 3914	15.9980 9339	21.8286 7627	27.7540 5190	33.7420 5957
−0.6	5.5218 6556	11.1477 3249	16.9911 7329	22.9469 7930	28.9703 0361	35.0384 2135
−0.4	6.1595 3442	12.0428 8636	18.0516 0729	24.1278 4699	30.2454 3063	36.3907 7149
−0.2	6.8751 0735	13.0069 8966	19.1734 8573	25.3647 3201	31.5725 5798	37.7920 4131
0.0	7.6634 1194	14.0311 7334	20.3469 3627	26.6473 8388	32.9412 6801	39.2317 1702
0.2	8.5142 1018	15.1029 6417	21.5590 7859	27.9628 4223	34.3385 7601	40.6968 5232
0.4	9.4140 5779	16.2082 5362	22.7959 6241	29.2973 5379	35.7509 1422	42.1739 8392
0.6	10.3489 2135	17.3336 0506	24.0447 2652	30.6384 5569	37.1660 9203	43.6511 3385
0.8	11.3063 8822	18.4679 5058	25.2951 0103	31.9763 7998	38.5746 8212	45.1191 1960
1.0	12.2767 8251	19.6032 3531	26.5398 8420	33.3044 4623	39.9703 5445	46.5718 6228

Table 2
First positive purely imaginary zeros $\zeta_{k,1}^{(3)}$ of $\Phi(1.5-\mathrm{j}k,3;\mathrm{j}z)$ and products $|k|\zeta_{k,1}^{(3)}$ and $|a|\zeta_{k,1}^{(3)}$ for large negative k

k	$\zeta_{k,1}^{(3)}$	$\lvert k\rvert\zeta_{k,1}^{(3)}$	$\lvert a\rvert$	$\lvert a\rvert\zeta_{k,1}^{(3)}$
−10000	0.00065 93654 06232	**6.59365 4**0623	10 000.00011 25000	**6.59365 41**365
−20000	0.00032 96827 04784	**6.59365 4**0956	20 000.00005 62500	**6.59365 41**142
−40000	0.00016 48413 52600	**6.59365 410**40	40 000.00002 81250	**6.59365 410**86
−60000	0.00010 98942 35093	**6.59365 410**55	60 000.00001 87500	**6.59365 410**76
−80000	0.00008 24206 76327	**6.59365 4106**1	80 000.00001 40625	**6.59365 410**72
−100000	0.00006 59365 41063	**6.59365 4106**2	100 000.00001 12500	**6.59365 410**70

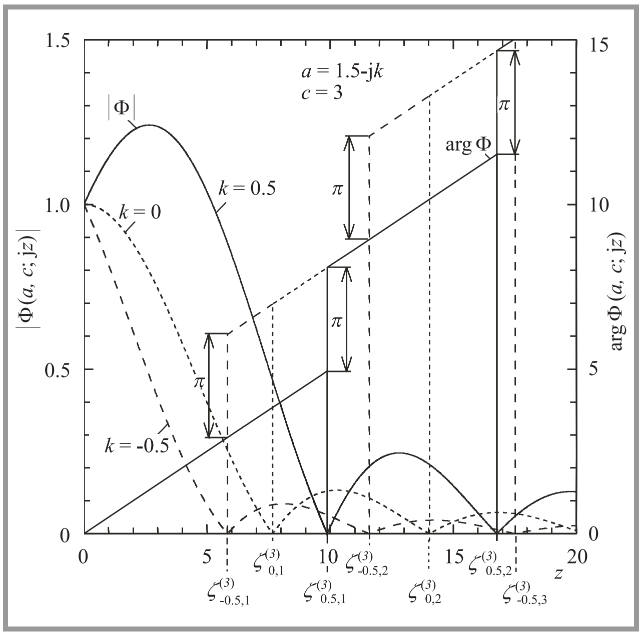


Fig. 3. Modulus and argument of Kummer function $\Phi(1.5-\mathrm{j}k,3;\mathrm{j}z)$ versus z for $k=+0.5$, 0 and -0.5.

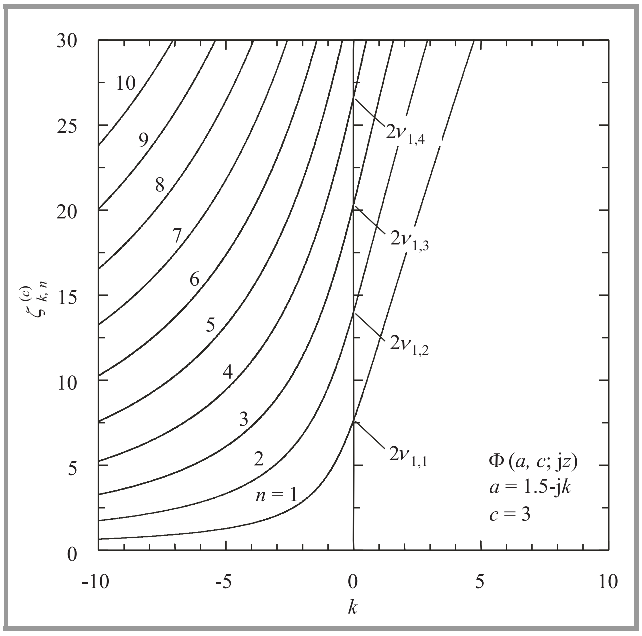


Fig. 4. Distribution of the first ten positive purely imaginary zeros of Kummer CHF $\Phi(1.5-\mathrm{j}k,3;\mathrm{j}z)$ with k.

spectively). The distribution of the first ten zeros of Φ with k is plotted in Fig. 4. The curves intersect the ordinate axis ($k=0$) at points $\zeta_{0,n}^{(3)}=2\nu_{1,n}$ [$\nu_{1,n}$ is the nth zero of Bessel function $J_1(x)$] which could be proved, using the second Kummer theorem [1–4, 7, 9, 10, 13, 16]. Values of $\zeta_{k,n}^{(c)}$ for small and large $|k|$ are listed in Tables 1 and 2. The analysis shows that it is true: $\lim\limits_{k\to-\infty}\zeta_{k,n}^{(c)}=0$ and $\lim\limits_{k\to+\infty}\zeta_{k,n}^{(c)}=+\infty$. The products $|k|\zeta_{k,n}^{(c)}$ and $|a|\zeta_{k,n}^{(c)}$ are of special interest, if k gets very large negative (see Table 2). It is valid [72]:

Lemma 5: If $\zeta_{k,n}^{(c)}$ is the nth positive purely imaginary zero of Kummer function $\Phi(a,c;x)$ in x $(n=1,2,3\ldots)$ provided $a=c/2-\mathrm{j}k$ – complex, $c=2\,\mathrm{Re}\,a$ – restricted positive integer, $x=\mathrm{j}z$ – positive purely imaginary, z – real, positive, $k=\mathrm{j}(a-c/2)$ – real, then the infinite sequences of positive real numbers $\{\zeta_{k,n}^{(c)}\}$, $\{|k|\zeta_{k,n}^{(c)}\}$ and $\{|a|\zeta_{k,n}^{(c)}\}$ are convergent for $k\to-\infty$ (c, n – fixed). The limit of the first sequence is zero and the limit of the second and third ones is the same. It equals the finite positive real number L, where $L=L(c,n)$. It holds:

$$\lim_{k\to-\infty}|k|\zeta_{k,n}^{(c)}=L(c,n)\,, \tag{7}$$

$$\lim_{k\to-\infty}|a|\zeta_{k,n}^{(c)}=L(c,n)\,. \tag{8}$$

For any $|k|$ and relevant $|a|$ it is true $|k|\zeta_{k,n}^{(c)}<L(c,n)<|a|\zeta_{k,n}^{(c)}$. In case $k\to+\infty$, $\{\zeta_{k,n}^{(c)}\}$, $\{|k|\zeta_{k,n}^{(c)}\}$, and $\{|a|\zeta_{k,n}^{(c)}\}$ also tend to $+\infty$. Results for complex Φ – function can be found in [55, 57–59, 72], too.

6. Some properties of the real Kummer function

6.1. Properties due to the analytical study by F. G. Tricomi

Tricomi has proved that if $\hat{a},\hat{c},\hat{x}$ are real, $\hat{x}>0$ and $\hat{c}>0$:

- the Kummer CHF $\Phi(\hat{a},\hat{c},\hat{x})$ has real positive zeros only if $\hat{a}<0$;
- the number of zeros $\hat{l}=abs[\hat{a}]$ is finite, $[\hat{a}]$ is the largest integer less or equal to $\hat{a}$, i.e., $[\hat{a}]\le\hat{a}$;
- at the point $\hat{a}=[\hat{a}]=-\hat{n}$ ($\hat{n}\le\hat{l}$ – a positive integer, $\hat{n}=1,2,\ldots,\hat{l}$) a new zero appears [1–3, 7, 9, 10, 44].

6.2. Properties due to the numerical study

The case $\hat{a}=\hat{c}/2+\hat{k}$ – real, $\hat{c}$ – positive integer, $\hat{k}$ – real ($\hat{k}=\hat{a}-\hat{c}/2$), $\hat{x}$ – real, positive, is treated. Computations of the function $\Phi(1.5+\hat{k},3;\hat{x})$ have been performed, making use of series (1). Figures 5 and 6 represent Φ versus $\hat{x}$ for $\hat{k}>0$ (solid lines), $\hat{k}=0$ (dotted curve) and $\hat{k}<0$ (dashed lines). The monotonous (oscillating) character of curves for $\hat{k}>-1.5$ ($\hat{k}<-1.5$) is in agreement with above analytical results. Values of the real zeros $\hat{\zeta}_{\hat{k},\hat{n}}^{(\hat{c})}$ of the same function are given in Tables 3–5 for different intervals of variation of $\hat{k}$. The distribution of the first eight zeros of Φ against $\hat{k}$ is drawn in Fig. 7. The numerical analysis indicates that it holds: $\lim\limits_{\hat{k}\to-\infty}\hat{\zeta}_{\hat{k},\hat{n}}^{(\hat{c})}=0$ and $\lim\limits_{\substack{\hat{k}\to-(\hat{n}-1)-\hat{c}/2\\ \hat{k}<-(\hat{n}-1)-\hat{c}/2}}\hat{\zeta}_{\hat{k},\hat{n}}^{(\hat{c})}=+\infty$.

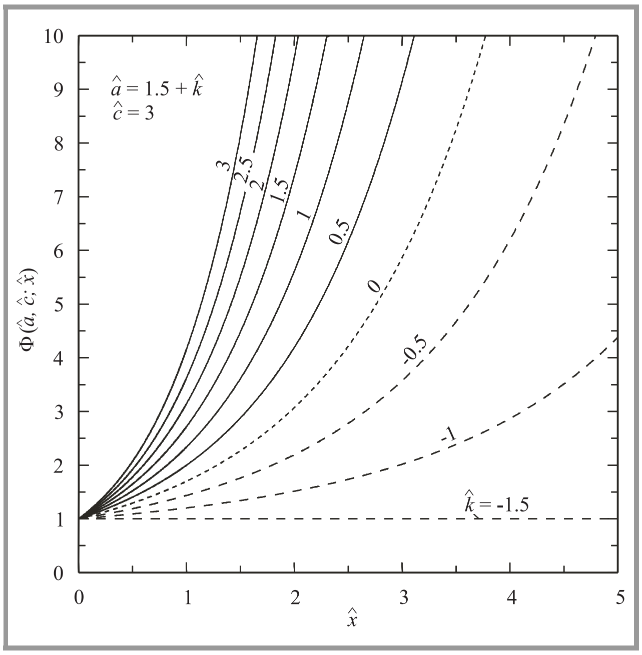


Fig. 5. Kummer function $\Phi(1.5+\hat{k},\ 3;\ \hat{x})$ against $\hat{x}$ for $\hat{k}=-1.5(0.5)3$.

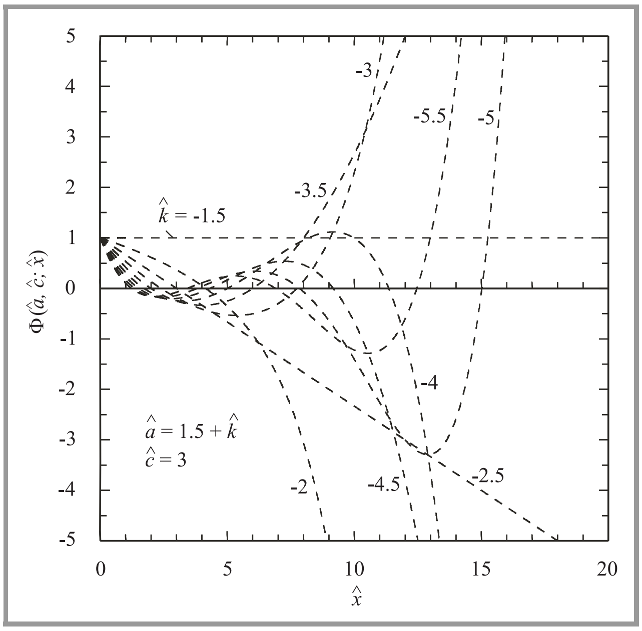


Fig. 6. Kummer function $\Phi(1.5+\hat{k},\ 3;\ \hat{x})$ versus $\hat{x}$ for $\hat{k}=-5(0.5)-1.5$.

The products $|\hat{k}|\hat{\zeta}_{\hat{k},\hat{n}}^{(\hat{c})}$ and $|\hat{a}|\hat{\zeta}_{\hat{k},\hat{n}}^{(\hat{c})}$ are of interest, if $\hat{k}$ is very large negative (Table 5). It is true:

Lemma 6: If $\hat{\zeta}_{\hat{k},\hat{n}}^{(\hat{c})}$ is the $\hat{n}$th positive real zero of Kummer function $\Phi(\hat{a},\hat{c};\hat{x})$ in $\hat{x}$ ($\hat{n}=1,2,\ldots,\hat{l}$, $\hat{l}=abs[\hat{a}]$) provided $\hat{a},\hat{c},\hat{x}$ are real, $\hat{c}$ – restricted positive integer and $\hat{k}=\hat{a}-\hat{c}/2$ – real ($\hat{a}=\hat{c}/2+\hat{k}$), then

Table 3
First six positive real zeros $\hat{\zeta}^{(3)}_{\hat{k},\hat{n}}$ of $\Phi(1.5+\hat{k},\,3;\,\hat{x})$ for $\hat{k} = -\left[(2\hat{n}+1)/2+1.10^{-\hat{s}}\right]$ and $\hat{s} = 2(1)10$

$\hat{s}$	$\hat{\zeta}^{(3)}_{\hat{k}(\hat{s}),1}$	$\hat{\zeta}^{(3)}_{\hat{k}(\hat{s}),2}$	$\hat{\zeta}^{(3)}_{\hat{k}(\hat{s}),3}$	$\hat{\zeta}^{(3)}_{\hat{k}(\hat{s}),4}$	$\hat{\zeta}^{(3)}_{\hat{k}(\hat{s}),5}$	$\hat{\zeta}^{(3)}_{\hat{k}(\hat{s}),6}$
10	32.6943 6952	39.2832 5273	45.2869 3680	50.9779 8646	56.4636 0593	61.8003 1150
9	30.1381 7435	36.6025 7840	42.4984 2791	48.0931 9869	53.4911 8502	58.7470 5165
8	27.5553 2227	33.8846 6661	39.6641 8688	45.1555 0085	50.4595 9370	55.6290 4210
7	24.9390 3482	31.1204 5777	36.7733 9084	42.1526 1734	47.3553 1077	52.4316 6489
6	22.2793 6643	28.2967 9449	33.8104 1524	39.0668 9583	44.1589 4739	49.1340 0339
5	19.5607 1308	25.3932 7731	30.7511 9691	35.8712 6945	40.8408 6789	45.7041 6336
4	16.7561 8418	22.3751 7209	27.5550 4993	32.5201 8558	37.3513 6214	42.0887 8082
3	13.8134 2126	19.1750 2405	24.1432 3161	28.9257 1322	33.5946 6329	38.1852 0648
2	10.6181 4852	15.6405 7545	20.3351 4451	24.8844 5526	29.3480 2383	33.7537 3550

Table 4
First six positive real zeros $\hat{\zeta}^{(3)}_{\hat{k},\hat{n}}$ of $\Phi(1.5+\hat{k},\,3;\,\hat{x})$ for $\hat{k} = -2(-1)-10$

$\hat{k}$	$\hat{\zeta}^{(3)}_{\hat{k},1}$	$\hat{\zeta}^{(3)}_{\hat{k},2}$	$\hat{\zeta}^{(3)}_{\hat{k},3}$	$\hat{\zeta}^{(3)}_{\hat{k},4}$	$\hat{\zeta}^{(3)}_{\hat{k},5}$	$\hat{\zeta}^{(3)}_{\hat{k},6}$
−2	4.1525 7778					
−3	2.3908 7384	7.7342 0261				
−4	1.7240 3430	4.9963 8913	11.3550 3906			
−5	1.3562 4234	3.8054 2722	7.8425 2881	15.0185 8200		
−6	1.1202 9295	3.0969 9425	6.1880 6299	10.8491 1987	18.7168 8187	
−7	0.9552 6444	2.6191 1978	5.1554 0981	8.7786 0273	13.9709 0761	22.4429 7395
−8	0.8330 6998	2.2725 0326	4.4346 8551	7.4417 8723	11.5221 5873	17.1799 4235
−9	0.7388 2652	2.0086 2555	3.8982 3676	6.4852 2265	9.9005 3270	14.3837 0603
−10	0.6638 7020	1.8005 8410	3.4814 0975	5.7592 2215	8.7176 8903	12.4948 1718

Table 5
First positive real zeros $\hat{\zeta}^{(3)}_{\hat{k},1}$ of $\Phi(1.5+\hat{k},\,3;\,\hat{x})$ and products $|\hat{k}|\hat{\zeta}^{(3)}_{\hat{k},1}$ and $|\hat{a}|\hat{\zeta}^{(3)}_{\hat{k},1}$ for large negative $\hat{k}$

$\hat{k}$	$\hat{\zeta}^{(3)}_{\hat{k},1}$	$\lvert\hat{k}\rvert\hat{\zeta}^{(3)}_{\hat{k},1}$	$\hat{a}$	$\lvert\hat{a}\rvert\hat{\zeta}^{(3)}_{\hat{k},1}$
−10000	0.00065 93654 15127	**6.59365 41**512	−9998.5	**6.59**266 51031
−20000	0.00032 96827 05895	**6.59365 41**179	−19998.5	**6.593**15 95938
−40000	0.00016 48413 52739	**6.59365 410**95	−39998.5	**6.593**40 68475
−60000	0.00010 98942 35134	**6.59365 410**80	−59998.5	**6.593**48 92666
−80000	0.00008 24206 76343	**6.59365 410**74	−79998.5	**6.593**53 04764
−100000	0.00006 59365 41072	**6.59365 410**72	−99998.5	**6.593**55 52023

the infinite sequences of positive real numbers $\left\{\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\right\}$, $\left\{|\hat{k}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\right\}$ and $\left\{|\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\right\}$ are convergent for $\hat{k} \to -\infty$ ($\hat{c}, \hat{n}$ – fixed). The limit of the first sequence is zero and the limit of the second and third ones is the same. It equals the finite positive real number $\hat{L}$, where $\hat{L} = \hat{L}(\hat{c}, \hat{n})$. It is valid:

$$\lim_{\hat{k}\to-\infty} |\hat{k}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}} = \hat{L}(\hat{c}, \hat{n})\,, \tag{9}$$

$$\lim_{\hat{k}\to-\infty} |\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}} = \hat{L}(\hat{c}, \hat{n})\,. \tag{10}$$

For any $|\hat{k}|$ and corresponding $|\hat{a}|$ it holds $|\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}} < \hat{L}(\hat{c}, \hat{n}) < |\hat{k}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}$. If $\hat{k} \to +\infty$, $\left\{\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\right\}$, $\left\{|\hat{k}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\right\}$ and $\left\{|\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\right\}$ go to $+\infty$, too.

Lemma 7: Let $\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}$ is the $\hat{n}$th positive real zero of Kummer function $\Phi(\hat{a}, \hat{c}; \hat{x})$ in $\hat{x}$ ($\hat{n} = 1, 2, \ldots, \hat{l}$, $\hat{l} = abs[\hat{a}]$) provided $\hat{a}, \hat{c}, \hat{x}$ are real, $\hat{c}$ – restricted positive integer and $\hat{k} = \hat{a} - \hat{c}/2$ – real ($\hat{a} = \hat{c}/2 + \hat{k}$). If $\hat{k} = -\left[(2\hat{n}+1)/2+1.10^{-\hat{s}}\right]$, and $\hat{s} = 1, 2, 3, \ldots$ is a positive integer, then the dif-

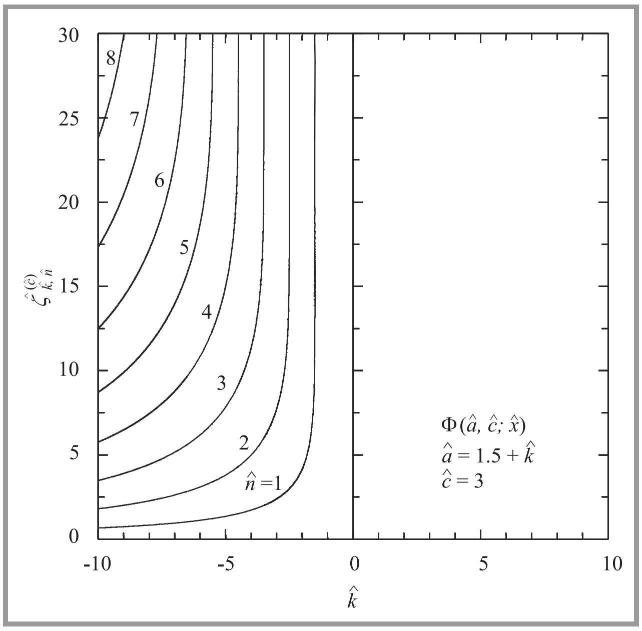


Fig. 7. Distribution of the first eight positive real zeros of Kummer function $\Phi(1.5+\hat{k},3;\hat{x})$ with $\hat{k}$.

ferences $\hat{\Delta}_{\hat{s}+1,\hat{s},\hat{n}}=\hat{\zeta}^{(\hat{c})}_{\hat{k}(\hat{s}+1),\hat{n}}-\hat{\zeta}^{(\hat{c})}_{\hat{k}(\hat{s}),\hat{n}}$ and $\hat{\Delta}^2_{\hat{s}+2,\hat{s}+1,\hat{s},\hat{n}}=\hat{\Delta}_{\hat{s}+1,\hat{s},\hat{n}}-\hat{\Delta}_{\hat{s}+2,\hat{s}+1,\hat{n}}$ where $\hat{k}(\hat{s}+1)$ and $\hat{k}(\hat{s})$ are any two neighbouring parameters for certain $\hat{n}$, tend to a finite real positive number and zero, respectively, especially if $\hat{s}$ gets large and $\hat{n}$ is small. Accordingly, the zeros $\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}$, situated close to the points $\hat{k}=-(2\hat{n}+1)/2$ can be computed from the approximate formula:

$$\hat{\zeta}^{(\hat{c})}_{\hat{k}(\hat{s}+2),\hat{n}}\approx\hat{\zeta}^{(\hat{c})}_{\hat{k}(\hat{s}+1),\hat{n}}+\hat{\Delta}_{\hat{s}+1,\hat{s},\hat{n}}=2\hat{\zeta}^{(\hat{c})}_{\hat{k}(\hat{s}+1),\hat{n}}-\hat{\zeta}^{(\hat{c})}_{\hat{k}(\hat{s}),\hat{n}}\,. \quad (11)$$

This relation permits to obtain the subsequent zero, if the values of the preceding two are known (Table 4). Results for real Φ – function are available also in [10, 11, 13, 14].

7. A theorem for the identity of zeros of certain Kummer functions

Theorem 1: If $\zeta^{(c)}_{k,n}$ is the nth positive purely imaginary zero of complex Kummer function $\Phi(a,c;x)$ in x ($n=1,2,3,\ldots$) provided $a=c/2-\mathrm{j}k$ – complex, $c=2\mathrm{Re}\,a$ – restricted positive integer, $x=\mathrm{j}z$ – positive purely imaginary, z – real, positive, $k=\mathrm{j}(a-c/2)$ – real, and if $\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}$ is the $\hat{n}$th positive real zero of real Kummer function $\Phi(\hat{a},\,\hat{c};\hat{x})$ in $\hat{x}$ ($\hat{n}=1,2,\ldots,\hat{l}$, $\hat{l}=abs[\hat{a}]$) provided $\hat{a},\hat{c},\hat{x}$ are real, $\hat{c}$ – restricted positive integer and $\hat{k}=\hat{a}-\hat{c}/2$ – real ($\hat{a}=\hat{c}/2+\hat{k}$), then the infinite sequences of positive real numbers $\{\zeta^{(c)}_{k,n}\}$, $\{|k|\zeta^{(c)}_{k,n}\}$ and $\{|a|\zeta^{(c)}_{k,n}\}$ are convergent for $k\to-\infty$ (c,n – fixed), and the infinite sequences of positive real numbers $\{\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$, $\{|\hat{k}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$ and $\{|\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$ are convergent for $\hat{k}\to-\infty$ ($\hat{c},\hat{n}$ – fixed). The limits of $\{\zeta^{(c)}_{k,n}\}$ and $\{\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$ equal zero. The limits of $\{|k|\zeta^{(c)}_{k,n}\}$ and $\{|a|\zeta^{(c)}_{k,n}\}$ coincide. They equal the positive real number L, where $L=L(c,n)$. The same is fulfilled for the limits of $\{|\hat{k}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$ and $\{|\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$ which equal the positive real number $\hat{L}$, where $\hat{L}=\hat{L}(\hat{c},\hat{n})$. On condition that $c=\hat{c}$ and $n=\hat{n}$, it is correct:

$$L(c,n)=\hat{L}(\hat{c},\hat{n})\,. \quad (12)$$

In addition, in case $k=\hat{k}$ – large negative, it is true:

$$\zeta^{(c)}_{k,n}\approx\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\,. \quad (13)$$

It holds $\zeta^{(c)}_{k,n}<\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}$ and $|\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}<|k|\zeta^{(c)}_{k,n}<L(c,n)<|a|\zeta^{(c)}_{k,n}<|\hat{k}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}$ for any $c=\hat{c}$, $n=\hat{n}$, $|k|=|\hat{k}|$ and $|a|\approx|\hat{a}|$, ($|\hat{a}|<|a|$). The rate of convergence decreases as follows $\{|k|\zeta^{(c)}_{k,n}\}$, $\{|a|\zeta^{(c)}_{k,n}\}$, $\{|\hat{k}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$ and $\{|\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$. When $c<<|k|$ ($\hat{c}<<|\hat{k}|$), the sequences $\{|\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$ and $\{|a|\zeta^{(c)}_{k,n}\}$ converge faster. For $c=3$ and $n=1,2,\ldots,10$, it is valid $L(c,n)=\hat{L}(\hat{c},\hat{n})$ =6.593654107, 17.71249973, 33.75517722, 54.73004731, 80.6387791, 111.48189218, 147.25958974, 187.9719664, 233.61907045, 284.20092871. Assuming that $k\to+\infty$ ($\hat{k}\to+\infty$), $\{\zeta^{(c)}_{k,n}\}$, $\{|k|\zeta^{(c)}_{k,n}\}$ and $\{|a|\zeta^{(c)}_{k,n}\}$, ($\{\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$, $\{|\hat{k}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$ and $\{|\hat{a}|\hat{\zeta}^{(\hat{c})}_{\hat{k},\hat{n}}\}$) tend to $+\infty$. The proof of Theorem 1 is based on the numerical study of the zeros of Kummer CHF, respectively on Lemmas 5 and 6 (Tables 2 and 5).

8. Azimuthally magnetized circular ferrite waveguide

An infinitely long, homogeneous, perfectly conducting circular waveguide of radius r_0, entirely filled with lossless ferrite, magnetized in azimuthal direction to remanence by an infinitely thin switching wire, is considered. The anisotropic load has a scalar permittivity $\varepsilon=\varepsilon_0\,\varepsilon_r$ and a Polder permeability tensor $\overleftrightarrow{\mu}=\mu_0\left[\mu_{ij}\right]$, $i,j=1,2,3$ of nonzero components $\mu_{ii}=1$ and $\mu_{13}=-\mu_{31}=-\mathrm{j}\alpha$, ($\alpha=\gamma M_r/\omega,\gamma$ – gyromagnetic ratio, M_r – remanent magnetization, ω – angular frequency of the wave). The propagation of normal and slow rotationally symmetric TE modes in the structure is examined. The following quantities are used in the study of the fields of first type: β – phase constant of the wave in the guide, $\beta_f=\beta_1\sqrt{\mu_{eff}}$, $\beta_1=\beta_0\sqrt{\varepsilon_r}$, $\beta_0=\omega\sqrt{\varepsilon_0\mu_0}$ – natural propagation constants of the unbounded azimuthally magnetized ferrite and dielectric media of relative permittivity ε_r and of free space, respectively, $\mu_{eff}=1-\alpha^2$ – effective relative permeability and $\beta_2=\left(\beta_f^2-\beta^2\right)^{1/2}$ – transverse distribution coefficient. The expressions: $\overline{\beta}=\beta/\left(\beta_0\sqrt{\varepsilon_r}\right)$, $\overline{\beta}_f=\beta_f/\left(\beta_0\sqrt{\varepsilon_r}\right)$, $\overline{\beta}_2=\beta_2/\left(\beta_0\sqrt{\varepsilon_r}\right)$ and $\overline{r}_0=\beta_0 r_0\sqrt{\varepsilon_r}$ provide universality of the results.

9. A microwave application of the complex Kummer function

9.1. Propagation problem for normal $\boldsymbol{TE_{0n}}$ *modes in an azimuthally magnetized circular ferrite waveguide*

The guided TE_{0n} waves in configuration described are normal, if $\bar\beta_2 = \left(\bar\beta_f^2 - \bar\beta^2\right)^{1/2}$ is real $\left(\bar\beta_f = \sqrt{\mu_{eff}}\right)$, i.e., $\bar\beta < \bar\beta_f$, $\left(\bar\beta > 0,\ \bar\beta_f > 0\right)$. They are governed by the following characteristic equation [54, 55, 57, 59–61, 63, 66, 69, 70, 72–74]:

$$\Phi\left(a, c; x_0\right) = 0\,, \tag{14}$$

where $a = 1.5 - \mathrm{j}k$, $c = 3$, $x_0 = \mathrm{j}z_0$, $k = \alpha\bar\beta/\left(2\bar\beta_2\right)$, $z_0 = 2\bar\beta_2\bar r_0$. It holds, provided $\bar\beta_2 = \zeta^{(c)}_{k,n}/\left(2\bar r_0\right)$ which defines the eigenvalue spectrum of the fields examined.

9.2. Phase characteristics

Using the roots $\zeta^{(3)}_{k,1}$ of Eq. (14) and the relations between barred quantities, the dependence of $\bar\beta$ on $\bar r_0$ with α as parameter for the normal TE_{01} mode in the ferrite waveguide is computed and plotted in Fig. 8. The solid (dashed) lines, corresponding to positive (negative) magnetization are of infinite (finite) length. Hence, transmission is possible for $\alpha_+ > 0$ $\left(\alpha_- < 0\right)$ in an unlimited from above (restricted from both sides) frequency band. The common starting point of the curves for the same $|\alpha|$ at the horizontal axis depicts the pertinent cutoff frequency $\bar r_{0cr} = \left[\zeta^{(3)}_{0,1}/2\right]/\left(1-\alpha^2\right)^{1/2}$. The ends of characteristics for $M_r < 0$ of co-ordinates $(\bar r_{0en-},\ \bar\beta_{en-})$ form an envelope (dotted line), labelled with En_{1-}, limiting from the right the area of propagation for negative magnetization. The curve, marked with $\alpha = 0$ (the ferrite degenerates into isotropic dielectric) is infinitely long (transmission takes place in an unlimited from above frequency range). The characteristics for $\alpha_+ > 0$ $(\alpha_- < 0)$ are single-valued (double-valued below cut-off, with an inversion point of abscissa $\bar r_{0i-}$). The envelope En_{1-} possesses a minimum $\min \bar r_{0right-} = 2L(3,1)$ at $\alpha_{-,\min} = -1\sqrt{2}$ and $\bar\beta_{right-,\min} = 1/\sqrt{2}$.

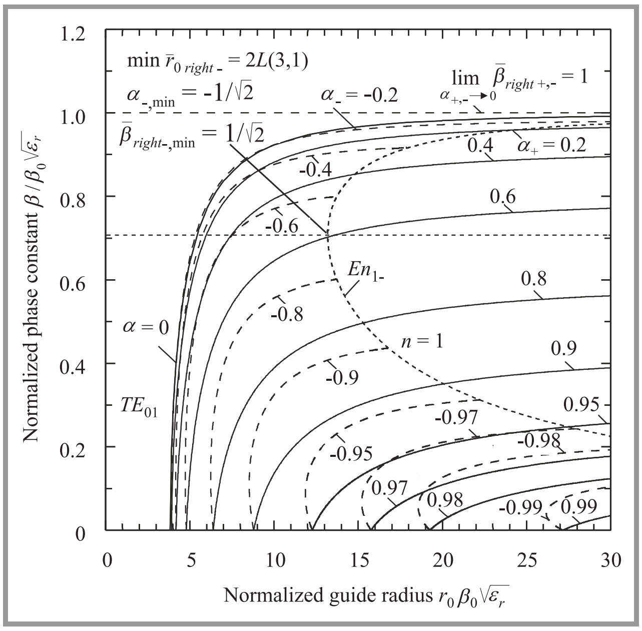


Fig. 8. Phase curves $\bar\beta(\bar r_0)$ of the normal TE_{01} mode in the circular ferrite waveguide.

9.3. Propagation conditions

Integrating the results of analysis of complex Kummer CHFs and of the problem studied, it turns out that the normal TE_{0n} waves propagate in one region whose boundaries for $M_r > 0$ and $M_r < 0$ are determined by the terms: $\alpha_{left+,-} < \alpha_{+,-} < \alpha_{right+,-}$, $k_{left+,-} < k_{+,-} < k_{right+,-}$, $\beta_{left+,-} < \bar\beta_{+,-} < \beta_{right+,-}$, $\bar r_{0left+,-} < \bar r_{0+,-} < \bar r_{0right+,-}$, where $\alpha_{left-} = -1$, $\alpha_{right-} = 0$, $\alpha_{left+} = 0$, $\alpha_{right+} = 1$, $k_{left-} = -\infty$, $k_{right-} = 0$, $k_{left+} = 0$, $k_{right+} = +\infty$, $\bar\beta_{left+,-} = 0$, $\bar\beta_{right+,-} = \left(1 - \alpha^2_{+,-}\right)^{1/2}$, $\bar r_{0left+} = \bar r_{0cr}$, $\bar r_{0right+} = +\infty$, $\bar r_{0left-} = \bar r_{0i-}$, $\bar r_{0right-} = \bar r_{0en-} = L(c,n)/\left[|\alpha_-|\left(1-\alpha_-^2\right)^{1/2}\right]$. Moreover $\bar\beta_{right} = \bar\beta_{en-}$. The subscripts "*left*", "*right*" designate the limits of domain in which certain quantity varies and the ones "+", "–" show the sign of magnetization to which the latter is relevant.

9.4. Phaser operation

The waveguide may provide differential phase shift $\Delta\bar\beta = \bar\beta_- - \bar\beta_+$ for TE_{01} mode when latching M_r in the area of partial overlapping $\Delta = \bar r_{0right-} - \bar r_{0left+} = \bar r_{0en-} - \bar r_{0cr}$ of the intervals $\Delta_- = \left(\bar r_{0left-},\ \bar r_{0right-}\right)$, and $\Delta_+ = \left(\bar r_{0left+},\ \bar r_{0right+}\right)$, pertinent to $\bar\beta_-(\bar r_{0-})$ and $\bar\beta_+(\bar r_{0+})$ curves for the same $|\alpha|$ ($\Delta = \Delta_- \cap \Delta_+$, Fig. 8 and Fig. 1 [74]). Hence, the condition for the geometry to operate as phaser at fixed $|\alpha|$ (the working point $\bar r_0$ to be part of Δ), is $\bar r_{0cr} < \bar r_0 < \bar r_{0en-}$, or [69]:

$$\zeta^{(3)}_{0,1}|\alpha| < 2\bar r_0|\alpha|\sqrt{1-\alpha^2} < 2L(3,1)\,. \tag{15}$$

Save from the graphs, $\Delta\bar\beta$ could be computed also directly from structure parameters, using the formulae $\Delta\bar\beta = A|\alpha|$, $\Delta\bar\beta = B/\bar r_0$, $\Delta\bar\beta = (C/\bar r_0)|\alpha|$ [66, 74]. The values of factors A, B, C are tabulated in [66, 74]. If $\bar r_0 > \bar r_{0en-}$, the configuration has potentialities as current controlled switch or isolator.

10. A microwave application of the real Kummer function

10.1. Propagation problem for slow $\boldsymbol{\widehat{TE}_{0\hat n}}$ *modes in an azimuthally magnetized circular ferrite waveguide*

The guided $\widehat{TE}_{0\bar n}$ waves examined are slow, if $\bar{\hat\beta}_2 = \left(\bar{\hat\beta}^2 - \hat\beta_f^2\right)^{1/2}$ is real ($\hat\beta_f^2 = \hat\mu_{eff}$, $\hat\mu_{eff} = 1 - \hat\alpha^2$), i.e., provided $\bar{\hat{\beta}}^2 > \bar{\hat{\beta}}_f^2$, ($\bar{\hat{\beta}} > 0$, $\bar{\hat{\beta}}_f^2 > 0$, $\bar{\hat{\beta}}_f^2 < 0$ or $\bar{\hat{\beta}}_f^2 = 0$). The solution of Maxwell equations subject to boundary condition at the wall $\bar{\hat{r}} = \bar{\hat{r}}_0$ yields the corresponding characteristic equation [69]:

$$\Phi\left(\hat{a}, \hat{c}; \hat{x}_0\right) = 0 \tag{16}$$

with $\hat{a} = 1.5 + \hat{k}$, $\hat{c} = 3$, $\hat{x}_0 = 2\bar{\hat{\beta}}_2\bar{\hat{r}}_0$, $\hat{k} = \hat{\alpha}\bar{\hat{\beta}}/\left(2\bar{\hat{\beta}}_2\right)$. It is valid in case $\bar{\hat{\beta}}_2 = \hat{\zeta}_{\hat{k},\hat{n}}^{(\hat{c})}/\left(2\bar{\hat{r}}_0\right)$, giving the eigenvalue spectrum looked for. (Equation (16) could be obtained from (14) putting $k = \mathrm{j}\hat{k}$ and $\bar{\beta}_2 = -\mathrm{j}\hat{\beta}_2$).

10.2. Propagation conditions

Combining the outcomes of the study of real Kummer CHFs and of the problem regarded, it is found that the slow $\widehat{TE}_{0\hat{n}}$ modes could be guided for $\hat{M}_r < 0$ solely in two areas, set by the criteria: $\hat{\alpha}_{left-}^{(1),(2)} < \hat{\alpha}_{-}^{(1),(2)} < \hat{\alpha}_{right-}^{(1),(2)}$, $\hat{k}_{left-}^{(1),(2)} < \hat{k}_{-}^{(1),(2)} < \hat{k}_{right-}^{(1),(2)}$, $\bar{\hat{\beta}}_{left-}^{(1)} < \bar{\hat{\beta}}_{-}^{(1)} < \bar{\hat{\beta}}_{right-}^{(1)}$, $\bar{\hat{\beta}}_{left-}^{(2)} > \bar{\hat{\beta}}_{-}^{(2)} > \bar{\hat{\beta}}_{right-}^{(2)}$, $\bar{\hat{r}}_{left-}^{(1),(2)} < \bar{\hat{r}}_{0-}^{(1),(2)} < \bar{\hat{r}}_{0right-}^{(1),(2)}$, with $\hat{\alpha}_{left-}^{(1)} = -1$, $\hat{\alpha}_{right-}^{(1)} = 0$, $\hat{\alpha}_{left-}^{(2)} = -\infty$, $\hat{\alpha}_{right-}^{(2)} = -(2\hat{n}+1)$, $\hat{k}_{left-}^{(1)} = -\infty$, $\hat{k}_{right-}^{(1)} = -(2\hat{n}+1)/2$, $\hat{k}_{left-}^{(2)} = \hat{\alpha}_{-}^{(2)}/2$, $\hat{k}_{right-}^{(2)} = -(2\hat{n}+1)/2$, $\bar{\hat{\beta}}_{left-}^{(1)} = \left[1 - \left(\hat{\alpha}_{-}^{(1)}\right)^2\right]^{1/2}$, $\bar{\hat{\beta}}_{right-}^{(1)} = \left\{\left[1 - \left(\hat{\alpha}_{-}^{(1)}\right)^2\right] / \left[1 - \left(\hat{\alpha}_{-}^{(1)}/(2\hat{n}+1)\right)^2\right]\right\}^{1/2}$, $\bar{\hat{\beta}}_{right-}^{(2)} = \left\{\left[\left(\hat{\alpha}_{-}^{(2)}\right)^2 - 1\right] / \left[\left(\hat{\alpha}_{-}^{(2)}/(2\hat{n}+1)\right)^2 - 1\right]\right\}^{1/2}$, $\bar{\hat{\beta}}_{left-}^{(2)} = +\infty$, $\bar{\hat{r}}_{0left-}^{(1)} = \hat{L}\left(\hat{c}, \hat{n}\right) / \left\{\left|\hat{\alpha}_{-}^{(1)}\right| \left[1 - \left(\hat{\alpha}_{-}^{(1)}\right)^2\right]^{1/2}\right\}$, $\bar{\hat{r}}_{0right-}^{(1)} = +\infty$, $\bar{\hat{r}}_{0left-}^{(2)} = 0$, $\bar{\hat{r}}_{0right-}^{(2)} = +\infty$. The superscripts (1), (2) designate the zone to which the corresponding quantity relates. Thus, the symbol $\widehat{TE}_{0\hat{n}}$ is a general notation for two waves $\widehat{TE}_{0\hat{n}}^{(1)}$ and $\widehat{TE}_{0\hat{n}}^{(2)}$, supported in the first and second regions, respectively.

10.3. Phase characteristics

Taking into account the propagation conditions and repeating the procedure, described in Subsection 9.2 with the roots $\hat{\zeta}_{\hat{k},1}^{(3)}$ of Eq. (16), the $\bar{\hat{\beta}}_{-}^{(1)}\left(\bar{\hat{r}}_{0-}^{(1)}\right)$ and $\bar{\hat{\beta}}_{-}^{(2)}\left(\bar{\hat{r}}_{0-}^{(2)}\right)$ – characteristics with $\hat{\alpha}_{-}^{(1)}$ and $\hat{\alpha}_{-}^{(2)}$ as parameters for the slow $\widehat{TE}_{01}^{(1)}$ and $\widehat{TE}_{01}^{(2)}$ modes, respectively in the structure are computed and presented with dashed curves of infinite length in Figs. 9 and 10, respectively. Thus, transmission takes place for $\hat{\alpha}_{-} < 0$ in two unlimited from above frequency bands. An envelope (dotted line), labelled with $\hat{E}n_{1-}$ (for the co-ordinates of the points of which $(\bar{\hat{r}}_{0en-}^{(1)}, \bar{\hat{\beta}}_{en-}^{(1)})$ it is valid $\bar{\hat{r}}_{0en-}^{(1)} = \bar{\hat{r}}_{0left-}^{(1)}$ and $\bar{\hat{\beta}}_{en-}^{(1)} = \bar{\hat{\beta}}_{left-}^{(1)}$), restricts from the left the area of propagation in case of weak anisotropy (Fig. 9). It has a minimum $\min \bar{\hat{r}}_{0left-} = 2\hat{L}\,(3, 1)$ at $\hat{\alpha}_{-,\min}^{(1)} = -1\sqrt{2}$ and $\bar{\hat{\beta}}_{left-,\min}^{(1)} = 1\sqrt{2}$. A comparison of both sets of curves shows that a large slowing down is provided if the anisotropy is strong, especially in case $\bar{\hat{r}}_0^{(2)}$ is small (see Fig. 10). Ferrite switches and isolators are the possible applications of the structure.

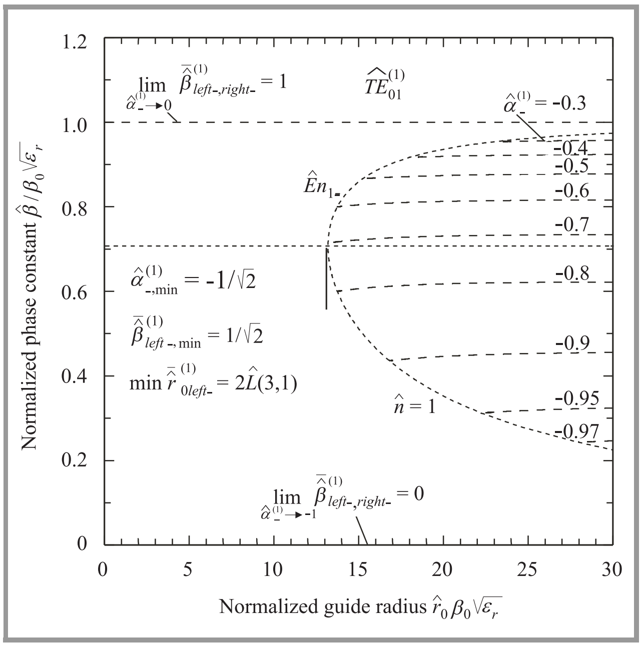


Fig. 9. Phase curves $\bar{\hat{\beta}}_{-}^{(1)}\left(\bar{\hat{r}}_{0-}^{(1)}\right)$ of the slow $\widehat{TE}_{01}^{(1)}$ mode in the circular ferrite waveguide for $-1 < \hat{\alpha}_{-}^{(1)} < 0$.

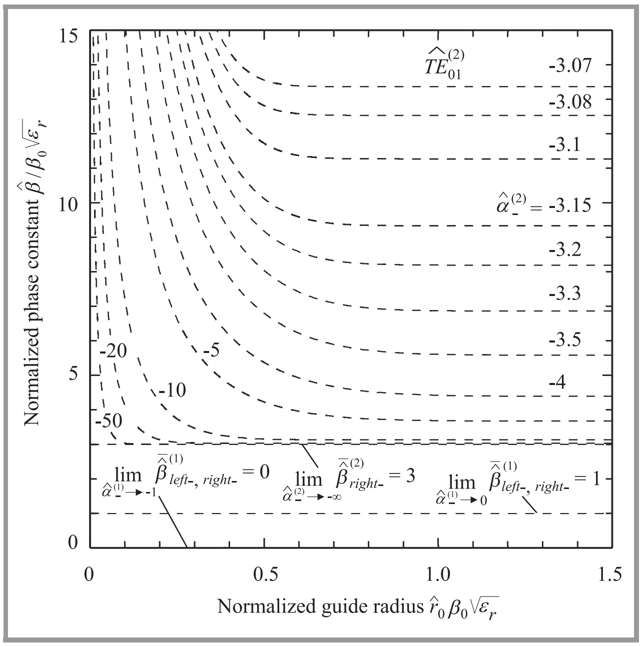


Fig. 10. Phase curves $\bar{\hat{\beta}}_{-}^{(2)}\left(\bar{\hat{r}}_{0-}^{(2)}\right)$ of the slow $\widehat{TE}_{01}^{(2)}$ mode in the circular ferrite waveguide for $\hat{\alpha}_{-}^{(2)} < -3$.

11. Areas of TE_{01} mode propagation

The joint consideration of the results of analysis of the anisotropic waveguide shows that in case of positive (counterclockwise) magnetization there is one (densely hatched) area in which normal TE_{01} mode is supported (Fig. 11). If the magnetization is negative (clockwise), the areas are already three: one (densely hatched) of normal and two (sparsely hatched) of slow wave propagation (Fig. 12).

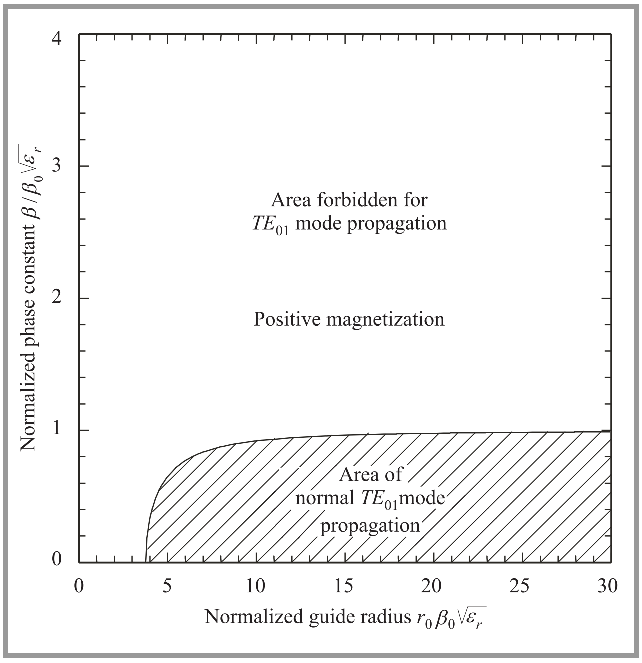


Fig. 11. Areas of TE_{01} mode propagation in case of positive magnetization.

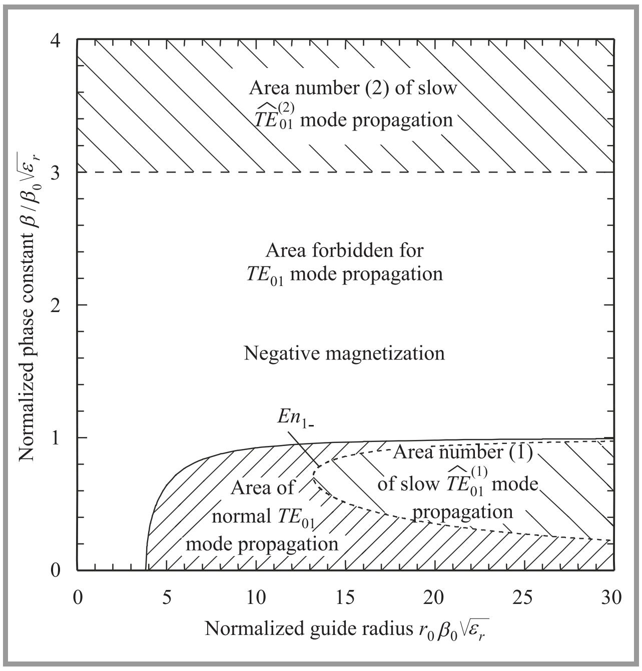


Fig. 12. Areas of TE_{01} mode propagation in case of negative magnetization.

An important corollary of Theorem 1 is the coincidence of the envelopes of characteristics for $M_r < 0$ of the normal (Fig. 8) and the slow mode (Fig. 9) in one curve (the dotted line in Fig. 12, labelled En_{1-}) which does not belong to any of the zones and serves as their common border, delimiting them. Indeed, since in view of Eq. (12) $L(3,1) = \hat{L}(3,1)$, the points $(\overline{r}_{0en-},\ \overline{\beta}_{en-})$ and $(\overline{\hat{r}}_{0en-}^{(1)},\ \overline{\hat{\beta}}_{en-}^{(1)})$ in the $\overline{r}_0 - \overline{\beta}$ phase plane, forming the En_{1-} and $\hat{E}n_{1-}$ – characteristics for the TE_{01} and $\widehat{TE}_{01}^{(1)}$ modes, respectively, are identical for all values of parameters $\alpha_- = \hat{\alpha}_-^{(1)}$ whose intervals of variation, determined by the corresponding propagation conditions in Sections 9.3 and 10.2, are the same. Area number (2) for the slow wave is separated from aforesaid two ones by a region where no transmission is allowed.

12. Conclusions

Some basic concepts of the theory, the special cases and examples of the use of the CHFs in different fields of physics are considered. The opinion is declared that a universal mathematical procedure, based on them would successfully substitute the methods for analysis of a large number of tasks, utilizing the numerous functions which are their special cases. This approach would make possible to reveal the interior connections among plenty of phenomena and would facilitate the physical interpretation of the results from their description, as well as the process of their generalization.

The problems for normal and slow rotationally symmetric TE modes in the circular waveguide, uniformly filled with azimuthally magnetized ferrite are threshed out as a sphere of microwave application of the complex, and real Kummer CHFs rescpectively. The propagation conditions and phase characteristics of the structure are obtained, using various properties of the wave function, established analytically and/or numerically. The main result of the study is that for positive (negative) magnetization one area of normal (three areas – one of normal and two of slow) TE_{01} mode propagation exists (exist). The region of normal and the first one of slow waves transmission in case of negative magnetization are demarcated by an envelope curve which can be traced by means of a numerically proved theorem for identity of the zeros of certain Kummer functions. The areas mentioned are separated from the second one for slow wave propagation by a zone in which no fields can be sustained. The phase behaviour reveals the potentialities of the structure as a remanent phaser (for normal waves) or as a current controlled switch and isolator (for both kinds of waves). The criterion for phaser operation of waveguide is deduced as a direct corollary of the aforesaid theorem for the zeros. A large number of configurations, containing a central ferrite rod of azimuthally magnetized ferrite, coated by an arbitrary number of dielectric layers could be described, extending the analysis method based on the Kummer CHFs.

References

[1] F. G. Tricomi, "Sulle funzioni ipergeometriche confluenti", *Ann. Math. Pura Appl.*, vol. 36, no. 4, pp. 141–175, 1947.

[2] F. G. Tricomi, *Lezioni sulle Funzioni Ipergeometriche Confluenti*. Torino: Editore Gheroni, 1952.

[3] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions*. Bateman project, vol. I, II. New York, Toronto, London: McGraw-Hill, 1953.

[4] F. G. Tricomi, *Equazioni Differenziali*. Torino: Edizioni Scientifiche Einaudi, 1953.

[5] H. Buchholz, *Die Konfluente Hypergeometrische Funktion mit Besonderer Berücksichtigung Ihrer Anwendungen*. Berlin, Göttingen, Heidelberg: Springer-Verlag, 1953.

[6] Ph. M. Morse and H. Feshbach, *Methods of Theoretical Physics*. Part I. New York: McGraw-Hill, 1953.

[7] F. G. Tricomi, *Funzioni Ipergeometriche Confluenti*. Rome: Edizioni Cremonese, 1954.

[8] E. L. Ince, *Ordinary Differential Equations*. New York: Dover Publications, 1956.

[9] F. G. Tricomi, *Fonctions Hypergéométriques Confluentes*. Paris: Gauthier-Villars, 1960.

[10] L. J. Slater, *Confluent Hypergeometric Functions*. Cambridge: Cambridge University Press, 1960.

[11] E. Janke, F. Emde, and F. Lösch, *Tafeln Höherer Funktionen*. Stuttgart: B.G. Teubner Verlagsgesellschaft, 1960.

[12] N. N. Lebedev, *Special Functions and Their Applications*. Moscow, Leningrad: State Publishing House of Physical and Mathematical Literature, 1963 (in Russian).

[13] *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. M. Abramowitz and I. Stegun, Eds. Applied Mathematics, Series 55. Washington: National Bureau of Standards, 1964.

[14] M. I. Zhurina and L. N. Osipova, *Tables of the Confluent Hypergeometric Function*. Moscow: Computational Center of the Academy of Sciences of USSR, 1964 (in Russian).

[15] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*. Cambridge: Cambridge University Press, 1965.

[16] W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*. Berlin, Heidelberg, New York: Springer-Verlag, 1966.

[17] A. R. Curtis, *Coulomb Wave Functions*. Moscow: Computational Center of the Academy of Sciences of USSR, 1969 (in Russian).

[18] L. N. Osipova, *Tables of the Confluent Hypergeometric Function of the Second Kind*. Moscow: Computational Center of the Academy of Sciences of USSR, 1972 (in Russian).

[19] E. Kamke, *Handbook of Ordinary Differential Equations*. Moscow: Nauka, 1976 (in Russian).

[20] S. Flügge, *Rechenmethoden der Quantentheorie*. Teil 1: *Elementare Quantenmechanik*. Band 53 der Grundlehren der mathematischen Wissenschaften, Berlin, 1947.

[21] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics. Nonrelativistic Theory*. Moscow: Nauka, 1974 (in Russian).

[22] A. Sommerfeld, *Atombau und Spektrallinien*. Band II, Braunschweig, 1951.

[23] J. Formánek: *Úvod do Kvantové Teorie*. Prague: Academia, 1983.

[24] R. E. Collin, *Foundations of Microwave Engineering*. New York, Toronto, London: Mc-Graw-Hill, 1966.

[25] A. E. Siegman, *Microwave Solid-State Masers*. New York, Toronto, London: Mc-Graw-Hill, 1964.

[26] R. Gran Olsson, "Biegung kreisformigen Platten von radial verändlicher Dicke", *Ingenieur-Arch.*, vol. 8, pp. 81–98, 1937.

[27] G. N. Watson, *A Treatise on the Theory of Bessel Functions*. Cambridge: Cambridge University Press, 1966.

[28] H. Buchholz, "Die Ausbreitung von Schallwellen in einem Horn von der Gestalt eines Rotationsparaboloids bei Anregung durch eine im Brennpunkt gelegene punktformige Schallquelle", *Ann. Phys.*, vol. 42, pp. 423–460, 1943.

[29] A. Erdélyi, "Inhomogene Saiten mit parabolischer Dichteverteilung", *S.-B. Akad. Wiss. Wien, Math.-Phys.*, Kl. IIa, vol. 146, pp. 431–467, 1937.

[30] H. A. Lauwerier, "The use of confluent hypergeometric functions in mathematical physics and the solution of an eigenvalue problem", *Appl. Sci. Res.*, vol. A2, pp. 184–204, 1950.

[31] W. Magnus, "Zur Theorie des zylindrisch-parabolischen Spiegels", *Z. Phys.*, vol. 11, pp. 343–356, 1941.

[32] M. Hashimoto, S. Nemoto, and T. Makimoto, "Analysis of guided waves along the cladded optical fiber: paraboloc-index core and homogeneous cladding", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-25, no. 1, pp. 11–17, 1977.

[33] J.-D. Decotignie and F. E. Gardiol, "Méthodes d'analyse de la propagation dans les fibres optiques", *Bulletin ASE/UCS*, vol. 70, no. 15, pp. 830–837, 1979.

[34] M. Hashimoto, "Asymptotic theory of vector modes in inhomogeneous optical fibres: uncladded fibres", *Proc. Inst. Elec. Eng.*, Pt. H *(Microwaves, Optics and Antennas)*, vol. 130, no. 4, pp. 261–275, 1983.

[35] R. G. Mirimanov, "Solution of a diffraction problem for a plane electromagnetic wave on a rotation paraboloid of infinite dimensions in terms of Laguerre functions", *Dokl. Acad. Nauk USSR*, vol. 60, no. 2, pp. 203–206, 1948.

[36] R. E. Collin, *Field Theory of Guided Waves*. New York, Toronto, London: Mc-Graw-Hill, 1960.

[37] L. B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves*. Englewood Cliffs: Prentice-Hall, 1973, vol. 1.

[38] S. N. Samaddar and C. J. Lombardo, "Differential equation associated with electromagnetic waves in an inhomogeneous medium where $\nabla\varepsilon(\mathbf{r})\cdot\mathbf{E}$ vanishes", *J. Franklin Inst.*, vol. 289, no. 6, pp. 457–468, 1970.

[39] S. N. Samaddar, "An approach to improve re-entry communications by suitable orientations of antenna and static magnetic field", *Radio Sci.*, vol. 69D, no. 6, pp. 851–863, 1965.

[40] L. N. Bol'shev and N. V. Smirnov, *Tables of Mathematical Statistics*. Moscow: Nauka, 1983.

[41] E. C. Titchmarsh, *Eigenfunction Expansions Associated with Second Order Differential Equations*. Oxford, 1946.

[42] W. B. Jones and W. J. Thron, *Continued Fractions. Analytic Theory and Applications*. Vol. 11. *Encyclopedia of Mathematics and Its Applications*. Reading, Massachusetts: Addison-Wesley, 1980.

[43] N. I. Vilenkin, *Special Functions and the Theory of Group Representations*. American Mathematical Society, Providence, R.I., 1968.

[44] E. Ya. Riekstiņš, *Asymptotics and Evaluations of Roots of Equations*. Riga: Zinatne, 1991 (in Russian).

[45] R. Burman, "Some electromagnetic wave functions for propagation along cylindrically stratified columns", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-16, no. 2, pp. 127–129, 1968.

[46] C. N. Kurtz and W. Streifer, "Guided waves in inhomogeneous focusing media". Part I: "Formulation, solution for quadratic inhomogeneity", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-17, no. 1, pp. 11–15, 1969.

[47] R. Yamada and Y. Inabe, "Guided waves along graded index dielectric rod", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-22, no. 8, pp. 813–814, 1974.

[48] O. Parriaux, "Propagation d'ondes éléctromagnetiques guidées dans des structures à symétrie cylindrique circulaire". Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 1975.

[49] H. Suhl and L. R. Walker, "Topics in guided wave propagation through gyromagnetic media." Part II. "Transverse magnetization and the non-reciprocal helix", *Bell System Tech. J.*, vol. 33, no. 4, pp. 939–986, 1954.

[50] N. Kumagai and K. Takeuchi, "Circular electric waves propagating through the circular waveguide containing a circumferentially magnetized ferrite cylinder", *IRE Wescon Convention Record*, Part I, pp. 123–130, 1958.

[51] A. L. Mikaelyan, *Theory and Application of Ferrites at Microwaves*, Moscow, Leningrad: State Energetic Publ. House, 1963 (in Russian).

[52] M. E. Averbuch, "Difference functions for evaluation of layered azimuthally magnetized ferrite waveguides", in *Proc. V Int. Conf. Microw. Ferr.*, Vilnius, USSR, 1980, vol. 4, pp. 126–133 (in Russian).

[53] T. Obunai and K. Hakamada, "Slow surface wave propagation in an azimuthally-magnetized millimeter-wave solid-plasma coaxial waveguide", *Jap. J. Appl. Phys.*, vol. 23, no. 8, pp. 1032–1037, 1984.

[54] K. P. Ivanov and G. N. Georgiev, "Asymptotic study of rotationally symmetric waves in a circular anisotropic waveguide", *Proc. Inst. Elec. Eng.*, Pt. H *(Microwaves, Antennas and Propagat.)*, vol. 132, no. 4, pp. 261–266, 1985 (special issue on microwave ferrite engineering).

[55] K. P. Ivanov and G. N. Georgiev, "On a class of electromagnetic wave functions for propagation along the circular gyrotropic waveguide", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-34, no. 8, pp. 853–862, 1986.

[56] V. I. Miteva and K. P. Ivanov, "Nonreciprocal effects in an azimuthally magnetised millimetre-wave solid-plasma circular guide", *Electron. Lett.*, vol. 23, no. 3, pp. 118–120, 1987.

[57] G. N. Georgiev, "Rotationally symmetric waves in a circular waveguide with azimuthally magnetized remanent ferrite." Ph.D. thesis, Institute of Electronics, Bulgarian Academy of Sciences, Sofia, Bulgaria, 1987 (in Bulgarian).

[58] K. P. Ivanov and G. N. Georgiev, "Kummer and Tricomi functions for analysis of circular gyrotropic waveguides with azimuthal magnetisation", *Proc. Inst. Elec. Eng.*, Pt. H *(Microwaves, Antennas and Propagat.)*, vol. 135, no. 2 pp. 125–128, 1988.

[59] K. P. Ivanov and G. N. Georgiev, "Azimuthally magnetized circular ferrite waveguides", in *Ferrite Phase Shifters and Control Devices*, J. Helszajn. London: McGraw-Hill, 1989, ch. 14, pp. 262–288.

[60] K. P. Ivanov and G. N. Georgiev, "Some properties of the circular waveguide with azimuthally magnetized ferrite", *J. Appl. Phys.*, vol. 67, no. 10, pp. 6529–6537, 1990.

[61] K. P. Ivanov and G. N. Georgiev, "Application of Tricomi functions in the eigenvalue analysis of the circular gyrotropic waveguide", *Eur. Trans. Telecommun. Relat. Technol. ETT*, vol. 2, no. 2, pp. 259–269, 1991.

[62] A. A. P. Gibson, R. Sloan, L. E. Davis, and D. K. Paul, "Double valued phase constants in gyrotropic waveguides", *Proc. Inst. Elec. Eng.*, Pt. H *(Microwaves, Antennas and Propagat.)*, vol. 138, no. 3, pp. 258–260, 1991.

[63] K. P. Ivanov, G. N. Georgiev, and M. N. Georgieva, "A simple formula for differential phase shift computation of ferrite-filled circular waveguide", in *Proc. 1992 Asia-Pacific Microw. Conf. APMC'92*, Adelaide, Australia, 1992, vol. 2, pp. 769–772.

[64] K. P. Ivanov, G. N. Georgiev, and M. N. Georgieva, "Propagation in a circular waveguide partially filled with azimuthally magnetized ferrite", in *Proc. 1992 URSI Int. Symp. Electromagn. Theory*, Sydney, Australia, 1992, pp. 167–169.

[65] L. E. Davis, R. Sloan, and D. K. Paul, "TM_{0m} modes in transversely magnetised semiconductor-filled coaxial waveguide and parallel plates", *Proc. Inst. Elec. Eng.*, Pt. H *(Microwaves, Antennas and Propagat.)*, vol. 140, no. 3, pp. 211–218, 1993.

[66] G. N. Georgiev and M. N. Georgieva-Grosse, "Formulae for differential phase shift computation in an azimuthally magnetized circular ferrite waveguide", in *Proc. Millenn. Conf. Anten. Propagat. AP-2000,* Davos, Switzerland, 2000, paper 1002.

[67] G. N. Georgiev and M. N. Georgieva-Grosse, "Some aspects of the slow waves in the circular waveguide with azimuthally magnetized ferrite rod", in *Proc. 2002 XIth Int. Conf. Math. Meth. Electromagn. Theory MMET'02*, Kiev, Ukraine, 2002, vol. 2, pp. 674–676.

[68] G. N. Georgiev and M. N. Georgieva-Grosse, "Some new properties of the circular waveguides with azimuthally magnetized ferrite", in *Proc. 25th ESA Anten. Worksh. Satell. Anten. Technol. ESA/ESTEC*, Noordwijk, The Netherlands, 2002, pp. 601–608.

[69] G. N. Georgiev and M. N. Georgieva-Grosse, "Some microwave applications of the Kummer confluent hypergeometric function", in *Proc. 5th Int. Conf. Transp. Opt. Netw. ICTON 2003*, Warsaw, Poland, 2003, vol. 2, pp. 14–21.

[70] G. N. Georgiev and M. N. Georgieva-Grosse, "Azimuthally magnetized circular ferrite phase shifter", in *Proc. 2003 ITG-Conf. Anten. INICA 2003*, Berlin, Germany, 2003, pp. 115–118.

[71] G. N. Georgiev and M. N. Georgieva-Grosse, "Conditions for phaser operation of the circular waveguides with azimuthally magnetized ferrite", in *Proc. 26th ESA Anten. Technol. Works. Satell. Anten. Modell. Des. Tools ESA/ESTEC*, Noordwijk, The Netherlands, 2003, pp. 359–366.

[72] G. N. Georgiev and M. N. Georgieva-Grosse, "A new property of complex Kummer function and its application to waveguide propagation", *IEEE Anten. Wirel. Propagat. Lett.*, vol. AWPL-2, pp. 306–309, 2003.

[73] G. N. Georgiev and M. N. Georgieva-Grosse, "Several hypotheses in the confluent hypergeometric functions based theory of the azimuthally magnetized circular ferrite waveguides", in *Proc. East-West Worksh. Adv. Techn. Electromagn.*, Warsaw, Poland, 2004, pp. 197–204.

[74] G. N. Georgiev and M. N. Georgieva-Grosse, "Some new simple methods for differential phase shift computation in the circular waveguides with azimuthally magnetized ferrite", in *Proc. 28th ESA Anten. Worksh. Space Anten. Syst. Technol. ESA/ESTEC*, Noordwijk, The Netherlands, 2005, part 2, pp. 1159–1166.

[75] D. M. Bolle and G. S. Heller, "Theoretical considerations on the use of circularly symmetric TE modes for digital ferrite phase shifters", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-13, no. 4, pp. 421–426, 1965. See also D. M. Bolle and N. Mohsenian, "Correction", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-34, no. 4, p. 427, 1986.

[76] P. J. B. Clarricoats and A. D. Olver, "Propagation in anisotropic radially stratified circular waveguides", *Electron. Lett.*, vol. 2, no. 1, pp. 37–38, 1966.

[77] R. R. Yurgenson and I. G. Teitel'baum, "Computation of a circular waveguide with azimuthally magnetized ferrite rod", *Antennas*, vol. 1, *Collection of papers*, Moscow, Sviaz', 1966, pp. 120–133 (in Russian).

[78] A. I. Potekhin and R. R. Yurgenson, "Computation of some microwave transmission lines with azimuthally magnetized ferrite", *Radiotekhnika i Electronika*, vol. 15, no. 3, pp. 456–464, 1970 (in Russian).

[79] W. J. Ince and G. N. Tsandoulas, "Modal inversion in circular waveguides." Part II. "Application to latching nonreciprocal phasers", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-19, no. 4, pp. 393–400, 1971.

[80] F. J. Bernues and D. M. Bolle, "A study of modes in circular waveguide with azimuthally magnetized ferrite", Div. of Engineering, Brown Univ., Providence, R.I., NSF Res. Grant NSF-GK 2351/1, Washington, 1971.

[81] S. G. Semenov, G. I. Vesselov, and N. I. Platonov, "On the solution of Maxwell equations for azimuthally magnetized gyromagnetic medium", *Radiotekhnika*, vol. 27, no. 8, pp. 47–49, 1972 (in Russian).

[82] O. Parriaux and F. E. Gardiol, "Propagation in circular waveguide loaded with an azimuthally magnetized ferrite tube", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-25, no. 3, pp. 221–224, 1977.

[83] R. S. Mueller and F. J. Rosenbaum, "Propagation along azimuthally magnetized ferrite-loaded circular waveguides", *J. Appl. Phys.*, vol. 48, no. 6, pp. 2601–2603, 1977.

[84] Y. Xu and J. Chen, "Electromagnetic waves in waveguides, containing azimuthally magnetized ferrite", *IEEE Trans. Magnet.*, vol. MAG-16, no. 5, pp. 1174–1176, 1980.

[85] G. A. Red'kin, A. E. Mudrov, and V. A. Meshcheriakov, "Phase shifter with azimuthally magnetized ferrite samples", in *Proc. Vth Int. Conf. Microw. Ferr.*, Vilnius, USSR, 1980, vol. 4, pp. 170–175 (in Russian).

[86] I. V. Lindell, "Variational methods for nonstandard eigenvalue problems in waveguide and resonator analysis", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-30, no. 8, pp. 1194–1204, 1982.

[87] S. Kal and N. B. Chakraborti, "Finite-element analysis of helical phase shifters", *Proc. Inst. Elec. Eng.*, Pt. H *(Microwaves, Antennas and Propagat.)*, vol. 132, no. 4, pp. 231–236, 1985 (special issue on microwave ferrite engineering).

[88] A. J. Baden-Fuller, *Ferrites at Microwave Frequencies*. IEE Electromagnetic Waves, Series 23. London: Peter Peregrinus, 1987.

[89] R. S. Mueller, "Microwave radiation from a magnetic dipole in an azimuthally magnetized ferrite cylinder", *IEEE Trans. Microw. Theory Techn.*, vol. MTT-34, no. 7, pp. 828–831, 1986.

Georgi Nikolov Georgiev was born on January 17, 1957 in Polikraishte, Bulgaria. In 1974 he graduated from the First Gymnasium "St. St. Cyril and Methodius", in Veliko Tirnovo, Bulgaria, with complete honours and a golden medal. In the same year he won the Bulgarian National Olympiade in Physics and took part in the 7th International Olympiade in Physics, in Warsaw, Poland. For his exceptional achievements he was awarded by the Minister of Education of Bulgaria. He received the diploma of radiophysics and electronics (honours) from the Faculty of Physics, University of Sofia "St. Clement Okhridski", Bulgaria, in 1979 and the Ph.D. degree in physics from the Institute of Electronics, Bulgarian Academy of Sciences, Sofia, in 1987 for a thesis on the theory of anisotropic waveguides. In 1988 he joined the Faculty of Mathematics and Informatics, University of Veliko Tirnovo "St. St. Cyril and Methodius", Bulgaria, as an Assistant Professor. Since 1994 he is an Associate Professor in classical electromagnetic theory. Currently he is teaching general physics, analytical mechanics and numerical methods. In 1991–1995 he was a Deputy Dean of the Faculty. His field of interests includes mathematical physics, microwave physics, analytical and computational electromagnetics, guided wave propagation through anisotropic media and ferrite control components. He is an author and co-author of over 50 publications in those fields. The main scientific activities and research results of Dr. Georgiev consist in the development of an approach based on the confluent hypergeometric functions for investigation of anisotropic circular waveguides with azimuthal magnetization. He is an IEEE Member since 1995 and an Advisor of the IEEE Student Branch which he created in 1998 at the University of Veliko Tirnovo. He is a Senior Member of the Chinese Institute of Electronics (CIE), Beijing, China, since 1996 (Member since 1995) and co-founder of the Trans Black Sea Region Union of Applied Electromagnetism (BSUAE), Metsovo, Greece, 1996. He is a Member of the co-ordination Committee of the Union and a founder of the BSUAE Bulgaria Chapter. He was co-organizer of the 1st and 2nd BSUAE Symposia (Metsovo and Xanthi, Greece, 1996 and 2000, respectively). He is a reviewer for several international scientific journals. He has been serving as a Member of the Editorial Board of the *Journal of Applied Electromagnetism*, published by the BSUAE in Athens, Greece, since 1997. Dr. Georgiev was also an invited lecturer at the Southwest Institute of Applied Magnetics of China, Mianyang, Sichuan Province, People's Republic of China (1994), National Technical University of Athens, Athens, Greece (1995), Belarussian State University, Minsk, Belarussia (1995), Dokuz Eylül University, Izmir, Turkey (2001) and Democritus University of Thrace, Xanthi, Greece (2001). In 2004 he was a Visiting Professor at the Democritus University of Thrace, Xanthi, Greece, and an INTAS supervisor of a personal communication systems antenna research project of the Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia, Swiss Federal Institute of Technology, ETH-Zürich, Switzerland, and the University of Veliko Tirnovo "St. St. Cyril and Methodius", Bulgaria. Languages spoken: English, Slovak, Russian, German, Czech, French, and Polish, beside the native Bulgarian.

e-mail: gngeorgiev@yahoo.com
Faculty of Mathematics and Informatics
University of Veliko Tirnovo "St. St. Cyril and Methodius"
BG-5000 Veliko Tirnovo, Bulgaria

Mariana Nikolova Georgieva-Grosse was born on November 8, 1964 in Polikraishte, Bulgaria. In 1982 she graduated from the First Gymnasium "St. St. Cyril and Methodius", in Veliko Tirnovo, Bulgaria, with complete honours and a golden medal. She received the diploma of radiophysics and electronics (honours) and the Ph.D. degree in physics both from the Faculty of Physics, University of Sofia "St. Clement Okhridski", Bulgaria, in 1987 and 1995, respectively. Her theses were on the theory of layered circular anisotropic waveguides of azimuthal magnetization and on the theory of nonlinear plasma waveguides, respectively. In 1995–1997 she was with the Faculty of Physics, University of Sofia as an Assistant Professor in physics. Simultaneously she taught courses in computer science at the Faculty of Mathematics and Informatics, University of Veliko Tirnovo "St. St. Cyril and Methodius", Bulgaria, as a Honorary Assistant Professor. In 1997 she was with the Bochum University, Germany, as a research Assistant. Since 2000 she is a Microsoft Certified Engineer. In 2000–2001 she was with the Duerr Systems GmbH, Stuttgart, Germany. In 2001 she was with Duerr Systems GmbH, at VW plant in Puebla, Mexico. Since 2001 she is with Robert Bosch GmbH, Stuttgart, Germany, in the development of advanced automotive electronic applications. Her field of interests includes computer science, mathematical physics, plasma physics, analytical and computational electromagnetics, guided wave propagation through anisotropic media and ferrite control components. She is an author and co-author of over 40 publications in those fields. The main scientific activities and research results of Dr. Georgieva-Grosse consist in the development of an analysis method based on the perturbation theory for investigation of surface waves in planar and cylindrical plasma waveguides with weak nonlinearity and on an approach based on the confluent hypergeometric functions for investigation by the study of anisotropic circular waveguides with azimuthal magnetization. She is a Senior Member of the Chinese Institute of Electronics (CIE), Beijing, China, since 1996 (Member since 1995) and a Member of the Trans Black Sea Region Union of Applied Electromagnetism (BSUAE), since 1996. Languages spoken: English, German, Slovak, Russian, and Czech, beside the native Bulgarian.

e-mail: Mariana.G@t-online.de
Meterstrasse 4/2
D-70839 Gerlingen, Germany

Intelligent decision system based on the evidential reasoning approach and its applications

Dong-Ling Xu and Jian-Bo Yang

http://doi.org/10.26636/jtit.2005.3.323

Abstract—Intelligent decision system (IDS) is a window-based software package that has been developed on the basis of the evidential reasoning (ER) approach, a recent development in handling hybrid multiple criteria decision analysis (MCDA) problems with uncertainties. In this paper, the evidential reasoning approach will be briefly described first, and its major differences from and the relationships with conventional MCDA methods will also be discussed. Then the main features, advantages and benefits of IDS will be demonstrated and explained using two application examples: supplier pre-qualification assessment and customer satisfaction survey analysis, which have been investigated as part of the research projects led by the authors and funded by the UK government and the EC. It is concluded in the paper that the ER approach can be used not only to deal with problems that traditional methods can solve, but also to model and analyse more complicated decision problems that traditional methods are incapable of handling.

Keywords— multicriteria decision support systems, knowledge management, intelligent decision system, the evidential reasoning approach.

1. Introduction

In increasingly competitive, demanding and hostile business environments, many organisations are under pressure to cut costs and improve quality of their services and products. During the past several years, we have been in close collaboration with a number of companies in applying multicriteria decision analysis methods to help them achieve those goals. Assessing suppliers systematically in e-procurement processes and conducting quality and service surveys among customers are two of the areas where many companies have asked us to provide support.

Such assessments and surveys are normally based on specially designed models and can be regarded as a typical type of multiple criteria decision analysis (MCDA) problems [1, 13], which normally include a large number of criteria having both a quantitative and qualitative nature. Traditional ways of conducting such assessments and surveys include the use of average scores as performance indicators. The advantage of such methods is their simplicity and practicality. However, an average score does not provide sufficient information on the diversity of the performances of a business, nor can it indicate where the business is doing well and where it needs to improve if its average performance is acceptable. Therefore strengths and weaknesses need to be identified separately to supplement average scores. However, questions have been raised as to the accuracy of average scores generated and the consistency between average scores and strengths and weaknesses identified [6, 8].

Recently, significant effort has been made by the authors and their colleagues to introduce a new MCDA method, the evidential reasoning (ER) approach into such assessment exercises [6, 12, 13]. Several projects have been funded by the UK Engineering and Physical Science Research Council (EPSRC) and the European Commission (EC) to conduct research in applying the ER approach to support such assessments. A number of papers and research reports have been generated and published as the results of the research projects. These results show that the ER approach can help to reduce subjectivity in the assessment processes and generate a range of useful information for an organisation in question. This paper will describe how the ER approach and its software realisation intelligent decision system (IDS) [9] can be applied to support supplier assessment and customer quality survey analysis.

In the following section, the ER approach and its development history will be described first and the IDS software will be introduced as well. A supplier pre-qualification assessment model and its implementation will then be discussed, followed by the description of a customer quality survey analysis using the IDS software. The paper will conclude in Section 5.

2. The evidential reasoning approach and its software realisation – IDS

The evidential reasoning approach uses an evidence-based reasoning process to reach a conclusion, which differs from traditional MCDA methods. The motivation of developing the ER approach originates from the authors' experiences of working with industry in developing decision support systems [16], in particular to deal with MCDA problems having both quantitative and qualitative information with uncertainties and subjectivity. The ER approach has been developed using the concepts from several disciplines, including decision sciences (in particular utility theory), artificial intelligence, statistical analysis, fuzzy set theory, and computer technology [10–12, 14–16].

The development of the ER approach has experienced five major stages. The first stage was the introduction of a belief structure into a decision matrix [16]. This provides a novel way to model MCDA problems, in particular those having both quantitative and qualitative criteria with uncertainties. In conventional methods, a MCDA problem is modelled using a decision matrix, with each criterion assessed at each alternative decision by a single value. In the ER approach, a MCDA problem is described using a belief decision matrix, with each criterion assessed at each alternative by a two-dimensional variable: possible criterion referential values (assessment grades) and their associated degrees of belief.

Mathematically, in the ER approach a MCDA problem with L criteria A_i $(i = 1,\dots,L)$, K alternatives O_j $(j = 1,\dots,K)$ and N evaluation grades H_n $(n = 1,\dots,N)$ for each criterion is represented using a belief decision matrix with $S\big(A_i(O_j)\big)$ as its element at the ith row and jth column, where $S\big(A_i(O_j)\big)$ is given as follows:

$$S\big(A_i(O_j)\big) = \Big\{\big(H_n, \beta_{n,i}(O_j)\big),\ n = 1,\dots,N\Big\}$$
$$i = 1,\dots,L,\ j = 1,\dots,K,$$

where $\beta_{n,i}(O_j)$ is the degree of belief to which the alternative O_j is assessed to the nth grade of the ith criterion. It should be noted that a criterion could have its own set of evaluation grades that may be different from those of other criteria and also criteria could consist of a hierarchy [12].

The above ER framework allows more information to be contained in the model where the decision maker is no longer forced to pre-aggregate decision information into a single value when the original information is truly two-dimensional. In this context, the ER framework not only provides flexibility in describing a MCDA problem, it also prevents any loss of information due to the conversion from two-dimensional to one-dimensional values in the modeling process.

The second stage was the introduction of the Dempster-Shafer theory [2, 5] into the ER approach so that the two-dimensional information contained in the belief decision matrix could be aggregated to produce rational and consistent assessment results. For years, the authors have been searching for appropriate theoretical frameworks to fulfil such a task and the Dempster-Shafer theory has been chosen because of its unique capacity of dealing with ignorance which is inherent in subjective assessments, its powerful evidence combination rules and the reasonable requirements to apply the rules [2, 3, 10, 11].

Instead of aggregating average scores, the ER approach employs an evidential reasoning algorithm to aggregate belief degrees, which has been developed on the basis of the belief decision matrix, decision theory and the evidence combination rule of the Dempster-Shafer theory [10–12, 14]. Thus, scaling grades is not necessary for aggregating criteria in the ER approach and it is in this way that the ER approach is different from other MCDA approaches, most of which aggregate average scores or utilities.

The ER aggregation process is briefly described as follows. The following descriptions are of a mathematical nature and may be skipped until the end of the last set of equations. First, the degrees of belief $\beta_{n,i}(O_j)$ (or $\beta_{n,i}$ for short) for all $n = 1,\dots,N$, $i = 1,\dots,L$ are transformed into basic probability masses [12, 14]. Let ω_i be the weight of the ith criterion, $m_{n,i}$ a basic probability mass representing the degree to which the ith criterion is assessed to the nth evaluation grade H_n. Let $m_{H,i}$ be a remaining probability mass unassigned to any individual grade after the ith criterion has been assessed; $m_{n,i}$ and $m_{H,i}$ are calculated as follows:

$$m_{n,i} = \omega_i \beta_{n,i} \quad n = 1,\dots,N,$$

$$m_{H,i} = 1 - \sum_{n-1}^{N} m_{n,i} = 1 - \omega_i \sum_{n=1}^{N} \beta_{n,i},$$
$$i = 1,\dots,L,$$

$$\overline{m}_{H,i} = 1 - \omega_i \quad \text{and} \quad \widetilde{m}_{H,i} = \omega_i \Big(1 - \sum_{n=1}^{N} \beta_{n,i}\Big)$$

with $m_{H,i} = \overline{m}_{H,i} + \widetilde{m}_{H,i}$ for all $i = 1,\dots,L$ and $\sum_{i=1}^{L} \omega_i = 1$. The probability mass assigned to the whole set of grades $H = [H_1, H_2, \dots, H_N]$, which is unassigned to any individual grade H_n, is split into two parts, one caused by the relative importance of the ith criterion or $\overline{m}_{H,i}$ and the other by the incompleteness of the ith criterion or $\widetilde{m}_{H,i}$.

Then, all L criteria are aggregated to generate the combined degree of belief for each possible grade H_n. Let $m_{n,I(1)} = m_{n,1}$ $(n = 1,\dots,N)$, $\overline{m}_{H,I(1)} = \overline{m}_{H,1}$, $\widetilde{m}_{H,I(1)} = \widetilde{m}_{H,1}$ and $m_{H,I(1)} = m_{H,1}$. The combined probability assignments $m_{n,I(L)}$ $(n = 1,\dots,N)$, $\overline{m}_{H,I(L)}$, $\widetilde{m}_{H,I(L)}$, and $m_{H,I(L)}$ can be generated by aggregating all the basic probability masses using the recursive evidential reasoning algorithm [14]:

$$\begin{aligned}
\{H_n\}: \quad & m_{n,I(i+1)} = K_{I(i+1)} \big[m_{n,I(i)} m_{n,i+1} + m_{H,I(i)} m_{n,i+1} \\
& \qquad + m_{n,I(i)} m_{H,i+1}\big] \\
& n = 1,2,\dots,N \\
\{H\}: \quad & m_{H,I(i)} = \widetilde{m}_{H,I(i)} + \overline{m}_{H,I(i)} \\
& \widetilde{m}_{H,I(i+1)} = K_{I(i+1)} \big[\widetilde{m}_{H,I(i)} \widetilde{m}_{H,i+1} + \overline{m}_{H,I(i)} \widetilde{m}_{H,i+1} \\
& \qquad + \widetilde{m}_{H,I(i)} \overline{m}_{H,i+1}\big] \\
& \overline{m}_{H,I(i+1)} = K_{I(i+1)} \big[\overline{m}_{H,I(i)} \overline{m}_{H,i+1}\big] \\
& K_{I(i+1)} = \Bigg[1 - \sum_{t=1}^{N} \sum_{\substack{i=1 \\ j \neq t}}^{N} m_{t,I(i)} m_{j,i+1}\Bigg]^{-1} \\
& i = \{1,2,\dots,L-1\} \\
\{H\}: \quad & \beta_H = \frac{\widetilde{m}_{H,I(L)}}{1 - \overline{m}_{H,I(L)}} \\
\{H_n\}: \quad & \beta_n = \frac{m_{n,I(L)}}{1 - \overline{m}_{H,I(L)}} \quad n = 1,2,\dots,N
\end{aligned}$$

Parameter β_n denotes the degree of belief to which the L criteria are assessed to the grade H_n and β_H represents the remaining belief degrees unassigned to any H_n. It has been proved that $\sum_{n=1}^{N}\beta_n+\beta_H=1$ [14]. The final distribution assessment for O_j generated by aggregating the L criteria can be represented as follows:

$$S(O_j)=\left\{\left(H_n,\beta_n(O_j)\right),\quad n-1,\ldots,N\right\}.$$

Suppose the utility (or score) of an individual output term H_n is denoted by $u(H_n)$. The average utility of $S(O_j)$ can be given as follows [12]:

$$u(O_j)=\sum_{n=1}^{N}\beta_n(O_j)u(H_n).$$

Note that β_n denotes the lower bound of the likelihood that the alternative O_j is assessed to H_n. The upper bound of the likelihood is given by $(\beta_n+\beta_H)$. Complementary to the above distribution assessment, a utility interval can also be established [12] if the assessment is incomplete or imprecise, characterized by the maximum, minimum and average utilities of $S(A^*)$ defined as follows given $u(H_{n+1})\geq u(H_n)$:

$$u_{\max}(O_j)=\sum_{n=1}^{N-1}\beta_n(O_j)u(H_n)+\big(\beta_N(O_j)+\beta_H(O_j)\big)u(H_N),$$

$$u_{\min}(O_j)=\big(\beta_1(O_j)+\beta_H(O_j)\big)u(H_1)+\sum_{n=2}^{N}\beta_n(O_j)u(H_n),$$

$$u_{avg}(O_j)=\frac{u_{\max}(O_j)+u_{\min}(O_j)}{2}.$$

Note that if all original assessments $S\big(A_i(O_j)\big)$ in the belief decision matrix are complete, then $\beta_H(O_j)=0$ and $u\big(S(O_j)\big)=u_{\max}(O_j)=u_{\min}(O_j)=u_{avg}(O_j)$. It should also be noted that the above utilities are only used for characterizing an assessment but not for criterion aggregation.

The computational complexity using the combination rule of the Dempster-Shafer theory could be one of the major points of criticism if the combination rule is not used properly. In fact, Orponen [4] showed that the combination of mass functions or basic probability assignments (BPAs) using Dempster's rule is #P-complete (the class #P is a functional analogue of the class NP of decision problems). But the computational complexity of reasoning using Dempster's rule based on the above specific ER framework becomes linear rather than #P-complete [10–12]. It should also be noted that conflicting information can be explicitly modelled using the ER framework with the normalized ω_k and logically processed using the ER algorithm, thereby overcoming another drawback of the original combination rule of the Dempster-Shafer theory in dealing with conflicting evidence.

The third stage was the development of the rule and utility-based information transformation techniques to transform various types of evaluation information to a unified framework so that all criteria of both a quantitative and qualitative nature can be assessed in a consistent and compatible manner in the ER framework [12]. This to certain extent mirrors the traditional normalisation techniques used to handle quantitative criteria with different units in MCDA problems. The key difference is that in the ER framework the new techniques can in a sense preserve the two-dimensional information represented in the belief structure. It has been proved that by using the developed information transformation techniques not only the expected utilities of the original and the transformed assessments are equivalent but the degrees of incompleteness or completeness in the original assessments are also preserved.

The fourth stage is the enhancement of the approximate reasoning process of the original ER approach. Although the Dempster-Shafer theory has been used as the theoretical framework for information aggregation in the ER approach, its original evidence combination rule would generate irrational synthesis results if there is conflicting evidence. Significant modifications have been made since the theory was first introduced into the ER approach to deal with MCDA problems. It is proved that the new reasoning process of the ER approach satisfies the following common sense synthesis rules (CSSR) [14]:

CSSR 1: *If no sub-criterion is assessed to an evaluation grade at all then the upper-level criterion should not be assessed to the same grade either.*

CSSR 2: *If all sub-criteria are precisely assessed to an individual grade, then the upper-level criterion should also be precisely assessed to the same grade.*

CSSR 3: *If all sub-criteria are completely assessed to a subset of grades, then the upper-level criterion should be completely assessed to the same subset as well.*

CSSR 4: *If sub-criterion assessments are incomplete, then an upper-level assessment obtained by aggregating the incomplete basic assessments should also be incomplete with the degree of incompleteness properly expressed.*

The fifth stage is the implementation of the ER approach by developing a Windows based software package, the intelligent decision system [9, 12, 14, 15]. As mentioned earlier, the ER approach models a MCDA problem using a belief decision matrix with two-dimensional values, so inevitably the calculations involved in the aggregation processes could be more complicated than some traditional methods such as the additive utility function approach. Without a user-friendly computer interface to facilitate information collection, processing and display, the task could be rather difficult to accomplish by hands, even for a relatively small scale MCDA problem.

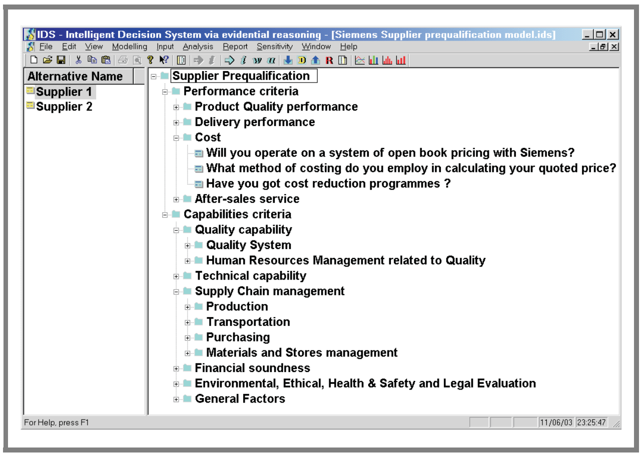
Although the ER approach involves relatively complicated calculations, its computational requirements are linearly proportional to the scale of a MCDA problem, namely the numbers of criteria and alternatives in a problem. IDS has been used in a variety of applications, such as motorcycle assessment [10], general cargo ship design selection (or assessment), marine system safety analysis and synthesis, executive car assessment, project management and organizational self-assessment [6, 13]. The experiences gained from these applications indicate that for MCDA problems with up to a few thousands of criteria and many alternatives, the calculation time using a PC is unnoticeable. It has also been proved in these applications that the ER approach not only produces consistent and reliable results for problems that can be solved using conventional MCDA methods, but also is capable of dealing with MCDA problems of the following features, which are difficult to handle using conventional methods without making further assumptions:

- mixture of quantitative and qualitative information,
- mixture of deterministic and random information,
- incomplete (missing) information,
- vague (fuzzy) information,
- large number (hundreds) of criteria in a hierarchy,
- large number of alternatives.

In addition to its mathematical functions, IDS is also a knowledge management tool. It records assessment information including evidence and comments in organised structures, provides systematic help at every stage of the assessment process including guidelines for grading criteria, and at the end of an assessment generates a tailored report with strengths and weaknesses highlighted at a click of a button. In the following sections, two application examples are to be examined to demonstrate some of the features of the IDS software package.

3. Supplier pre-qualification assessment

World markets have become increasingly competitive and integrated. In a global marketplace, a good supplier appears to be an invaluable resource for a buying organisation. The selection and management of right suppliers is the key element for a company to achieve its own performance targets.
Supplier pre-qualification assessment is considered as the critical step of a supplier selection process. Its objective is to screen out supply applicants who do not meet the basic requirements to such a degree that any further detailed assessment of their applications would be unnecessary. It also aims to provide feedback information to an applicant about where it should improve in order to be qualified as a supplier. It thus consists of both establishing minimal capacities below which a vendor will not be considered and determining whether an applicant can fulfil these basic requirements.
In recent years, the authors have established and supervised a number of summer consultancy projects for supplier assessment together with the Purchase Department of the Shared Service Ltd of Siemens UK, a global leading company in communications, electronics and electrical engineering. The objectives of such projects are to help investigate existing supplier assessment models, develop new models and realise them using the IDS software package both online and offline. One of such projects was dedicated to developing a supplier pre-qualification assessment model for the company [7]. The model has a hierarchy of criteria as shown in Fig. 1, which is the IDS main window for displaying an assessment model.

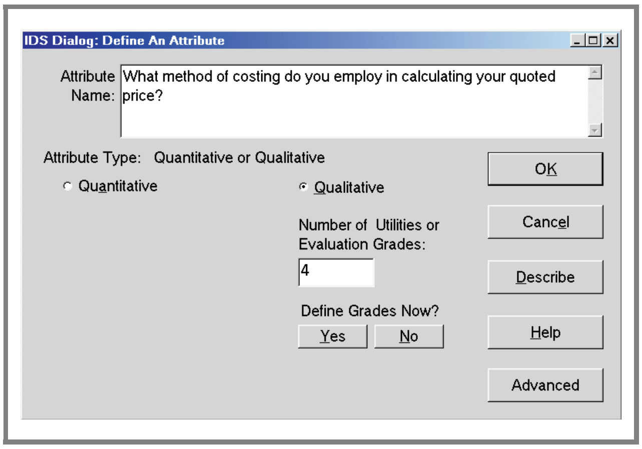

IDS - Intelligent Decision System via evidential reasoning - [Siemens Supplier prequalification model.ids]
File Edit View Modelling Input Analysis Report Sensitivity Window Help
For Help, press F1
11/06/03 23:25:47

Alternative Name
- Supplier 1
- Supplier 2

- Supplier Prequalification
  - Performance criteria
    - Product Quality performance
    - Delivery performance
    - Cost
      - Will you operate on a system of open book pricing with Siemens?
      - What method of costing do you employ in calculating your quoted price?
      - Have you got cost reduction programmes ?
    - After-sales service
  - Capabilities criteria
    - Quality capability
      - Quality System
      - Human Resources Management related to Quality
    - Technical capability
    - Supply Chain management
      - Production
      - Transportation
      - Purchasing
      - Materials and Stores management
    - Financial soundness
    - Environmental, Ethical, Health & Safety and Legal Evaluation
    - General Factors

Fig. 1. IDS main window for Siemens UK supplier pre-qualification model.

The IDS main window consists of a tree view on the right side to display the names of a hierarchy of criteria; a list view on the left side to show the names of alternative suppliers to be assessed; a menu bar where all IDS functions can be assessed for model building, data input, result analysis, reporting and sensitivity analysis; and a short cut bar for easy access to frequently used IDS functions.



Attribute Name: What method of costing do you employ in calculating your quoted price?

Attribute Type: Quantitative or Qualitative

○ Quantitative ● Qualitative

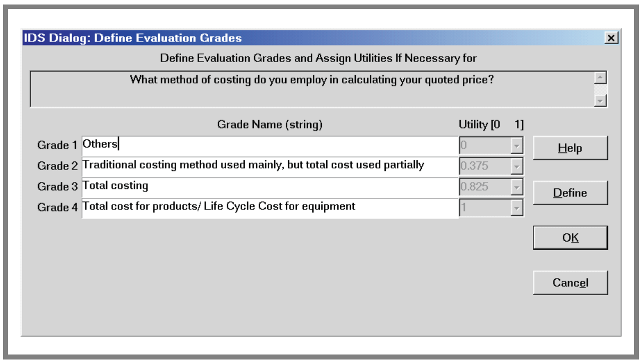
Number of Utilities or Evaluation Grades: 4

Define Grades Now?

Fig. 2. Define a qualitative criterion using IDS dialog.

The criteria hierarchy can be fully expanded in the same way as in Window Explorer. A criterion can be defined as a quantitative, qualitative or uncertain criterion using the IDS dialog windows [9]. For example, Fig. 2 shows the IDS dialog window for defining a qualitative criterion where the user can enter the name of the criterion, choose the number of assessment grades and provide a description about the criterion. Many of the criteria in the Siemens pre-qualification model are of a similar qualitative nature.

Not only can the user define the number of assessment grades, but they can describe and define each grade as well. Figure 3 shows the IDS dialog window for this purpose. Guidelines about how each grade could be chosen can be described by clicking the Define button. The utilities of grades are determined by both the utilities of the grades of high-level criteria and the propagation rules from lower level criteria to high level criteria.


IDS Dialog: Define Evaluation Grades

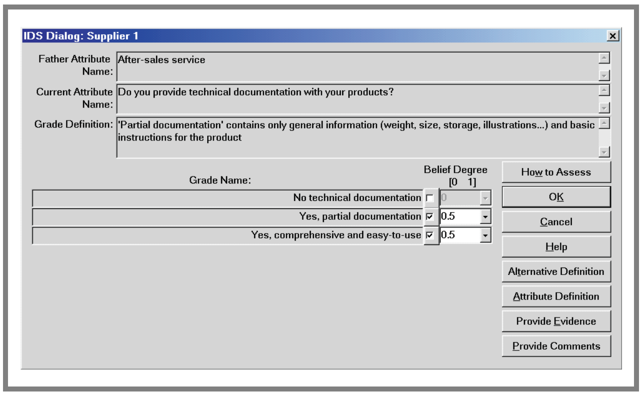
Define Evaluation Grades and Assign Utilities If Necessary for

What method of costing do you employ in calculating your quoted price?

	Grade Name (string)	Utility [0 1]
Grade 1	Others	0
Grade 2	Traditional costing method used mainly, but total cost used partially	0.375
Grade 3	Total costing	0.825
Grade 4	Total cost for products/ Life Cycle Cost for equipment	1

Fig. 3. Define assessment grades using IDS dialog.

A qualitative criterion can be assessed using the grades and a degree of belief to which each grade is assessed. Figure 4 shows an IDS input data dialog window where the user can choose one or more answers with different degrees of belief. The grade definition provides guidelines and/good practices about what a grade actually means, in what circumstances a grade (or answer) should be selected and to what degree a grade could be assessed to. Furthermore, the user can collect evidence to support an assessment and also provide comments on why the assessment is given this way. Such an assessment process is referred to as an evidence-based mapping process, which is designed to improve the objectivity and accuracy of the inherent subjective process. This also provides a structured knowledge base which is easy to access and could be used to support the assessment in subsequent discusses.


IDS Dialog: Supplier 1

Father Attribute Name: After-sales service

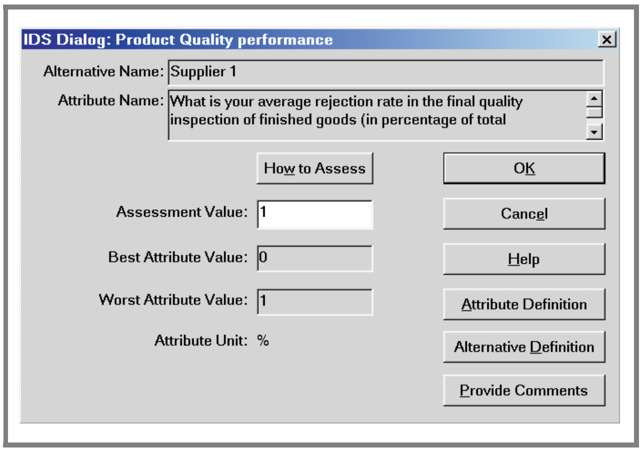
Current Attribute Name: Do you provide technical documentation with your products?

Grade Definition: 'Partial documentation' contains only general information (weight, size, storage, illustrations...) and basic instructions for the product

Grade Name:		Belief Degree [0 1]
No technical documentation	☐	0
Yes, partial documentation	☑	0.5
Yes, comprehensive and easy-to-use	☑	0.5

Fig. 4. Enter subjective assessment using IDS dialog.

Quantitative criteria can also be defined and used together with qualitative criteria for assessment. Figure 5 shows the numerical data input window. The best value and the worst value define the range of data that can be entered, which is defined by the user and between which an assessment figure can be assigned. Random numbers with various probabilities can also be defined and both the possible values and the likelihood can be entered as well, though this model does not contain such criteria.


IDS Dialog: Product Quality performance

Alternative Name: Supplier 1

Attribute Name: What is your average rejection rate in the final quality inspection of finished goods (in percentage of total

Assessment Value: 1

Best Attribute Value: 0

Worst Attribute Value: 1

Attribute Unit: %

Fig. 5. Enter numerical data using IDS dialog.

Apart from screening out poor supply applicants, the main purpose of such assessment includes the identification of strengths and weaknesses of an applicant, which could form a basis for subsequent detailed assessments and for creating action plans to address the weaknesses identified. As such, the concept of the distribution assessment developed in the ER approach would be helpful in identifying strengths and weaknesses. For example, Fig. 6 shows the final distribution assessment for a Siemens supplier "Supplier 1", which provides a panoramic view about the overall performance of the supplier in all areas. Clearly, the company has achieved the best performance in many areas, as over 40% of the areas are assessed to be "Best". However, the company does need to improve in nearly 21% of all

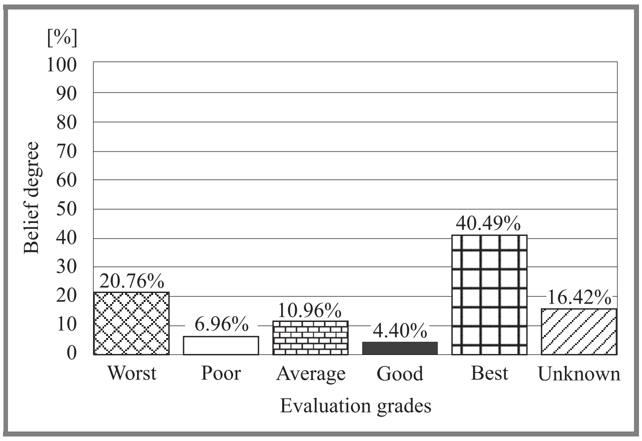


Fig. 6. Distribution assessment generated by IDS.

assessed areas. Also, the company was unable to answer some of the questions put forward by Siemens. In other words, over 16% of the areas need to be further investigated. On the whole, the average percentage score that the company has achieved is just below 60% with a variation between 51% and 57% (Fig. 7). The variation is caused due to the unanswered questions.

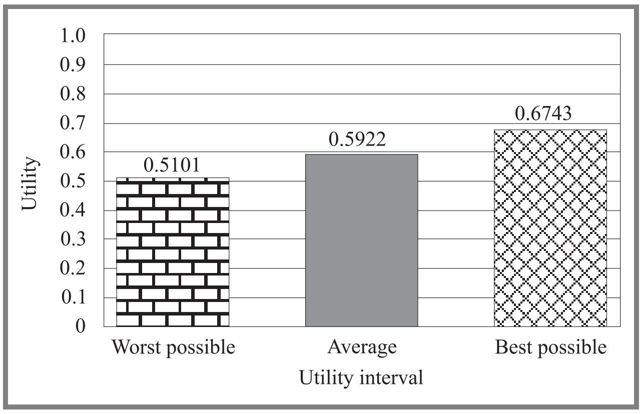


Fig. 7. Average assessment generated by IDS.

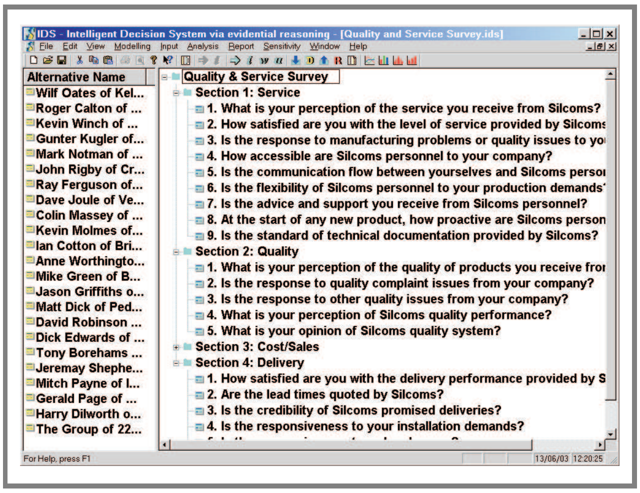
Using IDS, the performance distributions of the company on any criterion can be examined in a similar way. This enables Siemens to investigate the areas where the supplier has done well as well as the areas where the supplier has to improve. For example, the company received a zero score or the 100% "Worst" grading on the *product quality performance* criterion because of two problems. The first problem is that "*its average rejection rate in the final quality inspection of finished goods in percentage of total production*" is 3% whilst the lowest acceptable rate by Siemens is only 1%. The second problem is that "*its average return rate from customers in percentage of products or services delivered*" is 7% whilst the lowest acceptable rate by Siemens is only 1% as well. Such investigations provide both sides, Siemens on the customer side and Supplier 1 on the supplier side, with a clear objective view about what Supplier 1 needs to improve to achieve the standards required by Siemens.

The managers of Siemens UK and Supplier 1 both took part in the modeling process, the data collection and the result analysis. They are satisfied with the accuracy and objectivity of the investigation conducted using the ER approach supported by the IDS software.

4. Customer quality and service survey analysis

Customer quality and service survey can provide useful information for a company to improve the quality of its services and products. Silcoms is a medium manufacturing company, located in North West England and specialised in supplying components to aerospace industry among other businesses. The company faces tough competition from overseas in particular Asian companies which can supply cheap products. The management of Silcoms are aware of the competition and are totally committed to improving quality not only for the products they manufacture but also for the services they provide. The company has been given quality awards by the Excellence North West of England. The authors have collaborated closely with the company management and have been very much impressed by their desire to improve their products and services, which have already achieved high standards.

The company, together with the help of external consultants and academics including the authors, has developed a model for conducting quality and service survey among its customers. Figure 8 shows the model structure having four major areas each of which is addressed using a number of questions. To facilitate data collection, the answers to the questions adopt a five-grade scale.

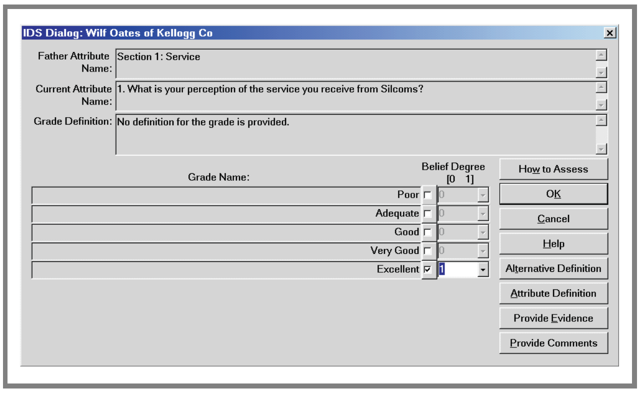


Alternative Name
Wilf Oates of Kel...
Roger Calton of ...
Kevin Winch of ...
Gunter Kugler of...
Mark Notman of ...
John Rigby of Cr...
Ray Ferguson of...
Dave Joule of Ve...
Colin Massey of ...
Kevin Molmes of...
Ian Cotton of Bri...
Anne Worthingto...
Mike Green of B...
Jason Griffiths o...
Matt Dick of Ped...
David Robinson ...
Dick Edwards of ...
Tony Borehams ...
Jeremay Shephe...
Mitch Payne of I...
Gerald Page of ...
Harry Dilworth o...
The Group of 22...

Quality & Service Survey
Section 1: Service
1. What is your perception of the service you receive from Silcoms?
2. How satisfied are you with the level of service provided by Silcoms
3. Is the response to manufacturing problems or quality issues to yo
4. How accessible are Silcoms personnel to your company?
5. Is the communication flow between yourselves and Silcoms persoı
6. Is the flexibility of Silcoms personnel to your production demands'
7. Is the advice and support you receive from Silcoms personnel?
8. At the start of any new product, how proactive are Silcoms person
9. Is the standard of technical documentation provided by Silcoms?
Section 2: Quality
1. What is your perception of the quality of products you receive froı
2. Is the response to quality complaint issues from your company?
3. Is the response to other quality issues from your company?
4. What is your perception of Silcoms quality performance?
5. What is your opinion of Silcoms quality system?
Section 3: Cost/Sales
Section 4: Delivery
1. How satisfied are you with the delivery performance provided by S
2. Are the lead times quoted by Silcoms?
3. Is the credibility of Silcoms promised deliveries?
4. Is the responsiveness to your installation demands?

Fig. 8. Questions numbered in four major sections.

Data were collected using a paper version of the model which was nicely bound together and individually sent to each customer. Figure 9 shows a typical answer window and no definition for the grades is provided as the question (criterion) and the answers (grades) are regarded to be straightforward. The customers often chose one answer and occasionally opted to not answer some questions either because the questions are irrelevant to their companies or there may be a lack of information.



IDS Dialog: Wilf Oates of Kellogg Co

Father Attribute Name: Section 1: Service

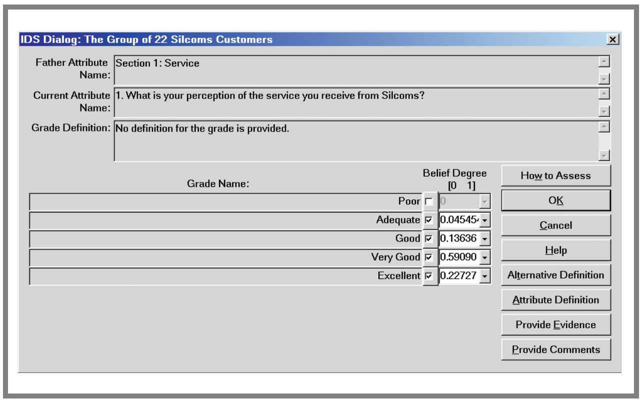
Current Attribute Name: 1. What is your perception of the service you receive from Silcoms?

Grade Definition: No definition for the grade is provided.

Grade Name:		Belief Degree [0 1]
Poor	□	0
Adequate	□	0
Good	□	0
Very Good	□	0
Excellent	☑	1

Fig. 9. Original answer provided by a customer.

In Fig. 8, the last alternative *The group of 22 customers* was generated by averaging the answers given by the 22 customers. In Fig. 10, the belief degree assigned to an answer was therefore the percentage of the 22 customers who had chosen this answer. IDS provides a function to combine the original answers provided by individual customers for group analysis. In Fig. 10 it is clear that nearly 80% of the customers have graded their perception of Silcoms service to be *Very good* or *Excellent*, which is an impressive result, considering that most customers were randomly selected with two known "critical" customers chosen deliberately.


IDS Dialog: The Group of 22 Silcoms Customers

Father Attribute Name: Section 1: Service

Current Attribute Name: 1. What is your perception of the service you receive from Silcoms?

Grade Definition: No definition for the grade is provided.

Grade Name:		Belief Degree [0 1]
Poor	☐	0
Adequate	☑	0.04545
Good	☑	0.13636
Very Good	☑	0.59090
Excellent	☑	0.22727

Fig. 10. Degrees of belief assigned by a group of customers.

The IDS provides a range of functions to support the analysis of such surveys, including the analyses of the individual customers' responses on any criteria and the comparison of results provided by the customers. Different groups of the customers can also be combined to show the collective opinions of these groups on any criteria. For example, Fig. 11 shows the collective assessment of the 22 customers on the quality and services provided by Silcoms. The distribution assessment shown in Fig. 11 provides a holistic view of the overall performances of Silcoms. The majority of the customers graded Silcoms at the *Very good* and *Excellent* grades in most areas with the combined belief degree of over 76%. This is a very good result for Silcoms, supporting the company's policy of placing services and quality in the first priority of their policies and strategies. However, there are a couple of customers who did provide critical assessment in some areas, which is clearly displayed. Unlike an average score, this panoramic view will not hide any unsatisfied areas for the good average assessment, thereby preventing the company from missing the opportunity of further improvement.

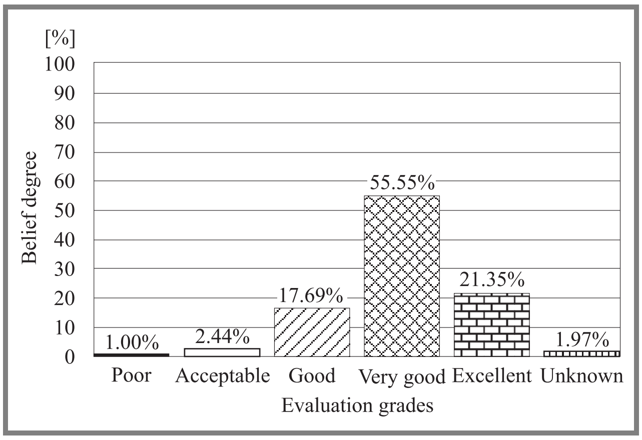


Fig. 11. Collective assessment of Silcoms quality and services.

5. Conclusions

In this paper, the evidential reasoning approach and the intelligent decision system were briefly introduced. Their applications to supplier assessments and the customer surveys of quality and service for two companies in the North West England were reported in some detail. The main feature of this kind of decision problems is that both quantitative and qualitative assessments are included and need to be treated both simultaneously and rationally. Using conventional decision methods, one may need to provide precise number for each assessment, which could be difficult from time to time. Also, assumptions may need to be made in cases where there are missing data or other uncertainties. Traditional methods may only be able to generate average numbers, where bad performances may be averaged out by good performance thereby missing opportunities to identify areas for improvement, which is indeed the very purpose of conducting such assessments in most cases. The IDS software provides easy to use functions to build assessment models, organise and manage knowledge, conduct analysis and generate results.

Acknowledgement

This work forms part of the projects supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under Grant No. GR/N65615 and the European Commission (EC) under Grant No. IPS-2000-00030. The authors are grateful to the managers of Shared Service Ltd of Siemens UK and Silcoms for supporting the research.

References

[1] V. Belton and T. J. Stewart, *Multiple Criteria Decision Analysis – an Integrated Approach*. Boston: Kluwer, 2002.

[2] B. G. Buchanan and E. H. Shortliffe, *Rule-Based Expert Systems*. Reading: Addison-Wesley, 1984.

[3] R. López de Mántaras, *Approximate Reasoning Models*. Chichester: Ellis Horwood, 1990.

[4] P. Orponen, "Dempster's rule of combination is # P-complete", *Artif. Intell.*, vol. 44, pp. 245–253, 1990.

[5] G. Shafer, *A Mathematical Theory of Evidence*. Princeton: Princeton University Press, 1976.

[6] C. H. R. Siow, J. B. Yang, and B. G. Dale, "A new modelling framework for organisational self-assessment: development and application", *AQS Qual. Manag. J.*, vol. 8, no. 4, pp. 34–47, 2001.

[7] J. L. Teng, "Development of a supplier pre-qualification model for a UK company". M.Sc. thesis, Manchester School of Management, UMIST, 2002.

[8] T. van der Wiele, A. R. T. Williams, F. Kolb, and B. G. Dale, "Assessor training for the European Quality Award", *Qual. World Techn. Suppl.*, pp. 12–18, March 1995.

[9] D. L. Xu and J. B. Yang, "Introduction to multi-criteria decision making and the evidential reasoning approach". Manchester School of Management, University of Manchester Institute and Technology, May 2001.

[10] J. B. Yang and P. Sen, "A general multi-level evaluation process for hybrid MADM with uncertainty", *IEEE Trans. Syst. Man Cyber.*, vol. 24, no. 10, pp. 1458–1473, 1994.

[11] J. B. Yang and M. G. Singh, "An evidential reasoning approach for multiple attribute decision making with uncertainty", *IEEE Trans. Syst. Man Cyber.*, vol. 24, no. 1, pp. 1–18, 1994.

[12] J. B. Yang, "Rule and utility based evidential reasoning approach for multiple attribute decision analysis under uncertainty", *Eur. J. Oper. Res.*, vol. 131, no. 1, pp. 31–61, 2001.

[13] J. B. Yang, B. G. Dale, and C. H. R. Siow, "Self-assessment of excellence: an application of the evidential reasoning approach", *Int. J. Prod. Res.*, vol. 39, no. 16, pp. 3789–3812, 2001.

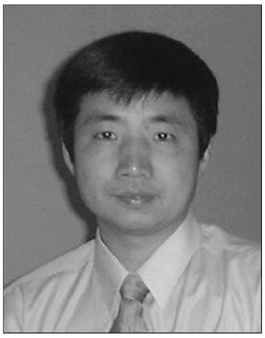
[14] J. B. Yang and D. L. Xu, "On the evidential reasoning algorithm for multiattribute decision analysis under uncertainty", *IEEE Trans. Syst. Man Cyber.*, Part A: *Systems and Humans*, vol. 32, no. 3, pp. 289–304, 2002.

[15] J. B. Yang and D. L. Xu, "Nonlinear information aggregation via evidential reasoning in multiattribute decision analysis under uncertainty", *IEEE Trans. Syst. Man Cyber.*, Part A: *Systems and Humans*, vol. 32, no. 3, pp. 376–393, 2002.

[16] Z. J. Zhang, J. B. Yang, and D. L. Xu, "A hierarchical analysis model for multiobjective decision making", in *4th IFAC/IFIP/ IFORS/IEA Conf. "Analysis, Design and Evaluation of Man-Machine System 1989"*, Xian, China, 1989. Oxford: Pergamon, 1990, pp. 13–18.

Dong-Ling Xu is a research fellow in the Manchester Business School, working in the area of decision analysis, decision support, and decision technology. As a co-designer, she developed two Windows based assessment tools called IDS multi-criteria assessor and IDS cost estimator, and several web based assessment and decision support tools. Dr. Dong-Ling Xu has published a book and over 60 papers in decision analysis and support, computerised traffic management and control, optimisation, control system design, statistical process control and statistical analysis. Her current research interests are in theoretical and applied research in multicriteria decision analysis under uncertainty.

e-mail: ling.xu@manchester.ac.uk
Manchester Business School (East)
The University of Manchester, PO Box 88
Manchester M60 1QD, United Kingdom



Jian-Bo Yang is Professor of decision and system sciences at the Manchester Business School, The University of Manchester, United Kingdom. He is also Visiting Professor of Huazhong University of Science and Technology of China. Prior to his current appointment, he was a faculty member of the University of Birmingham (1995–1997), the University of Newcastle upon Tyne (1991–1995), UMIST (1990) and Shanghai Jiao Tong University (1987–1989). In the past two decades, he has been conducting research in multiple criteria decision analysis under uncertainty, multiple objective optimisation, intelligent decision support systems, hybrid quantitative and qualitative decision modelling using techniques from operational research, artificial intelligence and systems engineering, and dynamic system modelling, simulation and control for engineering and management systems. His current applied research is in design decision-making, risk and safety analysis, production planning and scheduling, quality modelling and evaluation, supply chain modelling and supplier assessment, and the integrated evaluation of products, systems, projects, policies, etc. Professor Yang's research has been supported by EPSRC, EC, DEFRA, SERC, RGC of Hong Kong, NNSF of China and industry. He has published three books and about 100 papers in journals, presented over 90 papers in conferences, and developed several software packages in these areas including the Windows-based intelligent decision system (IDS) via evidential reasoning.

e-mail: Jian-Bo.Yang@manchester.ac.uk
Manchester Business School (East)
The University of Manchester, PO Box 88
Manchester M60 1QD, United Kingdom

Protocols for Wireless Sensor Networks: A Survey

Aarti Kochhar[1,2], Pardeep Kaur[1], Preeti Singh[1], and Sukesha Sharma[1]

[1] *University Institute of Engineering and Technology, Panjab University, Chandigarh, India*
[2] *Lovely Professional University, Phagwara, India*

http://doi.org/10.26636/jtit.2018.117417

Abstract—This paper presents a survey on the MAC and network layer of Wireless Sensor Networks. Performance requirements of the MAC layer are explored. MAC layer protocols for battery-powered networks and energy harvesting-based networks are discussed and compared. A detailed discussion on design constraints and classification of routing protocols is presented. Several routing protocols are compared in terms of such parameters as: energy consumption, scalability, network lifetime and mobility. Problems that require future research are presented. The cross-layer approach for WSNs is also surveyed.

Keywords*—*cross layer, Medium Access Control, protocols, Wireless Sensor Networks.

1. Introduction

Any Wireless Sensor Network (WSN) application requires the physical environment to be sensed for data transmitted over a channel to a base station. Power is required in order to sense data and send it to the base station. It can be obtained from a battery or may be harvested from a natural source. One of the basic architectures of a sensor node is shown in Fig. 1 [1]. It comprises 4 units responsible for power, processing and communications. Most energy is consumed by processing and communications.

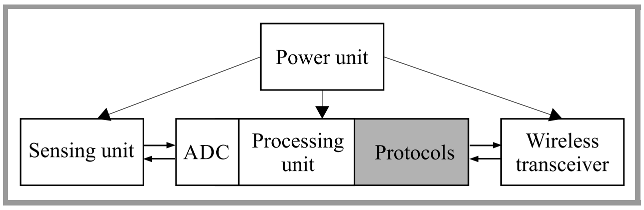


Fig. 1. Architecture of a typical sensor node.

In order to overcome energy, lifetime, traffic and mobility constraints, the communication protocol stack needs to be carefully designed.

The basic structure of a WSN protocol stack is discussed in Section 2. Transport and upper layers add reliability to the transmission of data only, which is not a key concern in the majority of WSN applications. Hence, only the data link layer (DLL) and the network layer are discussed in this paper. Energy consumption sources, classification, design constraints, respective protocols and outstanding research problems are discussed for the MAC and network layers. In Section 3, a sub-layer of DLL – Medium Access Control (MAC) layer – is surveyed. In Section 4, the network layer is examined. A comparison of both MAC and routing protocols has been tabulated in the respective sections. In Section 5, the cross layer approach, a technological advancement enhancing efficiency, is discussed.

2. Protocol Stack in WSNs

Proper design of the protocol stack is important for the overall efficiency of a WSN. WSN differs from conventional computer communication networks in the following ways:

- Contrary to computer network's well planned physical topology, the nodes are densely and randomly deployed in WSNs.
- Once designed, computer networks remain static, whereas WSNs are dynamic in nature. Failure of one node can change the entire topology. So, WSNs need to be self-configurable.
- Computer networks have IP addresses for their global identification. WSN nodes have no global identification because it creates a large overhead.
- Computer networks have a continuous supply of energy, whereas WSNs have limited resources. So, the WSN protocol stack needs to be energy-aware.

Protocol stacks in WSNs comprise five horizontal and five vertical levels. They have five layers and five management planes, as shown in Fig. 2 [2].

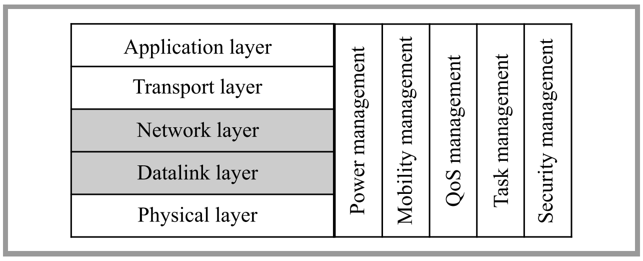


Fig. 2. Protocol stack of WSNs.

3. Data Link Layer

DLL has two sub-layers: MAC and Logical Link Control (LLC). LLC is used for link management, flow and error control. MAC is responsible for assembling data into frames and for disassembling frames to retrieve information. Nodes may be sharing a single channel for sending data over to the sink or to another node. Simultaneous transmission of data on a single channel will lead to a collision, causing loss of data and energy. To avoid this, nodes should agree on a time slot at which a particular node would be sending. To agree on timeslots, nodes need to communicate, which requires a channel too. Considering the propagation delay, it is difficult for a node to know the instantaneous status of another node. The transceiver also consumes a large amount of energy while accessing the media. MAC controls activity of the transceiver to conserve energy [3].

3.1. Energy Consumption Sources

There are a few energy consumption sources at the MAC layer [4]:

Collision – when two or more nodes try to send information on a single channel at the same time, the packets collide. Collided packets need to be discarded and retransmitted.

Overhearing – when a node receives a packet destined for another node, it consumes unnecessary energy.

Overhead – sending and receiving control information also requires energy, causing an additional overhead.

Idle listening – idle listening is listening to an idle channel on which traffic is expected.

Over-emitting – sending information to a node which is not ready to receive. Hence, packets are discarded and need to be retransmitted.

3.2. Performance Requirements for the MAC Layer

While designing MAC layer protocols, one needs to consider the following requirements [5]:

Throughput: Protocol efficiency is measured by its throughput. In the case of a wireless link, it may be related to capacity.

Scalability: Scalability refers to the protocol's adaptation to an increase in network size, traffic, overhead and load. One way to deal with this is to localize the interactions so that nodes need less global knowledge to operate.

Latency: Latency can be referred as the time delay between message transmission and message arrival. Latency is an important constraint for time-critical applications, and needs to be minimized.

Number of hops: It is the number of hops taken by packets to reach the sink. Operation of the MAC protocol varies between single-hop and multi-hop scenarios. In the case of multiple hops taken to reach the sink, data needs to be aggregated before sending it to the sink.

3.3. Classification of MAC Protocols

MAC protocols can be categorized into two types [6]:

- **schedule-based MAC** protocol in which nodes agree upon a fixed schedule to access the channel. So, each node has a fixed slot for communication. Outside their slots, nodes move into sleep mode, avoiding collision and overhearing. The lifetime of nodes is enhanced, as they do not communicate over the complete duty cycle;
- **random access-based protocol** in which nodes need to compete to reserve access to a channel. After collision, each node waits for a random time before accessing the channel again. Energy efficiency of random access-based protocol is low.

3.4. MAC Layer Protocols

A protocol for an application can be chosen based on performance and specific requirements. In the battery-powered area Sensor-MAC, a T-MAC is presented. Then, MAC protocols based on energy harvesting are presented (Fig. 3).

Sensor-MAC (S-MAC): In general, nodes are synchronized locally, to operate a periodic sleep-and-listen schedule. Each node belongs to a virtual cluster and each cluster has a common listen-and-sleep schedule, as shown in Fig. 4. This represents the basic idea of S-MAC [7]. Each node discovers its neighbors regularly and establishes a link with them. Then, it assigns a distinct frequency, time or

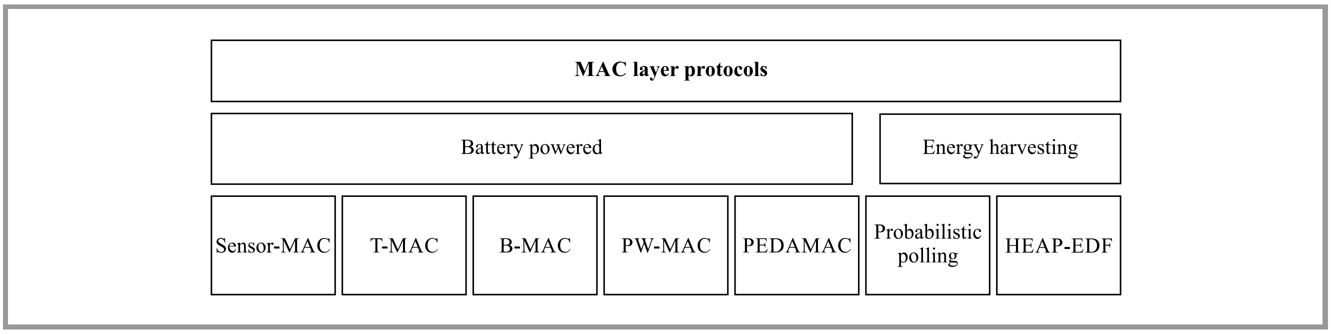


Fig. 3. MAC layer protocols.

code to each link. Long messages are divided before sending. Such a solution offer various advantages, as it self-organizes the network to a variation in topology. This change in topology can be the consequence of deaths or movement of a node. It also operates a lower duty cycle, so the consumption of power used for overhearing and idle listening is reduced. Network latency increases as nodes alternate between active and sleep mode. It can be avoided altogether if a node wakes up after sensing the wake-up of its neighbor. Since sleep-and-listen periods are predefined, efficiency of the protocol may decrease under variable traffic, as traffic may be forwarded to a sleeping node.

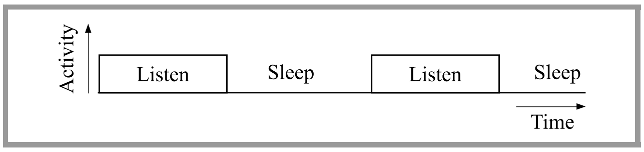


Fig. 4. Sleep-and-listen periods.

Timeout MAC (T-MAC): S-MAC has fixed listen and sleep periods, but applications with variable loads need dynamic listen and sleep periods. In T-MAC, the listen period ends when no event, such as receipt of data or sensing of activity has taken place for a threshold period (TP), as shown in Fig. 5 [8]. The listen period depends on current load. Transmission is based on Request-To-Send (RTS), Clear-To-Send (CTS) and acknowledgment (ACK) packets. Nodes close to the sink may have more data to send, so their listen periods are longer. The advantages are: RTS, CTS and ACK packets reduce collision rates and increase reliability. If listen periods are fixed, then nodes with less data will waste energy by idle listening. Energy consumption and idle listening are reduced as data can be sent in variable bursts. T-MAC has low sensitivity to latency, but it has a few drawbacks, such as it cannot support high data rate applications. Also, it has to trade-off throughput to maintain low energy consumption.

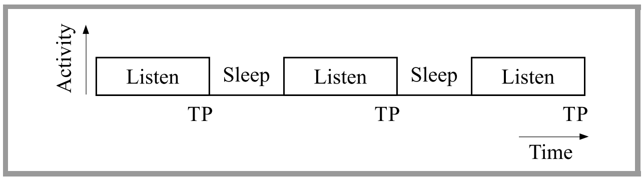


Fig. 5. Adaptive listen and sleep periods.

Berkeley MAC (B-MAC): B-MAC uses preamble sampling [9], [10]. Each time a node wakes-up, it checks for any activity before sending. The node is also waiting only for a certain period of time to receive data. After timeout, the node returns to sleep mode. B-MAC uses clear channel assignment, and makes local policy decisions to optimize network performance. Owing to preamble sampling, the duty cycle is reduced, which increases efficiency and throughput. Energy consumption is lower because of low-power listening. Also, it supports reconfiguration to improve latency. B-MAC has a few drawbacks, such as it has no ability to handle multi-packet environments and suffers form a hidden terminal problem. Also, overhead of the protocol increases. The protocol can be enhanced further using adaptive preamble sampling.

Predictive Wake-up MAC (PW-MAC): In PW-MAC, the wake-up schedule of nodes can be randomized [11], [12]. To inform the intended transmitters, the node will send a signal upon waking up. A sender can predict the receiver's wake-up time and can wake-up simultaneously to save energy. To address timing challenges, PW-MAC has an on-demand prediction-based error correction mechanism. PW-MAC has a reduced duty cycle, as it has a random node wake-up schedule. It has improved performance compared to S-MAC and B- MAC, as collisions can be avoided. Latency is less than 5% of that typical of other MAC protocols. A node needs only 10 bytes of memory to store the prediction state of other nodes. Each node has to send a signal on waking-up, so the overhead of the protocol is increased, although it is low compared to other protocols. Also, hardware can induce errors in predicting wake-up times of the receiver.

Power Efficient and Delay Aware MAC (PEDAMAC): To minimize energy consumption due to overhearing, PEDAMAC transmits data at more than one power level. The access points (also called sinks) coordinate sensor nodes. Access points are assumed to have no power constraints, while sensor nodes have limited power. PEDAMAC assumes that each node can reach the sink in one hop. It has four phases: topology learning, topology collection, scheduling and adjustment. The protocol allows the nodes to operate at different power levels, as per the requirement of the task being processed by the node. It has three power levels: maximum power P_m, medium P_x, and minimum P_s. Synchronization is done at P_m. The sink can broadcast topology-related information at P_x. Data is transmitted at P_s. Low transmission power saves energy and it is used in delay-bound applications, but it has a few drawbacks, such as the fact that protocol assumes a one hop distance to the sink, which may not always be the case. Distinct power levels increase the protocol overhead. Also, data may be dropped before delivered, if transmission power is too low, i.e. the range of radio is decreased because of power limitation. PEDAMAC can be enhanced by increasing the number of media or channels to further reduce the delay [12]–[14].

Energy harvesting is considered as the only energy source by Eu *et al.* [15]. It is not easy to predict the wake-up schedule of nodes powered by energy harvesters. Authors exploited the uncertain nature of energy harvesting sources to increase the performance of MAC protocols. MAC protocols based on battery-powered WSNs have different goals, such as increased lifetime compared to energy harvesting based WSN (EH-WSN). So, there is a need to have protocols designed specifically for EH-WSN.

Probabilistic polling: In probabilistic polling, the sink sets contention probability P_c in each node through a polling

Table 1
Performance evaluation of MAC protocols

Protocol	Throughput	Energy conservation	Maximum % of energy saved vs. S-MAC	Latency	Overhead	Scalability
S-MAC [7]	Low	Low	0	High	Low	High
T-MAC [9]	Low	High	85	N/A	Moderate	Low
B-MAC [9], [10]	High	Moderate	57	Moderate	High	Low
PW-MAC [11], [12]	High	High	80	Low	Moderate	High
PEDAMAC [13], [14]	Moderate	Moderate	38	Low	High	Low
Probabilistic polling [15]	High	N/A	N/A	Depends on energy harvesting rate	Low	High
HEAP-EDF [16]	Moderate	N/A	N/A	Depends on energy harvesting rate	Moderate	Low

packet [15]. Each node generates a random number ν, and when it is less than contention probability ($\nu < P_c$), the node is allowed to send. Otherwise, the node can go to the charging state. The sink keeps on changing contention probability depending on network response. If no sensor responds, the sink increases P_c. Also, when a node leaves the network, P_c is increased. In the case of collision and joining of new node, P_c is decreased by a larger amount. This approach is known as additive-increase and multiplicative-decrease. Contention probability P_c offers maximum throughput when it is equal to the inverse of the number of nodes receiving polling packets:

$$P_{opt} = \frac{1}{N_r}, \tag{1}$$

where P_{opt} is the optimal probability that maximizes throughput. N_r is the number of nodes receiving polling packets ($N_r \geq 1$).

This protocol can adapt to varying energy harvesting rates to ensure high throughput and the sink can also adjust P_c in the case of a collision. Hence, the protocol increases scalability of the network. Since P_c keeps changing due to collisions or when a node joins or leaves a network, it takes quite some time for the network to stabilize. This leads to increased network latency. Also, bandwidth is wasted until the network stabilizes at an appropriate P_c. Another drawback is that the protocol assumes a single hop distance to the sink, limiting protocol scalability.

HEAP-EDF: Power generated by ambient energy harvesting sources (HEAP), may vary, i.e. solar energy has different rates in the morning and in the afternoon. To overcome this, Earliest Deadline First (HEAP-EDF) uses a predict-and-update algorithm to reduce the temporal variations [16]. In HEAP-EDF, the sink polls the node with the minimum or the earliest wake-up time. The sensor will not poll the node whose energy has decreased below the transmission level because of previous polling. At the power-balance ratio of 0.5, HEAP-EDF offers the best fairness. The power-balance ratio is given as:

$$\emptyset = \sum_{n=1}^{N} \frac{T_c}{T_n}. \tag{2}$$

In Eq. (2), T_c is the duration of polling cycle, T_n is energy harvesting delay for n-th node and N is the number of sensor nodes. Simulations in [16] show that channel utilization reduces as the link error probability increases. HEAP-EDF performs worse in the case of large networks. Also, the single-hop approach is assumed, which limits application of the protocol to small networks.

3.5. Comparison of Protocols

Table 1 shows the performance of MAC protocols reviewed. B-MAC has a high throughput owing to preamble sampling, but this increases the overhead too. Since probabilistic polling and HEAP-EDF are based on an ambient energy harvesting source, energy consumption is not a relevant factor to be compared. In this case of HEAP-EDF, overhead can be decreased if energy harvesting rates are correlated. Protocols with high overheads cannot be scaled to a large network due to the increase in the number of control

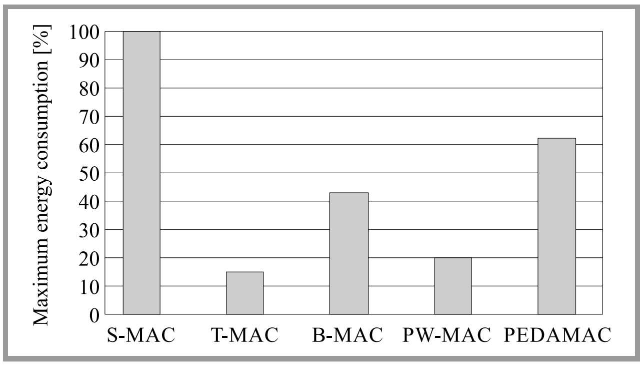


Fig. 6. Maximum energy consumption of protocols (considering S-MAC as a full-scale benchmark).

packets. In PEDAMAC, as transmission power decreases, the range of radio also decreases, which affects the network scalability. Column 4 represents the relative percentage of energy saved. In the reviewed papers, simulations are performed under different scenarios and with different considerations, so it is difficult to directly compare these protocols. Hence, the comparison values are presented as percentages of S-MAC serving as a benchmark. Figure 6 shows the energy consumption analysis. S-MAC consumes 2.8 mA/node and T-MAC consumes 0.4 mA/node [8]. B-MAC saves 57% more energy than S-MAC for a throughput of 240 b/s, because synchronization overhead increases in S-MAC [10]. PW-MAC protocol's duty cycle is only 11%, while duty-cycle of S-MAC is 50% [11]. Decreased duty-cycle leads to decreased energy consumption. Also, owing to operation at distinct power levels, PEDAMAC saves 38% more energy than S-MAC [14].

Table 2
Comparison of MAC protocols

Protocol	Type	Cross layer optimization	Energy conservation factors
S-MAC [7]	Single hop	No	Overhearing, idle listening
T-MAC [8]	Single hop		Idle listening, collision
B-MAC [9], [10]	Single hop		Overhearing, collision
PW-MAC [11], [12]	Multi hop		Idle listening, collision, retransmission
PEDAMAC [13], [14]	Single hop		N/A
Probabilistic polling [15]	Single hop		N/A
HEAP-EDF [16]	Single hop		N/A

Table 2 shows the comparison of MAC protocols. Column 4 represents the factors that were focused on while designing the respective protocols, in order to reduce energy consumption. The key consideration of probabilistic polling and HEAP-EDF is the optimal use of harvested energy rather than conservation of energy.

3.6. Open Research Problems

A number of MAC protocols have been proposed and designed for WSN, but there are still many open issues that need to be addressed. Cross-layer interaction and optimization are potential areas of research which can enhance the performance of MAC protocols. The MAC protocols available can be analyzed for various traffic generation and node distribution models. Development of multi-hop MAC protocols, in order to extend range and scalability, is another task to be considered in the future.

4. Network Layer

The main task of WSN is to sense and transmit data while using minimum resources. An efficient routing protocol is required at the network layer to choose a path with the minimum cost of delay, lifetime, energy or any other parameter that is more relevant to the application.

4.1. Energy Consumption Sources

Routing overhead is the main source of energy consumption at this particular level. Wang *et al.* presents one of the criteria to design the routing protocol with a minimum overhead to minimize energy consumption [17]. The overhead of a routing protocol varies with hop count and hop distance. In the case of small distances, single hop routing has less overhead. However, if the distance is long and cannot be covered with the available transmitted power, multi-hop routing is more efficient.

4.2. Design Constraints of a Routing Protocol

As compared to routing protocols designed for computer networks, WSN routing protocols need some distinctive features to handle a unique set of challenges [18], [19]:

Network scale: Node density may vary from hundreds to thousands, depending upon application. It is difficult to supervise such large, distributed structures. So, sensor nodes should be able to self-organize. The routing protocol should also deal with maintenance of global knowledge of such a large deployment.

Dynamics of node: WSNs are highly dynamic in nature. Owing to movement, power depletion and addition of new nodes, the topology of a WSN keeps changing. The routing protocol should be capable of adapting to frequent changes.

Resource constraints: A WSN need to operate on limited battery resources. Hence, the routing protocol should be able to transmit information over less than half a duty cycle. Some information-possessive applications need accuracy in data transmission. Therefore, the protocol should be able to trade-off energy consumption for accuracy.

Nature of node: Nodes operating over a certain coverage area may be homogenous or heterogeneous. Hence, the routing protocol needs to support nodes with unlike parameters and capabilities.

QoS: In a few applications delayed transmission of the sensed data may result in the loss of its significance. For such applications, delay is a critical parameter. Similarly, for a few other applications, other parameters – such as accuracy – may play a critical role. To maintain the quality of response of the application, these parameters need to be carefully traded-off for energy.

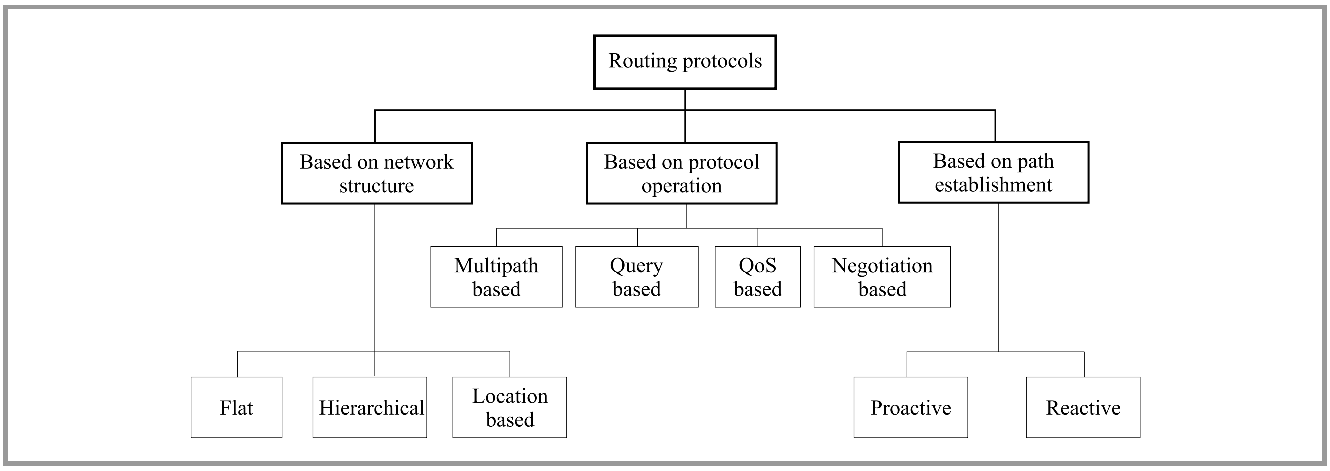


Fig. 7. Taxonomy of routing protocols.

4.3. Classification of Routing Protocols

Nodes can select from a number of available paths to transmit data to the sink or the base station. Routing protocols can be classified based on different criteria establishing the path to the sink, as shown in Fig. 7 [18], [20].

Routing protocols based on network structure:

Flat structure routing: All the nodes play the same role, i.e. each node is considered to be a base station and each node is provided with all information, so that the user can send a query to any node to get information.

Hierarchical structure routing: Not all nodes have the same capability. Higher capability nodes perform critical tasks, whereas less critical tasks are assigned to nodes with low capability. It is a two-level or multi-level structure.

Location-based routing: Nodes can be addressed based on their locations, whereas the location of a sensor can be detected using a satellite, provided the system is equipped with a low power GPS receiver. Another way is to measure the relative distance of the node from its neighbors based on strength of the signal received.

Routing protocols based on protocol operation:

Multipath routing protocol: In order to deliver data from source to destination, the protocol may rely on multiple paths. Multiple paths increase fault tolerance of the network, but also increase energy consumption and protocol overhead. An extension of the algorithm considers only the path with nodes having the highest energy. The path keeps changing whenever the protocol discovers a better path. By using the multipath routing protocol, reliability of the network can be increased in highly unreliable environments. A large packet can be divided into sub-packets and sent over different paths. A message can be reconstructed even if one of the sub-packets is lost due to link errors. Such an approach is known as multipath routing.

Query based routing: In query based routing, a node initiates a query and propagates it through the network. Each node receives the query and only the node having data that matches responds. Instead of propagating the query throughout the entire network, the node may send it in a random direction and wait for the response. If none of the nodes respond, then the node can propagate it through the whole network.

QoS based routing: In applications where parameters like delay, resources and bandwidth are critical, the routing protocol needs to maintain the quality and specifications of the critical parameter while delivering data. The routing protocol is responsible for maintaining a trade-off between energy and other metrics.

Negotiation based routing: Flooding and gossiping produce implosion and a single node may receive multiple data copies. The basic concept of the negotiation based protocol is to avoid propagation of duplicated packets. A sequence of negotiation messages is shared among the nodes to transmit redundant data to the next node. It reduces energy consumption and network congestion. The SPIN protocol discussed later is an example of the negotiation based protocol.

To deliver data from source to destination, the node initiating communication should know the path to the destination, i.e. path-based routing protocols are established in two ways:

Reactive path establishment: such protocols are event-driven. After a data packet has reached a node, the protocol decides the next node to be taken towards the destination. The decision about the next node may depend upon cache history, but in most cases the nodes have limited memory and low computational capability, hence no cache history is maintained. Another metric to decide the next hop can include distance, cost, bandwidth and energy of the node.

Proactive path establishment: under this scenario, the protocol decides the path to destination when the data packet is at the first hop or at the node where communication initiates. The path can be established based on

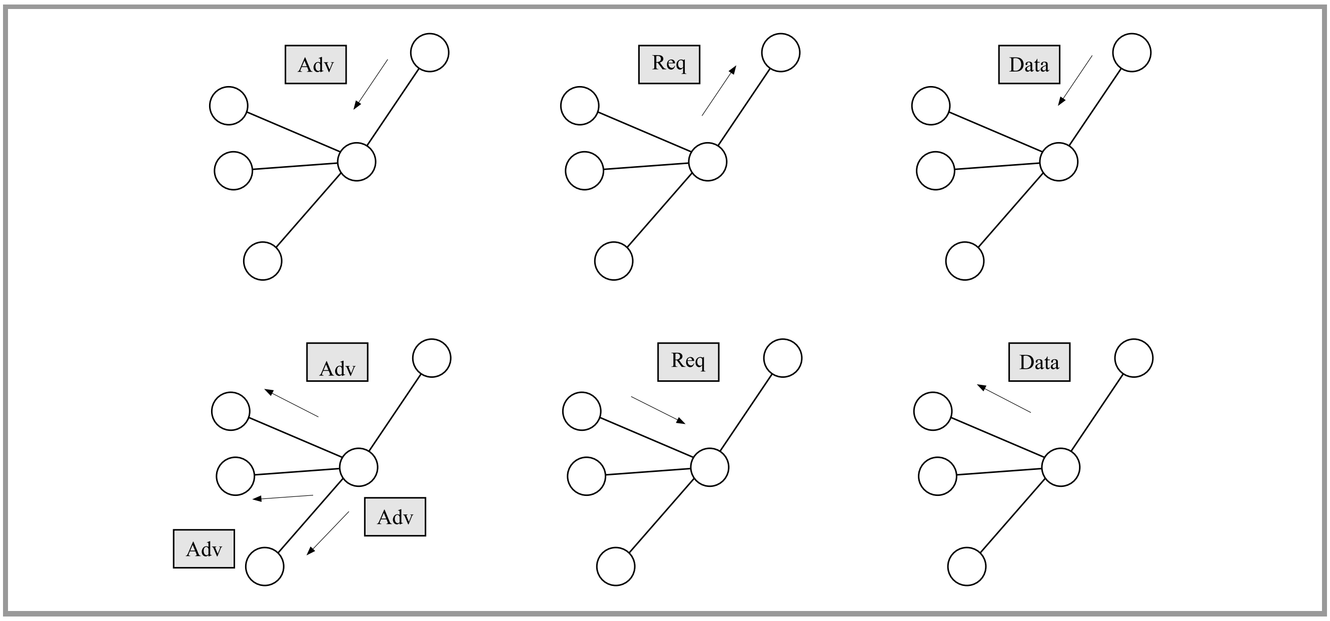


Fig. 8. SPIN protocol flowchart.

the minimum cost, maximum bandwidth or nodes with the highest energy levels. Once the path is established, all data packets propagate through the path selected. These protocols are not fault-tolerant, as data will be lost if the established link fails.

4.4. Network Layer Protocols

Flooding and gossiping: Flooding allows a node to send packets via all links. To avoid a packet looping indefinitely, the hop count or time-to-live is included in the packet. Another approach is gossiping, in which the packet is not sent via all outward links. The packet is transmitted only to a randomly selected neighbor. This saves bandwidth of the network but increases the delay from source to destination [21].

Sensor Protocols for Information via Negotiation (SPIN): SPIN overcomes the drawbacks of conventional dissemination protocols. SPIN is based on metadata [22]. Transmitter broadcasts the metadata of data. The receiver checks the information about data and sends the request to the transmitter if interested. Finally, the transmitter transmits the information to the interested receiver.

The SPIN protocol is presented in Fig. 9, where Adv are advertisement packets advertising metadata, Req are request for data packets from interested nodes to transmitter and Data are packets carrying data.

Directed diffusion: In directed diffusion, data packets are propagated through the network as interests, whereas the reverse reply link towards the transmitter is known as gradient [23]. Each node maintains a cache. When an event occurs, the node searches its cache. If the entry is not in the cache, it is added for future use. Caching increases efficiency and decreases energy consumption. Using the sequence of interests and gradients, the best path is reinforced between the transmitter and the receiver. Directed diffusion is based on the localized demand-driven query model. Receiver queries the sender node through interests for data and gets the response. The query-driven model increases the overhead.

Low-Energy Adaptive Clustering Hierarchy (LEACH): LEACH is a hierarchical cluster-based protocol. Nodes with higher energy are cluster heads [24] collecting information from all nodes in the cluster. Aggregated data is compressed and sent to the sink. LEACH reduces energy consumption, because cluster heads can be selected efficiently to increase network lifetime. The node generates a random number between 0 and 1, if the number generated is less than $T(n)$, the node can become a cluster head. The threshold ensures that the node has not become the cluster head in last $\frac{1}{p}$ rounds:

$$T(n) = \begin{cases} 0 & \text{if } n \notin G \\ \dfrac{p}{1-p\left(r \mod \left(\frac{1}{p}\right)\right)} & \forall\, n \in G \end{cases}, \tag{3}$$

where $T(n)$ is the threshold to choose the cluster head, G is the set of all nodes eligible for cluster head role, p is probability of being the cluster head and r is the current round number.

LEACH protocol is demonstrated in Fig. 9, where sensor nodes send data to cluster heads and cluster heads send aggregated data to the base station.

LEACH with Spare Management (LEACH-SM): It is a modification of the LEACH protocol. LEACH-SM has spare nodes which are normally in the sleep mode [25].

Table 3
Performance evaluation of MAC protocols

Protocol	Type	Energy consumption	Network lifetime	Mobility	Scalability
Flooding and gossiping [21]	Flat	High	Small	Yes	Low
SPIN [22], [28]	Negotiation based	Low	Small	Yes	Low
Directed diffusion [23]	Multipath	Moderate	Small	Limited	Low
LEACH [24], [29]	Hierarchical	High	Medium	No	Moderate
LEACH-SM [25]	Hierarchical	Moderate	Long	No	Moderate
DEEC [26]	Hierarchical, multilevel, heterogeneous	Low	Long	No	High
BLR [27], [28]	Location based	Low	Moderate	Limited	High

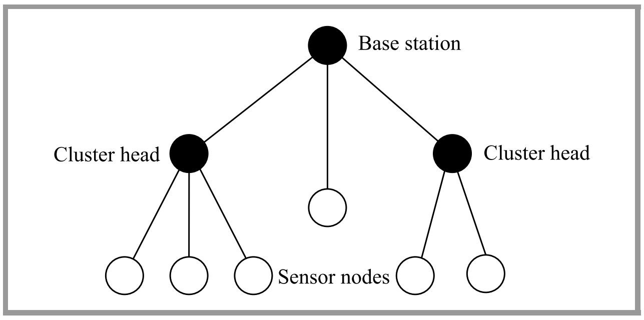


Fig. 9. Structure of a LEACH protocol network.

When the network is out of energy, spare nodes provide redundancy and increase network lifetime. LEACH-SM also has the capability to avoid deadlocks that may occur due to redundancy of nodes, and thus offers extended lifetime.

Distributed Energy Efficient Clustering (DEEC): DEEC was proposed for heterogeneous WSNs [26]. It considers multi-level heterogeneity. Other clustering protocols did not consider energy while choosing cluster heads. DEEC uses the knowledge about initial and residual energy of nodes while choosing cluster heads. DEEC used the same threshold in Eq. (3) to determine the cluster head, but threshold probability to select the cluster head depends on the heterogeneity of nodes.

$$p = \begin{cases} \dfrac{p'E(r)}{(1+am)E'(r)} & \text{if node is normal} \\ \dfrac{p'(1+a)E(r)}{(1+am)E'(r)} & \text{if node is advanced} \end{cases}, \qquad (4)$$

where p' is reference probability, $E(r)$ is residual energy and $E'(r)$ is average energy of the network, m is fraction of advanced nodes whose energy is a times higher than that of normal nodes. Normally, a cluster dies as its cluster head is out of energy. DEEC keeps on reassigning the role of the cluster head depending upon energy, to extend the lifetime of network. The DEEC stability period is 15% longer than in the stable election protocol. Also, it does not require any global knowledge to select the cluster head. Hence, it is more efficient than other clustering protocols.

Beacon-less Routing (BLR): In location based routing, nodes exchange a few messages called beacons to know the position of each other. These beacons create a large overhead and work inefficiently in erroneous wireless links. Therefore, BLR was proposed. BLR selects the next hop by computing the dynamic forwarding delay. A node broadcasts a data packet to all its neighbors but only the receiving node which is best positioned towards the destination, will forward the packet. Nodes within a certain area take part in forwarding. These areas are called forwarding areas and can be of any shape. The receiving node sends a passive acknowledgment back to the sending node [27], [28].

4.5. Comparison Table

Table 3 shows the comparison of the protocols' performance. Since the SPIN protocol is based on metadata, its energy consumption is low. The BLR protocol has a lower overhead. Hence, it is highly scalable and can be used for large networks. Figure 10 represents a lifetime analysis of the hierarchical protocols reviewed. The lifetime of LEACH-SM equals 183% of the lifetime of LEACH [25]. Also the lifetime of DEEC is 130% of the lifetime of LEACH [26].

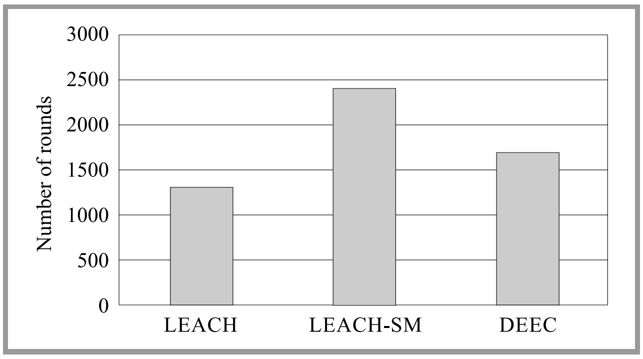


Fig. 10. Lifetime of hierarchical protocols.

4.6. Open Research Problems

Most routing protocols assume a channel to be loss-less. So, future work includes development of routing protocols for lossy wireless channels. Also, existing routing protocols can be evaluated for lossy channels. QoS can be enhanced to ensure latency-free routing. Security management can be explored to avoid such threats as sleep deprivation attacks, packet dropping attacks and collecting sensitive information [30]. The routing protocols considered can be upgraded to handle various types of traffic, and the effect of traffic on the lifetime of a network can be minimized. Different traffic profiles should be modeled and analyzed to design efficient routing protocols [31]. New routing protocols are required for mobile networks. Also, most routing protocols available assume that an ideal MAC protocol exists. Cross-layer optimization can be used to improve the performance of a network as a whole.

5. Cross Layer Approach

To improve network performance, interaction of parameters across the protocol stack is necessary. Energy is a parameter of the physical layer and routing is considered at the network layer. Layers need to interact to obtain the value of energy in a routing packet. This helps the routing protocol to choose an energy efficient path. Route energy packets which are used to exchange energy values among nodes are generated using the cross-layer design. Hoesel *et al.* presents a cross-layer approach in which the routing protocol uses topology and infrastructure information available at the MAC layer [32]. It reestablishes the route utilizing information at the MAC layer and outperforms S-MAC and Dynamic Source Routing (DSR) in mobile sensor networks [33]. Cross Layer MAC (CL-MAC) makes and optimizes scheduling decisions based on cross layer information [34]. Path-Loss Ordered Slotted Aloha (PLOSA) protocol is designed using cross-layer design for wireless data collection networks [35]. It helps in observing physical signals and orders the access of nodes accordingly. Nodes at a greater distance from the collector get an earlier chance to access the slot of the transmission channel. PLOSA has a high delivery rate and low latency.

A cross-layer approach has been presented by Catarinucci *et al.*, in which protocol solutions are integrated with hardware [36]. A new wake-up system consisting of a sensor node and a power meter circuit was suggested in the paper, where the suggested protocol exploits the hardware to reduce power consumption. It is indicated by Alrajeh *et al.* that to design a secure routing protocol, the cross-layer approach was necessary [37]. Most security attacks are multilayered. The sleep deprivation attack, for instance, occurs at the physical layer, whereas the packet dropping attack occurs at the network layer. The author proposed to keep an eye on the packet count, in order to prevent malicious nodes from sending unnecessary packets and creating congestion. Hence, in order to select an energy efficient and secure path, cross-layer communication is necessary.

6. Conclusions and Future Works

In this paper existing MAC protocols and routing protocols have been surveyed. MAC protocols have been reviewed for both type of nodes followed by their advantages and disadvantages. This paper suggests that because of its low latency, PEDAMAC can be used for delay sensitive applications. Owing to the random wake-up schedule, PW-MAC offers high throughput. It sends one update in 1400 s, so overhead is moderate.

Diversified routing protocols ranging from flat to multilevel have been discussed in this paper. This paper analyses that DEEC has 30% more rounds than LEACH, because of low energy consumption. LEACH-SM has 83% more rounds than LEACH. Then, the cross-layer approach has been presented that improves performance of the protocol stack as a system. Although the protocols discussed may seem promising, there are still many challenges that need to be faced in WSNs. The cross-layer approach is a research area that needs to be studied and analyzed widely. Traffic modeling is another prospective area which can be analyzed and studied for improving performance or security of networks. Energy harvesting algorithms and models for WSNs also are subject to great advancements in the future.

References

[1] M. Abozahhad, M. Farrag, and A. Ali, “A comparative study of energy consumption sources for wireless sensor networks”, *Int. J. of Grid Distrib. Comput.*, vol. 8, no. 3, pp. 65–76, 2015.

[2] J. Zheng and A. Jamalipour, *Wireless Sensor Networks: A Networking Perspective*. Wiley, 2014.

[3] A. S. Althobaiti and M. Abdullah, “Medium access control protocols for wireless sensor networks classifications and cross-layering”, *Procedia Comp. Sci.*, vol. 65, pp. 4–16, 2015.

[4] W. Ye, J. Heidemann, and D. Estrin, “An energy-efficient MAC protocol for wireless sensor networks”, in *Proc. of 21st Ann. Joint Conf. of the IEEE Comp. and Commun. Soc. IEEE INFOCOM 2002*, New York, NY, USA, 2002, pp. 1567–1576.

[5] K. Sohraby, D. Minoli, and T. Znati, *Wireless Sensor Networks Technology, Protocols and Applications*, 2nd ed. Wiley, 2013.

[6] U. Roedig, “f-MAC: A deterministic media access control protocol without time synchronization”, in *Wireless Sensor Networks – Third European Workshop, EWSN 2006 Zurich, Switzerland, February 13-15, 2006 Proceedings*, K. Römer, H. Karl, and F. Mattern, Eds., *LNCS* 3868. Zurich: Springer, 2006, pp. 276–291.

[7] M. A. Ameen, S. M. R. Islam, and K. Kwak, “Energy saving mechanisms for MAC protocols in wireless sensor networks”, *Int. J. of Distrib. Sensor Netw.*, vol. 6, no. 1, Article ID 163413, 2010 (doi: 101155/2010/163413).

[8] L. Wang and K. Liu, “An adaptive energy-efficient and low-latency MAC protocol for wireless sensor networks”, in *Proc. Int. Conf. on Wirel. Commun., Netw. and Mob. Comput. WiCOM 2007*, Shanghai, China, 2007, pp. 2440–2443.

[9] C. Cano, B. Bellalta, A. Sfairopoulou, and M. Oliver, “Low energy operation in WSNs: A survey of preamble sampling MAC protocols”, *Comp. Netw.*, vol. 55, no. 15, pp. 3351–3363, 2011.

[10] J. Polastre, J. Hill, and D. Culler, “Versatile low power media access for wireless sensor networks”, in *Proc. 2nd Int. Conf. on Embed. Netw. Sensor Syst. SenSys 2004*, Baltimore, MD, USA, 2004.

[11] D. B. Johnson, “PW-MAC: An energy-efficient predictive-wakeup MAC protocol for wireless sensor networks”, in *Proc. 30th IEEE Int. Conf. on Comp. Commun. IEEE Infocom 2011*, Shanghai, China, 2011, pp. 1305–1313.

[12] J. Kabara and M. Calle, “MAC protocols used by wireless sensor networks and a general method of performance evaluation”, *Int. J. of Distrib. Sensor Netw.*, vol. 8, no. 1, 2012 (doi: 10.1155/2012/834784).

[13] S. Coleri and P. Varaiya, “PEDAMACS: Power efficient and delay aware medium access protocol for sensor networks”, *IEEE Trans. on Mob. Comput.*, vol. 5, no. 7, pp. 920–930, 2006.

[14] M. G. C. Torres, “Energy consumption in wireless sensor networks using GSP”, Master’s Thesis, University of Pittsburgh, 2006.

[15] Z. A. Eu, H. P. Tan, and W. K. G. Seah, “Design and performance analysis of MAC schemes for Wireless Sensor Networks Powered by Ambient Energy Harvesting”, *Ad Hoc Netw.*, vol. 9, no. 3, pp. 300–323, 2011.

[16] Y. Jin and H. P. Tan, “Optimal performance trade-offs in MAC for wireless sensor networks powered by heterogeneous ambient energy harvesting”, in *Proc. IEEE IFIP Networking Conf. IFIP Networking 2014*, Trondheim, Norway, 2014
(doi: 10.1109/IFIPNetworking.2014.6857125).

[17] Q. Wang, M. Hempstead, and W. Yang, “A realistic power consumption model for wireless sensor network devices”, in *Proc. 3rd Ann. IEEE Commun. Soc. on Sensor and Ad Hoc Commun. and Netw. SECON 2006*, Reston, VA, USA, 2006, pp. 286–295
(doi: 10.1109/SAHCN.2006.288433).

[18] J. N. Al-Karaki and A. E. Kamal, “Routing techniques in wireless sensor networks: A survey”, *IEEE Wirel. Commun.*, vol. 11, no. 6, pp. 6–28, 2004 (doi: 10.1109/MWC.2004.1368893).

[19] K. Akkaya and M. Younis, “A survey on routing protocols for wireless sensor networks”, *Ad Hoc Netw.*, vol. 3, no. 3, pp. 325–349, 2005.

[20] S. P. Singh and S. C. Sharma, “A survey on cluster based routing protocols in wireless sensor networks”, *Procedia Comp. Sci.*, vol. 45, no. C, pp. 687–695, 2015.

[21] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: A survey”, *Comp. Netw.*, vol. 38, no. 4, pp. 393–422, 2002.

[22] W. R. Heinzelman, J. Kulik, and H. Balakrishnan, “Adaptive protocols for information dissemination in wireless sensor networks”, in *Proc. 5th Ann. ACM/IEEE Conf. on Mob. Comput. and Netw. MobiCom’99*, Seattle, WA, USA, 1999, pp. 174–85.

[23] C. Intanagonwiwat, R. Govindan, and D. Estrin, “Directed diffusion: A scalable and robust communication”, in *Proc. 6th Ann. Int. Conf. on Mob. Comput. and Netw. MobiCom’00*, Boston, MA, USA, 2000, pp. 56–67.

[24] M. M. Afsar and M. H. Tayarani-N, “Clustering in sensor networks: A literature survey”, *J. of Netw. and Comp. Appl.*, vol. 46, pp. 198–226, 2014.

[25] B. A. Bakr and L. T. Lilien, “Comparison by simulation of energy consumption and WSN lifetime for LEACH and LEACH-SM”, *Procedia Comp. Sci.*, vol. 34, pp. 180–187, 2014.

[26] L. Qing, Q. Zhu, and M. Wang, “Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks”, *Comp. Commun.*, vol. 29, no. 12, pp. 2230–2237, 2006.

[27] J. A. Sanchez, P. M. Ruiz, and R. Marin-Perez, “Beacon-less geographic routing made practical: Challenges, design guidelines, and protocols”, *IEEE Commun. Mag.*, vol. 47, no. 8, pp. 85–91, 2009.

[28] M. Heissenbüttel, T. Braun, T. Bernoulli, and M. Wälchli, “BLR: Beacon-less routing algorithm for mobile ad hoc networks”, *Comp. Commun.*, vol. 27, no. 11, pp. 1076–1086, 2004.

[29] L. J. G. Villalba, A. L. S. Orozco, A. T. Cabrera, and C. J. B. Abbas, “Routing protocols in wireless sensor networks”, *Sensors*, vol. 9, no. 11, pp. 8399–8421, 2009.

[30] K. Chelli, “Security issues in wireless sensor networks: Attacks and countermeasures”, in *Proc. of the World Congr. on Engin. WCE 2015*, London, United Kingdom, 2015, vol. I, pp. 519–524.

[31] Q. Wang, “Traffic analysis and modeling in wireless sensor networks and their applications on network optimization and anomaly detection”, *Netw. Prot. and Algorithms*, vol. 2, no. 1, pp. 74–92, 2010 (doi: 10.5296/npa.v2i1.328).

[32] L. Van Hoesel, T. Nieberg, J. Wu, and P. J. M. Havinga, “Prolonging the lifetime of wireless sensor networks by cross-layer interaction”, *IEEE Wirel. Commun.*, vol. 11, no. 6, pp. 78–86, 2004.

[33] D. B. Johnson and D. A. Maltz, “Dynamic source routing in ad hoc wireless networks”, *Mob. Comput.*, vol. 353, pp. 153–181, 1996.

[34] M. S. Hefeida, T. Canli, and A. Khokhar, “CL-MAC: A cross-layer MAC protocol for heterogeneous wireless sensor networks”, *Ad Hoc Netw.*, vol. 11, no. 1, pp. 213–225, 2013.

[35] D. Espes, X. Lagrange, and L. Suárez, “A cross-layer MAC and routing protocol based on slotted aloha for wireless sensor networks”, *Annales des Telecommun./Annals of Telecommun.*, vol. 70, no. 3–4, pp. 159–169, 2015.

[36] L. Catarinucci *et al.*, “A cross-layer approach to minimize the energy consumption in wireless sensor networks”, *Int. J. of Distrib. Sensor Netw.*, vol. 10, no. 1, 2014 (doi: 10.1155/2014/268284).

[37] N. A. Alrajeh, J. Lloret, and J. Loo, “Secure routing protocol using cross-layer design and energy harvesting in wireless sensor networks”, *Int. J. of Distrib. Sensor Netw.*, vol. 9, no. 1, 2013 (doi: 10.1155/2013/374796).

Aarti Kochhar is working as an Assistant Professor at Lovely Professional University, Phagwara, India. She has completed her M.E. degree in Electronics and Communication Engineering. She conducts research in the areas of wireless sensor networks.

E-mail: aarti.kochhar92@gmail.com
University Institute of Engineering and Technology
Panjab University
Chandigarh, India
Lovely Professional University
Phagwara, India

Pardeep Kaur is working as an Assistant Professor at the Electronics and Communication Engineering Department at U.I.E.T., Panjab University, Chandigarh, India. She received her B.Tech. and M.E. degrees in Electronics and Communication Engineering. She is pursuing her Ph.D. in wireless sensor networks. Her areas of interest include optical communication and wireless communication.

E-mail: pardeep.tur@gmail.com
University Institute of Engineering and Technology
Panjab University
Chandigarh, India

Sukesha Sharma is working as an Assistant Professor at the Electronics & Communication Engineering Department at U.I.E.T., Panjab University, Chandigarh, India. She has completed her B.Tech. and M.E. degrees in Electronics and Communication Engineering. Her research interests include embedded systems, automation and control, active vibration control and energy harvesting.
E-mail: er_sukesha@yahoo.com
University Institute of Engineering and Technology
Panjab University
Chandigarh, India

Preeti Singh is working as an Assistant Professor at the Electronics & Communication Engineering Department at U.I.E.T., Panjab University, Chandigarh, India. She has completed her B.Tech. and M.E. degrees in Electronics and Communication Engineering. She obtained her Ph.D. degree in 2013. Her areas of interest include optical communication (wired and wireless), optical biosensors and cognitive neuroscience.
E-mail: preets.singh.82@gmail.com
University Institute of Engineering and Technology
Panjab University
Chandigarh, India

Characterization of the indoor radio propagation channel at 2.4 GHz

Tadeusz A. Wysocki and Hans-Jürgen Zepernick

http://doi.org/10.26636/jtit.2000.3-4.23

Abstract — The unlicensed industrial, scientific, and medical (ISM) band at 2.4 GHz has gained increased attention recently due to the high data rate communication systems developed to operate in this band. The paper presents measurement results of fading characteristics, multipath parameters and background interference for these frequencies. Some statistical analysis of the measured data is presented. The paper provides information that may be useful in design and deployment of communication systems operating in the 2.4 GHz ISM band, like those compliant with IEEE 802.11 standard and Bluetooth open wireless standard.

Keywords — indoor radiocommunication, microwave propagation, fading channels, jamming.

1. Introduction

Recently a number of data communication systems has been developed to utilize the unlicensed industrial, scientific, and medical band at 2.4 GHz [1]. The two most prominent examples of such systems are IEEE 802.11 wireless LAN [2], and personal area network employing Bluetooth enabled devices [3]. To assist in deploying of those systems, characterization of the indoor radio propagation channel at 2.4 GHz is essential. Measurement results for the indoor radio propagation channel have been presented in various publications. However, they tend rather to focus on a single characteristic, e.g. pulse propagation characteristic, as in [4], or temporal fading caused by motion of people and other objects within the channel [5], or simply deal with different frequency bands, like in [6].
In this paper, we present measurement results for the three major channel characteristics in the 2.4 GHz ISM band, i.e. temporal fading, channel impulse response, and background noise. The paper is organized as follows. Section 2 deals with the temporal fading characteristics. Example measurement results together with the cumulative distribution functions for typical measurements fitted to those of Rician distributions [7] are presented there. Level crossing rates and average duration of fades extracted from the measurements are included, too. Section 3 is devoted to the measurements of multipath channel parameters for the 2.4 GHz ISM indoor channel. In Section 4, the results of background interference measurements are shown, with microwave ovens indicated as major sources of an electromagnetic pollution in this band. Section 5 concludes the paper.

2. Fading characteristics

The measurements reported here were conducted at a laboratory of the Cooperative Research Centre for Broadband Telecommunications and Networking, Curtin University of Technology, Perth, Western Australia. The room topology is shown in Fig. 1. This environment was of relatively small dimensions being rectangular in size 7.8 m by 9.95 m within a three-metre ceiling. The ceiling was located 1.5 m below the steel reinforced concrete floor for the second storey of the four-storey building. The laboratory had two doorways and no windows. It had steel-framed walls clad with plaster-glass, a dropped ceiling constructed with non-metallic acoustic tiles, and a carpeted concrete floor. The laboratory was heavy cluttered with test and measurement equipment located on the benches.

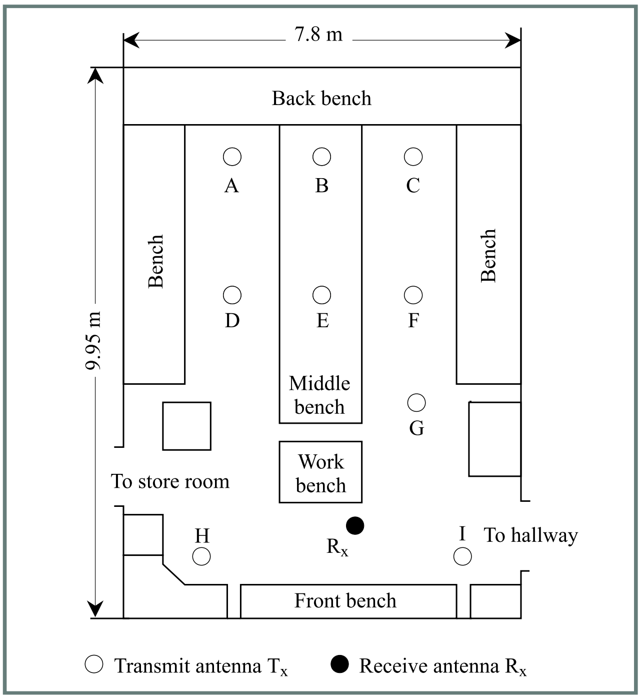


Fig. 1. Antenna placements for fading measurements.

2.1. Measurement procedure

We used the Hewlett Packard HP89441A to feed a quarter wave monopole antenna designed for the frequency range 2.3 ÷ 2.5 GHz. The transmitter and the identical receiving antennas were mounted on two separate identical PVC pipes of heights adjustable in the range 1.0 to 2.0 m. At the receiver, a Marconi TF2300A modulation analyzer was applied to perform an envelope detection of the down converted baseband signal. The resultant signal was sampled, and the results stored on a hard disk for post-processing.

For all fading measurements, receive antenna R_x remained fixed and transmit antenna T_x was moved to different placements (Fig. 1). For each T_x placement and with no movement of people, the received signal was calibrated to -65 dBm. This receive level provided a signal to measurement system noise ratio of 35 dB, thus allowing fade depths of this order to be identified. To determine impairments caused by sources from outside the laboratory, we initially kept all motion in the room at zero. The variations of the received signal amplitude were less than ± 0.2 dB and could be regarded as insignificant. Then, we collected measurements for nine different transmit antenna locations with three people moving around the receive antenna only, keeping them within a two metre radius. The duration of measurements for each pair of antenna locations was twenty seconds. As an example, Fig. 2 shows a typical fading pattern observed with transmitt antenna T_x at the placement C.

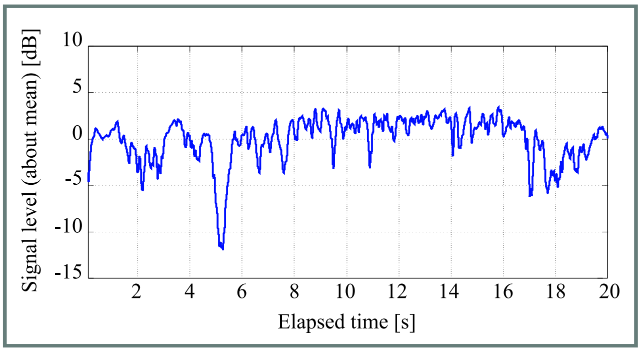


Fig. 2. Fading for transmit antenna at position C.

2.2. Fading distributions

In an indoor environment, we expect a line-of-sight path between transmit and receive antenna. Hence, the probability density function of the fast varying amplitude of the received instantaneous signal can be described by a Rician distribution. The probability that the received amplitude does not exceed a given threshold r is given by integration of the probability density function and is called cumulative distribution function $C(r)$. For curve fitting purposes, it is convenient to use the complementary cumulative distribution function $\overline{C}(r)$, which for a Rician distribution is given by [7]

$$\overline{C}(r) = 1 - C(r) =$$

$$= \exp\left[-\left(K + \frac{r^2}{2\sigma^2}\right)\right] \sum_{m=0}^{\infty} \left(\frac{\sigma\sqrt{2K}}{r}\right) I_m\left(\frac{r\sqrt{2K}}{\sigma}\right), \quad (1)$$

where $I_m(r)$ denotes the modified mth order Bessel function of the first kind, σ^2 is the local-mean scattered power, and K is called Rician factor specifying the ratio of power in the dominant path to power in the scattered path.

In calculating the empirical distribution functions, the measured data was classified into $\beta = 1.87(v-1)^{2/5}$ bins [8] where v is the number of samples collected in a measurement period. Then, a set of hypotheses for $\overline{C}(r)$ with $K = 0$ to $K = 15$ in 0.1 increments were tested to match the measured function. We applied the Kolmogorov-Smirnov goodness-of-fit technique for testing the relevance of match between measurement and hypothesis. Since the power level was normalized about median, a respective Rician factor K fully specifies a particular fading distribution. Table 1 shows the obtained Rician factors K.

Table 1
The Rician factors at various transmit antenna placements

Placement	A	B	C	D	E	F	G	H	I
K [dB]	2.79	8.19	9.86	5.95	8.80	3.80	8.26	7.85	5.79

Figure 3 shows curve fitting results for T_x at C. The large Rician factor of K = 9.86 dB indicates a strong line-of-sight path. Note also that the deviation between measured and matched distribution appears large for small power levels but is actually small due to scaling.

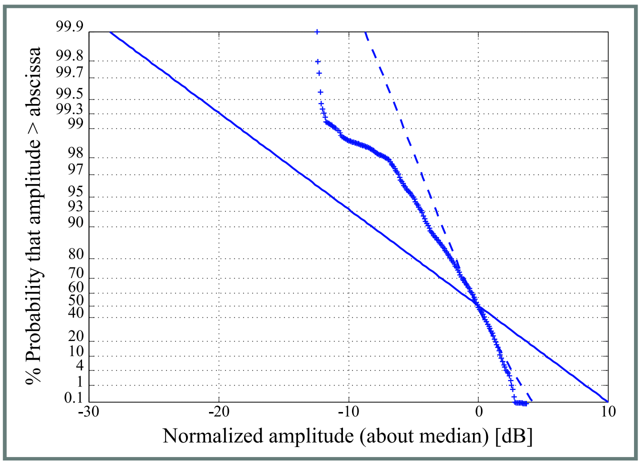


Fig. 3. Complementary cumulative distribution function $\overline{C}(r)$ for transmit antenna T_x at placement C; „+" fading measurement, „- -" Rician fading with K = 9.86 dB, „—" Rayleigh fading as reference.

2.3. Fading statistics

To design data and signaling formats for a wireless system, we require statistics which quantify the number of times a given threshold is crossed and the duration of time for which the signal is below that threshold. By counting all N crossings with positive slope at a given level L for measurement period T, the level crossing rate can be computed as [9]

$$N_L = \frac{N}{T} . \qquad (2)$$

The average duration of fades $\bar{t}_L$ in respect to level L and measurement period T is given by [9]

$$\bar{t}_L = \sum_{i=1}^{N} \frac{t_i}{N} , \qquad (3)$$

where t_i is an ith fade duration, i.e. time for which the received signal is below a given level L. Table 2 displays fading statistics of selected fading measurements. Obviously, the average duration of very deep fades is rather short.

Table 2
Fading statistics for transmit antenna at selected placements

L [dB]	N_L [s^{-1}]			$\bar{t}_L$ [s]		
	C	E	I	C	E	I
6			0.101			9.868
3	0.554	2.195	2.526	1.765	0.363	0.290
0	1.308	2.147	3.183	0.295	0.141	0.118
-3	0.755	0.927	1.768	0.150	0.161	0.080
-6	0.151	0.781	0.859	0.158	0.121	0.092
-9	0.050	0.488	0.354	0.275	0.129	0.143
-12	0.050	0.439	0.303	0.008	0.092	0.131
-15		0.293	0.202		0.085	0.147
-18		0.195	0.202		0.091	0.110
-21		0.195	0.152		0.062	0.115
-24		0.098	0.152		0.069	0.079
-27		0.049	0.051		0.032	0.079

3. Multipath channel parameters

Delay profile measurements were conducted in the same laboratory as the fading measurements. However, this time the location of transmit antenna T_x remained fixed for all measurements, and the receive antenna R_x was placed at different locations within the laboratory. A plan view of the laboratory with the four antenna positions used for delay profile measurements is depicted in Fig. 4.

3.1. Delay profile measurements

Because of the confined environment, we could use the standard channel impulse response measurement system [10] based on the vector network analyzer HP 8753C equipped with an S-parameter test set HP 85046A. We used the same antennas as for the fading measurements. The results were recorded using HP VEE user interface and stored on a hard drive for post-processing. Albeit the 2.4 GHz ISM band spans from 2.4 to 2.485 GHz, we performed measurements in the band 2.1 to 2.8 GHz which gave us a resolution of 837.05 mm, when normal window [11] was used. This translates into time resolution of 2.7921 ns.

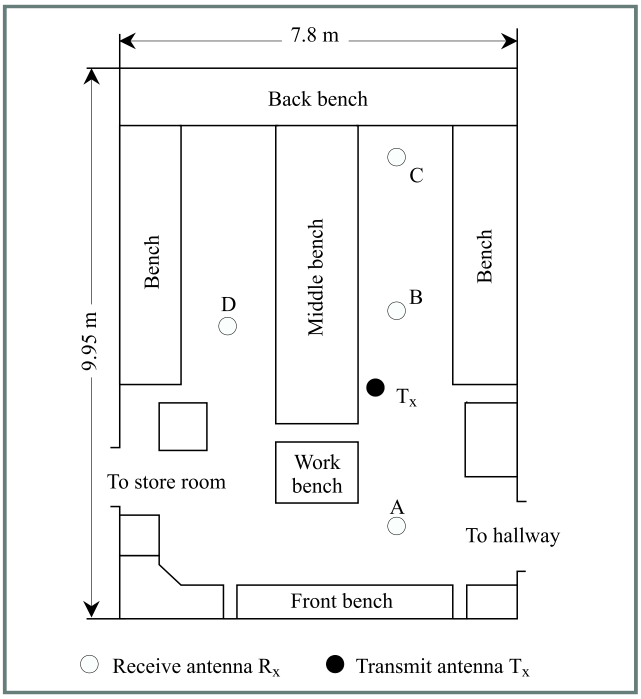


Fig. 4. Antenna placements for delay profile measurements.

During the measurement periods, three people kept moving in a similar manner around the receive antenna R_x, staying all the time within a two metre radius. Receive positions A, B and C were located within the left-hand aisle of the laboratory but with different distances to transmit antenna T_x. The specific characteristic of the receive position A was its direct vicinity to the nearby hallway which could cause additional scattering. In other words, time dispersion parameters at location A were expected to be of higher value than those observed at the three other locations. The particular feature of the receive location D was that the direct line-of-sight to the transmit antenna T_x did not exist because of the middle bench obstruction.

Figures 5 to 8 show the measured power delay profiles for the receive antenna R_x located at positions A, B, C, and D. The largest amplitude of each profile has been normalized to 0 dB.

3.2. Time dispersion parameters

It can be seen from the plots presented in Figs. 5 to 8 that a received wide-band signal will suffer spreading in time compared to the transmitted signal. This effect is called delay spread. Several delay-related parameters used for channel classification can be extracted from the measured power delay profile $P(\tau_\nu)$ where τ_ν denotes νth time instant.

The most commonly used time dispersion parameters are [12]:

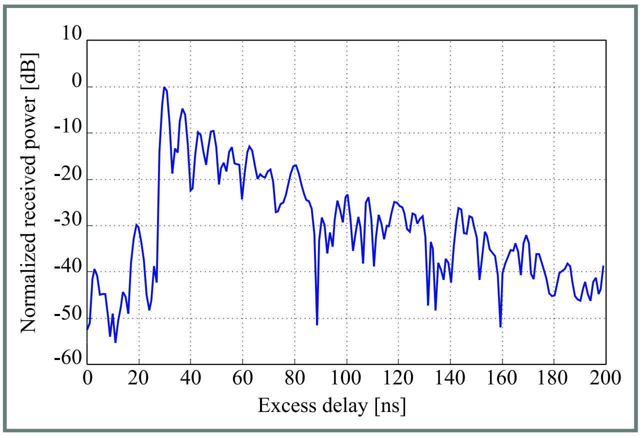


Fig. 5. Power delay profile for receive position *A*.

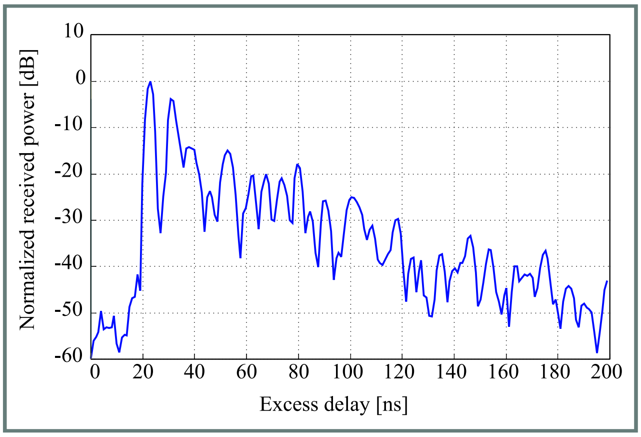


Fig. 6. Power delay profile for receive position *B*.

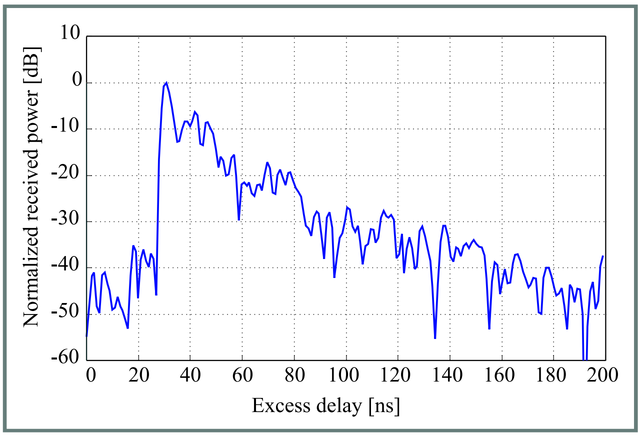


Fig. 7. Power delay profile for receive position *C*.

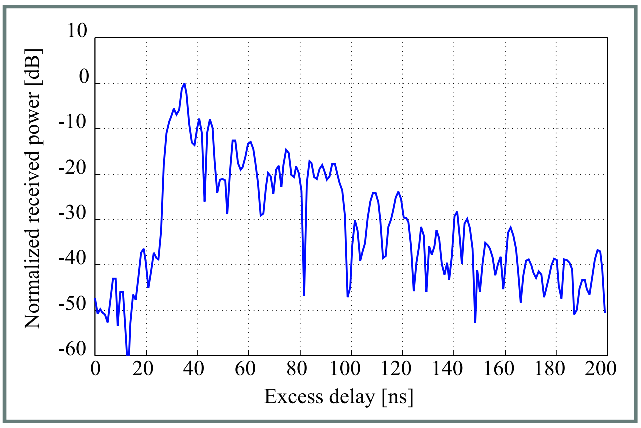


Fig. 8. Power delay profile for receive position *D*.

- Mean excess delay, which is the first moment of the power delay profile defined by:

$$m_\tau = \frac{\sum\limits_\nu P(\tau_\nu)\tau_\nu}{\sum\limits_\nu P(\tau_\nu)} \,. \tag{4}$$

- Root mean square (rms) delay spread, which is the square root of the second central moment of the power delay profile and is given by:

$$\sigma_\tau = \sqrt{\frac{\sum\limits_\nu \left[\tau_\nu - m_\tau\right]^2 P(\tau_\nu)}{\sum\limits_\nu P(\tau_\nu)}} \,. \tag{5}$$

- Maximum excess delay *X* [dB], defined as the time period during which the power delay profile falls to *X* [dB] below its maximum [10].

Note that mean excess delay and rms delay spread have to be computed with respect to a reasonable threshold for the multipath noise floor. If this threshold were set too low, it would result in too high values for these dispersion parameters. Our statistical analysis is based on a noise threshold set to four times of the noise standard deviation, which is known as a rule of thumb (as used in [13]). Numerically, the noise threshold was set to −34 dB with respect to the normalized receiver power.

A dual representation of delay spread in terms of a frequency domain parameter is given by the coherence bandwidth B_c. This parameter specifies the frequency range over which a transmission channel affects the signal spectrum nearly in the same way, giving an approximately constant attenuation and a linear change in phase. The coherence bandwidth is inversely proportional to rms delay spread. Assuming frequency correlation between amplitudes of frequency components being above 0.9, the coherence bandwidth can be approximated by [9]

$$B_c \approx \frac{1}{50\,\sigma_\tau} \,. \tag{6}$$

The time dispersion parameters extracted from the measured power delay profiles are summarized in Table 3. As we expected, the mean excess delay at receive position *A* is higher than at *B* and *C* which was assumed to be caused by multipath scattering into the nearby hallway. It is interesting to note that at the receive position *D*, we obtained the highest value of maximum excess delay in respect to 20 dB threshold from 0 dB power level. This is likely a result of the missing line-of-sight path at receive position *D*.

Table 3
Time dispersion parameters for receive antenna at various placements

Placement	A	B	C	D
Mean excess delay [ns]	57.04	43.40	49.10	54.85
Rms delay spread [ns]	30.55	24.12	22.19	26.82
Max excess delay 20dB [ns]	53.73	59.73	50.74	66.67
Coherence bandwidth [kHz]	654.66	829.19	901.31	745.71

4. Background interference

One of the important aspects of channel characterization is classification of noise and interference sources that need to be taken to account in the given bandwidth. This is particularly important in the case of unlicensed bands, as is the case of the 2.4 GHz ISM band, where users of wireless technologies operating in these bands are not required to obtain operating licenses provided that higher gain antennas are not used. On the other hand, there are no guarantees that the band is free of interference.

To identify the possible noise and interference sources that can affect use of the 2.4 GHz ISM band for communication purposes, we performed a series of measurements in some typical operating environments, including a university campus, a large shopping centre, and an industrial workshop. We used Hewlett Packard Spectrum Analyzer HP 8595E with a colinear antenna for omnidirectional measurements and a corner reflector antenna for directional measurements, i.e. when the particular source had been identified.

After conducting those measurements, we identified one major class of interference sources radiating in the 2.4 GHz ISM band, which were microwave ovens. Apart from that, the band was almost clear. The only other interference sources were building alarms as those used typically in businesses. Although only discovered in two cases from various measurement sites, they were found to output discrete frequency signals of moderate to low levels. The radiation came from the alarm units themselves (the most likely from oscillator circuits), and not the sensors, as they operated at infrared frequencies. A typical spectrum analyzer trace taken at a distance of 1 m from the alarm unit, with the directional corner reflector antenna is given in Fig. 9. The reading was obtained at Wilson's Engraving Works in Perth, Western Australia.

Microwave ovens were identified as major sources of interference at both the university environment and the shopping mall. The spectrum analyzer trace obtained from the 2 m measurements, using the directional corner reflector antenna and „peak-and-hold" function of the analyzer, is shown in Fig. 10. The plot presents raw data only. To derive the exact values of interference level from the trace it is necessary to account for system gains and losses.

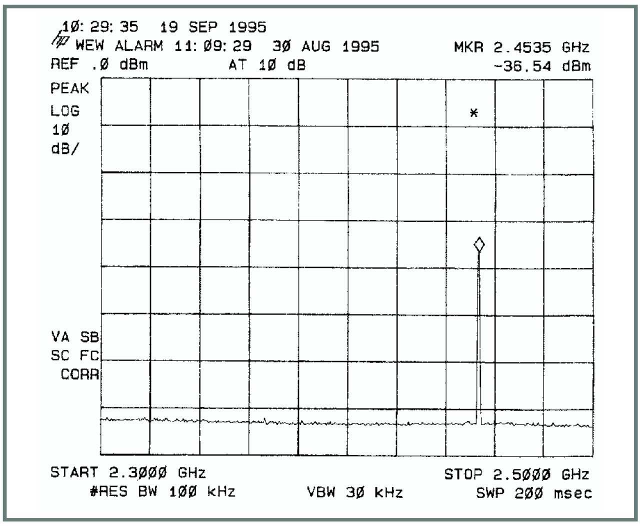


Fig. 9. Sample spectrum analyzer output for an alarm unit.

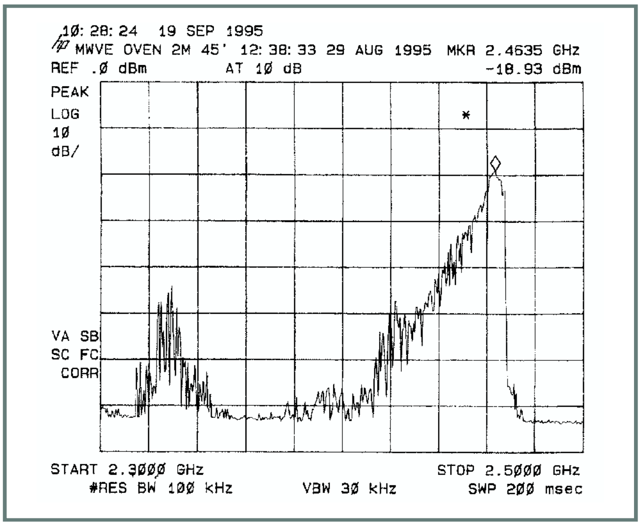


Fig. 10. Sample spectrum analyzer output for a single microwave oven at 2 m using the corner reflector antenna.

The plot of Fig. 10 shows a typical signature of the leakage output spectrum for a microwave oven. Brief analysis of two other microwave ovens, demonstrated that they at least (and presumably most others) share a similar signature of leakage output spectrum, both in intensity (related to efficiency and sealing of microwave) and spectrum occupation. Considering microwave ovens operating in multiplicity, the omnidirectional measurements were taken. This was deemed so, because of the multitude of signals and many reflective surfaces from which they could bounce in the Whitford City Shopping Centre Food Hall. It was impossible to selectively determine the source of any one signal since there were many food outlets operating microwave ovens simultaneously. Rather, the various microwave oven output signals combined producing a composite interference environment, that could only be measured with relevance with an omnidirectional antenna.

Figure 11 shows a typical output trace for the food hall environment. Once again, the trace provides us with a spectral signature of the environment. It has the usefulness of providing a good indication of what spectrum one can expect in such a topology, with respect to frequency variation and interference intensity. The trace is actually similar to the one presented in Fig. 10, except that the noise is spread more evenly over the entire ISM band, as a result of the differing characteristics of the various microwave ovens. However, it is still possible to notice the peak in noise power between 2.4 ÷ 2.47 GHz, which is directly coincident with the band of interest for microwave wireless LAN's.

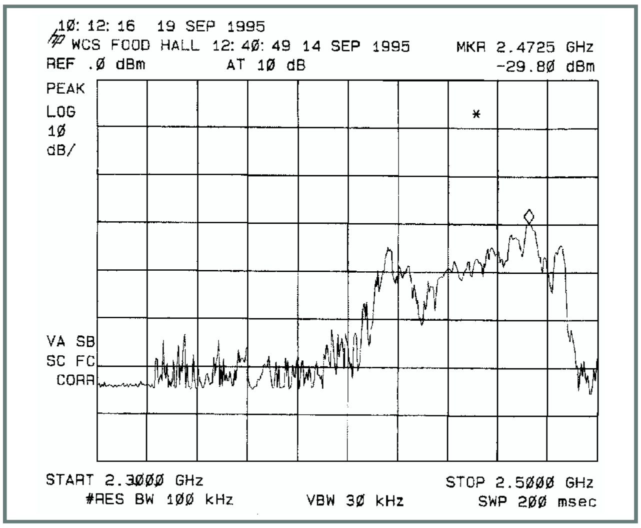


Fig. 11. Sample spectrum analyzer output for multiple microwave oven environment using omnidirectional antenna.

Both plots presented in Figs. 10 and 11 indicate that microwave ovens generate interference of a considerable level in the whole 2.4 GHz ISM band. The plots were obtained using „peak-and-hold" function of the analyzer. The instantaneous output of the analyzer revealed that the single operating microwave oven generated a single tone signal with the frequency hopping within the whole bandwidth. Therefore, one can expect the direct sequence (DS) spread spectrum (SS) devices, like DS mode of IEEE 802.11 wireless LAN, should perform quite well even in the close proximity to operating microwave ovens. On the other hand, the frequency hopping (FH) SS devices, like the Bluetooth compliant ones might experience difficulties in such an environment.

5. Conclusions

In this paper, we presented the measurement results, which can be used to characterize the 2.4 GHz unlicensed ISM band from the viewpoint of its usefulness for high data rate communications. The reported measurements deal with characteristics of fades and multipath channel parameters. In addition, special considerations are given to identifying possible sources of interference present in this band. The fading characteristics as well as multipath parameters were measured at one location that was a heavy cluttered laboratory. Thus, the presented results can be used to characterize indoor environment. The background interference measurements were performed at several different locations. The only major source of interference identified were microwave ovens, which based on the results, can have significant impact even on some SS systems. Because of the nature of this interference, it can be expected to particularly impair signal quality in FH SS systems, like Bluetooth enabled devices.

References

[1] V. K. Garg, K. Smolik, and J. E. Wilkes, *Applications of CDMA in Wireless/Personal Communications*. Upper Saddle River: Prentice Hall, 1997.

[2] „IEEE 802.11 Standard for Wireless LAN", IEEE Standards Department, New York, 1997.

[3] J. Haartsen *et al.*, Bluetooth Specification Version 1.0, Part B, Baseband Specification, 1999.

[4] S.-C. Kim, H. L. Bertoni, and M. Stern, „Pulse propagation characteristics at 2.4 GHz inside buildings", *IEEE Trans. Veh. Technol.*, vol. VT-45, no. 3, pp. 579–592, 1996.

[5] E. Walker, H.-J. Zepernick, and T. Wysocki, „Fading measurements at 2.4 GHz for the indoor radio propagation channel", in *1998 Int. Zurich Sem. Broadband Commun.*, Zurich, Switzerland, Feb. 1998, pp.171–176.

[6] H. Hashemi, M. McGuire, T. Vlasschaert, and D. Tholl, „Measurements and modelling of temporal variations of the indoor radio propagation channel", *IEEE Trans. Veh. Technol.*, vol. VT-43, no. 3, pp. 733–737, 1994.

[7] R. Steele, *Mobile Radio Communications*. New York: IEEE Press, 1994.

[8] J. S. Bendat and A. G. Piersol, *Measurement and Analysis of Random Data*. New York: Wiley, 1966.

[9] W. C. Y. Lee, *Mobile Communication Design Fundamentals*. New York: Wiley, 1993.

[10] T. S. Rappaport, *Wireless Communications – Principles and Practice*. New York: IEEE Press, 1996.

[11] „HP 9753C Network Analyzer Operating Manual", Hewlett-Packard, USA, 1990.

[12] T. Wysocki, H.-J. Zepernick, and R. Weber, „Mobile Communications", in *Wiley Encyclopedia of Electrical and Electronics Engineering*, J. G. Webster, Ed. New York: Wiley, 1999, vol. 13, pp. 343–353.

[13] D. Lacroix, C. L. Despins, G. Y. Delisle, P. Marinier, and P. Luneau, „Experimental characterization of outdoor microcellular quasi-static channels in the UHF and SHF bands", in *Proc. IEEE ICCC'97*, Montreal, CD, June 1997.

Tadeusz A. Wysocki received the M.Sc.Eng. degree with the highest distinction in telecommunications from the Academy of Technology and Agriculture, Bydgoszcz, Poland, in 1981. In 1984, he received his Ph.D. degree, and in 1990, was awarded a D.Sc. degree (habilitation) in telecommunications from the Warsaw University of Technology. In 1992, Dr. Wysocki moved to Perth, Western Australia to work at Edith Cowan University. He spent the whole 1993 at the University of Hagen, Germany, within the framework of Alexander von Humboldt Research Fellowship. After returning to Australia, he was appointed a Project Leader, Wireless LANs, within Cooperative Research Centre for Broadband Telecommunications and Networking. Since December 1998 he has been working as an Associate Professor at the University of Wollongong, NSW, within the School of Electrical, Computer and Telecommunications Engineering, being a Research Coordinator of the Switched Networks Research Centre. The main areas of Dr. Wysocki's research interest include: indoor propagation of microwaves, code division multiple access (CDMA), digital modulation and coding schemes as well as mobile data protocols including those for ad-hoc networks. He is the author or co-author of three books, over 100 research publications and nine patents. He is a Senior Member of IEEE.

e-mail: Wysocki@uow.edu.au
School of Electrical
Computer and Telecommunications Engineering
University of Wollongong
NSW 2251

Hans-Jürgen Zepernick received the Dipl.-Ing. degree in electrical engineering from the University of Siegen, Germany, in 1987. He then was with the Radio and Radar Department of Siemens AG, Munich, Germany. From August 1989 until April 1995, he was with the Department of Communications Engineering at the University of Hagen, Germany, researching into the areas of mobile communications and error control coding. In 1994, he received the Dr.-Ing. degree. In 1995, he joined the Cooperative Research Centre for Broadband Telecommunications and Networking in Perth, Australia, as a Research Fellow working in the Wireless ATM project on physical layer topics. He is an author or co-author of some 45 technical papers. His research interests include radio channel characterisation and modelling, coding and modulation, equalisation, spread-spectrum systems, wireless networks and third generation wireless systems.

e-mail: Hans@atri.curtin.edu.au
Australian Telecommunications Research Institute
Curtin University of Technology

Extensions of the minimum labelling spanning tree problem

Raffaele Cerulli, Andreas Fink, Monica Gentili, and Stefan Voß

http://doi.org/10.26636/jtit.2006.4.396

Abstract— In this paper we propose some extensions of the minimum labelling spanning tree problem. The main focus is on the minimum labelling Steiner tree problem: given a graph *G* with a color (label) assigned to each edge, and a subset *Q* of the nodes of *G* (basic vertices), we look for a connected subgraph of *G* with the minimum number of different colors covering all the basic vertices. The problem has several applications in telecommunication networks, electric networks, multimodal transportation networks, among others, where one aims to ensure connectivity by means of homogeneous connections. Numerical results for several metaheuristics to solve the problem are presented.

Keywords— network design, metaheuristics, spanning trees, labelling trees, Steiner tree problem.

1. Introduction

Many real-world problems can be modelled by means of graphs where a label or a weight is assigned to each edge and the aim is to optimize a certain function of these weights. In particular, one can think of problems where the objective is to find homogeneous subgraphs (respecting certain connectivity constraints) of the original graph. This is the case, e.g., for telecommunication networks (and, more generally, any type of communication networks) that are managed by different and competing companies. The aim of each company is to ensure the service to each terminal node of the network by minimizing the cost (i.e., by minimizing the use of connections managed by other companies).

This kind of problem can be modelled as follows. The telecommunication network is represented by a graph $G = (V,E)$ where with each edge $e \in E$ is assigned a set of colors L_e and each color denotes a different company that manages the edge. The aim of each company is to define a spanning tree of G that uses the minimum number of colors. When the graph represents a transportation network and the colors, assigned to each edge, represent different modes of transportation, then looking for a path that uses the minimum number of colors from a given source s to a given destination t means to look for a path connecting s and t using the minimum number of different modes of transportation.

We focus on the *minimum labelling Steiner tree problem* (MLSteiner): given a graph $G = (V,E)$, with a label (color) assigned to each edge and a subset $Q \subseteq V$ of nodes of G (basic vertices or nodes), we look for an acyclic connected subgraph of G spanning all basic nodes and using the minimum number of different colors. This problem is an extension of the *minimum labelling spanning tree problem* (MLST): given a graph G with a label (color) assigned to each edge we look for a spanning tree of G with the minimum number of different colors.

In this paper, first we review the earlier results existing in the literature to solve the MLST. Then we discuss how these approaches can be easily extended to efficiently solve the MLSteiner and present a comprehensive study of experimental results.

The sequel of the paper is organized as follows. Section 2 summarizes existing approaches for the MLST as well as some important references on the Steiner problem in graphs. In Section 3 we sketch some extensions of the MLST related to the MLSteiner which is the focus of this study. Section 4 presents our modifications of the solution approaches for the MLST to solve the MLSteiner. In Section 5 we present experimental results, and, finally, Section 6 gives some further research options.

2. Literature review

2.1. Earlier approaches to solve the MLST

The MLST was initially addressed by Broersma and Li [2]. They proved, on the one hand, that the MLST is $\mathcal{NP}$-hard by reduction from the minimum dominating set problem, and, on the other hand, that the "opposite" problem of looking for a spanning tree with the maximum number of colors is polynomially solvable. Independently, Chang and Leu [6] provided a different $\mathcal{NP}$-hardness proof of the problem by reduction from the set covering problem. They also developed two heuristics to determine feasible solutions of the problem and tested the performance of these heuristics by comparison with the results of an exact approach based on an A^* algorithm.

Krumke and Wirth [14] formulated an approximation algorithm (in the sequel referred to as maximum vertex covering algorithm – MVCA) with logarithmic performance guarantee and showed also that the problem cannot be approximated within a constant factor. Wan *et al.* [19] provided a better analysis of the greedy algorithm given in [14] by showing that its worst case performance ratio is at most $\ln(n-1)+1$ where n denotes the number of nodes of the given graph, i.e., $n = |V|$. Recently, Xiong *et al.* [23] obtained the better bound $1+\ln b$ where each color appears at most b times. Moreover, Xiong *et al.* [21] proposed a genetic algorithm to solve the MLST and provided some experimental results.

In [5] we presented several metaheuristic approaches to solve the MLST (namely, simulated annealing, reactive tabu search, the pilot method and variable neighborhood search) and compared them with the results provided by the MVCA heuristic presented in [14, 21]. Recently, a modification of our pilot method combined with the genetic algorithm of Xiong *et al.* [21] was shown to be effective by [22].

A variant of the problem has been studied by Brüggemann *et al.* [3], where the MLST with bounded color classes has been addressed. In this variant, each color of the graph is assumed to appear at most r times. This special case of the MLST is polynomially solvable for $r = 2$, and $\mathcal{NP}$-hard and APX-complete for $r \geq 3$. Local search algorithms for this variant, that are allowed to switch up to k of the colors used in a feasible solution have been studied, too. For $k = 2$, the authors showed that any local optimum yields an $\frac{(r+1)}{2}$-approximation of the global optimum, and this bound is best possible. For every $k \geq 3$, there exist instances for which some local optimum is a factor of $\frac{r}{2}$ away from the global optimum.

2.2. The Steiner tree problem

Consider an undirected connected graph $G = (V, E)$ with node set V, edge set E, and nonnegative weights associated with the edges. Given a set $Q \subseteq V$ of specified vertices (called terminals or basic vertices) *Steiner's problem* in graphs (SP) is to find a minimum cost subgraph of G such that there exists a path in the subgraph between every pair of basic vertices. In order to achieve this minimum cost subgraph additional vertices from the set $S := V \setminus Q$, called Steiner vertices, may be included. Since all edge weights are assumed to be nonnegative, there is an optimal solution which is a tree, called Steiner tree.

Correspondingly, *Steiner's problem in directed* graphs (SPD) is to find a minimum cost directed subgraph of a given graph that contains a directed path between a root node and every basic vertex. Applications of the SP and the SPD are frequently found in many problems related to network design and telecommunications. Beyond that, SP and SPD have equal importance also for the layout of connection structures in networks as, e.g., in topological network design, location science and VLSI (very large scale integrated circuits) design.

The SP is a well-studied problem and there is a wealth of excellent reference providing information on Steiner problems, such as [11]. Additional surveys on quite broad aspects of Steiner tree problems are provided by [10, 16, 20] as well as, most recently, [17].

3. Extension of the MLST

Steiner tree problems refer to important problem classes in graphs. The SP may be called one of the most important combinatorial optimization problems. Modifications and generalizations of Steiner tree problems will certainly arise and become a core focus of research and telecommunications applications including additional online optimization problems as well as stochastic optimization approaches. In that sense we have defined the MLSteiner as an extension of both, the MLST as well as the SP. But in this section we go beyond this.

Examples for possible generalizations may include, e.g., a weighted labelling Steiner tree problem with budget constraints. Here we are given a graph G with a label (color) assigned to each edge and we look for a spanning tree with respect to a given subset Q of the nodes of G with the minimum number of different colors. Furthermore, one may incorporate some weights on the edges and define a budget constraint on the sum of the weights of included edges while still minimizing the number of labels or more versatile capacitated Steiner tree problems.

Other ideas on generalizations of the MLST refer to certain ring network design problems with or without budget constraints (see, e.g., [8]) that may be formulated in terms of minimizing the number of labels once they are to be defined and considered. While these generalizations may prove to be important, subsequently we focus on the MLSteiner.

4. Different metaheuristic approaches to solve the MLSteiner

We aim for a Steiner tree which connects all required or basic nodes with a minimum number of colors/labels. Within the steps of the search process only those edges that are colored according to the currently activated colors may be used. For describing different metaheuristic approaches for the MLSteiner we heavily rely on our previous approaches for the MLST described in [5]. We have adapted our code for the MLST so that the algorithm checks whether the resulting subgraph (restricted to the edges with actually used or activated colors) connects all required nodes. If there are disconnections, large penalty values are added to the objective function so that the search process is directed towards feasibility.

Before going into detail let us introduce some notation. Given an undirected graph $G = (V, Q, E)$ with V being the set of nodes, $Q \subseteq V$ the subset of basic nodes and E denoting the set of edges, let c_e be the color (label) associated with edge $e \in E$ and $L = \{c_1, c_2, \dots, c_l\}$ be the set of all colors. We denote by $C(F) = \bigcup_{e \in F} c_e$ the set of colors assigned with edges in $F \subseteq E$. Any subgraph T of G can be represented by the set of its colors $C(T)$. Given a set of colors C, we define by $V(C)$ the subset of nodes of G covered by the edge set defined by C, i.e., $V(C) = \{i \in V : e \in E$ is incident to i and $c_e \in C\}$. A set of colors C is *feasible* for the MLSteiner if and only if the corresponding set of edges defines a connected subgraph $G_C = (V', E')$ that spans all the basic nodes of G, i.e., $V' \cap Q = Q$. (Moreover, we note in passing that we assume $|L_e| = 1$ throughout the remainder of this paper. That is, exactly one color is assigned to each edge.).

4.1. Greedy

The algorithm starts with an empty set of edges. Then, it iteratively selects one color among the unused ones and inserts all edges of that color in the graph until all the basic nodes are connected. At each iteration it tests all the unused colors and chooses a color in that way that the decrease in the number of Steiner connected components is as large as possible, where we define a Steiner connected component as a connected component $H = (V', E')$ of the graph that contains at least one basic node, e.g., $V' \cap Q \neq \emptyset$. The proposed algorithm is illustrated below.

Algorithm: The greedy heuristic

```
Let $C = \emptyset$ be the set of used colors.
Repeat
    let $H$ be the subgraph of $G$ restricted to edges
      with colors from $C$;
    let $H'$ be the subgraph of $H$ restricted to
      the Steiner connected components of $H$;
    for all $c_i \in L \setminus C$ do
      determine the number of Steiner connected
      components when inserting all edges with
      color $c_i$ in $H$;
    end for
    choose color $c_i$ with the smallest resulting number
      of Steiner connected components and do:
      $C = C \cup \{c_i\}$;
until $H'$ is connected.
```

The greedy strategy we adopt differs from the MVCA heuristic since it carries out operations on the Steiner connected components of subgraph H, while MVCA considers all the connected components of such a graph.
The running time of the proposed greedy strategy is $O(l^2 n)$, where l is the total number of different colors in G. Indeed, the *repeat* loop will take $O(l)$ steps and we have $O(ln)$ to carry out the *for*-loop.
Since the MLST is a special case of the MLSteiner, then, by applying the same reasoning introduced in [19], we can derive the following approximation result.

Theorem 1: Given any MLSteiner instance with n nodes and q basic nodes ($q < n$, $n > 1$), the greedy algorithm provides an $(\ln(q-1)+1)$-approximation.

4.2. Variable neighborhood search

Variable neighborhood search (VNS) goes back to Mladenović and Hansen [15]. The underlying idea of VNS is to generalize the classical local search based approaches by considering a multi-neighborhood structure, i.e., a set of pre-selected neighborhood structures $\mathcal{N} = \{N_1, N_2, \dots, N_s\}$ such that $N_j(C)$, $j = 1, 2, \dots, s$ is the set of solutions in the jth neighborhood of C. The basic VNS algorithm, applied to solve the MLSteiner, is described below.

Algorithm: The basic VNS algorithm

```
Step 1. Consider an initial feasible solution $C \subseteq L$
        and set $k \leftarrow 1$.
Step 2. Generate at random a solution $C' \in N_k(C)$.
Step 3. Apply a local search algorithm, starting from
        the initial solution $C'$, to obtain a local optimum $C''$.
Step 4. If $|C''| < |C|$ then: $C \leftarrow C''$ and set $k \leftarrow 1$
        otherwise $k \leftarrow k+1$ .
Step 5. If $k \leq k_{\max}$ then go to Step 1, else Stop.
```

We implemented VNS by using three different neighborhood structures, in order to check whether one neighborhood is better than the other. In particular, given a feasible color set C, we consider the following neighborhood structures:

- ***k* – switch neighborhood** $N_k^1(C)$
  A set $C' \in N_k^1(C)$ if and only if we can get the color set C' from the color set C by removing up to k colors from C and adding up to k new colors. That is, $N_k^1(C) = \{C' \subseteq L : |C' \setminus C| \leq k \text{ and } |C \setminus C'| \leq k\}$.
- ***k* – covering neighborhood** $N_k^2(C)$
  A set $C' \in N_k^2(C)$ if and only if the common colors between C and C' cover at least k basic nodes. That is, $N_k^2(C) = \{C' \subseteq L : |V(C' \cap C) \cap Q| \geq k$ and $|C'| \leq |C|\}$.
- ***k* – mixed neighborhood** $N_k^3(C)$
  A set $C' \in N_k^3(C)$ if and only if C' contains exactly $|C| - k$ colors in common with C and all the remaining different colors cover a greater number of basic vertices. That is, $N_k^3(C) = \{C' \subseteq L : |C \setminus C'| = k$, $|C' \setminus C| \leq k$ and $|V(C \setminus C') \cap Q| \leq |V(C' \setminus C) \cap Q|\}$.

For each of the neighborhood structures described above, the procedure starts from an initial feasible solution C provided by the greedy algorithm described in Subsection 4.1. At each generic iteration the VNS:

- selects at random a feasible solution C' in the neighborhood $N_k^i(C)$;
- applies a local exchange strategy that, for a maximum number $h_{\max}$ of iterations, tries to decrease the size of C' to obtain a possible better solution C'' by removing π labels and adding up to π new labels, where $\pi = 2, 3, \dots, |C'|$;
- defines the new neighborhood to be explored in the next iteration.

In our implementation of VNS, we let parameter $k_{\max}$ vary during the execution, that is $k_{\max} = \min\{|C|, \frac{l}{4}\}$, where $|C|$ is the "size" of the current feasible solution whose neighborhood is being explored.

In the sequel we refer to the implementations of VNS using $N_k^1()$, $N_k^2()$ and $N_k^3()$ as VNS1, VNS2 and VNS3, respectively.

4.3. Simulated annealing

Simulated annealing (SA) extends basic local search by allowing moves to worse solutions [13]. The basic concept of SA is the following: starting from an initial solution (in our implementation from an empty set of activated colors as in the greedy heuristic), successively, a candidate move is randomly selected. This move is accepted if it leads to a solution with a better objective function value than the current solution, otherwise the move is accepted with a probability that depends on the deterioration Δ of the objective function value. The acceptance probability is computed according to the Boltzmann function as $\mathrm{e}^{-\Delta/T}$, using a temperature T as control parameter.

Following [12], the value of T is initially high, which allows many worse moves to be accepted, and is gradually reduced through multiplication by a parameter *cooling factor* according to a geometric cooling schedule. Given a parameter *size factor*, *size factor* $\times\ l$ candidate moves are tested (note that l denotes the neighborhood size) before the temperature is reduced. The starting temperature is determined as follows: given a parameter *initial acceptance fraction* and based on an abbreviated trial run, the starting temperature is set so that the fraction of accepted moves is approximately *initial acceptance fraction*. A further parameter, *frozen acceptance fraction* is used to decide whether the annealing process is *frozen* and should be terminated. Every time a temperature is completed with less than *frozen acceptance fraction* of the candidate moves accepted, a counter is increased by one, while this counter is re-set to 0 each time a new best solution has been obtained. The whole procedure is terminated when this counter reaches a parameter *frozen limit*.

For our implementation we follow the parameter setting of [12], which was reported to be robust for various problems. That is, we use $\alpha = 0.95$, *initial acceptance fraction* = 0.4, *frozen acceptance fraction* = 0.02, *size factor* = 16 and *frozen limit* = 5.

4.4. Reactive tabu search

The basic paradigm of *tabu search* (TS) is to use information (in the sense of an adaptive memory) about the search history to guide local search approaches to overcome local optimality (see [9] for a survey on tabu search). In general, this is done by a dynamic transformation of the local neighborhood. Based on some sort of memory certain moves may be forbidden, they are defined tabu (and appropriate move attributes such as a certain index indicating a specific color put into a list, called tabu list). As for SA, the search may imply acceptance of deteriorating moves when no improving moves exist or all improving moves of the current neighborhood are set tabu. At each iteration a best admissible neighbor may be selected. A neighbor, respectively a corresponding move, is called admissible, if it is not tabu.

Reactive TS (RTS) aims at the automatic adaptation of the tabu list length [1]. The idea is to increase the tabu list length when the tabu memory indicates that the search is revisiting formerly traversed solutions. A possible specification is the following. Starting with a tabu list length s of 1, it is increased to $\min\{\max\{s+2, s\times 1.2\}, b_u\}$ every time a solution has been repeated, taking into account an appropriate upper bound b_u (to guarantee at least one admissible move). If there is no repetition for some iterations, we decrease it to $\max\{\min\{s-2, s/1.2\}, 1\}$. To accomplish the detection of a repetition of a solution, one may apply a trajectory based memory using hash codes.

For RTS, it is appropriate to include means for diversifying moves whenever the tabu memory indicates that one is trapped in a certain basin of attraction. As a trigger mechanism one may use, e.g., the combination of at least three solutions each having been traversed three times. A very simple escape strategy is to perform randomly a number of moves (depending on the average of the number of iterations between solution repetitions). For our implementation of RTS we consider as initial solution (as for the SA and the greedy heuristic) an empty set of activated colors. As termination criterion we consider a given time limit.

4.5. Pilot method

Using a greedy construction heuristic such as the MVCA as a building block or application process, the pilot method is a metaheuristic with the primary idea of performing repetition exploiting the application process as a look ahead mechanism [7, 18]. In each iteration (of the pilot method) one tentatively determines for every possible local choice (i.e., move to a neighbor of the current solution, called master solution) a look ahead or pilot solution, recording the best results in order to extend at the end of the iteration the master solution with the corresponding move. This strategy may be applied by successively performing, e.g., a construction heuristic for all possible local choices (i.e., starting a new solution from each incomplete solution that can result from the inclusion of any not yet included element into the current incomplete solution).

We apply the pilot method in connection with a greedy local search strategy operating on a solution space that includes incomplete (infeasible) solutions and a neighborhood that considers the addition of colors (see MVCA). We take into account infeasibilities by adding appropriate penalty values. The pilot method successively chooses the best local move (regarding the additional activation of one color) by evaluating such neighbors with a steepest descent until a local optimum, and with that a feasible solution, is obtained. (Note that as for the MVCA, at the end it may be beneficial to greedily drop colors while retaining feasibility.)

5. Experimental results

In this section we report some of our computational results. We considered different groups of instances in order to evaluate how the performance of the algorithms is influenced by both

– the number and distribution of the basic nodes;

– the distribution of the labels on the edges.

In particular, we defined different scenarios based on different parameter settings: n – number of nodes of the graph; l – total number of colors assigned to the graph; m – total number of edges of the graph computed by $m = \frac{d(n-1)n}{2}$, where d is a measure of density of the graph, and, q – the number of basic nodes of the graph. Parameter settings are: $n = 50$, 100, $l = 0.25n, 0.5n, n, 1.25n$, $d = 0.2, 0.5, 0.8$ and $q = 0.2n, 0.4n$, for a total of 48 different scenarios. For each scenario we generated ten different instances. All the generated data are available upon request from the authors.

Our results are reported in Tables 1–4. In each table the first three columns show the parameters characterizing the different scenarios (n, l, d, while the values of q determine the different tables). The remaining columns give the results of the greedy heuristic and of our metaheuristics: variable neighborhood search (VNS1, VNS2 and VNS3), simulated annealing, reactive tabu search and the pilot method, respectively. For the results reported on the greedy heuristic we note that we have implemented the idea to try at the end to greedily drop colors while retaining feasibility. In two of all 480 cases this reduced the objective value by 1.

In general we can say that the pilot method behaves best with respect to solution quality. The SA is usually outperformed by all other metaheuristics and RTS overall behaves a bit better than the VNS for VNS2 and VNS3. The RTS and VNS1 are somewhat incomparable as there does not seem to be a clear picture which method behaves best with respect to solution quality. Among the VNS implementations the first version seems to provide better results than the other two implementations.

Closer inspection of the results reveals a few probably unusual behaviours. First of all, we encountered a considerable role of how ties are broken. Assuming that a tie is broken unfavorably and one encounters an increase or decrease of the objective function by 1, the percentage deviation is affected considerably as most problem instances tend to have very small objective function values (considering the pilot method, only in 6 of the cases with $q = 0.2n$

Table 1
Computational results for $n = 50$ and $q = 0.2n$

n	l	d	Greedy	VNS1	VNS2	VNS3	SA	RTS	Pilot
50	12.5	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0
50	12.5	0.5	1.1	1.1	1.1	1.1	1.1	1.1	1.1
50	12.5	0.2	2.1	2.0	2.0	2.1	2.0	2.0	2.0
50	25	0.8	1.2	1.2	1.2	1.2	1.2	1.2	1.2
50	25	0.5	1.9	1.9	1.9	1.9	1.9	1.9	1.9
50	25	0.2	2.9	2.9	2.9	2.9	2.9	2.9	2.9
50	50	0.8	2.0	2.0	2.0	2.0	2.0	2.0	2.0
50	50	0.5	2.8	2.8	2.6	2.7	2.7	2.8	2.6
50	50	0.2	4.0	3.9	3.9	4.0	4.0	4.0	3.9
50	62.5	0.8	2.3	2.0	2.0	2.0	2.2	2.2	2.0
50	62.5	0.5	3.1	2.8	2.9	3.0	3.0	3.0	2.8
50	62.5	0.2	4.4	4.3	4.4	4.5	4.4	4.4	4.3
Sum			28.8	27.9	27.9	28.4	28.4	28.5	27.7

Table 2
Computational results for $n = 100$ and $q = 0.2n$

n	l	d	Greedy	VNS1	VNS2	VNS3	SA	RTS	Pilot
100	25	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0
100	25	0.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
100	25	0.2	2.2	2.1	2.1	2.1	2.1	2.1	2.1
100	50	0.8	1.9	1.9	1.9	1.9	1.9	1.9	1.9
100	50	0.5	2.0	2.0	2.0	2.0	2.0	2.0	2.0
100	50	0.2	3.3	3.2	3.2	3.3	3.5	3.3	3.2
100	100	0.8	2.4	2.1	2.2	2.2	2.5	2.4	2.0
100	100	0.5	3.0	3.0	3.0	3.1	3.4	3.0	3.0
100	100	0.2	4.9	4.6	5.1	5.0	5.2	4.8	4.6
100	125	0.8	2.8	2.8	2.8	2.8	3.0	2.8	2.8
100	125	0.5	3.5	3.4	3.5	3.6	4.0	3.5	3.4
100	125	0.2	5.7	5.3	5.9	5.7	6.2	5.6	5.4
Sum			34.2	32.9	34.2	34.2	36.3	33.9	32.9

Table 3
Computational results for $n = 50$ and $q = 0.4n$

n	l	d	Greedy	VNS1	VNS2	VNS3	SA	RTS	Pilot
50	12.5	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0
50	12.5	0.5	1.4	1.4	1.4	1.4	1.4	1.4	1.4
50	12.5	0.2	2.6	2.6	2.6	2.6	2.6	2.6	2.6
50	25	0.8	2.0	2.0	2.0	2.0	2.0	2.0	2.0
50	25	0.5	2.0	2.0	2.0	2.0	2.0	2.0	2.0
50	25	0.2	4.3	3.9	3.9	3.9	3.9	4.0	3.9
50	50	0.8	2.4	2.2	2.3	2.2	2.5	2.4	2.1
50	50	0.5	3.2	3.0	3.0	3.0	3.1	3.0	3.0
50	50	0.2	5.9	5.5	5.9	5.8	5.8	5.8	5.3
50	62.5	0.8	2.7	2.7	2.7	2.8	2.9	2.7	2.6
50	62.5	0.5	3.6	3.3	3.2	3.5	3.7	3.6	3.2
50	62.5	0.2	6.2	6.6	6.7	6.5	6.4	6.2	6.0
Sum			37.3	36.2	36.7	36.7	37.3	36.7	35.1

Table 4
Computational results for $n = 100$ and $q = 0.4n$

n	l	d	Greedy	VNS1	VNS2	VNS3	SA	RTS	Pilot
100	25	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0
100	25	0.5	1.9	1.9	1.9	1.9	1.9	1.9	1.9
100	25	0.2	3.1	3.0	3.0	3.0	3.0	3.0	3.0
100	50	0.8	2.0	2.0	2.0	2.0	2.0	2.0	2.0
100	50	0.5	2.4	2.2	2.2	2.2	2.4	2.3	2.2
100	50	0.2	4.8	4.4	4.5	4.8	4.4	4.6	4.3
100	100	0.8	3.0	3.0	3.0	3.0	3.0	3.0	3.0
100	100	0.5	3.9	3.6	3.6	3.6	3.8	3.9	3.6
100	100	0.2	6.6	6.9	7.6	7.0	7.4	6.6	6.5
100	125	0.8	3.0	3.0	3.0	3.0	3.4	3.0	3.0
100	125	0.5	4.1	4.1	4.2	4.3	4.9	4.1	4.0
100	125	0.2	7.6	8.1	9.2	8.2	8.1	7.6	7.0
Sum			43.4	43.2	45.2	44.0	45.3	43.0	41.5

and 32 of the cases with $q = 0.4n$ the objective turned out to be larger than 5, i.e., 6, 7, or 8). In this sense, a random neighbor selection within the SA and the VNS implementations may already lead to an unfavorable objective function value that is difficult to be overcome which explains the few cases where the SA results are even worse than those of the greedy approach.

For RTS the approach first mimics the behaviour of a steepest descent like the greedy heuristic. Based on the way infeasibilities are penalized, the method usually stays within the feasible region so that the method may be caught within some basin of attraction related to the first local optimum found. That is, the RTS does not really work as expected, since in most cases (about 95%) the best results have been obtained within the first second of the computation. After that the method did not find improvements quite often even if they would have been possible.

The computations for our methods have been made on a Pentium IV 1.8 GHz. The termination criteria for the different methods follow the descriptions given above. The RTS is terminated after a time limit of 10 seconds for instances with $n = 50$ and after 40 seconds for $n = 100$. In general the computational times are moderately increasing for decreasing values of d, they are also increasing for increasing values of q and l, and they are considerably increasing for an increasing number of nodes. While RTS has a given time limit, computational times for the other methods tend to be below those numbers for larger values of d for all methods while they become slightly larger than those for RTS in case of the VNS implementations for $d = 0.2$. Computational times for the VNS implementations mainly depend on graph density: the more sparse the graph the larger the times. The computational times for the pilot method mainly depend on the number of nodes and the value of l. If l increases then the times for the pilot method may easily become considerably larger than those of SA and RTS but also larger than those of the VNS. Detailed computational times are reported in Table 5 for the largest instances to get a feeling about the general behaviour of our methods.

Table 5
Computational times [s] for $n = 100$ and $q = 0.4n$

n	l	d	VNS1	VNS2	VNS3	SA	RTS	Pilot
100	25	0.8	0.4	0.2	4.6	11.1	40.0	1.0
100	25	0.5	0.6	0.6	2.8	9.7	40.0	1.2
100	25	0.2	19.2	23.4	46.0	6.6	40.0	1.4
100	50	0.8	1.1	1.2	7.0	22.1	40.1	6.5
100	50	0.5	12.0	5.2	28.2	17.4	40.0	6.1
100	50	0.2	49.9	81.6	77.0	12.3	40.0	9.1
100	100	0.8	15.3	10.9	57.8	43.5	40.1	43.3
100	100	0.5	44.8	101.9	50.3	33.7	40.0	43.0
100	100	0.2	72.0	128.2	96.2	21.7	40.0	63.5
100	125	0.8	29.0	66.0	36.8	54.0	40.1	73.1
100	125	0.5	50.5	75.5	53.4	40.5	40.1	76.7
100	125	0.2	94.3	174.8	128.0	27.3	40.0	115.1

We should note that a detailed analysis of the results reveals that the pilot method usually does not need as much time as shown to find the indicated solutions, since in almost all cases the best result has been obtained in the first few seconds of the computations. This gives a strong hint that a small evaluation depth (see [18]) may be used to reduce the computation times without discarding solution quality.

6. Conclusions and further research

In this paper we have considered a generalization of the minimum labeling spanning tree problem to the case where not necessarily all but only a subset of required nodes need to be spanned. Common metaheuristics have successfully been applied to this generalization and the results are in line with our expectation gained from experimentation with the original labeling spanning tree problem. The most visible result is that the pilot method outperforms the other approaches with respect to solution quality while the computation times of the pilot method can be considerably larger than those of reactive tabu search or simulated annealing especially for larger problem instances. The computation times of our implementations for the pilot method and the variable neighborhood search are somewhat comparable with some exceptions for smaller densities of the given graphs where the pilot method may be faster. This motivates one direction of our further research consisting in the combination of the two metaheuristics that seem to behave better, the pilot method and the VNS1 [4]. Moreover, the results have been obtained for one generalization of the labeling spanning tree problem and future research refers also to extending those ideas to other generalizations such as the one also proposed in this paper considering additional budget constraints. Moreover, allowing for more than one color assigned to each edge poses an interesting case motivated by some applications.

Another step in our research refers to developing various mathematical programming formulations for the MLST as well as the MLSteiner to obtain optimal solutions to better judge on the quality of our heuristic solutions at least for small and moderately sized problem instances.

References

[1] R. Battiti, “Reactive search: toward self-tuning heuristics”, in *Modern Heuristic Search Methods*, V. J. Rayward-Smith, I. H. Osman, C. R. Reeves, and G. D. Smith, Eds. Chichester: Wiley, 1996, pp. 61–83.

[2] H. Broersma and X. Li, “Spanning trees with many or few colors in edge-colored graphs”, *Discus. Math. Graph Theory*, vol. 17, pp. 259–269, 1997.

[3] T. Brüggemann, J. Monnot, and G. J. Woeginger, “Local search for the minimum label spanning tree problem with bounded color classes”, *Oper. Res. Lett.*, vol. 31, pp. 195–201, 2003.

[4] R. Cerulli, A. Fink, M. Gentili, and S. Voß, “Applications of the pilot method and VNS to hard modifications of the minimum spanning tree problem”, in *Mini Euro Conf. VNS*, Tenerife, Spain, 2005.

[5] R. Cerulli, A. Fink, M. Gentili, and S. Voß, “Metaheuristics comparison for the minimum labelling spanning tree problem”, in *The Next Wave on Computing, Optimization, and Decision Technologies*, B. L. Golden, S. Raghavan, and E. A. Wasil, Eds. New York: Springer, 2005, pp. 93–106.

[6] R.-S. Chang and S.-J. Leu, "The minimum labeling spanning trees", *Inform. Proces. Lett.*, vol. 63, pp. 277–282, 1997.

[7] C. W. Duin and S. Voß, "The pilot method: A strategy for heuristic repetition with application to the Steiner problem in graphs", *Networks*, vol. 34, pp. 181–191, 1999.

[8] A. Fink, G. Schneidereit, and S. Voß, "Solving general ring network design problems by metaheuristics", in *Computing Tools for Modeling, Optimization and Simulation*, M. Laguna and J. L. González Velarde, Eds. Boston: Kluwer, 2000, pp. 91–113.

[9] F. Glover and M. Laguna, *Tabu Search*. Boston: Kluwer, 1997.

[10] F. K. Hwang and D. S. Richards, "Steiner tree problems", *Networks*, vol. 22, pp. 55–89, 1992.

[11] F. K. Hwang, D. S. Richards, and P. Winter, *The Steiner Tree Problem*. Amsterdam: North-Holland, 1992.

[12] D. S. Johnson, C. R. Aragon, L. A. McGeoch, and C. Schevon, "Optimization by simulated annealing: an experimental evaluation", Part I, "Graph partitioning", *Oper. Res.*, vol. 37, pp. 865–892, 1989.

[13] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, "Optimization by simulated annealing", *Science*, vol. 220, pp. 671–680, 1983.

[14] S. O. Krumke and H.-C. Wirth, "On the minimum label spanning tree problem", *Inform. Proces. Lett.*, vol. 66, pp. 81–85, 1998.

[15] N. Mladenović and P. Hansen, "Variable neighbourhood search", *Comput. Oper. Res.*, vol. 24, pp. 1097–1100, 1997.

[16] S. Voß, "Modern heuristic search methods for the Steiner tree problem in graphs", in *Advances in Steiner Trees*, D.-Z. Du, J. M. Smith, and J. H. Rubinstein, Eds. Boston: Kluwer, 2000, pp. 283–323.

[17] S. Voß, "Steiner tree problems in telecommunications", in *Handbook of Optimization in Telecommunications*, M. Resende and P. M. Pardalos, Eds. New York: Springer, 2006, pp. 459–492.

[18] S. Voß, A. Fink, and C. Duin, "Looking ahead with the pilot method", *Ann. Oper. Res.*, vol. 136, pp. 285–302, 2004.

[19] Y. Wan, G. Chen, and Y. Xu, "A note on the minimum label spanning tree", *Inform. Proces. Lett.*, vol. 84, pp. 99–101, 2002.

[20] P. Winter, "Steiner problem in networks: a survey", *Networks*, vol. 17, pp. 129–167, 1987.

[21] Y. Xiong, B. Golden, and E. Wasil, "A one-parameter genetic algorithm for the minimum labeling spanning tree problem". Tech. Rep., University of Maryland, 2003.

[22] Y. Xiong, B. Golden, and E. Wasil, "Improved metaheuristics for the minimum labeling spanning tree problem". Tech. Rep., University of Maryland, 2005 (also in *Metaheur. Int. Conf.*, Vienna, Austria, 2005).

[23] Y. Xiong, B. Golden, and E. Wasil, "Worst-case behavior of the MVCA heuristic for the minimum labeling spanning tree problem", *Oper. Res. Lett.*, vol. 33, pp. 77–80, 2005.

Raffaele Cerulli is currently Associate Professor in operations research at the Department of Mathematics and Computer Science of the University of Salerno, Italy. He holds laurea degree in computer science from University of Salerno. His current research interests are in combinatorial optimization problems arising from real applications in the field of telecommunications and logistics. In particular, he focused on network flow problems as well as covering problems on graphs. He has several papers in various journals and he was Director of the International School on "Pattern Analysis" as well as the organizer of workshops on Graph Theory in Italy.

e-mail: raffaele@unisa.it
Department of Mathematics and Computer Science
University of Salerno
Via Ponte Don Melillo
84084, Fisciano (Salerno), Italy

Andreas Fink is Professor and the Chair of Business Administration, in particular Information Systems, at the Helmut-Schmidt-University in Hamburg, Germany. He holds diploma degrees in business administration and computer science from the University of Technology Darmstadt and the Ph.D. in economics from the University of Technology Braunschweig. His research is mainly concerned with the use of information technology to support decision-making in fields such as telecommunications, logistics, and supply chain management. His publications have appeared in various journals.

e-mail: andreas.fink@hsu-hamburg.de
Department of Economics
Helmut-Schmidt-University / UniBw Hamburg
Holstenhofweg 85
22043 Hamburg, Germany

Monica Gentili is currently Assistant Professor in operations research at the Department of Mathematics and Computer Science of the University of Salerno, Italy. She received the laurea degree in statistics in 1998 and the Ph.D. in operations research in 2003 at the University of Rome "La Sapienza". Her current research interests are concentrated on combinatorial optimization problems, mainly on mathematical models and algorithm design for planning and control of traffic flows on networks as well as covering problems on graphs. They include sensor location problems on networks, routing and distribution problems, graph theory.

e-mail: mgentili@unisa.it
Department of Mathematics and Computer Science
University of Salerno
Via Ponte Don Melillo
84084, Fisciano (Salerno), Italy

Stefan Voß – for biography, see this issue, p. 20.

Stackelberg Security Games: Models, Applications and Computational Aspects

Andrzej Wilczyński[1,2], Agnieszka Jakóbik[2], and Joanna Kołodziej[2]

[1] *AGH University of Science and Technology, Cracow, Poland*
[2] *Tadeusz Kościuszko Cracow University of Technology, Cracow, Poland*

http://doi.org/10.26636/jtit.2016.3.749

Abstract—Stackelberg games are non-symmetric games where one player or specified group of players have the privilege position and make decision before the other players. Such games are used in telecommunication and computational systems for supporting administrative decisions. Recently Stackleberg games became useful also in the systems where security issues are the crucial decision criteria. In this paper authors briefly survey the most popular Stackelberg security game models and provide the analysis of the model properties illustrated in the realistic use cases.

Keywords*—*Bayesian games, game theory, leadership, Nash equilibrium, normal form games, security games, Stackelberg equilibrium, Stackelberg games.

1. Introduction

Game theory is the formal, mathematical methodology for analyzing interactions between intelligent players: people, corporations, software agents, or making decisions robots. The theory is useful for solving problems in many disciplines, from economics, business, and law, public policy to telecommunication. Game theory provides the tools for determining optimal behavior in competitive environments. Formally, *a game refers to all the situations involving two or more intelligent individuals making rational decisions* [1]. The players are making decisions consistently to obtain the assumed target. The player is considered intelligent, if he knows the game rules and can make decisions based on his knowledge.

The basic examples of game theoretical modeling include the simulations of the competitive processes in economics, political science, psychology, or biology. The players are interest groups, politicians, or competing animal species. Computer science uses game theory during modeling multi-agent systems, online algorithms or processes in computer networks [2].

Game theory is also useful in the cases where security is important: in everyday life and security of the large-scale IT systems such as computational grids and clouds. The airport police behavior as one side of the conflict playing against thieves or terrorists was modeled. Randomizing schedules for patrolling, checking, or monitoring is typical outcome of the models [3].

In this paper, authors focus on Stackelberg security models, where one or group of players are the privilege in the game. They play first, and the rest of the players follow the leader(s) and make their decisions based on the leader's actions. Such games can be a good proposal for supporting the decisions in the cloud systems, where security remains a challenging research and engineering task. The existing Stackelberg models related to the security aspects in high performance computing telecommunication and transportation systems are surveyed and the models properties from the implementation perspective are analyzed. The effectiveness of the models has been justified in realistic use cases.

The paper is organized as follows. In Section 2 the basic definitions and backgrounds of the game-theoretical models are explained together with the definition of the generic Stackelberg game. In Sections 3 and 4 the secure Stackelberg game is defined and the most popular Stackelberg security models are reviewed. The computational and implementation aspects of the analyzed Stackelberg models are discussed in Section 5. In Section 6 the realistic use cases for Stackelberg security games are presented. Section 7 concludes the paper.

2. Game Theory – Backgrounds and Game Models

Game theoretical models are very useful in the formal analysis of task, data and information management and decision-like processes in highly distributed large-scale computational environments mainly because of the strict mathematical formalism. Although, there are many types of games and also many formal models of such games, the most commonly used and known is the *normal-form game model* introduced by Tadelis *et al.* [4] as follows:

Normal-form game consists of three sets: players, strategies and payoff functions specified for each player in order to define the solution of the game for each combination of the players' actions.

Formally, the n-player normal game Γ_n can be defined by the following rule:

$$\Gamma_n = (N, \{S_i\}_{i \in N}, \{Q_i\}_{i \in N}) \,, \qquad (1)$$

where:

- $N = \{1,\dots,n\}$ is the set of players,
- $\{S_1,\dots,S_n\}$ $(card\ S_i \geq 2; i = 1,\dots,n)$ is the set of strategies for the players,
- $\{H_1,\dots,H_n\}; H_i : S_1 \times \cdots \times S_n \to \mathbb{R}; \forall_{i=1,\dots,n}$ is the set of payoff functions of the players.

The strategy of the player in the game can be defined as a plan of actions of that player to make the game beneficial for him. Two classes of strategies are defined, namely pure strategies and mixed strategies [4].

Definition 1. Pure strategy of the player i is the deterministic plan of player's actions during the game. The set of all pure strategies specified for player i is denoted by S_i. A profile of pure strategies in the n-players game Γ_n is defined by the following vector of the players' strategies:

$$s = [s_1, s_2, \dots, s_n], s_i \in S_i; (i = 1,2,\dots,n). \tag{2}$$

Such strategy profile can be defined for any combination of the players' pure strategies in the game Γ.

Definition 2. Let us denote by $S_i = s_{i1}, s_{i2}, \dots, s_{im}$ the finite set of m pure strategies of the player i. Let us also denote by ΔS_i the simplex over S_i. ΔS_i is the set of all probability distributions over S_i.
The mixed strategy of player i is denoted by $\sigma_i \in S_i \subset \Delta S_i$ and is defined as follows [4]:

$$\sigma_i = \{\sigma_i(s_{i_1}), \sigma_i(s_{i_2}), \dots, \sigma_i(s_{i_m})\}, \tag{3}$$

where $\sigma_i(s_i)$ is the probability that the player i plays according to the strategy s_i.

One can conclude from the above definition that $\sigma_i(s_i) \geq 0$ for all $i = 1, \dots N$ and

$$\sigma_i(s_{i_1}) + \sigma_i(s_{i_2}) + \dots + \sigma_i(s_{i_m}) = 1. \tag{4}$$

It can be also observed that the mixed strategy becomes pure if $\sigma_i(s_{i_j}) = 1$ for some j $\sigma_i(s_{i_k}) = 0$ for all $k \neq j$.
In the mixed strategy model, the decisions of each player are randomized according to the probability distribution $\sigma_i(s_i)$. In such a case, the payoffs are also non-deterministic.

Definition 3. Tadelis *et al.* [4] the expected payoff of player i in 2-players game is defined as:

$$H_i(s_i, \sigma_{-i}) := \sum_{s_{-i} \in S-i} \sigma_{-i}(s_{-i}) H_i(s_i, s_{-i}), \tag{5}$$

where $H_i(s_i, s_{-i})$ is the payoff function calculated for the player i. It is assumed in that game, that player i chooses the pure strategy $s_i \in S_i$ and his opponents plays the mixed strategy $\sigma_{-i} \in \Delta S_{-i}$.

Similarly:

Definition 4. The expected payoff of player i when he chooses the mixed strategy $\sigma_i \in \Delta S_i$ and his opponents plays the mixed strategy $\sigma_{-i} \in \Delta S_{-i}$ is defined in the following way:

$$H_i(\sigma_i, \sigma_{-i}) = \sum_{s_i \in S_i} \sigma_i(s_i) H_i(s_i, \sigma_{-i}) =$$
$$= \sum_{s_i \in S_i} \Big(\sum_{s_{-i} \in S_{-i}} \sigma_i(s_i) \sigma_{-i}(s_{-i}) H_i(s_i, s_{-i}) \Big). \tag{6}$$

The main aim of each player during the game it to maximize his expected payoff by defining the optimal strategy. The most commonly encountered concept of the game solution is an equilibrium point defined as follows:

Definition 5. An n-dimensional vector $(\bar{s}_1, \dots, \bar{s}_n)$ of strategies is called an equilibrium point or Nash equilibrium, if:

$$H_i(\bar{s}_1, \dots, \bar{s}_n) = \max_{s_i \in S_i} H_i(\bar{s}_1, \dots, \bar{s}_{i-1}, s_i, \bar{s}_{i+1}, \dots, \bar{s}_n)$$
$$\text{for all } i = 1, \dots, n. \tag{7}$$

The Nash equilibrium [5] can be interpreted as a steady state of the play of a strategic game, in which each player holds correct expectations concerning the other players' behaviors. If the strategies chosen by all players are Nash equilibrium, no player is interested in changing his strategy.
An n-vector $\bar{H} = \big(H_1(\bar{s}_1, \dots, \bar{s}_n), \dots, H_n(\bar{s}_1, \dots, \bar{s}_n)\big)$ is called a value of the game. The strategies $(\bar{s}_1, \dots, \bar{s}_n)$ are the pure strategies (see Def. 1). It means that they are never changed during the game.
Some equilibrium points cannot be accepted as solutions of the game. It is usually required that the solution should not satisfy the following condition:

Definition 6. An n-dimensional vector of strategies $(\hat{s}_1, \dots, \hat{s}_n)$ is Pareto non-optimal, if there exists another n-vector $(\check{s}_1, \dots, \check{s}_n)$, for which the following two conditions hold:

$$\forall_{i \in \{1,\dots,n\}} H_i(\hat{s}_1, \dots, \hat{s}_n) \leq H_i(\check{s}_1, \dots, \check{s}_n), \tag{8}$$

$$\exists_{i \in \{1,\dots,n\}} H_i(\hat{s}_1, \dots, \hat{s}_n) < u_i(\check{s}_1, \dots, \check{s}_n). \tag{9}$$

One can say that the n-vector $(\check{s}_1, \dots, \check{s}_n)$ dominates $(\hat{s}_1, \dots, \hat{s}_n)$.
It can be observed, that vector $(s_1, \dots, s_n)$ cannot be accepted as the solution of the game, if it is Pareto non-optimal (even if it is the Nash equilibrium).

2.1. Minimization of the Game Multi-loss Function

The problem of detecting the Nash equilibrium of a finite strategic non-cooperative game can be also formulated as a global optimization problem with loss instead of payoff functions.
Let us define a set of loss (cost) functions for the players:

$$\{Q_1, \dots, Q_n\}; Q_i : S_1 \times \cdots \times S_n \to \mathbb{R}; \forall_{i=1,\dots,n}. \tag{10}$$

Each player tends to the minimization of his loss function in the game, which is equivalent with the maximization of the payoff function. Let us define a set of *players' response functions* $\{r_i\}_{i=1,\dots,n}$; $r_i : S_1 \times \cdots \times S_n \rightarrow \mathbb{R}$ where:

$$r_i(\hat{s_i}) = \arg\min_{s_i \in S_i} \{Q_i(s_1,\dots,s_n)\}, \tag{11}$$

where $\hat{s_i} = (s_1,\dots,s_{i-1},s_{i+1},\dots,s_n)$. The response function defines an optimal strategy for the player.

We can define now *a multi-loss function* $Q : S_1 \times \cdots \times S_N \rightarrow \mathbb{R}$ for the game by the formula:

$$Q(s_1,\dots,s_n)=\sum_{i=1}^{n}\left[Q_i(s_1,\dots,s_n)-\min_{s_i \in S_i} Q_i(s_1,\dots,s_n)\right]. \tag{12}$$

Note that the multi-loss function has non-negative values. In such a case, the Nash equilibrium is the result of the global minimization of the function Q. The players' strategies are called the decision variables and the players' loss functions are called players' objective functions.

It follows from the definition of the function Q that is needed to minimize first the loss functions of the players and then to compute the values of the multi-loss function. Thus the detection procedure of the Nash equilibrium is a parallel algorithm composed of two cooperated units:

- **main unit** – which solves the problem of global minimization of the function Q,
- **subordinate unit** – which solves the problems of minimization of the players' loss functions Q_i.

The subordinate unit could be a parallel algorithm designed for the numerical optimization of the real functions of several variables.

2.2. Stackelberg Games

In all game scenarios considered in the Section 2, it was assumed that the games are symmetric. It means that all players have the same privileges and knowledge about the game conditions and the other players' strategies and actions. However, that assumption may never occur in the real situations, where usually there is a player (or group of players) with the deeper knowledge of the game conditions. Cloud administrators and local cloud service providers can be good examples of the realistic potential players in non-symmetric resource allocation decision making game model. In grid and cloud computing, Stackelberg games are the most popular non-symmetric game models used for supporting the decisions of various system users.

In Stackelberg game [6], one user acts as a *leader* and the rest are his *followers*. The leader may keep his strategy fixed while the followers react independently subject to the leader's strategy. Formally, the N-players Stackelberg game can be defined as two-level game model, where the players act sequentially as follows: (i) the leader is the only player active at the first level, he chooses his best-response strategy; (ii) at the second level, the followers react rationally to the leader's action. It means that they try to minimize their game cost functions subject to the leader's choice. Finally, the leader updates his strategy to minimize the total game cost.

The solution of the Stackelberg game is called *Stackelberg equilibrium*. In such a case, each follower observes the leader's strategy x and responds with strategy $f(x) : x \rightarrow y$ that is optimal with respect to his expected payoff. Two types of Stackelberg equilibrium points can be defined, namely Strong Stackelberg Equilibrium (SSE) and Weak Stackelberg Equilibrium (WSE). SSE assumes that the follower breaks ties in favor of the defender. It means that he chooses his optimal strategy, which is also optimal from the leader's perspective. WSE assumes that the follower chooses the worst strategy from the leader's perspective [7]. Formally, both scenarios can be defined in the following way:

Definition 7. A pair of strategies $\big(x, f(x)\big)$ is defined as Strong Stackelberg Equilibrium if the following conditions are satisfied [7]:

1. The leader plays his best-response strategy:
$$H_l\big(x, f(x)\big) \geq H_l\big(x', f(x')\big), \tag{13}$$
for all leader's strategies x'.
2. The follower plays his best-response strategy:
$$H_f\big(x, f(x)\big) \geq H_f\big(x, y'\big), \tag{14}$$
for all follower's strategies y'.
3. The follower breaks ties in favor of the leader:
$$H_l\big(x, f(x)\big) \geq H_l(x, y'), \tag{15}$$
for all optimal follower's strategies y'.

2.3. Bayesian Stackelberg Games

In Bayesian Stackelberg Game, the *type* of player must be specified for each of N players. In two players game, there is only *one leader type*, although there are multiple follower types, denoted by $l \in L$. Authors define the probability p^l that a follower of type l will appear in the game. The leader does not know the follower's type. For each player type (leader or follower) n, there is a set of strategies σ_n and a utility function of the game $Q_n : L \times \sigma_1 \times \sigma_2 \rightarrow \mathbb{R}$, which is usually defined as the game cost function of the given player n [8].

Bayesian game can be transformed into a normal-form game using Harsanyi transformation. Let us assume there are two follower types 1 and 2. Type 1 will be active with probability α, and follower type 2 will be active with proba-

Table 1
Payoff tables for a Bayesian Stackelberg game with 2 follower types

	c	d	c'	d'
a	2.1	4.0	1.1	2.0
b	1.0	3.2	0.1	3.2

Table 2
Harsanyi transformed payoff table

	cc'	cd'	dc'	dd'
a	$2\alpha+(1-\alpha),1$	$2,\alpha$	$4\alpha+(1-\alpha),(1-\alpha)$	$4\alpha+2(1-\alpha),0$
b	$\alpha,(1-\alpha)$	$\alpha+3(1-\alpha),2(1-\alpha)$	$3\alpha,2\alpha+(1-\alpha)$	$3,2$

bility $1-\alpha$. A chance node must be specified for Harsanyi transformation. That node is required for the specification of the follower's type. It transforms the leader's incomplete information regarding the follower into an imperfect information game. In the transformed game, the leader still has two strategies while there is a single follower type with four $(2\cdot 2)$ strategies [8]. That scenario is illustrated in Tables 1 and 2.

3. Security Stackelberg Games

Decision processes of users, administrators and resource owners in high performance computational systems are very complex especially in the case, where security and data protection are the important decision criteria. Game models and Stackelberg games in particular, can be very useful in supporting such difficult decisions. The game models used in security applications are called *security games*.

Security game is a game between defender and attacker. The attacker may pick any target from the target set:

$$Targets = \{t_1,\dots,t_n\}\,. \qquad (16)$$

The defender may cover targets by available resources from the set of resources:

$$Resources = \{r_1,\dots,r_K\}\,. \qquad (17)$$

Tambe *et el.* [9] defined the compact security game model. In this model, all resources are identical and may be assigned to any target and payoffs depend only on the identity of the attacked target and whether or not it is covered by the defender.

Any security game represented in this compact form can also be represented in normal form. The attack vector A maps directly to the attacker's pure strategies, with one strategy per target. For the defender, each possible allocation of resources corresponds to a pure strategy in the normal form. A resource allocation maps each available resource to a target, so there are *n Choose m* ways to allocate *m* resources to *n* targets [9].

Let us denote the defender utility if t_i is attacked when it is covered by $U_d^c(t_i)$, and defender utility if t_i is attacked when it is uncovered by $U_d^u(t_i)$, and attacker utility $U_a^c(t_i)$ and $U_a^c(t_i)$, respectively. Then, during the game, it is assumed that adding the resource to cover targets benefits the defender and operates to the detriment of attacker:

$$U_d^c(t_i) - U_d^u(t_i) > 0, U_a^u(t_i) - U_a^c(t_i) > 0\,. \qquad (18)$$

For each resource r_i there is a subset S_i of of the schedules S that r_i can cover. The example of such a situation is marshal's fly tours. In security game, the defender may play best-response strategy, however, it depends on the attacker's behavior.

In normal representation of security game, the attacker's pure strategy is specified as a set of targets. The attacker's mixed strategy is defined by the following vector $\mathbf{a} = [a_1,\dots,a_n]$ representing the probability of attacking the targets. The defender's pure strategy is defined by the coverage vector $\mathbf{d} \in \{0,1\}^n$, where d_i represents if target t_i is covered or not. Let us denote by $D \in \{0,1\}^n$ the set of possible coverage vectors, and by $\mathbf{c}$ the vector of coverage probabilities. The defender's mixed strategy C is defined as the vector of probabilities of playing each $\mathbf{d} \in D$. For strategy C, the defenders utility is defined as:

$$U_d(C,a) = \sum_{i=1}^{n} a_i\big(c_i U_d^c(t_i) + (1-c_i)U_d^u(t_i)\big)\,, \qquad (19)$$

and attacker's utility is defined in the following way:

$$U_a(C,a) = \sum_{i=1}^{n} a_i\big(c_i U_a^c(t_i) + (1-c_i)U_a^u(t_i)\big)\,. \qquad (20)$$

In symmetric security games, the Nash equilibrium can be also estimated. In such a case, the defender plays his best-response strategy C, such that for any other strategy C', his utility is the most beneficial:

$$U_d(C,a) > U_d(C',a)\,. \qquad (21)$$

The attacker plays also his best-response strategy a, such that for any other strategy a', his utility is the most beneficial:

$$U_a(C,a) > U_a(C,a')\,. \qquad (22)$$

The game model, in which the defender makes his decision first and attacker chooses his strategy based on the results of the defender's action, is called *security Stackelberg game*. In that game, $g(C) = \mathbf{a}$ is the attacker response function. Strong Stackelberg Equilibrium (SSE) can be found by:

- the defender plays the best-response strategy C, such that $U_d(C,g(C)) >= U_d(C',g(C'))$ for all C',
- the attacker plays the best-response strategy C, such that $U_a(C,g(C)) >= U_a(C,g'(C))$ for all g',C,
- the attacker breaks ties optimally for the leader: $U_d(C,g(C)) >= U_d(C,\tau(C))$ for all C, where $\tau(C)$ is the set of followers best responses to C.

The basic version of the game assumes that utility functions are common knowledge. In SSE (see Def. 7), the attacker must know the defender's utility, in order to compute his own strategy. In Nash equilibrium, the attacker does not follow the defender's actions. In real life applications, defender does not know the attacker's utility function and the game may be defined by using the Bayesian model. The assumption that attacker responds optimally (selects the best-response strategy) may not happen either (imperfect follower case) [10].

4. Secure Stackelberg Game-based Models

In this section, the most popular Stackelberg security game models are surveyed. Presented models were selected due to the increasing limitations on resources and growing attackers' number, incorporating uncertainty about the optimal behavior of attackers, uncertainty about the observation possibility.

4.1. DOBSS Model

Paruchuri *et al.* in [11] considered the Bayesian Stackelberg security game for one leader, multiple independent followers and the situation when the leader does not know the follower type. For leader strategy vector $x = [x_1, \ldots, x_n] \in [0,1]$ represents the proportion of times when pure strategy $i = 1, \ldots, n$ was chosen. The authors proposed the algorithm for finding the optimal mixed strategy for the leader, under the assumption that the follower (attacker) knows this mixed strategy choosing his own. The authors defined the following two utility matrices for the leader $U_d^{i,j} = R_{i,j}$ and attacker $U_a^{i,j} = C_{i,j}$. It is assumed that the leader plays pure strategy i and attacker plays pure strategy j.

Let us denote by $q = [q_1, \ldots, q_n] \in \{0,1\}$ the mixed strategy for the follower, X – leader pure strategies index set, and by Q – the pure follower's strategy indexes. The algorithm is implemented in the following steps (one follower is considered):

- for fixed leader strategy X the follower solves the linear problem to find his optimal response:

$$\max_q \sum_{j\in Q}\sum_{i\in X} C_{i,j} x_i q_j \,, \tag{23}$$

  with constraints that means that every pure strategy is possible:

$$\sum_{j\in Q}^{q_j>=0} q_j = 1 \,; \tag{24}$$

- the leader finds the strategy x that maximizes his utility, under the assumption that the follower used optimal response $a(x)$:

$$max_q \sum_{i\in X}\sum_{j\in Q} R_{i,j} q(x) x_i \,, \tag{25}$$

  with assumption that each pure strategy is possible:

$$\sum_{i\in X}^{x_i\in[0,1]} x_i = 1 \,. \tag{26}$$

The authors proposed also the model for multiple followers, with specified recognition probability of the follower's type. Let us denote by $U_d^{i,j,l} = R_{i,j}^l$ and $U_a^{i,j,l} = C_{i,j}^l$ the utility matrices of the leader's respectively. Leader plays pure strategy i and attacker plays pure strategy j, and the follower type is l. Let us also denote by p^l the probabilities of playing with the follower of type l. The solution of such a game can be defined as quadratic programming problem (specified for the leader) with the following distribution over the follower type p^l:

$$\max_{x,q,a} \sum_{i\in X}\sum_{l\in L}\sum_{j\in Q} p^l R_{i,j}^l q_j^l x_i \,, \tag{27}$$

with the following leader's and follower's strategies:

$$\sum_{i\in X}^{x_i\in[0,1]} x_i = 1, \ \sum_{j\in Q}^{q_j^l\in[0,1]} q_j^l = 1 \,. \tag{28}$$

It can be observed that $q_j^l = 1$ only for a strategy that is optimal for follower l:

$$0 =< \left(a^l - \sum_{i\in X} C_{i,j} x_i <= (1 - q_j^l) M\right) , \tag{29}$$

where M is the fixed large positive number, and $a \in \mathbf{R}$.

In the above models, the players are completely rational (they play according to the concrete calculated strategy) and followers can follow the leader's strategy. The quadratic problem given by Eqs. (27)–(29) may be linearized by defining the new variables $z_{i,j}^l := x_i q_j^l$.

4.2. BRASS, BOSS and MAXMIN Models

Pita *et al.* in [12] proposed three mixed-linear program algorithms for solving the Bayesian Stackelberg games. They considered the following two game scenarios:

- bounded rationality of the followers scenario – the leader cannot be sure that he will play the game according to the calculated strategy with the selection ε-optimal response strategy – the follower may choose any response,
- uncertainty scenario – the recognition of the leader's strategy by the follower can be incorrect.

In the first case, the problem of solving the game was defined as the following BRASS linear programming problem:

$$\max_{x,q,h,a,\gamma} \sum_{l\in L} p^l \gamma^l \,, \tag{30}$$

where the leader's and follower's strategies can be specified as:

$$\sum_{i\in X}^{x_i\in[0,1]} x_i = 1 \,, \tag{31}$$

allowing to select more than one policy per follower type

$$\sum_{j\in Q} q_j^l >= 1, \sum_{j\in Q} h_j^l = 1 \,, \tag{32}$$

and the condition that ensure that $q^l_j = 1$ only for a strategy that is optimal for follower l:

$$0 =< \big(a^l - \sum_{i\in X} C_{i,j}x_i <= (1-h^l_j)M\,, \qquad (33)$$

$$\varepsilon(1-q^l_j) =< a^l - \sum_{i\in X} C^l_{i,j}x_i <= \varepsilon + \big(1-q^l_j\big)M\,, \qquad (34)$$

$$(1-q^l_j)M + \sum_{i\in X} R^l_{i,j}x_i >= \gamma_l\,, \qquad (35)$$

where $h^l_j =< q^l_j$, $h^l_j, q^l_j \in \{0,1\}$, for the fixed large positive number M and $a \in \mathbf{R}$.

In the uncertainty scenario model (BOSS), developed by Jain *et al.* [12], the follower may not change the optimal calculated strategy, but deviate from it. Instead of x_i, the follower plays $x_i + \delta_i$.

The authors proposed also the third MAXMIN model, which is a simple combination of BRASS and BOSS models. The main aim in this model is to maximize the minimal reword γ irrespective of the followers' action:

$$\max_{\gamma} \sum_{l\in L} p^l \gamma^l\,, \qquad (36)$$

where the leader's and follower's are defined in the following way:

$$\sum_{i\in X}^{x_i\in[0,1]} x_i = 1\,, \qquad (37)$$

$$\sum_{i\in X}^{x_i\in[0,1]} R^l_{i,j}x_i >= \gamma^l\,. \qquad (38)$$

4.3. COBRA Models

Pita *et al.* in [12] defined following three game models: (i) COBRA(0, ε) model (bounded rationality), (ii) COBRA(α, 0) model (observational uncertainty), and (iii) COBRA(α, ε) model as the combination of (i) and (ii). Parameters α and ε are two main parameters of the games. For the real leader's strategy x and follower's strategy x', the problem of solving the game is defined as the linear problem $x'_i = \alpha(1/|X|) + (1\alpha)x_i$. The value of α=1 indicates the player's behavior in the situation of no knowledge about the other strategies – any strategy is uniformly probable. For $\alpha = 0$ (full information available), $x'_i = x_i$ is the optimal strategy played by the follower. For $\alpha = 1$, $x'_i = (1/|X|)$ is the probability of playing the strategy x'_i.

Using that model, the following problem as the game solution was formulated:

$$\max_{x,q,h,a,\gamma} \sum_{l\in L} p^l \gamma^l\,, \qquad (39)$$

under the following constrains:

$$\sum_{i\in X}^{x_i\in[0,1]} x'_i = 1, \sum_{j\in Q} q^l_j >= 1, \sum_{j\in Q} h^l_j >= 1\,, \qquad (40)$$

$$0 =< (a^l - \sum_{i\in X} C_{i,j}x'_i <= (1-h^l_j)M\,, \qquad (41)$$

$$\varepsilon(1-q^l_j) =< a^l - \sum_{i\in X} C^l_{i,j}x_i <= \varepsilon\,, +(1-q^l_j)M\,, \qquad (42)$$

$$(1-q^l_j)M + \sum_{i\in X} R^l_{i,j}x_i >= \gamma_l\,, \qquad (43)$$

where $x'_i = \alpha(1/|X|) + (1\alpha)x_i$, $h^l_j =< q^l_j$, $h^l_j, q^l_j \in \{0,1\}$, for M being the large positive number, and $a \in \mathbf{R}$.

4.4. ORIGAMI Model

Kiekintveld *et al.* in [13] defined the model in which the attack set can be computed directly for the attacker in order to cover target benefits of defender and for the detriment of attacker. Let us denote by C the coverage vector for the defender selected the optimal strategy, and by c_t the probabilities that t-th target is covered. It is assumed, that including any additional target to the attack set cannot increase the players' payoffs in the equilibrium states of the game. Using indifference equation if $U_a(C) = x$ then:

$$c_t >= \frac{x - U^u_a(t_i)}{U^c_a(t_i)U^u_a(t_i)}\,, \qquad (44)$$

for each target t_i, such that

$$U^u_a(t_i) > x\,. \qquad (45)$$

In the algorithm defined for solving the ORIGAMI game models, the target has maximal $U^u_a(t_i)$, and the attack set is updated in each algorithm iteration for decreasing $U^u_a(t_i)$. After each update of the attack set, the coverage of each target is updated to reach the indifference of attacker payoffs in the attack set.

4.5. SU-BRQR Model

Nguyen *et al.* in [14] modified the standard Stackelberg security model by introducing the following subjective utility function:

$$a_i = w_1c_i + w_2U^u_a(t_i) + w_3U^c_a(t_i)\,, \qquad (46)$$

where w, w_2, w_3. The optimal strategy is calculated as follows:

$$\max_{c} \sum_{i=1}^{n} \frac{e^{(w_1c_i + w_2U^u_a(t_i) + w_3U^c_a(t_i))}}{\sum_{j'=1}^{n} e^{(w_1c_j + w_2U^u_a(t_j) + w_3U^c_a(t_j))}} \cdot \ldots \qquad (47)$$

$$\ldots \cdot \big(c_jU^c_a(t_j) + (1-c_j)U^u_a(t_j)\big)\,,$$

where

$$\sum_{i=1}^{n} c_t <= K, \quad 0 =< c_t <= 1\,.$$

In this model, the adversary has his own preferences according to the importance of the rewords, penalties, and probabilities. The authors recommended the maximum likehood estimation method for estimating the game parameters w_1, w_2, w_3.

4.6. Eraser-C Model

Tsai *et al.* in [15] tried to simplify the standard security Stackelberg game model. In their model, the payoffs depend on the structure of the coverage set (the attacked target can be its element or not). In this model the actions of the players are defined by the targets instead of coverage sets.

4.7. ASPEN Model

Jain *et al.* in [16] considered large, arbitrary schedules in the Stackelberg security game. The main idea of their model is to represent strategy space for defender using column generation, subcompositions into smaller problems, and a technique for searching the space of attacker strategies. The solution is dedicated for large number of defenders of different types.

4.8. GUARDS Model

Bo An *et al.* in [17] defined the model for massive scale games with hundreds of heterogeneous security activities, reasoning over different kind of potential threats. They considered the situation when the defender has the possibility of protecting targets by different heterogeneous security activities for each potential target, and an adversary can execute heterogeneous attacks on a target. In addition, the defender is able to allocate more than one resource for covering a given target. Moreover, the authors defined the defender's uncertainty regarding the payoff values of the attacker, and uncertainty in the attackers' observation of the defender's strategy. Pita *et al.* proposed model for heterogeneous security activities for each target and heterogeneous threats for each target.

4.9. Multiple SSE Case

Tambe *et al.* in [18] defined the game scenario, where the attacker deviates from optimal strategy, with unknown capability constraints that may restrict the attack set. Authors introduced equilibrium refinement algorithm. In the case of multiple SSE states, the developed algorithm is able to choose the robust equilibrium for the most efficient utilization of the available resources. The idea is based on the fact that if the vector of coverage $c = [c_1, \ldots, c_n]$ generates the SSE, then it is possible to find another SSE by reducing coverage of targets outside the attack set. The authors defined the maximum attack set (MSSE) as:

$$M = \{t \in Target\ s : U_a^u(t) >= U_a(c,a)\}. \tag{48}$$

They proved that any security game could not have two maximum attack sets with different attack sets. The authors sorted target set using the values of utility function $U_a^u(t)$ in the following way:

$$Target\ s_{sorted} = \{t_1, \ldots, t_n\}. \tag{49}$$

The authors also developed the concrete algorithm for computing the unique maximum attack set. It starts with $M = t_1$ and generates new targets in each iterated loop (Algorithm 1).

Algorithm 1: Computing the unique maximum attack set

```
$i \leftarrow 0, M \leftarrow, Target\ s_{sorted}$
while $i <= n$ do
    if $M = Target\ s_{sorted}$ then return $M$
    $j \leftarrow i+1,\ M' \leftarrow M \cup \{t_j\},\ Target\ s_{sorted}$
            while $j > n$ and $U_a^u(t_{j+1} = U_a^u(t_j))$ do
                $M' \leftarrow M' \cup \{t_{j+1}\}, j++$
            end do
if Condition C1 is true or C2 is violated for attack set $M'$
then return $M$
    $M \leftarrow M', i \leftarrow j$
end do
```

The following conditions were defined for the above model:

- C1 – $\sum_{t \in M} c_t <= m$,
- C2 – $c_t <= 1$ for each $t \in M$.

4.10. Multi-step Attack MILP Model

Vorobeychik *et al.* in [19] considered the game scenario when each attack may be realized in many steps and to be completed it requires an arbitrary number (h) of such steps. Mixed integer linear programming (MILP) formulation for defender was proposed by discretizing the time unit interval defender probabilities was spited into L intervals. Authors proposed $d_{i,j,l}$ as the binary variables such equals 1 indicates a particular discrete probability choice $p_l \in [0,1]$ for $l = 1\ldots,L-1$ with $p_0 = 0$ and $P_L = 1$, such that only one chose is possible, that is $\sum_l d_{i,j,l} = 1$. Based on this idea, new set of variables $w_{i,j,l} = d_{i,j,l} v_j$ was introduced, where v_j is the expected attacker value of starting in state j. The model includes the probability that a target j is visited in exactly t steps, starting from i and the probability that j is visited in $1\ldots h$ steps.

5. Computational Aspects

All secure Stackelberg game models surveyed in the previous section can be solved by the global optimization of the game utilization function (loss or game payoff) in the same way it was defined in Section 2 for the generic game models. Such global optimization problems for Stackelberg security games can be defined usually as special cases of mixed-integer linear problems (MILP) or mixed-integer-quadratic programming problems (MIQP). Depending on the type of the game, such problems are of different computational complexity (Table 3). Such complexity can be expressed by the number of control variables (strategies), the number of leaders and followers and the number of uncertainty parameters in the game, which are estimated by using the likelihood methods.

Table 3
The characteristics of surveyed Stackelberg models

Reference	Size	Value examined
[17]	5 / 20	Runtime, memory usage
[12]	50 / 200	Runtime, memory usage
[13]	3 / 8	Defender expected utility
[14]	9 / 24	Defenders expected utility
[20]	3 / 3	Pure strategies behavior
[12]	3 / 10	Runtime, expected rewards
[14]	3 / 10	Runtime
[11]	2 / 14	Runtime, speed up

The following theorem was proof according to the computational complexity of the problem [5]. In 2-player normal-form games, an optimal mixed strategy to commit to can be found in polynomial time using linear programming, in 3-player normal-form games, finding an optimal mixed strategy to commit to is NP-hard. Moreover, finding an optimal mixed strategy to commit to in 2-player Bayesian games is NP-hard, even when the leader has only a single type and the follower has only two actions.

5.1. Equilibrium Points

SSE and NE equilibrium states (defined in Section 2) are the typical solutions for Stackelberg and non-cooperative symmetric games. In Stackelberg security game, there is however, the third type of equilibrium state, which can be the most beneficial solution of such game in many practical applications.

Let us denote by $\Omega_{NE} :=$ a set of strategies played for reaching the Nash equilibrium, and by $\Omega_{SSE} :=$ a set of strategies for reaching strong Stackelberg equilibrium.

Definition 8. For a defender's mixed strategy C and attacker's best response strategy $E(C) = \max_{i=1}^{n} U_a(c,t_i)$, a set of defender's minimax strategies is defined as:

$$\Omega_M := \{C : E(C) = E^*\}, \tag{50}$$

where $E^* = \min_C E(C)$ is the minimum of attacker's best response utilities over all defender's strategies.

The following relations among these three types of equilibrium states can be specified:

- in a security game the set of defenders minimax strategies is equal to the set of defenders NE strategies, that is $\Omega_M = \Omega_{NE}$,
- if C is the SSE strategy in a security game that satisfies the property that for any recourse and any subset of a schedule is also a possible schedule then $\Omega_{SSE} \subset \Omega_M = \Omega_{NE}$.

Solving MILP and MIQP problems may be done by one of traditional methods: simplex method, interior-point methods, Conic linear programming, descent methods, conjugate direction methods or Quasi-Newton methods [21]. In addition a lot of new methods were developed recently, from among them: relaxation method [22], Dantzig-Wolfe decomposition [23], primal nested-decomposition method [24].

5.2. Time of Solution Finding

All the Stackelberg security game models presented in Section 3 cannot be compared to each other in the straightforward way, because they differ according to the assumptions. A simple summative analysis have been performed with runtime, memory usage expected utility values, strategies behavior and speed up as the main criteria. The results of such analysis are presented in Table 4. The time that is necessary for computing proper strategies depends on the characteristics of the machine that was used for computation.

We can conclude from conducted simple analysis of the surveyed Stackelberg game models that the strategy space may exponentially increase with the number of security activities, attacks, resources, and the time necessary for finding the game solution.

Table 4
The time a for finding solution to the maximum problem

Reference/model	Time [min]	Size [targets]
[17]	8.2	250
[12]	116	200
[13] DOBSS	4.5	20
[13] ERASER	10.5	3000
[13] ORIGAMI	10.2	3500
[12] COBRA	7.5	8 followers
[12] DOBSS	11	8 followers
[14] BOSS	16.5	200
[11]	16.5	4

6. Use Cases

Stackelberg security games have been successfully implemented in realistic large-scale IT systems for supporting the system management and users and administrators decisions. In this section the most interesting use cases for such game models are reported.

The most spectacular implementation of the security Stackelberg game model is the security system at the Los Angeles International Airport. Randomizing schedules in such systems for monitoring the system performance is a critical issue. The main reason for that is the importance of the knowledge about the possible patrolling that may cause terrorist attacks. This use case was realized as a software-assistant multi-agent system called ARMOR (Assistant for Randomized Monitoring over Routes). This model supports the administrators and users decisions about the location of the checkpoints in the physical environment or canine patrol routes. The decision model is based on the Bayesian Stackelberg games, in which the optimal mixed strategy is generated for the leader (patrol) and the follower (terrorist) may know this mixed strategy when choosing his own strategy in the game [9].

The next example of the practical Stackelberg game is the strategic security allocation system in transportation networks (IRIS) used by Federal Air Marshal Service (FAMS). In transportation networks with hundreds thousands of vehicles, police has to create patrolling schedules in order to ensure safety. Aggressors can observe the law-enforcement patterns and try to exploit generated schedule. IRIS systems use the fastest known solver for this class of security games, namely ERASER-C [9].

Another Stackelberg use case is the United States Transportation Security Administration system (TSA). The transportation systems are very large and protecting them requires many personnel and security activities. System supported the decisions how properly divide resources between layers of security activities. In this type of game, TSA acts as a defender who has a set of targets to protect, a number of security activities and a limited number of resources. The name of dedicated software system is Game-theoretic Unpredictable and Randomly Deployed Security (GUARDS) [9].

There are many applications of game theory in communications and networking. Using a variety of tools from game theory, there was possible to find new solutions in areas related to cellular and broadband networks such as uplink power control in CDMA networks, resource allocation in OFDMA networks, deployment of femtocell access points, IEEE 802.16 broadband wireless access, and vertical handover in heterogeneous wireless networks [25].

7. Conclusions

Security Stackelberg games presented in this paper are very promising tools for modeling the data and user managements, as well as supporting complex decision processes in competitive computational environments with possible conflicts of interests of the users and system administrators and service and resource providers. All surveyed models were based on the realistic characteristics of the systems, namely existing limitations in access to the resources, uncertainty about follower types, non-optimal behavior of the players or limited knowledge of the opponents' actions and strategies. Increasing the efficiency of the game model is strictly connected with the increase of the calculated number of parameters in the game and equations to solve in the game optimization models, which makes of course the all implementation of such models more complex.

Although, all optimization problems related to solving the presented Stackelberg security games are NP-hard, the practical use cases reported in this paper show the high potential practical benefits of using the presented games in transportation systems in USA. It makes such models a potential efficient tool for supporting the complex decisions in large-scale cloud environments, which will be the next step of authors' research on security aspects in cloud computing.

References

[1] R. B. Myerson, *Game Theory: Analysis of Conflict*. Harvard University Press, 1991).

[2] N. Nisan *et al.*, Ed., *Algorithmic Game Theory*. Cambridge University Press, 2007.

[3] J. Pita *et al.*, "Using game theory for Los Angeles Airport security", *Artif. Intell.*, vol. 30, no. 1, pp. 43–57, 2009 (doi: http://dx.doi.org/10.1609/aimag.v30i1.2173).

[4] S. Tadelis, *Game Theory: An Introduction*. Princeton University Press, 2013.

[5] J. F. Nash, "Equilibrium points in *n*-person games", *Proc. of the National Academy of Sciences of the United States of America*, vol. 36, no. 1, pp. 48–49, 1950.

[6] B. von Stengel and S. Zamir, "Leadership with commitment to mixed strategies", Tech. Rep. LSE-CDAM-2004-01, CDAM Research Report, 2004 [Online]. Available: http://www.cdam.lse.ac.uk/Reports/Files/cdam-2004-01.pdf

[7] J. Gan and B. An, "Minimum support size of the defender's strong Stackelberg equilibrium strategies in security games", in *Proc. AAAI Spring Symp. on Appl. Computat. Game Theory*, Stanford, CA, USA, 2014 [Online]. Available: http://www.ntu.edu.sg/home/boan/papers/AAAISS14b.pdf

[8] P. Paruchuri *et al.*, "Efficient Algorithms to Solve Bayesian Stackelberg Games for Security Applications", in *Proc. 23rd Nat. Conf. on Artificial Intelligence AAAI'08*, Chicago, IL, USA, 2008, vol. 3, pp. 1559–1562.

[9] M. Tambe, *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*, 1st ed. Cambridge University Press, 2011.

[10] D. Korzhyk, Z. Yin, C. Kiekintveld, V. Conitzer, and M. Tambe, "Stackelberg vs. Nash in security games: an extended investigation of interchangeability, equivalence, and uniqueness", *J. Artif. Intell. Res.*, vol. 41, no. 2, pp. 297–327, 2011.

[11] M. Jain *et al.*, "Bayesian Stackelberg games and their application for security at Los Angeles international airport", *ACM SIGecom Exchan.*, vol. 7, no. 2, article no. 10, 2008 (doi: 10.1145/1399589.1399599).

[12] J. Pita, M. Jain, M. Tambe, F. Ordoñez, and S. Kraus, "Robust solutions to Stackelberg games: Addressing bounded rationality and limited observations in human cognition, *Artif. Intell.*, vol. 174, no. 15, pp. 1142–1171, 2010, (doi.org/10.1016/j.artint.2010.07.002).

[13] R. Yang, C. Kiekintveld, F. Ordoñez, M. Tambe, and R. John, "Improving resource allocation strategies against human adversaries in security games: An extended study", *Artif. Intell.*, vol. 195, pp. 440–469, 2013 (doi: 10.1016/j.artint.2012.11.004).

[14] A. Tambe and T. Nguyen, "Robust resource allocation in security games and ensemble modeling of adversary behavior", in *Proc. 30th Ann. ACM Symp. Appl. Comput. SAC'15*, Salamanca, Spain, 2015, pp. 277–282 (doi: 10.1145/2695664.2695686).

[15] J. Tsai, S. Rathi, C. Kiekintveld, F. Ordoñez, and M. Tambe, "IRIS – A tool for strategic security allocation in transportation networks", in *Proc. 8th International Conference on Autonomous Agents and Multiagent Systems AAMAS 2009*, Budapest, Hungary, 2009, vol. 2, pp. 1327–1334.

[16] M. Jain, E. Kardes, C. Kiekintveld, F. Ordoñez, and M. Tambe, "Security games with arbitrary schedules: A Branch and price approach", in *Proc. 24th AAAI Conf. on Artif. Intell. AAAI-10*, Atlanta, GE, USA, 2010, pp. 792–797.

[17] B. An, J. Pita, E. Shieh, M. Tambe, C. Kiekintveld, and J. Marecki, "GUARDS and PROTECT: next generation applications of security games", *SIGecom Exch.*, vol. 10, no. 1, pp. 31–34, 2011 (doi: 10.1145/1978721.1978729).

[18] B. An, M. Tambe, F. Ordoñez, E. A. Shieh, and C. Kiekintveld, "Refinement of strong Stackelberg equilibria in security games", in *Proc. 25th AAAI Conf. on Artif. Intell.*, San Francisco, CA, USA, 2011 [Online]. Available: www.aaai.org/OCS/index.php/AAAI/AAAI11/paper/view/3461/3928

[19] J. Letchford and Y. Vorobeychik, "Computing optimal security strategies in networked domains: a cost-benefit approach", in *Proc. 11th Int. Conf. on Autonom. Agents and Multiagent Syst. AAMAS'12*, Valencia, Spain, 2012, vol. 3, pp. 1303–1304.

[20] J. B. Clempner and A. S. Poznyak, "Stackelberg security games: Computing the shortest-path equilibrium", *Expert Syst. with Applications*, vol. 42, no. 8, pp. 3967–3979, 2015 (doi: 10.1016/j.eswa.2014.12.034).

[21] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, 3rd ed. Springer, 2008 (doi: 10.1007/978-0-387-74503-9).

[22] M. Held, R. M. Karp, and P. Wolfe, "Large scale optimization and the relaxation method", in *Proc. of the ACM Annual Conference ACM'72*, Boston, MA, USA, 1972, vol. 1, pp. 507–509 (doi: 10.1145/800193.569964).

[23] J. Rios, "Algorithm 928: A general, parallel implementation of Dantzig-Wolfe decomposition", *ACM Trans. Mathem. Softw.*, vol. 39, no. 3, article no. 21, 2013 (doi: 10.1145/2450153.2450159).

[24] J. K. Ho and R. P. Sundarraj, "Distributed nested decomposition of staircase linear programs", *ACM Trans. Mathem. Softw.*, vol. 23, no. 2, pp. 148–173, 1997 (doi: 10.1145/264029.264031).

[25] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game Theory in Wireless and Communication Networks*, 1 ed. Cambridge University Press, 2012.

Andrzej Wilczyński is an Assistant Professor at Cracow University of Technology and Ph.D. student at AGH University of Science and Technology. The topics of his research are multiagent systems and cloud computing.

E-mail: and.wilczynski@gmail.com
AGH University of Science and Technology
Mickiewicza av 30
30-059 Cracow, Poland

Tadeusz Kościuszko Cracow University of Technology
Warszawska st 24
31-155 Cracow, Poland

Agnieszka Jakóbik (Krok) received her M.Sc. in the field of stochastic processes at the Jagiellonian University, Poland and Ph.D. degree in the field of neural networks at Cracow University of Technology, Poland, in 2003 and 2007, respectively. She is an Assistant Professor at Cracow University of Technology. Her main scientific interests are cryptography, cloud systems, including cloud security, big data systems, modeling and simulation using artificial intelligences.

E-mail: agneskrok@gmail.com
Faculty of Physics, Mathematics and Computer Science
Tadeusz Kościuszko Cracow University of Technology
Warszawska st 24
31-155 Cracow, Poland

Joanna Kołodziej is an Associate Professor in Department of Computer Science of Cracow University of Technology. She is a vice Head of the Department for Sciences and Development. She serves also as the President of the Polish Chapter of IEEE Computational Intelligence Society. She published over 150 papers in the international journals and conference proceedings. She is also a Honorary Chair of the HIPMOS track of ECMS. The main topics of here research is artificial intelligence, grid and cloud computing, multiagent systems.

E-mail: jokolodziej@pk.edu.pl
Faculty of Physics, Mathematics and Computer Science
Tadeusz Kościuszko Cracow University of Technology
Warszawska st 24
31-155 Cracow, Poland

Mitigation of Scintillation Effects in WDM FSO System using Multibeam Technique

Marvi Grover, Preeti Singh, and Pardeep Kaur

Department of Electronics and Communication Engineering, University Institute of Engineering and Technology, Panjab University, Chandigarh, India

http://doi.org/10.26636/jtit.2017.111917

Abstract—Free Space Optical communication (FSO) has engrossed a large section of researchers in recent times due to its wide bandwidth, effortless deployment and immune links making it appropriate for communication purposes. This wireless optical technique requires clear and non-turbulent atmospheric conditions for efficient transmission. In this paper, authors aim at reducing the effect of turbulent atmospheric conditions like scintillation effect on FSO. Multibeam technique, which uses spatially diverse transmitters for transmission, has been used for increasing the achievable link distance of the FSO system. Parameters like quality factor and bit error rate have been used to check the received signal quality.

Keywords*—*laser, link length, multibeam, scintillation.

1. Introduction

Free Space Optical communication (FSO) or sometimes addressed as laser communication is a technology that uses laser beams through free space to reach the receiver. This technology owes its growing importance to the incredible increase in the volume of data transfer all over the world and the resultant increase in bandwidth requirements. FSO's key attributes like rapid data transfer, quicker deployment, cost effective infrastructure and data rates as high as tens of gigabytes per 1 second make it a viable alternative for the short-range radio frequency (RF) links [1], [2]. Licensed frequency bands, spectrum congestion and lesser data rates as compared to FSO, are some of the demerits of RF. Nowadays, FSO is finding its application in almost every stratum of daily life, ranging from ship to ship communication to enterprise connectivity [3].

Like every other technology FSO also has some limitations and some design considerations which need to be contemplated. Light beam carrying the information travels through air and is encumbered by the atmospheric effects, like rain, fog, snow, haze, and the atmospheric turbulences due to temperature and pressure fluctuations in the atmosphere [4]. Absorption, scattering and scintillation of light are consequences of turbulent atmospheric conditions [5]. Line of sight (LOS) is an imperative requirement in FSO communication, but sometimes physical objects like birds or poles temporarily obstruct it, making the link unachievable.

This paper focuses on the impairments caused by atmospheric effects on an FSO link. When considering the atmospheric effects, scintillation effect is the most detrimental one, so the authors here have tried to reduce this effect using some techniques described in this paper.

A brief description of the harm caused by scintillation on the light beam is given below.

1.1. Scintillation Effect

Scintillation refers to the turbulence caused by thermal inhomogeneities along the path of light beam. Wind velocity is always variable, which transfers heat and water vapors in the form of eddies. Temperature changes in the atmosphere caused by these eddies lead to heating up of air pockets called Fresnel zones having different temperatures and different densities, which lead to refractive index differences [5]. Turbulences are random, which means that these pockets are continuously being created and destroyed. Fluctuations in the refractive index of air deform the laser beam causing "beam dancing" at the receiver. Figure 1 shows the scintillation effect with air pockets having different refractive indices. Randomly formed pockets refract the optical wavefront of the incoming beam due to which the signal cannot be received properly [6]. The refractive index structure parameter C_n^2, accounts for the strength of fluctuations. C_n^2 varies from 10^{-16} m$^{-2/3}$ (weak scintillation) to 10^{-12} m$^{-2/3}$ (strong).

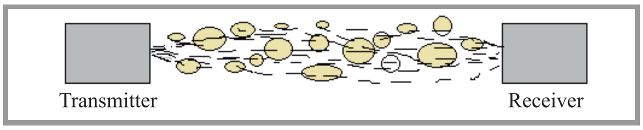


Fig. 1. Heated air pockets which lead to scintillation of light.

Two common effects of scintillation on the optical beam are:

- Beam Wander – the refractive index fluctuations are due to turbulent eddies of size varying from few millimeters to hundred meters. Beam wander means that the beam is deflected from its original path and loses its los. It happens when the size of refractive index inhomogeneity is greater than the beam diameter;

- Beam Spreading – when the inhomogeneities are lower than the size of beam diameter, they tend to broaden the beam but do not deflect it. This is called beam spreading. It defocuses the beam reducing its intensity.

In communication systems, bandwidth is always a factor that needs deliberation, so only the mitigation of channel turbulence like scintillation effect does not solve the purpose. It should be combined with efficient bandwidth utilization in order to make it a quintessential system. One of the best techniques used here is Wavelength Division Multiplexing (WDM).

2. WDM Systems

WDM allows multiplying data streams over optical carriers having different wavelengths called channels and sent as a single signal. WDM FSO systems use a single light beam to transmit the multiplexed signal through free space [7]. A multiplexer is used at the transmitter to combine different modulated carriers and a demultiplexer at the receiver to restore each one (Fig. 2).

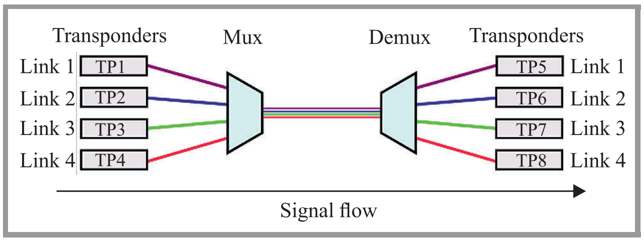


Fig. 2. WDM technology.

WDM system used in combination with FSO are called WDM FSO and can be classified into two types: single beam and multibeam systems.

Single beam system uses one pair of transmitter and receiver. Only one beam carrying the information travels through the channel. In case of FSO systems, if the light beam is obstructed by an object, which prevents it from reaching the receiver, the signal is lost and communication stops.

The multibeam WDM uses more than one beams of the multiplexed signal. Each beam travels a different path, and thus its attenuation is different. This technique uses spatially diverse transmitters and so it is also called Spatial Diversity Technique [8], [9]. At the receiver, the beam that has undergone least attenuation is selected and processed for data extraction. This technique serves as a solution for various FSO limitations like physical obstructions, scintillation effect, weather effects, etc. Multibeam system improves the link achievability and reduces the probability of link failure to a large extent [10]–[12]. When WDM FSO system uses multiple beams for transmission, they are called "Hybrid multibeam WDM FSO systems". Figure 3 shows a hybrid multibeam WDM FSO, which combines the advantages of WDM and spatial diversity to increase the system capacity and link reliability.

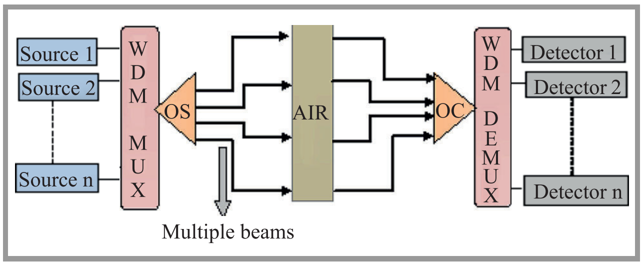


Fig. 3. Multibeam WDM FSO system block diagram.

3. System Design and Analysis

In this paper two WDM FSO systems are used and analyzed under scintillation effect. First is WDM FSO, which uses single beam technology and system 2 uses the multibeam technology. System 1 has been already used by the authors in [14]. System 2 has been designed with an aim to improve the efficiency of system 1 under identical atmospheric conditions. Quality factor (Q) and bit error rate (BER) have been used as the measures of received signal quality. Comparative analysis of both systems has been done in terms of link distance and received power for best values of quality factor and BER. The software used for analysis are OptiSystem v12 and Matlab.

3.1. System Model

Figure 4 shows the layout of system 1 designed in OptiSystem software. The transmitter section consists of continuous wave (CW) laser source. The fork component is used to copy the signal generated by the laser source so that it can be given to the multiplexer, which separates it into carriers differing in wavelength. The pseudo-random bit sequence (PRBS) source is used to generate codes corresponding to the information signal. It is followed by non-return to zero (NRZ) pulse generator, which gives the electrical pulses for the signal generated by the PRBS using NRZ pulse generation format. The Mach-Zehnder modulator (MZM) does the modulation and next the modulated signal is transmitted through the free space channel. At the receiver, a demultiplexer is installed with signal carrier selects then the photodetector for conversion to electrical signal. In next block the signal is filtered, regenerated and sent to the corresponding user. BER analyzer is used to view the quality factor, BER value calculation and eye diagram of the received signal.

System 2 differs from system 1 only in the way of transmission after modulation. As this is a multibeam system, it uses spatially diverse transmit apertures to transmit the signal. As shown in the Fig. 5, a fork is used after MZM modulator to simulate four different transmit apertures and a single receiver lens, which make it a 4×1 WDM FSO system. The four beams transmitted are identical but travel

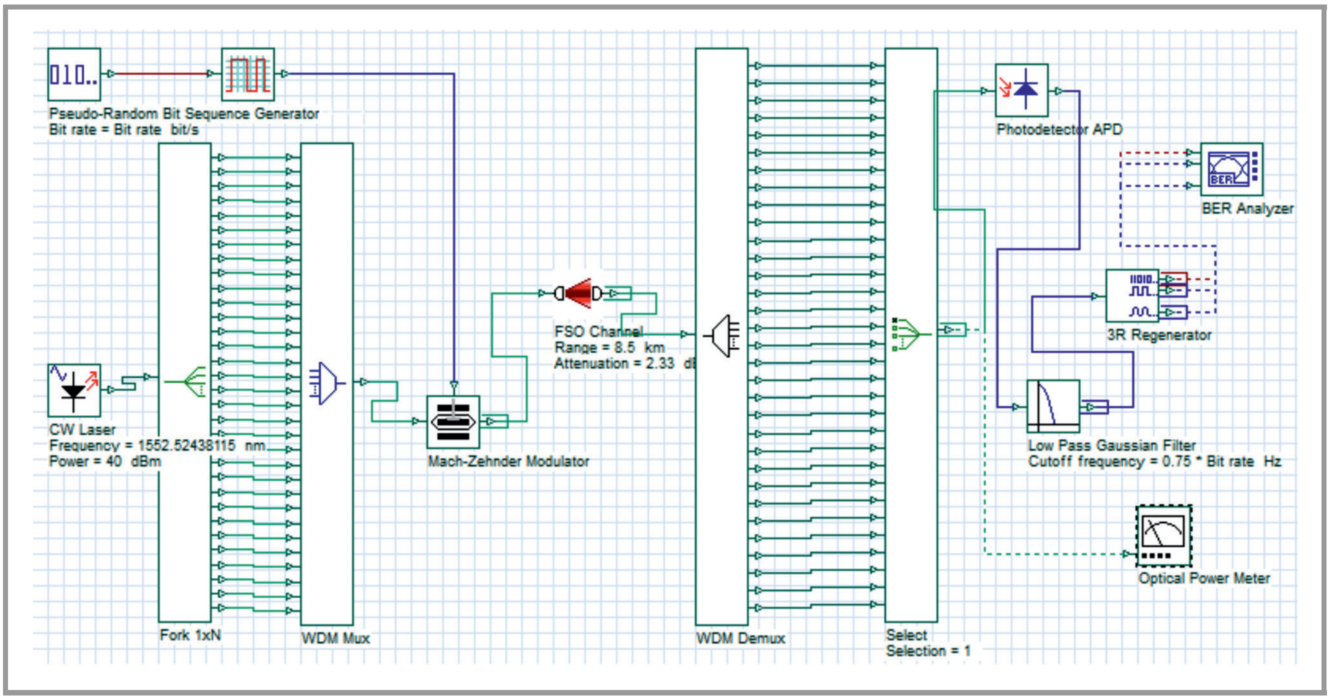


Fig. 4. System 1 designed in OptiSystem.

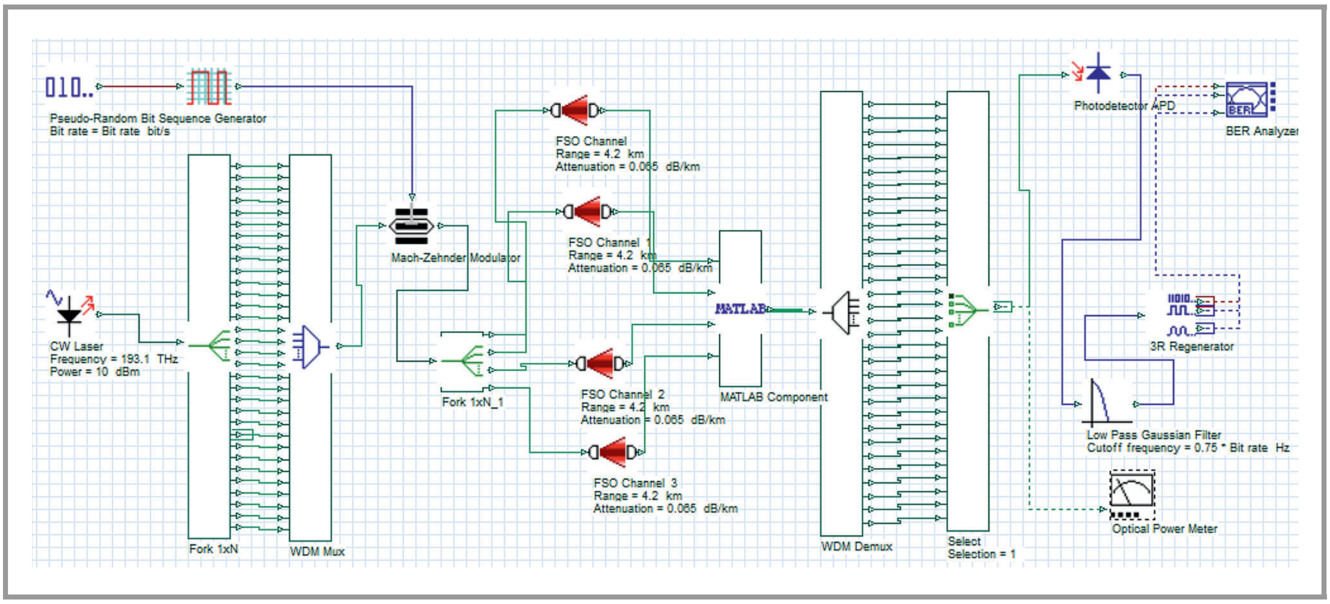


Fig. 5. System 2 schematic.

different paths to the receiver and thus undergo different amount of scintillation. Figure 5 shows a MATLAB component, which intakes the incoming signals at the receiver, selects the least attenuated out of the four, and sends it to the demultiplexer.

The CW laser operates at 1550 nm, used data rate is 10 Gb/s, transmitter and receiver lens apertures are kept as 15 cm. Geometrical loss has also been considered in the analysis, so the beam divergence is taken to be 2 mrad. There are various models available, for mathematical modeling of the turbulence affected FSO channel. These models give the probability density function (PDF) of the received signal after passing through the turbulent atmospheric conditions. When the channel is affected by weak turbulence it is modeled using log-normal model. In case of strong turbulence in the channel negative exponential model and K-turbulence model are used [13]. This work has been done using "gamma-gamma" turbulence model [14], which is used when the turbulence varies from moderate to strong.

The gamma-gamma model is used to model the irradiance of optical channels for moderate to strong turbulence channels resulting from small scale and large scale refractive index fluctuations due to temperature and pressure inhomogeneities. The PDF of the turbulent channel is given by:

$$P(I) = \frac{2(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{\Gamma(\alpha)\Gamma(\beta)} I^{\frac{\alpha+\beta}{2}-I} K_{\alpha-\beta}\left(2\sqrt{\alpha\beta I}\right), \qquad (1)$$

where $\frac{1}{\alpha}$ and $\frac{1}{\alpha}$ are the variances of the small scale and large scale eddies respectively, Γ is the gamma function and $K_{\alpha-\beta}(\ldots)$ is the modified second order Bessel function. I is the intensity of the received signal.

Equations (2)–(3) give the values of α and β respectively:

$$\alpha = \mathrm{e}^{\frac{0.49\sigma_r^2}{\left(1+1.11\sigma_r^{\frac{12}{5}}\right)^{\frac{5}{6}}}} - 1\,, \qquad (2)$$

$$\beta = \mathrm{e}^{\frac{0.51\sigma_r^2}{\left(1+0.69\sigma_r^{\frac{12}{5}}\right)^{\frac{5}{6}}}} - 1\,, \qquad (3)$$

where σ_r^2 is the Rytov variance, which characterizes the strength of turbulence and is calculated by:

$$\sigma_r^2 = 1.23 C_n^2 k^{\frac{7}{6}} z^{\frac{11}{6}}\,, \qquad (4)$$

where k is the wave number, z is the range of the link, and C_n^2 is the refractive index structure parameter, which is the qualitative measure of optical turbulence.

4. Performance Analysis

With the effect of scintillation depends on the refractive index structure parameter C_n^2, which is given as a parameter to the FSO channel and the signal is attenuated according to the value of C_n^2. For system 1, C_n^2 is taken to 10^{-13} m$^{-\frac{2}{3}}$, which corresponds to strong turbulence. When simulated for refractive index structure parameter, the maximum link distance achieved with acceptable quality factor is 1.9 km. The Q factor value for this distance was recorded to be 5.96 and the BER was $1.21 \cdot 10^{-9}$. Weather is assumed to be clear to see the effect of scintillation, so in the attenuation specification of the FSO channel, the value given is 0.065 dB/km.

Multibeam WDM FSO system uses four beams of the system propagate independently hence, suffer different amount of scintillation, which depends upon the refractive index structure parameter. The value of C_n^2 used for the four beams is 10^{-16}, 10^{-15}, 10^{-14}, 10^{-13} m$^{-\frac{2}{3}}$ to represent that the beams undergo different scintillation eddies due to their different propagation paths. This system works efficiently up to 4.2 km with the Q factor of 5.94 and BER of $1.44 \cdot 10^{-9}$. If the distance is further increased, the Q factor falls below its value for successful communication.

5. Results Discussion

Both systems have been compared in terms of Q factor value and received optical power. Figures 6 and 7 present systems performance in terms of Q factor variation with respect to link distance and illustrates the difference in quality of received signal at various link lengths. Graph shows that system 1 works till around 1.9 km whereas for system 2, signal quality is acceptable up to 4.2 km.

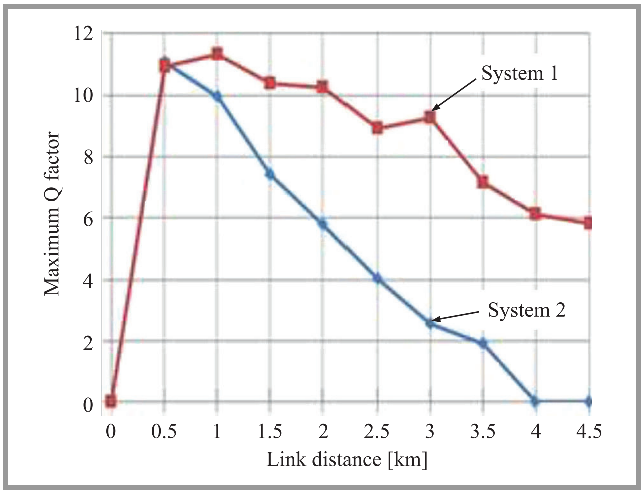


Fig. 6. Comparison of system 1 and 2 under scintillation effect in terms of maximum Q factor.

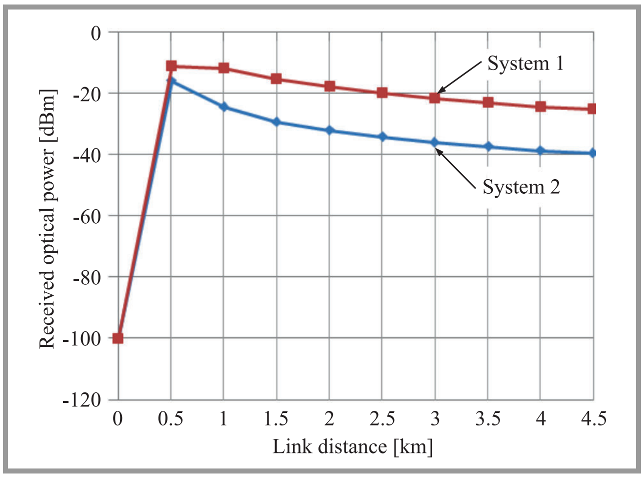


Fig. 7. Comparison of system 1 and 2 under scintillation effect in terms of received optical power.

Graph comparing the received power for both the systems (Fig. 7) shows that the received optical power of system 2 is always greater than that of system 1 when plotted against the link distance. Both the graphs clearly favor the performance of system 2, when analyzed under scintillation effect. Both systems have also been compared using the eye diagrams. Figure 8 shows the eye diagrams of both the systems at 1.9 km and show that the Q factor of system 2 (the red curve in the diagram) is much higher than that of system 1, also the eye height for system 2 is 110, whereas that for system 1 is around 20. This difference in the eye heights also indicates better signal reception of system 2.

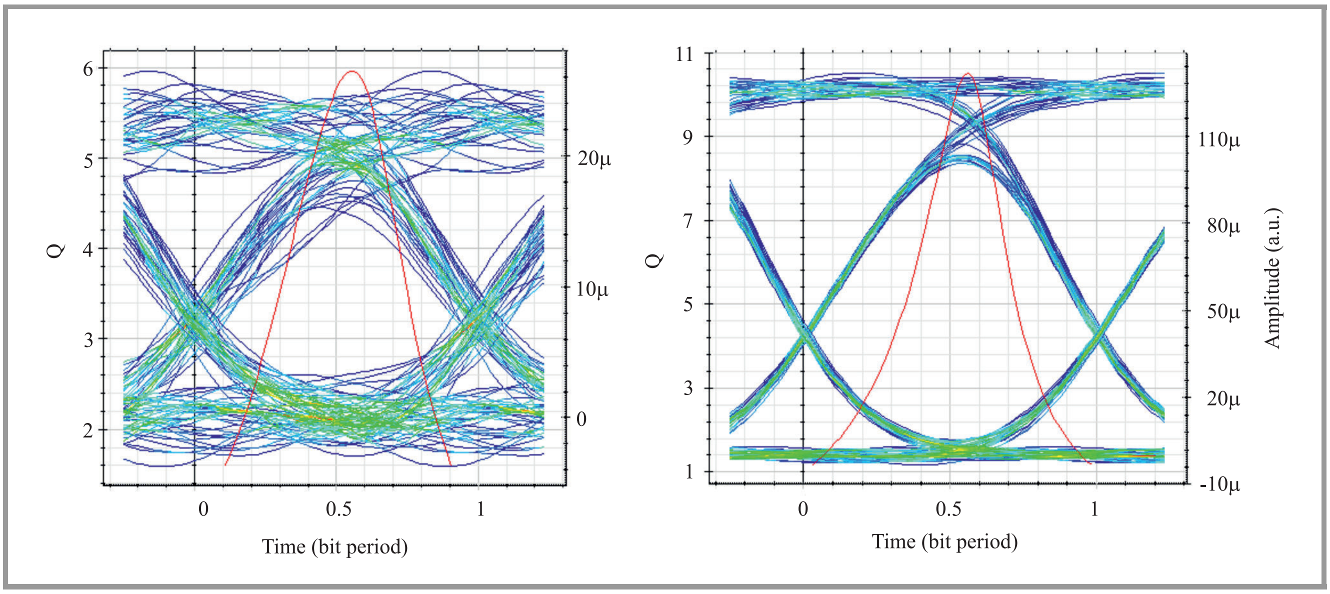


Fig. 8. Eye diagrams for system 1 (left) and 2 (right) at 1.9 km under scintillation effect.

Table 1
Comparison of system 1 and 2 under scintillation effect

	$C_n^2\ [\mathrm{m}^{-\frac{2}{3}}]$	Max. link distance [km]	Q factor	Min. BER
System 1	10^{-13}	1.9	5.94	$1.2 \cdot 10^{-9}$
System 2	10^{-16}, 10^{-15}, 10^{-14}, 10^{-13}	4.2	5.95	$1.34 \cdot 10^{-9}$

Both system performances and difference in the FSO link distance in Table 1 is summarized.

5.1. Validation of Results Using Matlab

To check the credibility of the above results, multibeam FSO links have been simulated using Matlab. The PDF of received power has been plotted against the received power I using MATLAB as well as OptiSystem. Figure 9 shows the comparison graph obtained by using PDF and I values both software tools.

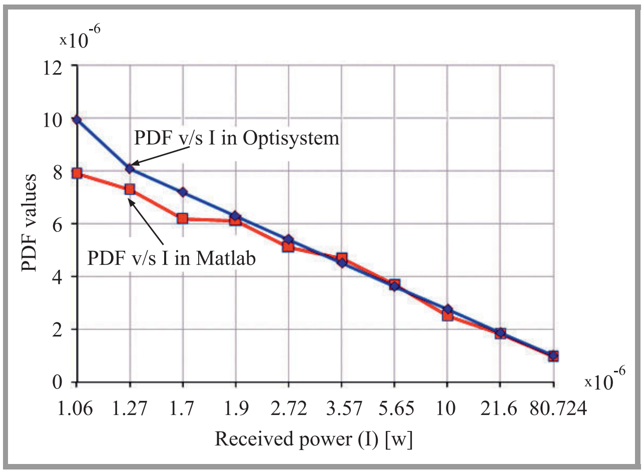


Fig. 9. Comparison of PDF vs. I curves obtained by using OptiSystem and Matlab under identical FSO channels.

There is a big similarity between results given by both software. Thus, it can be inferred, that the analysis done is a valid one.

6. Conclusion

Analysis shows that when simulated under scintillation effect, multibeam system transmits successfully up to 4.2 km. It is much greater than that achieved by the single beam system, which transmits only up to 1.9 km, under same atmospheric conditions. Multibeam system outperforms single beam system taking into account scintillation effect. Thus it can be used in the FSO applications where the signal reliability is important.

References

[1] D. Kedar and S. Arnon, "Urban optical wireless communication networks: The main challenges and possible solutions", *IEEE Commun. Mag.*, vol. 42, no. 5, pp. S2–S7, 2004 (doi: 10.1109/MCOM.2004.1299334).

[2] J. Kaufmann, "Free space optical communications: An overview of applications and technologies", in *Boston IEEE Communications Society Meeting CommSoc 2011*, Boston, MA, USA, 2011.

[3] A. Malik and P. Singh, "Free space optics: Current applications and future challenges", *Int. J. of Optics*, vol. 2015, article ID 945483 (doi: 10.1155/2015/945483).

[4] V. Sharma and G. Kaur, "Degradation measures in free space optical communication (FSO) and its mitigation techniques – A review", *Int. J. of Computer Appl.*, vol. 55, no. 1, pp. 23–27, 2012.

[5] M. Abtahi, P. Lemieux, W. Mathlouthi, and L. A. Rusch, "Suppression of turbulence-induced scintillation in free space optical communication systems using saturated optical amplifiers", *Lightwave Technol.*, vol. 24, no. 12, pp. 4966–4973, 2006.

[6] M. Ali and A. Ali, "Atmospheric turbulence effect on free space optical communication", *Int. J. of Emerging Technol. in Comput. and Appl. Sci.*, vol. 5, no. 4, pp. 345–351, 2013.

[7] A. Malik and P. Singh, "Comparative analysis of point to point FSO system under clear and haze weather conditions", *Wirel. Personal Commun.*, vol. 80, no. 2, pp. 483–492, 2014.

[8] A. B. Mohammad, "Optimization of FSO system in tropical weather using multiple beams", in *Proc. 5th Int. Conf. on Photonics ICP 2014*, Kuala Lumpur, Malaysia, 2014
(doi: 10.1109/ICP.2014.7002326).

[9] Y. Zhao, D. Xu, and X. Zhong, "Scintillation reduction using multibeam propagating technique in atmospheric WOCDMA system", *Chinese Optics Lett.*, vol. 4, no. 11, pp. 110602–110605, 2011 (doi: 10.3788/COL201109.110602).

[10] T. A. Tsiftsis *et al.*, "Optical wireless links with spatial diversity over strong atmospheric turbulence channels", *IEEE Trans. on Wirel. Commun.*, vol. 8, no. 2, pp. 951–957, 2009.

[11] S. A. Al-Gailani, A. B. Mohammad, and R. Q. Shaddad, "Enhancement of free space optical link in heavy rain attenuation using multiple beam concept", *Optik*, vol. 124, no. 21, pp. 4798–4801, 2013.

[12] N. H. M. Noor, W. Al Khateeb, and A. W. Naji, "Experimental evaluation of multiple transmitters/receivers on free space optics link", in *Proc. IEEE Student Conf. on Res. and Develop. SCOReD 2011*, Cyberjaya, Malaysia, 2011 (doi: 10.1109/SCOReD.2011.6148721).

[13] X. Zhu and J. M. Kahn, "Free-space optical communication through atmospheric turbulence channels", *IEEE Trans. on Commun.*, vol. 50, no. 8, pp. 1293–1300, 2002.

[14] D. Shah, B. Nayak, and D. Jethawani, "Study of different Atmospheric channel models", *Int. J. of Electron. and Commun. Engin. & Technol.*, vol. 5, no. 1, pp. 105–112, 2014.

Marvi Grover has completed her M.Sc. in Electronics and Communication Engineering from U.I.E.T., Panjab University, Chandigarh, India. She has done her research in wireless optical communication. Her area of interest is free space optical communication.

E-mail: marvi310191@gmail.com
Department of Electronics and Communication Engineering
University Institute of Engineering and Technology
Panjab University
Chandigarh, India

Preeti Singh is an Assistant Professor in Electronics and Communication Engineering Department in U.I.E.T., Panjab University, Chandigarh, India. She has done her B.Sc. and M.Sc. degree in Electronics and Communication Engineering. She got her Ph.D. in the 2013. Her areas of interest are optical communication (wired and wireless), optical biosensors and cognitive neuroscience.

E-mail: preeti_singh@pu.ac.in
Department of Electronics and Communication Engineering
University Institute of Engineering and Technology
Panjab University
Chandigarh, India

Pardeep Kaur is working as Assistant Professor in Electronics and Communication Engineering Department in U.I.E.T., Panjab University, Chandigarh, India. She has done her B.Sc. and M.Sc. degree in Electronics and Communication Engineering. She is perusing her Ph.D. in wireless sensor networks. Her areas of interest are optical and wireless communication.

E-mail: pardeep.tur@gmail.com
Department of Electronics and Communication Engineering
University Institute of Engineering and Technology
Panjab University
Chandigarh, India

Comparative Study of Wireless Sensor Networks Energy-Efficient Topologies and Power Save Protocols

Ewa Niewiadomska-Szynkiewicz, Piotr Kwaśniewski, and Izabela Windyga

http://doi.org/10.26636/jtit.2009.3.942

Abstract— Ad hoc networks are the ultimate technology in wireless communication that allow network nodes to communicate without the need for a fixed infrastructure. The paper addresses issues associated with control of data transmission in wireless sensor networks (WSN) – a popular type of ad hoc networks with stationary nodes. Since the WSN nodes are typically battery equipped, the primary design goal is to optimize the amount of energy used for transmission. The energy conservation techniques and algorithms for computing the optimal transmitting ranges in order to generate a network with desired properties while reducing sensors energy consumption are discussed and compared through simulations. We describe a new clustering based approach that utilizes the periodical coordination to reduce the overall energy usage by the network.

Keywords— ad hoc network, energy conservation protocols, topology control, wireless sensor network.

1. Introduction to Ad Hoc and Wireless Sensor Networks

An ad hoc network is a wireless decentralized structure network comprised of nodes, which autonomously set up a network. No external network infrastructure is necessary to transmit data – there is no central administration. Freely located network nodes participate in transmission. Network nodes can travel in space as time passes, while direct communication between each pair of nodes is usually not possible. Generally, ad hoc network can consist of different types of multi functional computation devices.

Wireless sensor network (WSN) is most often set up in an ad hoc mode by means of small-size identical devices grouped into network nodes distributed densely over a significant area. These devices, each equipped with central processing unit (CPU), battery, sensor and radio transceiver networked through wireless links provide unparalleled possibilities for collection and transmission of data and can be used for monitoring and controlling environment, cities, homes, etc. In most cases WSNs are stationary or quasi-stationary, while node mobility can be ignored. There is no prearrangement assumption about specific role each node should perform. Each node makes its decision independently, based on the situation in the deployment region, and its knowledge about the network. In the case of networks comprising several hundreds or thousands of nodes, it is necessary to choose an architecture and technology which will enable relatively cheap production of individual devices. For this reason, WSNs need some special treatment as they have unavoidable limitations, for example, limited amount of power at their disposal. Each battery powered device, participating in WSN needs to manage its power in order to perform its duties as long, and as effective as possible. Wireless sensors are thus characterized by low processing speed, limited memory and communication range.

Wireless sensor networks [1]–[3] can be used in different environments and situations and perform tasks of different kinds. Their application will condition the network topology and the choice of technology for its production. The network protocols used in the case of networks whose operating range covers a single building will differ from those operating within large space areas. The construction of a network capable of performing its task requires obtaining information on the devices (nodes) it comprises. The crucial data is the following: geographical location of network nodes, admissible power of radio transmitter and options for control of signal power, estimated number of network nodes, number of nodes that can be lost before the network is declared non-operational, assumed network functionality (maximization of nodes operational time, maximization of throughput, etc.).

In our paper we discuss the approaches to design the optimal w.r.t. minimal energy consumption WSN topologies. The short description of communication methods, energy conservation techniques (power save protocols) and algorithms for computing the optimal transmitting ranges in order to generate a network with desired properties while reducing sensors energy consumption (topology control protocols) is provided. Power save protocols attempt to save nodes energy by putting its radio transceiver in the sleep state. Topology control protocols are responsible for providing the routing protocols with the list of the nodes' neighbors, and making decisions about the ranges of transmission power utilized in each transmission. We analyze the properties of two location based distributed topology control protocols, and report the results of simulation experiments covering a wide range of network system configurations. Finally, we discuss the idea of our novel location based power save scheme utilizing hierarchical structure with periodic coordination of network nodes activity.

2. Communication Methods

Communication protocols used in modern wireless networks like IEEE 802.11 or Bluetooth (IEEE 802.15.1) enable ad hoc mode operation. However, for the protocols to operate in this mode in practice, several basic issues must be solved [2]–[4]. The most important ones are:

- **Limited resources**. Nodes comprised by the network are often small battery-fed devices, which means their power source is limited. The network's throughput is also limited.
- **Poor quality of connection**. The quality of wireless transmission depends on numerous external factors, like weather conditions or landform features. Part of those factors change with time.
- **Small communication range**.

Small communication range in WSN networks results in communication limitations. Each node communicates only with the nodes present in its closest vicinity – the neighbors. For this reason, the natural communication method in wireless sensor networks is the multihop routing. When using multihop routing, it is assumed that the receiving node is located outside the transmitter's range. Contrary to single-hop networks, the transmitter must transmit data to the receiver by means of intermediate nodes. This is a certain limitation that hinders the implementation of routing algorithms but enables the construction of network of greater capacity. Multihop network enables simultaneous transmission via many independent routes. Independence of routes reduces the interference between individual nodes, which additionally enhances the wireless transmission speed in comparison to single-hop networks, where devices share a common space.

Individual WSN network node can collect data recorded by sensors but do not have enough power to process it. Moreover, analyzes require collection of information from many points. Therefore, efficient inter-node communication is necessary in order to transfer data to the base station.

3. Topology Control

Transmission of data package between two network nodes x_i and x_j requires power proportional to d_{ij}^2, where d_{ij} denotes the Euclidean distance between sender and receiver. Lets assume that instead of performing direct transmission, a relay node x_k is used. In such case two transmissions need to be performed: from a source node x_i to a relay node x_k (distance d_{ik}) and from the node x_k to the destination node x_j (distance d_{kj}). Lets consider a triangle $x_i x_k x_j$, also let α be an angle at vertex x_k. By elementary geometry we have:

$$d_{ij}^2 = {d_{ik}}^2 + {d_{kj}}^2 - 2 d_{ik} d_{kj} \cos\alpha, \tag{1}$$

when $\cos\alpha \leq 0$, total amount of energy spent to transmit a data package is smaller when a relay node is used. Generally, short transmissions in the network are desired. They involve smaller power consumption and cause less interference in a network, simultaneously effected, transmissions, thus increasing the network throughput. In general, the goal of topology control (TC) [3] is to identify the situation when the using of the relay node is more energy-efficient than direct transmission and create the network topology accordingly. Topology control assumes that the nodes have impact on the power used to transmit a message. The basic task of TC algorithm consists in attributing the level of power used to send messages to every node in order to minimize the amount of power received from the power source, while at the same time maintaining the coherence of the network.

3.1. Topology Control Protocols

Topology control protocols are responsible for providing the routing protocols with the list of nodes' neighbors, and making decisions about the ranges of transmission power utilized in each transmission. The open systems interconnection (OSI) network model assumes that routing task is dealt with the network layer. On the other hand all functions and procedures required to send data through the network are stored in the OSI data link layer. Therefore the topology control layer is placed partially in the OSI network layer and the OSI data link layer, as presented in Fig. 1.

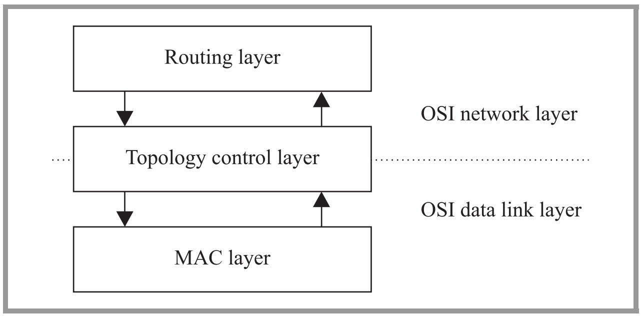


Fig. 1. Placement of topology control layer in the OSI stack.

Topology control protocols may utilize various information about a network, nodes localization and resources [3]–[5]. We can divide these protocols into several groups.

- **Homogeneous topology control protocols** assume that each node uses the same value of transmission power, which reduces the problem to simpler task of finding the minimal level of transmit power such that certain network property is achieved.
- **Location based topology control protocols** utilize the information about geographical location of nodes in the deployment area.
- **Neighbor based topology control protocols** assume that no information about location of nodes is available but each node can determine set of its neighbors and build an order on this set. Order may be based on round trip time, link quality or signal strength.

3.2. Location Based Protocols

We implemented and tested two location based protocols: R&M developed by Rodoplu and Meng, described in [6] and LMST (local minimum spanning tree) proposed by Li, Wang and Song in [7]. The short description of these techniques is provided.

The R&M and LMST protocols. Let N be a set of n wireless nodes deployed in the certain region and forming WSN. Assuming that R_i denotes the maximal transmission range assigned to ith node we can generate the communication graph $G = (N,E)$ induced by R on a given WSN. The E denotes a set of directed edges, and the directed edge $[x_i,x_j]$ exists if x_i and x_j are neighbors, i.e., $d_{ij} \leq R_i$, where d_{ij} denotes the Euclidean distance between sender and receiver. The communication graph G obtained when all the nodes transmit at maximum power is called *maxpower graph*.

Let us consider the situation when all nodes transmit the collected data to one (or more) master node(s) x_m – a base station(s). We can formulate the minimum energy all-to-one communication problem of calculating the optimal reverse spanning tree T of maxpower graph G rooted at x_m:

$$\min_{T} \sum_{x_i \in N, i \neq m} C(x_i, Pred_T(x_i)), \tag{2}$$

where $Pred_T(x_i)$ denotes the predecessor of ith node in the spanning tree T and $C(\cdot)$ the energy cost of transmission from x_i to its predecessor.

The R&M protocol calculates the most energy-efficient path from any node to the master node. It is composed of two phases.

- **Phase 1**. The goal is to compute the enclosure graph of all nodes in WSN. Each node sends a broadcast message, at maximum power, containing its ID and location information. As such message is received by x_i from any neighbor node, x_i identifies the set of nodes locations for which communicating through relay node is more energy efficient than direct communication (the relay region of x_i). Next, x_i checks if the newly found node is in the relay region of any previously found neighbors. A node is marked *dead* if it lays in the relay region of any neighbor of x_i, and *alive* otherwise. After receiving broadcast messages from all neighbors, the set of nodes marked with *alive* identifier creates the enclosure graph of x_i.

- **Phase 2**. In the second phase the optimal, i.e., minimum-energy reverse spanning tree rooted at the master node is computed. The Bellman-Ford algorithm [8] for shortest path calculation is used on the enclosure graph that was determined in the phase 1. Each node computes the minimal cost, i.e., minimal energy to reach the master node given the cost of its neighbors, and broadcasts the message with this value at its maximum power. The operation is repeated every time a message with a new cost is received. After all nodes determine the minimum energy neighbor link, the optimal topology is computed.

The second considered protocol LMST can be used to WSN with nodes equipped with transceivers with the same maximum power. LMST operates in three phases.

- **Phase 1**. Each node sends a broadcast message, at maximum transmit power, containing its ID and location information to its one hop neighbor in the maxpower graph.

- **Phase 2**. The topology is generated. Each node determines a set of its neighbors, calculates Euclidean distance to every neighbor, and finally creates a minimum spanning tree based on its neighbors and computed distances (edge weights in the MST). Final network topology is derived from local MST created by all nodes. Neighbor set of each node consists of nodes, which are its direct neighbors in its local MST. Unfortunately, created topology may contain unidirectional links. Two approaches are proposed to solve this problem: it is assumed that all of them are bidirectional links or all unidirectional links are removed.

- **Phase 3**. Transmission power required to reach every neighbor in a given topology is calculated based on the broadcast messages transmitted in the first phase. Based on the measurements of power of the broadcast messages and knowledge about power level used when transmitting the message, it is possible to compute power level needed to reach the target neighbor.

Simulation results. The performance of R&M and LMST in terms of energy conservation was investigated through simulation. We carried out a set of experiments for various wireless sensor network topologies. It was assumed that all data collected in sensors were transmitted to one base station. We compared the results obtained using both algorithms with those when energy consumption was not considered while routing calculation. The key metric for evaluating the listed methods was the energy consumption used for data transmission. All experiments were conducted using the popular software for network simulation – ns-2 [9]. We implemented R&M and LMST protocols based on modules provided in ns-2 library of classes. The sensor networks with 50 – 300 nodes simulating the commercially available MICA2 sensors [10] with randomly generated positions in a square regions 400×400 to 3000×3000 were considered in our experiments. The technical parameters of sensors were taken from [11], i.e., the radio power consumption for transmission was from 8.6 mA (RF transmission power -20 dBm) to 25.4 mA (RF transmission power 5 dBm), the initial energy resource of each node was assumed to be equal to 21 kJ.

The objective of the first series of simulations was to compare the topologies calculated using described algorithms. The results are presented in Figs. 2 – 4. The base station is marked with the bold dot in presented figures. Figures 3 and 4 show the topologies formed using the LMST and R&M protocols. The obtained results can be compared with the topology generated without utilizing any TC algorithm (Fig. 2).

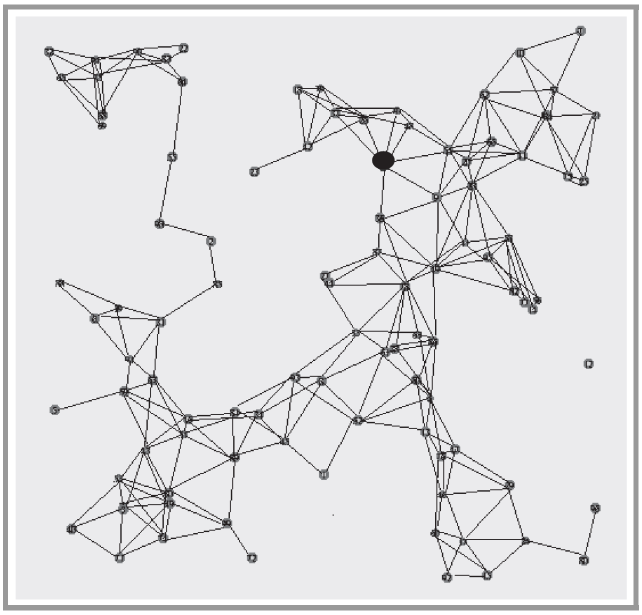


Fig. 2. Topology calculated without TC protocols.

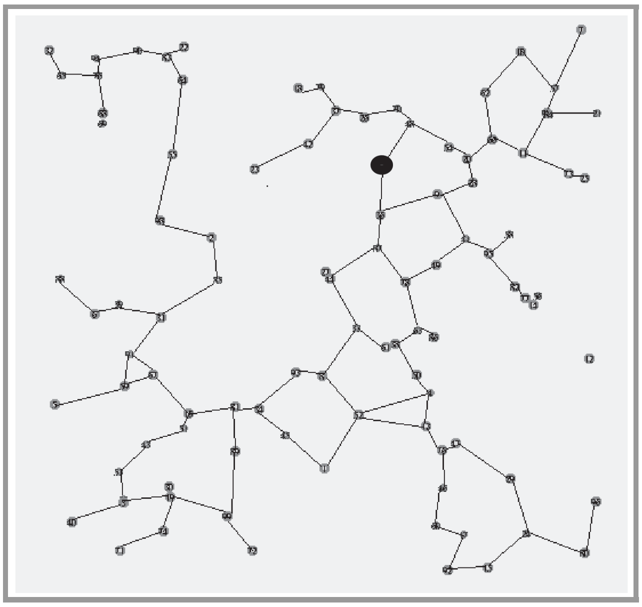


Fig. 3. Topology calculated using LMST method.

The second case study was related to simulation of data transmission in WSNs. Different sizes of networks were examined. In this experiment it was assumed that each node in WSN generates a single message that has to be delivered to the base station. In addition all nodes could play the role of relay nodes. The shortest path from each node to the destination was calculated taking into account topologies generated using R&M and two versions of LMST: LMST0 (topology can contain unidirectional links), LMST1 (topology contains only bidirectional links). The total energy consumed by all nodes for data transmission was divided by the number of nodes.

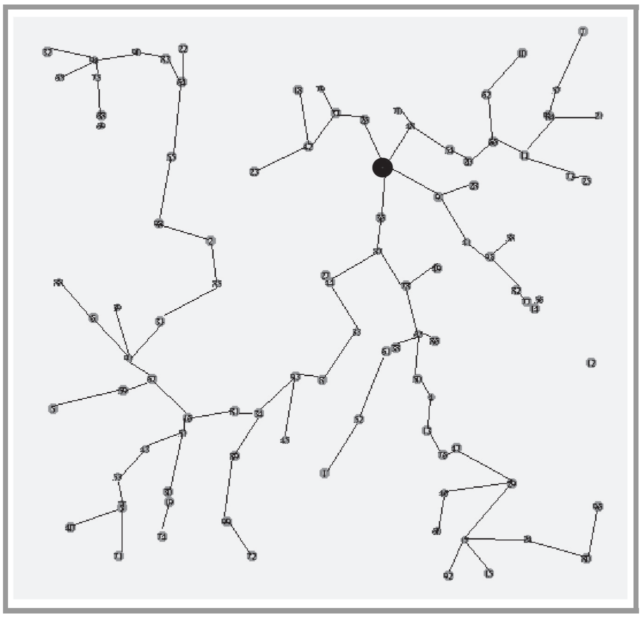


Fig. 4. Topology calculated using R&M method.

Figure 5 depicts the results of calculations, i.e., the average energy used by one node in WSN for data transmission.

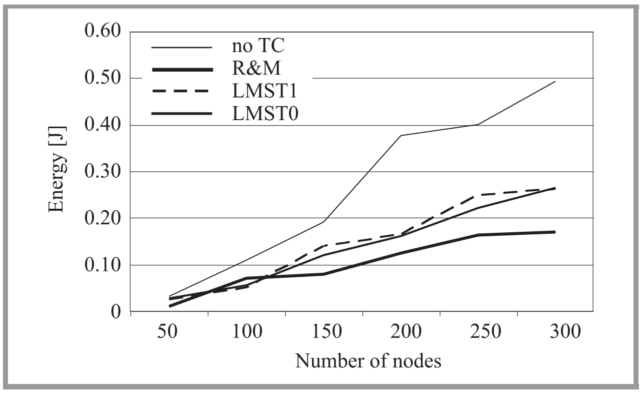


Fig. 5. Average energy consumption by one node for single transmission to the base station; different TC methods and network size.

Figures 6 and 7 show the average amount of energy used by one node for data transmission in case of different TC protocols, number of relay nodes transmitting to the base station and distance to the base station. WSN with 150 nodes was considered. It can be observed that in case of R&M and LMST protocols the energy usage for transmission in the whole network decreases while increasing the number

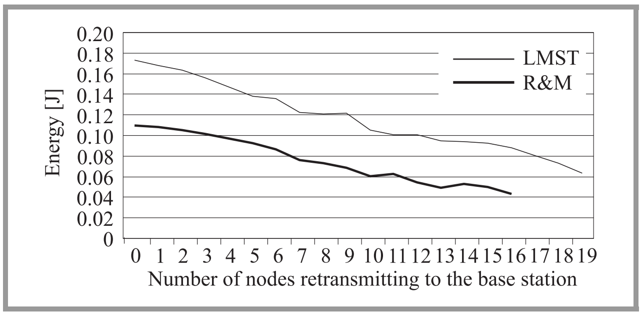


Fig. 6. Average energy usage for transmission w.r.t. the number of relay nodes; different TC methods.

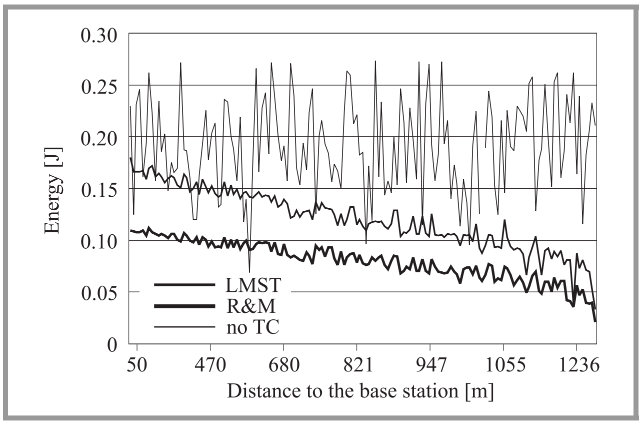


Fig. 7. Average energy usage for transmission w.r.t. the distance to the base station; different TC methods.

of relay nodes transmitting to the base station. It is obvious that the energy used for data transmission by nodes located far from the base station is smaller than those used by nodes closed to the master node, which have to retransmit a lot of messages (Fig. 7).

Table 1 contains the average number of messages generated by one node in WSN that can be transmitted to the base station up to its batteries are dead. The results obtained for different networks and topologies are compared.

Table 1
Average number of messages transmitted by one node to the base station

TC methods	Network size			
	150	200	250	300
Without TC	109 950	55 633	52 380	42 543
R&M	261 241	167 177	127 328	123 549
LMST0	173 893	130 485	94 130	78 850
LMST1	150 233	126 389	84 181	80 001

Discussion. The R&M and LMST protocols can be successfully used to calculate optimal topology in many WSN application scenarios. Both methods have to spent some energy to build the topology, which is concerned with beacon messages broadcasting in the first phase of their operation. However, the energy consumption for topology generation is small, i.e., LMST – 0.0011 J and R&M – 0.052 J for WSN of 50 nodes and energy resource of each node equal 21 kJ. Both protocols generate energy-efficient topologies (see Fig. 5). The energy consumption for data transmission in case of small size of the network (less than 150 nodes) is similar, while using topologies formed by R&M and LMST. In case of large size networks the R&M protocol seems to be much more efficient.

In summary, both techniques generate different topologies and have some advantages and drawbacks. In case of R&M we obtain more energy-efficient topologies but two potential drawbacks of the algorithm can be observed. The computation performed in the second phase of R&M requires the exchange of global information, which induces message overhead, and the explicit radio propagation model is used to compute the optimal topology. Hence, the calculated topology strongly depends on the accuracy of the channel model. Data transmission while applying the LMST protocol is more energy-intensive, but created topology is more robust and preserves connectivity in the worst case. In addition, it can be computed in a fully distributed fashion.

4. Energy Conservation

4.1. Power Consumption

The handling of the wireless transceiver contributes significantly to the node's overall energy consumption. Depending on the state of the transceiver, different levels of power consumption are being observed. Table 2 summarizes the sample power consumption of some 802.11 wireless interfaces.

In order to extend the working time of individual devices, it is frequent practice that some node elements are deactivated, including the radio transceiver. They remain inactive for most time and are activated only to transmit or receive messages from other nodes. Radio transceiver in WSN network node can operate in one out of four modes, which differ in the consumption of power necessary for proper operation: transmission – signal is transmitted to other nodes (greatest power consumption), receiving – message from other node is received (medium power consumption), stand-by (idle) – transceiver inactive, turned on and ready to change to data transmission or receiving (low power consumption), sleep – radio transceiver off.

Table 2
Aspiration and reservation levels

Interface	Power consumption [W]			
	transmit	receive	idle	sleep
Aironet PC4800	1.4–1.9	1.3–1.4	1.34	0.075
Lucent Bronze	1.3	0.97	0.84	0.066
Lucent Silver	1.3	0.90	0.74	0.048
Cabletron Romabout	1.4	1.0	0.83	0.13
Lucent WaveLAN	3.10	1.52	1.5	–

4.2. Power Save Protocols

The power-saving protocols used in sensor networks impose reduced consumption by putting the radio transceiver into the sleep mode. The use of such protocols involves the limitation of accessible band, and can also interrupt the data transfer in the network. Adequate choice of radio transceiver's switch-off time introduces further difficulty in the implementation of network protocols. The literature (e.g., [3]) present algorithms designed to limit the power consumption while simultaneously minimizing the negative impact on the network throughput and on the efficiency of data transmission routing. Different types of protocols are used depending on the application of the network. Two categories can be distinguished.

- **Synchronous power save protocols**, where it is assumed that nodes periodically wake up to exchange data packets. The sleep cycles of all nodes are globally synchronized. The main issue is to adjust length of sleep and wake phases that will minimize energy consumption and impact on a given network's throughput.
- **Topology based power save protocols**, where a subset of nodes which topologically covers whole network is selected. Nodes belonging to this set are not allowed to operate in the sleep mode. Other nodes are required to be periodically awake in order to receive incoming traffic.

Power save protocols should be capable of buffering traffic destined to the sleeping nodes and forwarding data in partial network defined by the covering set. The covering set membership needs to be rotated between all nodes in the network in order to maximize the life time of the network.

It was observed that grouping sensor nodes into clusters can reduce the overall energy usage in a network. Clustering based algorithms seems to be the most efficient routing protocols for wireless sensor network. Abbasi and Younis in the paper [12] present a taxonomy and general classification of clustering schemes. The survey of energy-efficient clustering based protocols can be found in [13]–[16].

We developed a new clustering based power save protocol that utilizes the periodical coordination mechanism to reduce the energy consumption of a network. The proposed algorithm is an extension of the popular geographic adaptive fidelity (GAF) protocol.

The GAF protocol. The GAF protocol described in [17] is a power save protocol that utilizes the information about the geographical location of the nodes. It relies heavily on the concept of *node equivalence*. The nodes A and B are *equivalent* with regard of data transmission between nodes C and D if and only if it is possible to use either node A or node B as a relay for the transmission between nodes C and D. The *node equivalence* is a feature that is not easily discovered. It is easy to notice, that nodes A and B, *equivalent* with regard of data transmission between nodes C and D do not have to be *equivalent* with regard of transmission between nodes D and E.

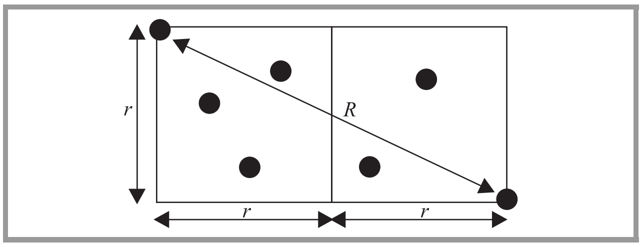


Fig. 8. Network grid construction for GAF protocol.

In order to solve this problem, the GAF protocol partitions the network using a geographic grid. The grid size r is defined such that each node in one grid square is in transmission range of all nodes within adjacent grid squares. The sample construction of such a grid is depicted in Fig. 8. With elementary geometry we have grid size of $R/\sqrt{5}$, where R denotes the maximal transmission range assigned to each node. The construction of such a grid allows the GAF protocol to preserve the original network connectivity.

The sole concept of the GAF protocol is to maintain only one node with its radio transceiver turned on per grid square. Such a node is called an *active* node and is responsible for relaying all the network traffic on behalf of its grid square. When there are more nodes in a grid square, the function of an *active* node is rotated between all the nodes in a grid square. The full graph of state transitions in the GAF algorithm is depicted in Fig. 9.

Each node starts operation in the *discovery* mode, meaning the node has its radio transceiver turned on and is pending to switch to *active* state. The node spends a fixed amount of time T_D in discovery state, when the time has passed, the node switches to the *active* state. After spending a fixed amount of time T_A in *active* state, the node switches back to the discovery state. Whenever a node changes state to *discovery* or *active*, it sends a broadcast message containing node ID, grid ID and the value of a *ranking function*. If a node in *discovery* or *active* state receives a message from a node in the same grid and a higher value of the ranking function, it is allowed to change its state to *sleep* and turn its radio transceiver off for T_S. The ranking function and timers T_D, T_A, T_S can be used to tune the algorithm. Usually the ranking function selects nodes with "longest ex-

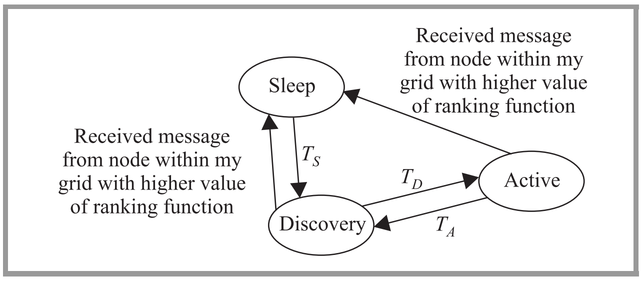

Sleep
Received message from node within my grid with higher value of ranking function
Received message from node within my grid with higher value of ranking function
T_S
T_D
Active
Discovery
T_A

Fig. 9. State transitions in GAF protocol.

pected life time" as the active nodes. The GAF protocol can be easily adapted to mobility scenarios, in such a case the ranking function utilizes information about the time, when a node will leave the grid square.

The coordination-based power save protocol (CPSP). The typical wireless sensor network consists of large quantity of sensor nodes and a base station – a dedicated node which serves as a destination for messages generated by the sensor nodes. The base station is responsible for relaying information gathered by the network to the network operator. It can be assumed that the base station has significantly more resources than the sensor nodes and is directly connected to the power grid. The wireless sensor network is utilized to deliver messages generated by the sensor nodes to the base station. From the operator's point of view there is no difference between not having any nodes in the network and the nodes not being able to deliver their messages to the base station.

We propose to utilize the dedicated network node (or nodes) as a network coordinator (or coordinators) in order to ensure that the base station is able to receive messages from the network nodes for as long period of time as possible. The base station is a natural candidate to play a role of the coordinator. Our protocol assumes that the network is partitioned by a geographical grid in the same manner as in the GAF protocol. In addition we assume that not every network grid needs to maintain an active node. The cells that do not need to establish an active node are determined by the coordinator. The grids that must maintain an active node operate similarly to the grids in the GAF protocol. In remaining grids all nodes are put to sleep state until the next topology update.

The coordinator views the network grids as a graph. The nodes periodically send to the coordinator information about amount of power available to them, which enables the coordinator to assign weights to the edges in the graph. Periodically, the coordinator calculates minimum spanning tree on the graph with itself as the root of the tree. The leaves of the tree are network grids that do not need to maintain an active node. The structure of spanning tree was chosen in order to preserve the original network connectivity. The calculated network topology is sent to all network nodes using a dedicated broadcast algorithm.

The CPSP broadcast algorithm. The CPSP broadcast algorithm relies heavily on the structure of the network and the information it is supposed to deliver to all network nodes. In order to perform the broadcast transmission, extended GAF discovery messages are utilized. Each discovery message contains the sequence number of latest transmitted network map. Since each network grid is able to receive discovery messages originating from neighboring grids, it is able to determine whether it is necessary to broadcast the latest received packet. If the grid determines that the neighboring grid has newer information, it sends a discovery message for neighboring grids to hear it. The size of broadcasted messages is kept as small as possible, information which cells should maintain an active node is sent as a bitmap – one bit represents one network grid.

Simulation results. The coordinated power save protocol was implemented in the environment of the ns-2 network simulator [9]. The proposed protocol was compared with the plain GAF protocol and a network with no power save capabilities at all. Figure 10 shows the performance of examined algorithms on a network with 60 stationary nodes distributed uniformly over a 800 x 800 meter region. Figure 11 presents the performance of the proposed broadcast algorithm against the plain GAF protocol.

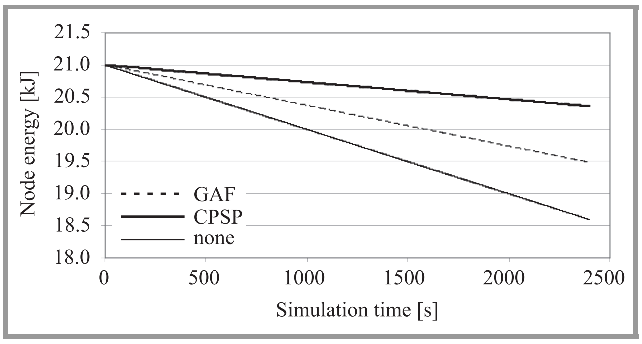


Fig. 10. Average energy consumption, various power save methods.

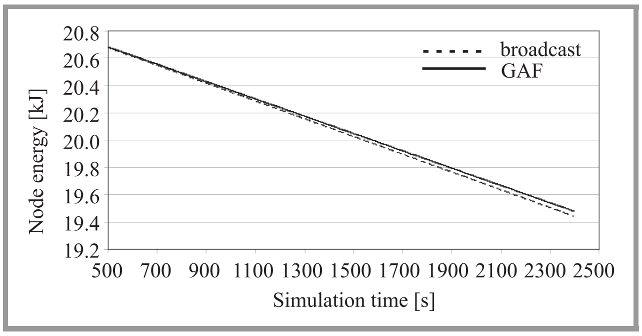


Fig. 11. Average energy consumption; CPSP broadcast and GAF comparison.

The initial energy resource of each node was assumed to be 21 kJ. Additionally it was assumed that the nodes utilize standard 802.11 radio transceiver. The traffic scheme utilized during simulation assumed random nodes sending messages to the base station at random moments of time. The messages sent to the base station were batches of 512 byte packets. The map of the network and the traffic scheme were generated using standard utilities shipped with the ns-2 network simulator.

The metric for evaluating the GAF and CPSP methods was the average amount of energy left in the node during the time of simulation. Although the main objective of CPSP algorithm is to optimize the lifetime of the network and the utilized metric does not directly show the performance of protocols in that area, it was chosen in order to be able to compare the proposed CPSP protocol with other power save solutions.

Discussion. The proposed coordinated power save protocol in its current state allows greater average energy savings than plain GAF protocol. The amount of energy saved is greater than in the GAF protocol due to larger number of sleeping nodes. The use of CPSP protocol introduces a slight overhead caused by the necessity of transmitting messages containing current statuses of nodes to the coordinator and broadcasting coordinator decisions to all nodes in the network. The proposed mechanism can be easily adapted to introduce a coordinator in a wireless sensor networks for other purposes than power saving.

5. Summary and Conclusions

The paper provides the short overview of the energy conservation techniques and algorithms for calculating energy-efficient topologies for WSNs. The efficiency of four location based approaches, i.e., two schemes for topology control and two power save algorithms are discussed based on the results of simulation experiments. The energy efficient method of introducing a coordinator to a WSN is presented. We show that our algorithm outperforms the results obtained for popular clustering based power save protocol GAF.

In general, the simulation results presented in the paper show that topology control and power save protocols effect the scheduling transmissions in a wireless sensor network, and confirm that all discussed approaches to reduce the energy consumption improve the performance of this type of network.

Acknowledgement

This work was supported by Warsaw University of Technology Research Program grant 2008.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirici, “A survay on sensor networks”, *Commun. ACM*, pp. 102–114, Aug. 2002.

[2] A. Hac, *Wireless Sensor Network Design*. New York: Wiley, 2003.

[3] P. Santi, *Topology Control in Wireless Ad Hoc and Sensor Networks*. Chichester: Wiley, 2005.

[4] P. Kwaśniewski and E. Niewiadomska-Szynkiewicz, “Optimization and control problems in wireless ad hoc networks”, in *Evolutionary Computation and Global Optimization*. Warsaw: WUT Publ. House, 2007, no. 160, pp. 175–184.

[5] J. Branch, G. Chen, and B. Szymański, “ESCORT: Energy-efficient sensor network communal routing topology using signal quality metrics”, Lecture Notes in Computer Science, vol. 3420, Berlin/Heidelberg: Springer, 2005, pp. 438–448.

[6] V. Rodoplu and T. Meng, “Minimum energy mobile wireless networks”, *IEEE J. Selec. Areas Mob. Comp.*, vol. 4, no. 3, pp. 310–317, 1999.

[7] N. Li, J. Hou, and L. Sha, “Design and analysis of an MST-based topology control algorithm”, in *Proc. IEEE Infocom’03 Conf.*, San Francisco, USA, 2003, pp. 1702–1712.

[8] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs: Pentice-Hall, 1992.

[9] Ns2. The network simulator, http://www.isi.edu/nsnam/ns/

[10] MICA2, Crossbow Technology Inc., http://www.xbow.com/Products/productdetails.aspx?sid=174

[11] MPR/MIB user’s manual, Crossbow Technology Inc., 2007, http://www.xbow.com/support/support_pdf_files/mpr-mib_series_users_manual.pdf

[12] A. A. Abbasi and M. Younis, “A survey on clustering algorithms for wireless sensor networks”, *Comput. Commun. Arch.*, vol. 30, no. 14–15, pp. 2826–2841, 2007.

[13] D. J. Dechene, A. E. Jardali, M. Luccini, and A. Sauer, “Wireless sensor networks – a survey of clustering algorithms for wireless sensor networks”, Project Report, Department of Electrical and Computer Engineering, The University of Western Ontario, Canada, Dec. 2006.

[14] N. Israr and I. Awan, “Energy efficient intra cluster head communication protocol”, in *Proc. 6th Ann. Postgrad. Symp. Converg. Telecommun. Netw. Broadcast.*, Liverpool, UK, 2006.

[15] N. Israr and I. Awan, “Coverage based inter cluster communication for load balancing in heterogeneous wireless sensor networks”, *J. Telecommun. Syst.*, vol. 38, no. 3–4, pp. 121–132, 2008.

[16] O. Younis and S. Fahmy, “HEED: a hybrid, energy-efficient, distributed clustering aproach for ad hoc sensor networks”, *EEE Trans. Mob. Comp.*, vol. 3, no. 4, pp. 366–379, 2004.

[17] Y. Xu, J. Heidemann, and D. Estrin, “Geography-informed energy conservation for ad hoc routing”, in *Proc. IEEE Ann. Int. Conf. Mob. Comp. Netw.*, Rome, Italy, 2001.

Piotr Kwaśniewski received his M.Sc. in computer science from the Warsaw University of Technology, Poland, in 2005. Currently he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2007 he works at the Polish Airports State Enterprise. His research area focuses on wireless sensor networks, mobile networks, topology control, energy efficient protocols.
e-mail: P.Kwasniewski@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Izabela Windyga received her B.Sc. in computer science from the Warsaw University of Technology, Poland, in 2008. Currently she is a M.Sc. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Her research area focuses on wireless sensor networks and topology control.
e-mail: I.Windyga@stud.elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Ewa Niewiadomska-Szynkiewicz – for biography, see this issue, p. 67.

Analysis of Burst Ratio in Concatenated Channels

Jakub Rachwalski and Zdzisław Papir

AGH University of Science and Technology, Krakow, Poland

http://doi.org/10.26636/jtit.2015.4.988

Abstract—Burst ratio is a parameter that quantifies packet loss patterns in transmission networks. It has been defined for an end-to-end scenario, therefore burst ratio can be determined only if the characteristics of the whole transmission path are known. In this paper, the burst ratio parameter applicability to cases when the transmission path consists of a series of transmission channels with known packet loss rate and burst ratio values is extended. The paper also presents the results of simulations performed with NS2 software, demonstrating the validity of the burst ratio analysis. Consequently, the research makes it possible to determine the value of the burst ratio parameter in concatenated packet networks, which in turn supports delivering higher quality VoIP services.

Keywords—bursty packet loss, E-model, quality of experience, voice over IP.

1. Introduction

Voice over Internet Protocol (VoIP) applications play a crucial role in connecting people and businesses around the world. It is a huge business for hardware manufacturers, network operators and service providers. In order to assure end customer satisfaction, the transmission networks must be designed well, and the quality of the provided VoIP service must be constantly monitored and maintained. In order to achieve this, all factors that affect the application quality of experience (QoE) [1] must be recognized.

The quality of VoIP carried over packet networks is influenced by multiple factors [2]. They include user-dependent aspects (e.g. user expectations), terminal quality (e.g. microphone sensitivity) and application settings (e.g. audio codec). The quality is also affected by transmission network-dependent factors, which include throughput, round-trip time and packet loss. To some extent, they can be controlled by network design and maintenance.

One of the transmission network-dependent factors that influences the perceived quality of VoIP transmissions is the burst ratio parameter [3]. It quantifies the packet loss pattern by describing the extent to which the packets were lost in bursts. The burstiness of packet loss affects the perceived media quality. If the number of audio packets lost sequentially is low enough not to be noticed by the human cognitive system, or it can be concealed by the packet loss concealment (PLC) technique [4], then the event has no impact on the perceived quality. In contrast, long sequences of lost packets can be easily perceived as an annoying quality deterioration. Therefore, the burstiness (burst ratio) of packet loss can be correlated with the perceived quality of VoIP service [5].

In order to provide a VoIP service of the best possible quality, the burst ratio parameter needs to be well recognized and analyzed. Thus far, it has only been defined for end-to-end transmission scenarios. In this case, in order to calculate the burst ratio of a transmission, the characteristics of the complete, end-to-end transmission path must be measured. This article describes the research into defining the end-to-end value of the burst ratio parameter, when the transmission is carried over multiple concatenated transmission channels and only the characteristics of each individual intermediate channels are determined.

Although extensive research on the influence of bursty packet loss on the QoE of VoIP has been carried out [6], [7], the authors are the first to analyze burst ratio in concatenated channels. In work [8], the results of theoretical studies are presented in which the formula for burst ratio in the concatenated scenario is derived. This article presents results of NS2 simulations [9] performed in order to validate the equations in a real environment. The results demonstrate the validity of the aforementioned theoretical considerations.

The results help control the burst ratio parameter by describing the impact of individual transmission channels on the burst ratio of the complete transmission path. The results will improve the quality and reliability of VoIP applications, thus improving end user satisfaction.

The remainder of this paper is structured as follows. In Section 2 the burst ratio parameter is presented and described in detail. In Section 3 we describe the methodology and features of the simulations that were carried out to validate the theoretical studies. Section 4 presents the results of the validation of the equation for Burst Ratio in concatenated channels. In Section 5 the verification of the simplified form of the equation is presented. Potential applications of the results are presented in Section 6. Finally, the conclusions are given in Section 7.

2. Burst Ratio Overview

This section presents the definition and application of burst ratio. It also contains results of our previous studies in the field of extending the burst ratio parameter applicability to multi-channel scenarios.

In order to describe packet loss of a communication channel, the packet loss rate Ppl is used. It indicates the probability of losing a packet during transmission over the channel. However, it is not a complete channel description as it does not capture packet loss patterns. Under the same packet loss rate, the loss can be evenly distributed over the whole transmission, or take place in bursts if multiple consecutive packets are lost.

The parameter that describes the packet loss pattern is burst ratio (denoted as $BurstR$). It is defined in [3] as the average length of observed bursts in a packet arrival sequence (average burst length) normalized over the length of burst expected for purely random loss (μ):

$$BurstR = \frac{\text{Average measured burst length}}{\mu}. \tag{1}$$

Burst ratio describes the packet loss pattern by expressing how much longer or shorter the measured bursts were than in the hypothetical case when all the packets were lost randomly under the same packet loss rate. Therefore, the burst ratio quantifies the observed packet loss as:

- bursty if $BurstR > 1$,
- random if $BurstR = 1$,
- scattered if $BurstR < 1$.

The length of packet loss burst expected for purely random loss (μ) is given as [10]:

$$\mu = \frac{1}{1 - Ppl}, \tag{2}$$

where Ppl stands for the probability of packet loss. The formula shows that even for purely random loss the observed burst length increases with higher packet loss, in the multiplicative inverse way. This is why the $BurstR$ value can differ dramatically for the same observed packet loss burst length, depending on the packet loss rate μ.

Generally speaking, for the same packet loss rate, higher values of burst ratio indicate that the packets are being lost in series. Conversely, lower values of the parameter mean that the packet loss was distributed more evenly over the transmission.

It is common to model packet loss in digital transmission channels with time-discrete state models, Markov chains [11], [12]. The approaches include two-state Markov chain, Gilbert or Gilbert-Elliot models. When examining the lossy transmission, authors are focusing on two-state Markov chain due to its simplicity and flexibility. In two-state Markov chain the successful transmission of a packet over a channel and losing a packet are marked with two different transmission channel states (Markov chain states). An example of the chain is shown in Fig. 1. In this case, if the channel successfully transmits a packet, it is in the F (found) state. If the packet is lost, the channel is in the L (lost) state. At any given time, the channel can only be in one of these two states.

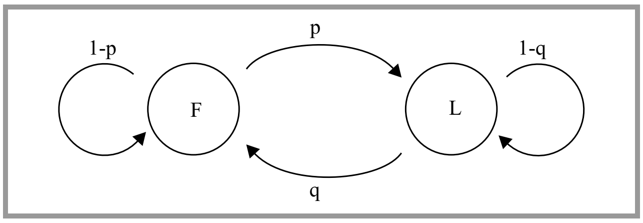


Fig. 1. In two-state Markov loss model F and L represent the found and lost states of a channel, while p and q describe the probabilities of switching the F and L states.

The two-state Markov chain is described with two parameters: p and q probabilities. The probability of losing a packet if the previous packet was successfully transmitted (transition from F to L) is described by p. Similarly, q defines the probability of successfully transmitting a packet if the previous one was lost (transition from L to F). Consequently, probability $1-p$ describes the probability of losing packets in series.

In two-state Markov chains a packet may be lost if the previous packet was successfully transmitted (with probability p) or if the previous packet was lost (with probability $1-q$). Therefore, for two-state Markov chains the probability of losing a packet is determined as:

$$Ppl = \frac{p}{p+q}. \tag{3}$$

For random loss, $q = 1-p$, the probability of losing a packet is equal to p:

$$Ppl = p. \tag{4}$$

A transmission channel modeled with the two-state Markov chain exhibits the burst ratio following the formula [13]:

$$BurstR = \frac{1}{p+q}. \tag{5}$$

Burst ratio is used in E-model [13], a commonly used analytical method of voice quality assessment. E-model uses numerous transmission parameters in order to calculate the transmission ratio factor R, which can then be used to obtain an estimated mean opinion score for the conversational scenario.

Figure 2 presents how the estimated mean opinion score value changes when the burst ratio parameter value varies between 1 and 4. The figure was created with an assumption that the G.711 codec without packet loss concealment (PLC) was used, a 1% packet loss rate was observed and other E-model parameters were used at their default values [14].

Figure 2 shows that there is a clear correlation between the application quality and the burst ratio value. Therefore, in order to calculate estimated mean opinion scores using the E-model, the burst ratio parameter must be accurately determined.

Originally, burst ratio was defined only for scenarios where the transmission is monitored and analyzed end-to-end. In [8] authors studied the burst ratio in a situation

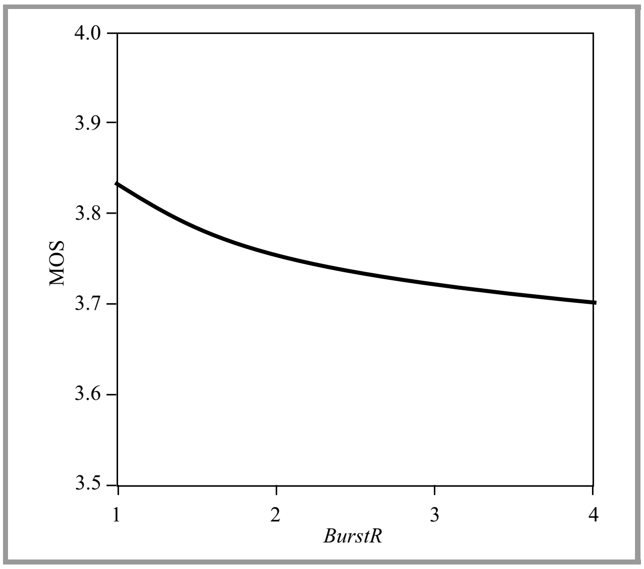


Fig. 2. Based on the E-model relationship between the estimated mean opinion score (MOS) and burst ratio parameter (*BurstR*) for 1% packet loss and the G.711 codec without PLC.

where the transmission path consists of a series of channels, and each is monitored separately. In this case, the burst ratio of the complete path must be calculated using the measured characteristics of separate channels, as presented in Fig. 3.

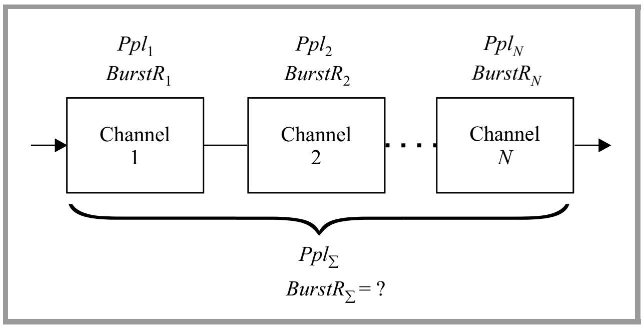


Fig. 3. The problem of burst ratio in concatenated channels network.

It was shown in [8] that if each channel can be modeled with a two-state Markov chain, the burst ratio of the complete transmission path consisting of N channels is described by the formula:

$$BurstR_\Sigma = \frac{1 - \prod\limits_{n=1}^{N}(1 - Ppl_n)}{1 - \prod\limits_{n=1}^{N}\left(1 - \frac{Ppl_n}{BurstR_n}\right)}, \tag{6}$$

where Ppl_n and $BurstR_n$ are the parameters of the n-th channel.

The exact value of the burst ratio can be determined with the regular burst ratio equation. However, for channels characterized by low packet loss the following formula can be assumed:

$$\prod_{n=1}^{N} Ppl_n = 0. \tag{7}$$

In this case the packet loss of multiple concatenated channels is as follows:

$$Ppl_\Sigma = \sum_{n=1}^{N} Ppl_n. \tag{8}$$

Based on this assumption, the burst ratio value of concatenated channels can be presented with the following, simpler equation.

$$BurstR'_\Sigma = \frac{\sum\limits_{n=1}^{N} Ppl_n}{\sum\limits_{n=1}^{N} \frac{Ppl_n}{BurstR_n}} \tag{9}$$

Analysis performed in [8] shows that this simplification is a reliable approximation of Eq. (6). The error introduced by the simplification depends on the characteristics of each channel and increases with increasing packet loss rate and burst ratio values.

As the assumption of modeling the channels with two-state Markov chains is a simplification, the authors verified the formula in a simulated network using Network Simulator 2 (NS2). The results of this verification are shown below.

3. Simulation Environment

In this section the methodology used to verify the accuracy of Eqs. (6) and (9) is described. The verification has been performed by running extensive simulations in NS2 [9].

The fundamental part of the simulation environment was designed during a seminar in Telekom Innovation Laboratories [15], which is a recognized research and develop-

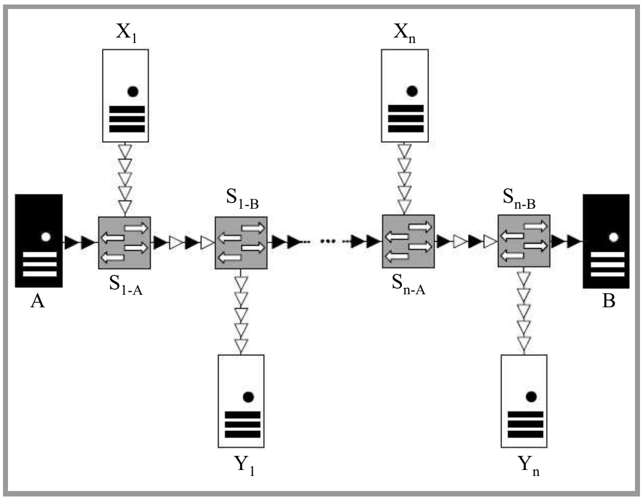


Fig. 4. The generic topology used in the simulations.

Table 1
Simulations parameters

Object	Parameter	Value	Comment
VoIP traffic	Transport protocol	UDP	
	Traffic generator	CBR	
	Packet size	50–1500 bytes	Value selected randomly (uniform distribution)
	Inter-packet interval	0.002–0.06 s	
	Start time delay	0.5–1 s	
Backgroud traffic	Number of streams transported by a single switch	1–10	Value selected randomly (uniform distribution)
	Transport protocol of a stream	TCP, UDP	
	TCP packet size	1000 bytes	
	TCP window size	2–20	
	TCP congestion control algorithm	Tahoe	
	TCP application	FTP	
	UDP traffic generator	Pareto	
	UDP Pareto shape parameter	1.4	
	UDP Pareto burst time	50–5000 ms	Value selected randomly (uniform distribution)
	UDP Pareto idle time	30000–375000 ms	
	UDP Pareto sending rate in burst	400–700 kb/s	
	UDP Pareto packet size	50–1500 bytes	
	Start time delay for each stream	0.5–1 s	
Switches	Number of intermediate switches	2–10	Each simulation repeated for every value
	Queuing scheme of each switch	DropTail, RED, FQ, SFQ	Value selected randomly (uniform distribution)
	Buffer size of each switch	2–20 packets	
Links	Capacity	500–1000 kb/s	Value selected randomly (uniform distribution)
	Propagation delay	0–200 ms	
Simulation	Duration	10, 100, 1000 s	Each simulation repeated for every value

ment institute in the field of quality of audio and multimedia applications.

NS2 is a commonly used [16] simulation environment for testing and studying communication protocols and networks. It can be used to simulate TCP/IP protocol stacks, traffic sources of various distributions and packet queuing and dropping mechanisms.

The release NS2 2.35 was used in this research in order to simulate packet transmission over a series of switches and to analyze packet loss. Each switch serves a number of packet streams and drops packets in case of a buffer overflow. After each simulation the burst ratio calculated at the end of the transmission path using Eq. (1) is compared with the burst ratio value calculated from the transmission parameters of each intermediate switch using Eq. (6). The calculations are performed by analyzing the NAM trace files generated by each NS2 simulation.

The topology used in the simulations is a path presented in Fig. 4. It contains two endpoints (A and B) responsible for a VoIP transmission, n pairs of background traffic servers $(X_1, Y_1, \dots, X_n, Y_n)$ and n pairs of switches $(S_{1-A}, S_{1-B}, \dots, S_{n-A}, S_{n-B})$. VoIP traffic, marked with black arrows, is sent from server A to server B. n background traffic streams, marked with white arrows, are sent between servers X_1 and $Y_1, \dots, X_n$ and Y_n. VoIP traffic and background traffic compete for resources of shared links, which are built up by pairs of switches $S_{1-A} \longleftrightarrow S_{1-B}, \dots, S_{n-A} \longleftrightarrow S_{n-B}$. Consequently, at switches $S_{1-A}, \dots, S_{n-A}$ the VoIP packets and the background transmission compete for access to the shared links. If not enough bandwidth is available to serve both streams, the switches drop packets. Therefore, in the simulation the transmission path of the VoIP application consists of a series of links. However, packets may be dropped at shared links only. Other links do not drops packets because they always have enough bandwidth due to transmitting either VoIP or background traffic only. At the end of the simulation, the packet loss analysis of each switch which drops packets is performed.

During the analysis the VoIP application packet loss rate and the burst ratio value are calculated. Using these values and Eq. (6), the burst ratio of the whole transmission (from node A to B) is calculated. The calculated value is compared with the value calculated at node B based on the analysis of VoIP stream packets that were not successfully delivered, Eq. (1). The result of the comparison quantifies the accuracy of Eq. (6).

It should be noted that in the simulations the packet loss takes place in shared links only. Therefore, in the remaining sections the terms "channel" and "shared link" are used interchangeably.

The results of the simulations may depend on the topology as well as transmission and network parameters. The complete list of parameters identified and analyzed during the simulations is presented in Table 1. The parameters were randomly altered within a range of values during each simulation in order to reduce the influence of a specific parameter value on the results. The parameter values and ranges of values were adjusted so the results of the simulations were relevant for the study of burst ratio parameter.

In order to obtain meaningful results it was important that the VoIP traffic was constantly generating packets. Therefore VoIP traffic utilized the user datagram protocol (UDP) with a constant bit rate. Additionally, the randomization of the background traffic was of crucial importance in order to assure a full spectrum of simulation conditions. Therefore, the background traffic used UDP (with the Pareto distribution) and TCP protocols, both selected randomly for each simulation. Moreover, the start time and the total number of transmitted packets within each transmission were also randomized. As a result the VoIP traffic faced different conditions in each simulation run. The wide spectrum of conditions meant the VoIP traffic was characterized by a wide range of parameters values *BurstR* and packet loss rate *Ppl*.

This paper presents the results of a total 250,000 simulations, each representing different network conditions. They were carried out in order to demonstrate the validity of the equations. As a result, the validation contains relevant and fully conclusive results.

4. Accuracy of Burst Ratio Calculation

In this section the simulation results run in order to validate Eq. (6) are presented. The equation was numerically verified by the authors in [8], where a transmission channel was modeled by a two-state Markov channel. This section contains simulations results, where the transmission environment was modeled with real networks characteristics, simulated using NS2.

The verification has been performed by comparing two burst ratio values:

- the *BurstR* value measured at the end of the transmission path using Eq. (1),
- the value calculated using Eq. (6), which incorporates the characteristics of each intermediate transmission channel, denoted below as $BurstR_\Sigma$.

The comparison is presented as relative error δ_Σ, defined as follows:

$$\delta_\Sigma = \frac{BurstR_\Sigma - BurstR}{BurstR}. \tag{10}$$

If δ_Σ is equal to 0, Eq. (6) is perfectly accurate. A positive value of δ_Σ means that the experienced packet loss is less bursty than that estimated using Eq. (6). A negative value of δ_Σ means that the burst ratio value calculated with Eq. (6) underestimated the burstiness of the analyzed traffic.

The number of shared links may have an impact on the final results, because the VoIP traffic needs to compete for resources in each link. The more shared links, the more VoIP packets may be lost. In order to study this impact, each simulation was rerun with two, six and ten shared links.

The results published in this section present the relationship between relative error δ_Σ (in %) and packet loss *Ppl*, number of transmitted packets or *BurstR* of the complete transmission. The error is analyzed in the form of a mean and its confidence intervals. The mean value of the relative error is shown using black lines. The 95% confidence intervals of the mean are marked with gray areas.

Figure 5 presents the relationship between the relative error δ_Σ of the burst ratio calculation using Eq. (6) and packet

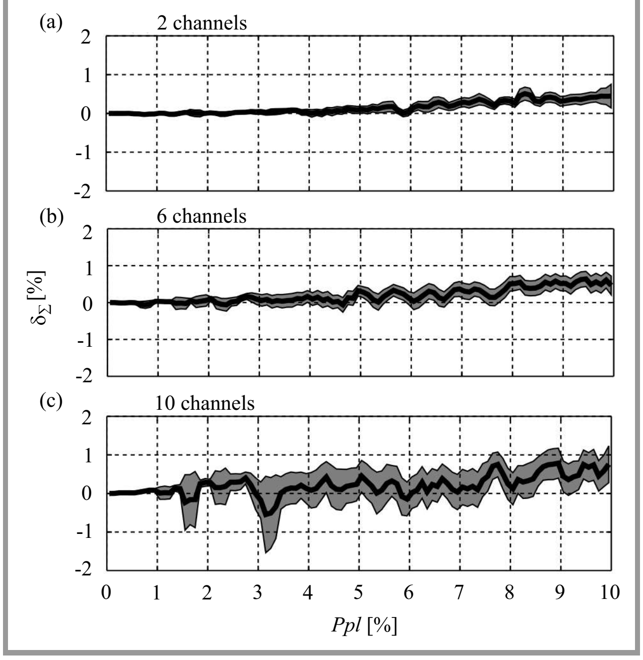


Fig. 5. Relationship between the burst ratio calculation error δ_Σ and the packet loss rate *Ppl* of the whole transmission. The solid line represents mean relative error while the gray areas present the 95% confidence intervals of the mean. The figures were created with a packet loss range of 0–10%. The subplots presents results for simulations of two, six and ten intermediate channels.

loss Ppl of the whole transmission. It can be observed that for values of packet loss lower than 1%, the relative error is negligible, regardless of how many intermediate channels the transmission contains. As the packet loss increases, the mean error and its confidence interval increase slightly as well. The observed increase is dependent on the number of intermediate channels. The higher the number of channels, the higher the error for the same value of packet loss. However, the relative error never reaches 2%, which indicates a high accuracy of the equation.

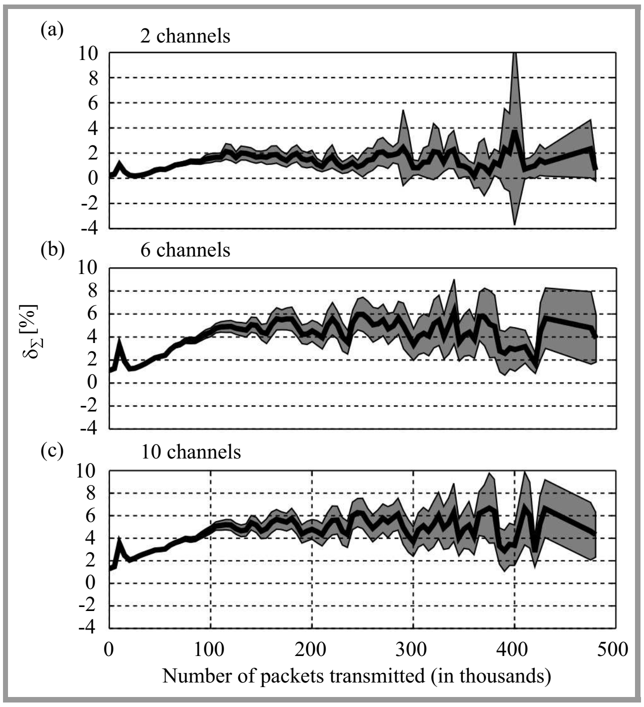


Fig. 6. Relationship between the burst ratio calculation error δ_Σ and the number of transmitted packets. The solid lines represent mean relative error while the gray areas present the 95% confidence intervals of the mean. The subplots presents results for simulations of two, six and ten intermediate channels.

Figure 6 presents the relationship between the burst ratio calculation error δ_Σ and the number of transmitted packets during measurement. The figure shows that the mean error initially slightly increases for the shorter observations and then stabilizes at a level of 2% for two intermediate channels or 5% for ten channels. Figure 6 presents results for up to $500,000$ transmitted packets, which corresponds to approximately 2 hours 45 minutes observation of a transmission. Such a long observation is unrealistic and its results are presented only for reference. More reasonable duration of observation is up to 5 minutes, which corresponds to 0–15,000 of transmitted packets. In this range the error never exceeds 4%, regardless of the number of intermediate channels.

Figure 7 presents the relationship of the relative error δ_Σ of the burst ratio calculation using Eq. (6) and burst ratio value *BurstR* of the complete transmission. It can be seen that regardless how many intermediate channels are used the relative error is low around $BurstR = 1$. For two channels, the error value is negligible, regardless of the burst ratio value. In the case of several intermediate channels, as the burst ratio increases, the error decreases and for $BurstR > 1.5$ the error becomes negative. In the worst case, for the scenario of ten intermediate channels the error reaches -9%. It can be seen that for fewer channels, *BurstR* of the complete path reaches higher values. For ten intermediate channels the highest value of *BurstR* slightly exceeds 2.5, while for two channels it is over 3.5. This effect can be explained by analyzing Eq. (9). The formula shows that *BurstR* value of the complete path is approximately equal to the weighted harmonic mean of all intermediate channels' *BurstR* values. As the result, the more channels are involved in the transmission, the lower probability that end-to-end burst ratio reaches high values.

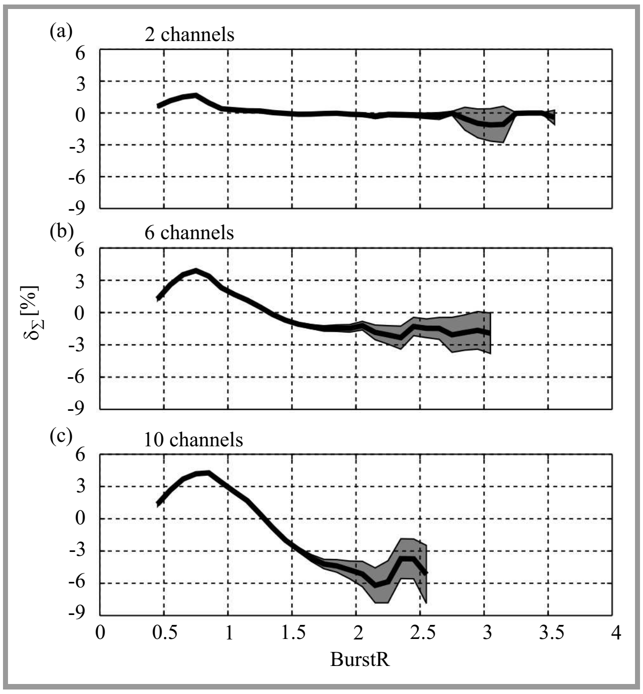


Fig. 7. Dependency of the burst ratio calculation error δ_Σ on the *BurstR* value of the complete transmission. The solid lines represent mean relative error while the gray areas present the 95% confidence intervals of the mean. The subplots presents results for simulations of two, six and ten intermediate channels.

All these results show that when Eq. (6) is used it provides reliable results and a high precision of the measurement. The accuracy of the calculation is always very high, but the most precise results are achieved in the two-channel scenario, when packet loss of the complete transmission path is limited or the burst ratio of the complete transmission path is not higher than $BurstR = 1.5$.

5. Accuracy of the Simplified Equation

As well as the regular burst ratio equation, validated above, we also show a simplified version of the equation, Eq. (9).

This simplification reveals that the burst ratio of the whole transmission path can be approximated with a weighted harmonic mean of properties of individual channels. This equation was verified numerically in [8]. The results indicate that the simplified equation's inaccuracy increases with higher values of packet loss and burst ratio of the whole transmission. However, the verification was performed with the assumption that the transmission channels can be modeled with two-state Markov chains, which is a form of simplification. This section presents the results of equation validation performed in an environment that simulates real characteristics of transmission channels.

The verification of the simplified burst ratio equation – Eq. (9) is performed by calculating the simplification error Δ_{BurstR}. It expresses the difference between the error of the simplified equation and the error of the regular burst ratio equation – Eq. (6). The values that are compared are mean relative error (in %) and the 95% confidence interval of the mean. Both were introduced in Section 4. The comparison of mean error is performed by calculating the difference between absolute values of mean error δ'_{Σ} of the simplified burst ratio equation (Eq. 9) and mean error of the regular burst ratio equation δ_{Σ}, as described in Eq. 10. The comparison is presented below:

$$\Delta_{BurstR} = \left|\delta'_{\Sigma}\right| - \left|\delta_{\Sigma}\right| . \tag{11}$$

If the calculated difference of the mean error is equal to 0, both Eq. (6) and Eq. (9) are equally accurate. When Δ_{BurstR} is positive, Eq. (6) is more accurate, while if Δ_{BurstR} is negative, the simplified equation is more accurate.

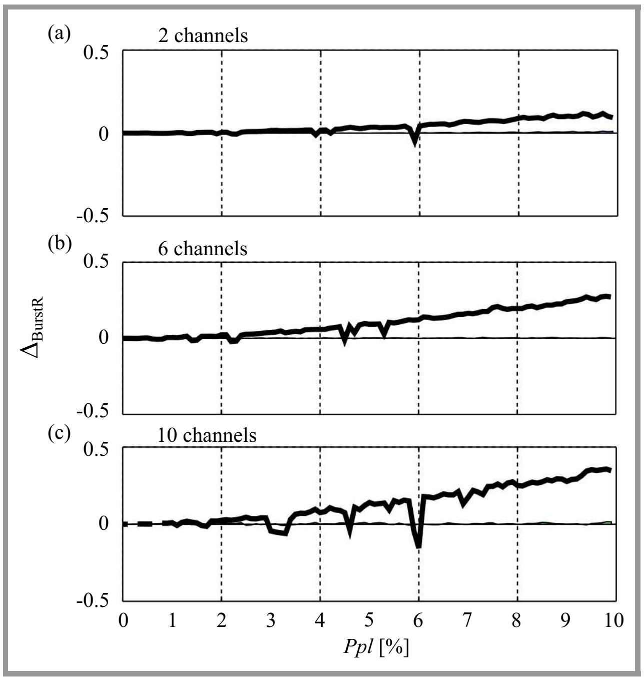


Fig. 8. Difference Δ_{BurstR} between the calculation errors of the regular equation δ_{Σ} and the simplified equation δ'_{Σ} in the domain of packet loss Ppl of the whole transmission. The solid lines represent the difference between mean relative errors while the gray areas represent the difference between the 95% confidence intervals. The subplots presents results for simulations of two, six and ten intermediate channels.

The comparison of the 95% confidence intervals of the mean is performed in a similar way, by subtracting the value of the confidence interval for the regular equation from the value of the confidence interval for the simplified equation.

The figures published in this section present the calculated differences of mean error using black lines. The gray areas in the figures correspond to the confidence interval differences of the means.

Figure 8 presents the differences of mean errors and confidence intervals in the domain of packet loss in the range of 0–10%. It can be seen that regardless how many intermediate channels are used, the difference is negligible in that it never exceeds 0.5%. However, it should be noted that there is almost no difference in the confidence interval width (marked with gray fields).

The results clearly show the validity of the simplified equation. The difference in performance, compared with the regular equation, is almost indistinguishable. However, the regular equation almost always performs slightly better than the simplified formula. Therefore, when the highest accuracy of the measurements is required, the regular equation is used. However, when the top priority is ease of calculation, the simplified equation is applied.

6. Applications

As mentioned above, burst ratio is one of the parameters used in the ITU-T E-model, which is used to assess the quality of VoIP. Therefore, the formula presented has a wide spectrum of potential applications, mainly facilitating the VoIP MOS level assessment.

The formulas can be used during network planning. When a network is being designed, a set of technical requirements is specified for the network. They include packet loss, round trip time and mean opinion score (MOS) of VoIP transmission. When network topology is defined, the characteristics of all the network elements are assumed. Even if the topology is complex and the network contains hundreds of elements, the VoIP transmission MOS assessment between any endpoints may be required. Without proper calculation of the burst ratio value between the endpoints, a precise assessment of application quality is not possible. Using the formulas presented and the E-model, MOS can be easily and precisely assessed between any endpoints of the designed network. Therefore, during the network design phase, corrections may be applied to the network topology to help provide the best quality of the VoIP service.

Another application of the formula is when a network is already operating and a re-design of the topology or routing is required. In this case the formula may help assess the impact of the changes on the quality of the VoIP transmission. A good example would be a network that contains multiple elements which introduce packet loss. If only one of them could be upgraded, it would be important to select the optimal element to upgrade. By using the formula, the network administrator can easily assess how end-to-end VoIP quality would be affected, depending on which elements are upgraded.

The formulas can also be successfully used during monitoring of networks. The measurements, as described in [17], need specially configured environments. Therefore they can only be performed within a single network, owned by a single company. If a VoIP transmission path is established via several different networks, which are administered by different companies, the complete path monitoring is not possible. In this case, the formulas can be used in order to calculate the VoIP transmission MOS using monitoring logs of the individual networks.

7. Conclusions

The results clearly show that the equations presented can be successfully used to calculate the burst ratio parameter, when the complete transmission path consists of multiple concatenated channels. Although the equation has been derived theoretically using two-state Markov models, in real-life scenarios, simulated here using NS2, the equation is still valid. Its accuracy is the highest when the number of concatenated channels is limited to two, when the packet loss of the complete transmission path is low, or the burst ratio of the complete transmission path is not higher than $BurstR = 1.5$.

Moreover, the results show that the simplified version of the equation is almost as accurate as the regular equation, therefore it can be used as an engineering tool. The simplified formula reveals that the burst ratio value of the complete transmission path can be regarded as a harmonic mean of the individual channels burst ratio values, weighted with the their packet loss probabilities.

The results also demonstrate that the equation is valid and therefore can be used in QoE measurements and network performance assessment. Moreover, the formula has a wide spectrum of potential application. As such, it would be useful in improving the quality of VoIP applications.

Acknowledgments

This work has been supported by the AGH University of Science and Technology under contract no. 11.11.230.018. The authors would like to thank Prof. Sebastian Moeller, Prof. Alexander Raake, Dr. Błażej Lewcio and Dennis Guse for consultations and assistance in designing the validation methodology presented in this paper.

References

[1] "Vocabulary for Performance and Quality of Service", ITU-T Rec. P.10/G.100 (incl. Amendment 2), Tech. Rep., 2008.

[2] A. Raake, *Speech Quality of VoIP: Assessment and Prediction*. Wiley, 2006.

[3] J. W. McGowan, "Burst ratio: a measure of bursty loss on packet-based networks", 16 2005, US Patent 6,931,017.

[4] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio", *IEEE Network*, vol. 12, no. 5, pp. 40–48, 1998.

[5] A. Raake, "Short-and long-term packet loss behavior: towards speech quality prediction for arbitrary loss distributions", *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 14, no. 6, pp. 1957–1968, 2006.

[6] A. D. Clark, P. D. F. Iee, and M. Ieee, "Modeling the effects of burst packet loss and recency on subjective voice quality", in *Proc. IP Telephony Worksh.*, New York, 2001.

[7] W. Jiang and H. Schulzrinne, "Perceived quality of packet audio under bursty losses", in *Proc. 21st Ann. Joint Conf. IEEE Comp. Commun. Soc. InfoCom 2002*, New York, USA, 2002.

[8] J. Rachwalski and Z. Papir, "Burst ratio in concatenated markovbased channels", *J. Telecommun. Inform. Technol.*, no. 1, pp. 84–90, 2014.

[9] "The Network Simulator NS-2" [Online]. Available: http://www.isi.edu/nsnam/ns/

[10] S. M. Ross, *Introduction to Probability Models*, 11th ed. Academic Press, 2014.

[11] J.-C. Bolot, "Characterizing end-to-end packet delay and loss in the interne", *J. of High Speed Netw.*, vol. 2, no. 3, pp. 305–323, 1993.

[12] H. A. Sanneck and G. Carle, "Framework model for packet loss metrics based on loss runlengths", *Proc. of SPIE*, vol. 3969, pp. 177–187, 1999.

[13] "The E-Model, a computational model for use in transmission planning", "ITU-T Rec. G.107, Tech. Rep., 2014.

[14] "Transmission impairments due to speech processing", ITU-T Rec. G.113, Tech. Rep., 2007.

[15] J. Rachwalski, "Bursty loss modelling in E-model", Semminar in Telekom Innovation Laboratories, Berlin, 25 Apr. 2013.

[16] S. Kurkowski, T. Camp, and M. Colagrosso, "MANET simulation studies: the incredibles", *ACM SIGMOBILE Mob. Comput. and Commun. Rev.*, vol. 9, no. 4, pp. 50–61, 2005.

[17] "Cisco IOS IP SLAs Configuration Guide", Cisco Systems, Inc., Tech. Rep., 2008.

Jakub Rachwalski obtained his M.Sc. from the University of Science and Technology in Krakow, Poland in 2009. He is currently a Ph.D. student under the supervision of Prof. Zdzisław Papir. His current research focuses on the Quality of Experience in VoIP applications, with special interest in the influence of the packet loss distribution on the perceived quality.

jrachwal@agh.edu.pl
Department of Telecommunications
Faculty of Computer Science, Electronics
and Telecommunications
AGH University of Science and Technology
Mickiewicza av. 30
30-059 Krakow, Poland

Zdzisław Papir is Professor and a deputy chair at Department of Telecommunications, AGH University of Science and Technology in Krakow, Poland. During 1994–1995 he was serving for the Polish Cable Television as a Network Design Department Manager. Between 1999–2006 he was a guest co-editor for IEEE Communications Magazine responsible for the Broadband Access Series. He has been participating in several R&D IST European projects being responsible for Network performance evaluation and quality assessment of communication services. He has also been appointed as an ICT expert by the European Commission. His current research interests include modeling of telecommunication networks/services and measuring quality of experience.

E-mail: papir@kt.agh.edu.pl
Department of Telecommunications
Faculty of Computer Science, Electronics
and Telecommunications
AGH University of Science and Technology
Mickiewicza av. 30
30-059 Krakow, Poland

FDTD modeling and experimental verification of electromagnetic power dissipated in domestic microwave ovens

Paweł Kopyt and Małgorzata Celuch-Marcysiak

http://doi.org/10.26636/jtit.2003.1.151

Abstract — The FDTD (Finite Difference Time Domain) method has proven to be effective in modeling high-frequency electromagnetic problems in telecommunications industry. Recently it has been successfully applied in microwave power engineering. In order to accurately model scenarios typical in this field one has to deal with the movement of objects placed inside cavities. This paper describes a simple algorithm that makes it possible to take into account object rotation – important in simulations of domestic microwave ovens. Results of example simulations are presented and an experimental verification of the simulation tool is performed.

Keywords — electromagnetic simulations, FDTD algorithm, microwave heating.

1. Introduction

Microwave power engineering community has only recently discovered the FDTD and other modeling methods successfully used in telecommunications for years. Like in any other field, also in this application modeling is very beneficial as it cuts down the design costs. In the literature one can find various approaches to model heating devices. The general possibilities and limitations were described in [1]. In [4] the authors used FDTD to model microwave power distribution in a food object placed inside a microwave oven. Simulations and optimisation results of power uniformity in objects passing through a tunnel industrial oven have been presented in [2].

One important feature has been so far overlooked. In real-life heating devices the greater uniformity of power distribution inside the heated objects is often achieved with the object movement inside the cavity while heating is on. In industrial ovens the foodstuffs are simply passing through a cavity on a conveyor belt [2], while in small-scale simpler devices – like domestic microwave oven – the object is placed on a rotating shelf.

The accurate modeling of microwave heating scenarios requires that movement of the objects be taken into account. This paper introduces a simple but effective method to conduct an accurate modeling of an object heated inside a domestic microwave oven with the object rotation. The algorithm itself is presented in Section 2. Section 3 describes various power uniformity criteria. One of those presented is chosen for further use. Section 4 presents the results of simulations and comparison of the introduced method with results obtained without its application. Section 5 discusses comparisons of the FDTD simulations with real-life measurements.

2. Method of simulation

The basic electromagnetic simulation software has been developed for years and currently on the market one can find ready packages of proven reliability and accuracy, successfully used in various problems. Especially, the software based on FDTD method has proven advantageous over other solutions employing FEM method [2]. One of those – QuickWave 3D developed by QWED – with implemented conformal FDTD method [5, 6] is known in microwave power industry. Several papers presenting results obtained with its help have been published and it ranks high on the list of available software suited for microwave power simulation [7].

An additional feature makes it even more suitable to the task – the possibility to define scenario geometry using parametrized macros in the so-called UDO language. This feature together with the possibility to define scripts (tasker files) instructing QuickWave to dump chosen data at a given moment facilitates effective simulation and optimisation of complex problems. It has been successfully used in [3].

That way of conducting simulation has been adopted in our case. The experiments have been conducted with a model of an example microwave oven. The whole geometry of the model, together with excitation and the sample object being heated, have been prepared with a single macro. By modifying its parameters one can place the sample object at any angle inside the oven. Next, using an external routine written in Matlab [8], one can trigger QuickWave simulation several times, each time for a different object position. Calculated dissipated power pattern in the object cross-section will be different for each angle. It means that by collecting the data on dissipated power for each angle and summing them up one can arrive at more accurate information on how evenly the dissipated power is distributed inside the object during heating and rotation.

Table 1
Relative error of the averaging (for sets 7 and 8 the mirroring was used)

Set number	1	2	3	4	5	6	7	8
Step [°]	5	10	20	30	45	60	*)	**)
No. angles	72	36	18	12	8	6	6	7
Effective no. angles	72	36	18	12	8	6	10	12
Error [%]	0	0.0916	0.3043	1.1609	4.0086	5.1448	1.9044	1.1267

*) Angles in set no. 7: 0°, 36°, 72°, 180°, 216°, 252°.

**) Angles in set no. 8: 0°, 30°, 60°, 90°, 180°, 210°, 240°.

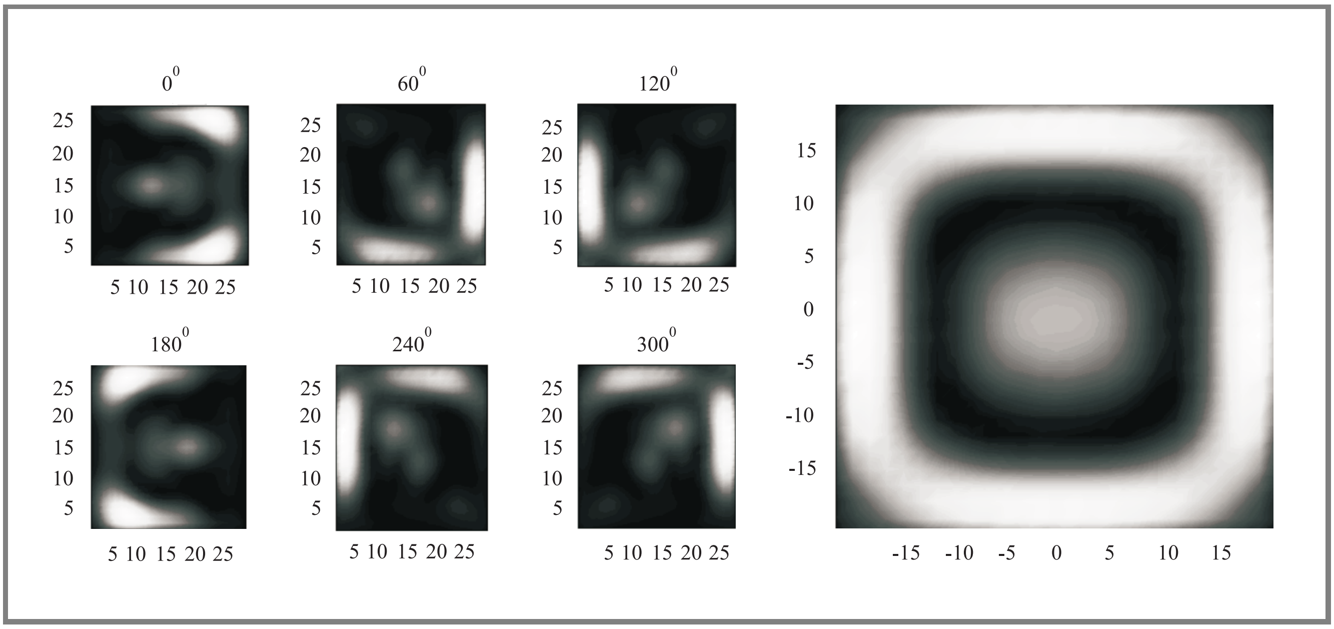


Fig. 1. Dissipated power pattern for six angular positions of the object (from 0 to 300° with the step of 60°).

The FDTD method requires that the model be discretized with small parallelepipeds (or cells). Fields and resulting power distribution are calculated within each such cell. The collected information comes in the form of snapshots of dissipated power in the object's cross-sections. If the object is rotated, the snapshot is rotated as well. Obviously direct summing of the data for different rotation angles is not possible. This task requires that the snapshots of dissipated power be first brought back to their original position – 0 degrees. Since in each case the angle of rotation is known one can achieve this using standard formula for coordinate system rotation applied to each cell:

$$\begin{aligned} x' &= x\cos\alpha + y\sin\alpha\,, \\ y' &= -\sin\alpha + y\cos\alpha\,, \end{aligned} \tag{1}$$

where x, y are coordinates of the cell center of the rotated snapshot, while x' and y' are coordinates of the cell center after the back-rotation.

After the back-rotation it is possible to average the dissipated power for all the angular positions of the object. Figure 1 presents six matrices that have been rotated back and are ready for averaging. During all the experiments the object (a parallelepiped 40 × 40 × 20 mm) has been placed centrally on the shelf and rotated around its center. The object is made of meat modeled as material of relative electric permitivity $\varepsilon = 50 - j20$.

The accuracy of the averaging procedure for different number of distinct positions has been tested. As a reference we have used averaged pattern P^5 obtained by summing up power snapshots taken every 5°. Table 1 contains results of comparison between the reference pattern and patterns based on a smaller number of positions. The relative error values have been obtained with the following formula:

$$e(a) = \frac{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{M}\left(P_{ij}^{5} - P_{ij}^{a}\right)^2}{NM}\,, \tag{2}$$

where P^a is a pattern whose accuracy we verify against the reference pattern P^5, N and M are the dimensions of the patterns (given in cells).

The error grows as the number of angular positions decreases. There is a way to lower the number of angles without deteriorating the accuracy. We have presented 6 templates in Fig. 1. It is clear that due to the symmetry of the problem one can find pairs of snapshots that are mirrored copies of each other (60° and 300° or 120° and 240°). It is enough to get one result and sum it up twice with and without mirroring. Sets 7 and 8 of data in Table 1 contain angles only from the first and third quarters of the coordinate system, with subsequent mirroring. It is clear that error is low despite the relatively small number of angles, for which the simulations were performed.

3. Power uniformity criteria

Effective comparison of the uniformity of power distribution within different objects requires a carefully chosen criterion. A criterion like that is also necessary in order to conduct an optimization process. We have browsed the available literature and chosen two criteria for further consideration. We have also proposed a new one which seems to be most useful.

- **Differential power difference** – difference between the highest and the lowest value of dissipated power found inside the object (or its cross-section in the case of analyzing the object as a set of two-dimensional layers) [2].
- **Statistical criterion** – standard deviation of the dissipated power distribution normalized by the mean value [4].
- **Integral power criterion** – maximum value of power dissipated in a small volume (surface when we deal with single layers of object) normalized by the total power dissipated in the object (one layer).

The differential power criterion constructed as a simple difference between power dissipated in hottest and coldest spot and used in [2] is not suitable in our case. It takes into account only two points out of many available in object's volume which does not fully describe power distribution in complex two- or three-dimensional geometries.

The statistical criterion seems to be a better choice as all the points (or cells) of the object give contribution to the total value of the criterion. We have conducted some tests of the criterion trying to confirm that it gives results, which agree with intuition. The tests have shown that the criterion cannot be used in our case despite the fact that it has been successfully applied in [3, 4].

The third proposed criterion – the integral power criterion – has been eventually chosen for further experiments. It has been confirmed in tests that it can be treated as accurate description of the power distribution. What is also important is that this criterion is completely based on physical concepts and phenomena (e.g. power dissipated in a specified region) that we want to measure. The criterion is defined with the following formulae:

$$f_p(v) = \max\left(\frac{\int\limits_{S_A(x_0,y_0)} SAR(x,y)\,ds}{\int\limits_{S} SAR(x,y)\,ds}\right), \qquad (3)$$

$$f_p(v) = \max\left(\frac{\int\limits_{S_A(x_0,y_0)} -SAR(x,y)+2E\big(SAR(x,y)\big)\,ds}{\int\limits_{S} SAR(x,y)\,ds}\right), \qquad (4)$$

where S is the area of the object's cross-section (if whole volume of the object is taken into account then the integral is calculated over the volume), S_A is the small fraction of the whole area of the object's cross-section, SAR is function describing the distribution of the specific absorption rate (in reality it is discretized due to the nature of the applied simulation method) and E is the symbol of calculating expected value.

It also has one important advantage – it allows taking into consideration not only the value of power dissipated in cold (or hot) spots but also the shape of the power distribution. A sharp peak in power over a small region is not so important as a peak similar in value but spread over a larger region. The tuning of the criterion means changing the S_A area. If it is small in comparison to S then the criterion is tuned to sharp peaks. By making the S_A larger we average the value of power peaks over greater area thus making the influence of such peaks smaller.

There is also another advantage of the integral power criterion. A typical goal of the optimization process is a uniform power distribution within the heated object. Yet from a practical viewpoint, this is usually a secondary goal as it is more important to assure that in the volume of the heated object none of the regions will be heated too much (hot spots elimination) or all the regions' temperature will be high enough (cold spots elimination) to kill pathogens. One can easily modify the criterion for each case. The (3) criterion is to be used in hot spot elimination while (4) should be used to eliminate cold spots. The modification in (4) is simply turning the power distribution around its mean value. Thus all the peaks become valleys and all the valleys (or cold spots) are peaks which the optimization algorithm will try to eliminate.

4. Results of simulation

Using the algorithm described in Section 2 together with the integral power criterion presented in Section 3 two numerical experiments have been conducted. The results show that taking into account the object rotation can change the resulting dissipated power distribution in heated objects, and consequently the optimum object design.

The experiments have been performed with a modified domestic microwave oven model. On the shelf a sample object has been placed whose shape can be modified with a single parameter. In order to maintain simplicity of the experiments and keep the computation time reasonably short the object's shape can be changed by rounding its corners. The parameter is defined with the following formula:

$$n = \frac{r_c}{0.5a}, \tag{5}$$

where r_c is the radius of the curvature of the object's corners while a is the length of the object's side (the assumption is that the object is equilateral). One can easily introduce other parameters (e.g. ratio of the object's sides). Two example objects of different shapes are presented in Fig. 2.

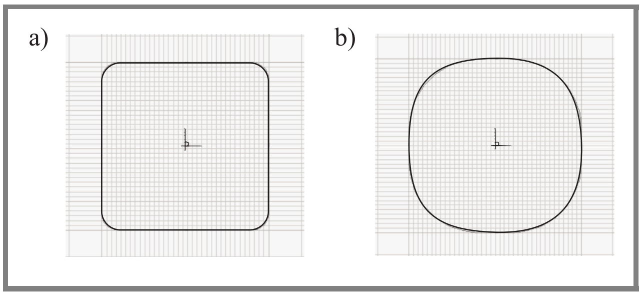


Fig. 2. Two example objects: (a) $n = 0.2$; (b) $n = 0.8$.

The sample object used in the experiments has been made of meat (material density is 1 g/cm^3), its volume has been 300 cm^3, its height – 20 mm. The experiment goal has been to average dissipated power over 6 angles (angle set 7 listed in Table 1) in the plane cutting the sample object at the level of 2, 10 and 18 mm from its base. Then the integral power criterion has been used to calculate the uniformity of the distribution in each layer. The calculations have been repeated for different shapes of the object (different values of n parameter in Eq. (5)). The results are presented in Fig. 3 together with similar data obtained by calculating the uniformity of power distribution for one angle only (0°).

Clearly the object rotation contributes to higher heating uniformity. It is also important that object rotation makes the power distribution less sensitive to the shape of the heated object. From Fig. 3 one can see that when the object rotation has been taken into account, the shape factor influences the uniformity to a much lesser extent as compared to the case in which the rotation has not been employed. Without rotation, we observe one local minimum at $n = 0.8$ and a global minimum at $n = 0.1$. With rotation, the local minimum disappears, and the global one shifts to $n = 0.2$. This means that the optimum design of a food package would be different in the two cases.

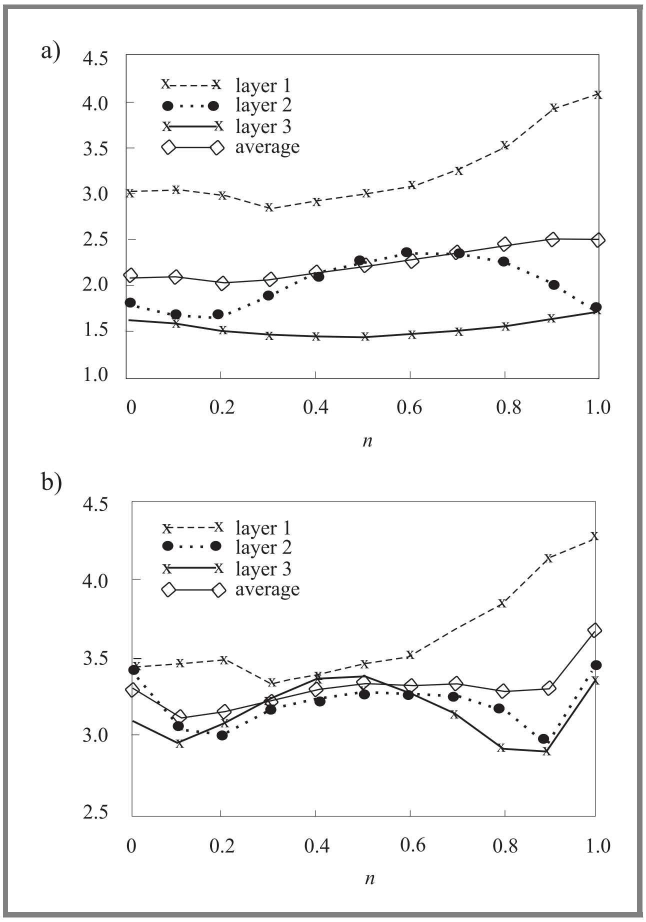


Fig. 3. Comparison of the uniformity (integral power criterion) of dissipated power distribution with (a) object rotation; (b) without rotation.

5. Experimental validation of the accuracy of simulation

The accuracy of the simulation software employed to obtain the results described in the previous section has been verified experimentally. The measurement equipment recently acquired by the Institute of Radioelectronics has been employed in the task and presented in Fig. 4. It consists of a microwave oven (by Plazmatronika, www.plazmatronika.pl) with adjustable power level controlled with a PC, and a signal conditioner with set of eight thermometers (by Fiso Systems, www.fiso.com) that can register temperature changes simultaneously.

A set of temperature measurements has been conducted. A sample of bread 60(w) × 60(d) × 20(h) mm has been

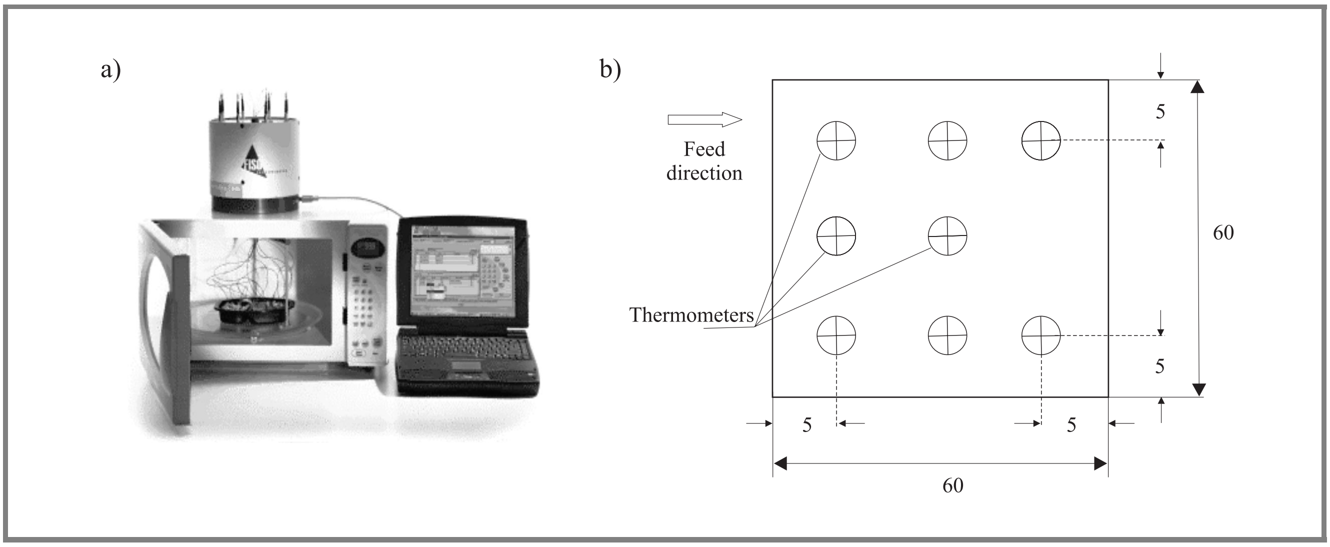


Fig. 4. Measurement system used in verification of the simulation results (a); placement of thermometers in sample (b).

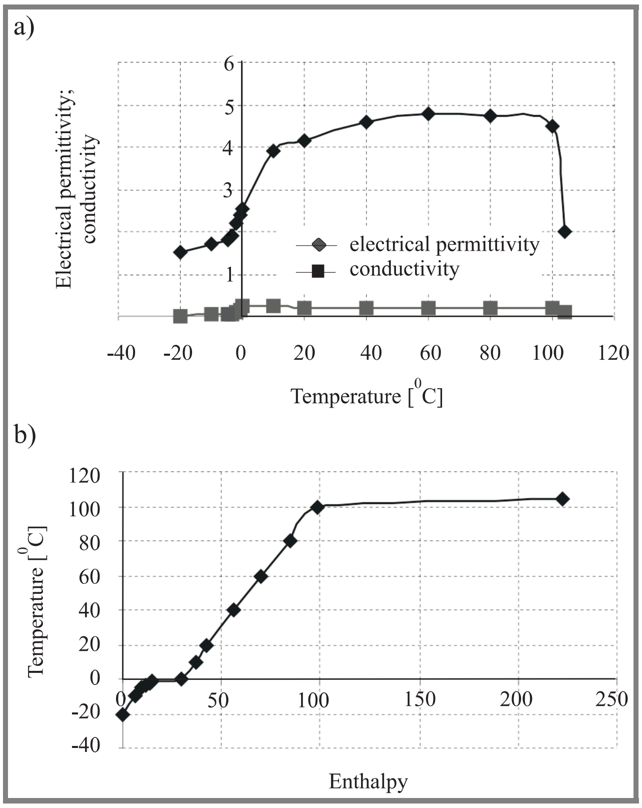


Fig. 5. Data contained in the file describing the media: (a) dependence of the relative electrical permittivity and conductivity [S/m] on the temperature; (b) dependence of the temperature in a cell on the enthalpy in this cell.

placed centrally on the oven turntable. The power level has been set to 100 W and the heating lasted around 5 minutes. Temperature has been registered at 8 locations shown in Fig. 4b, in the middle layer of the sample. Since the main goal of the experiment has been to check the accuracy of the software, we have performed the heating without sample rotation. It has made the computation time much shorter as there has been no need for repeated simulations of the oven with sample at various angular positions.

After the measurements the computer model of the microwave oven has been prepared and a set of simulation results obtained. An additional module – called BHM (basic heating module) [9] – has been used that takes into account the temperature-induced changes of the media parameters. The changes are based on the data stored in an external file containing enthalpy and corresponding media parameters: electrical permittivity, conductivity (losses) as well as temperature. The data stored in the file used in the bread simulation have been presented in Fig. 5.

The BHM module repetitively modifies the media parameters according to the dissipated power and data provided in the file. The total heating time has to be divided into smaller timesteps in order to accurately model the gradual changes of the media parameters. The operations performed in each step have been illustrated in Fig. 6. First, the steady-state needs to be reached in order to calculate the dissipated power envelope in the simulated circuit. Then the lossy materials are being heated for the time equal to the timestep length. After the heating has been done the enthalpy is obtained and the media parameters are modified according to the file, and temperature in each cell is changed.

The comparison of the experimentsal data and the simulation results has been presented in Fig. 7a. The comparison has been made in the hot spot which is just off the sample center. The curves are close to each other with the biggest discrepancy occuring at the beginning of heating process. This can be ascribed to the inertia of the thermometer. The

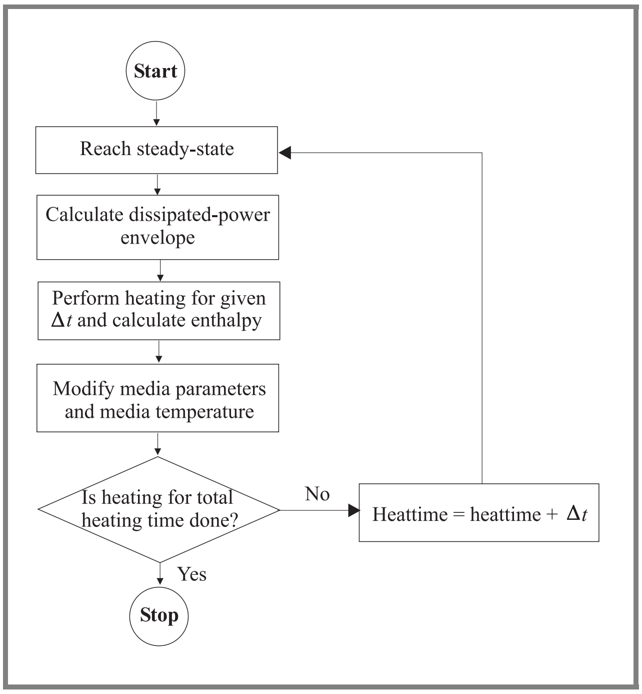

Start
Reach steady-state
Calculate dissipated-power envelope
Perform heating for given Δt and calculate enthalpy
Modify media parameters and media temperature
Is heating for total heating time done?
No
Heattime = heattime + Δt
Yes
Stop

Fig. 6. Operation of the BHM module.

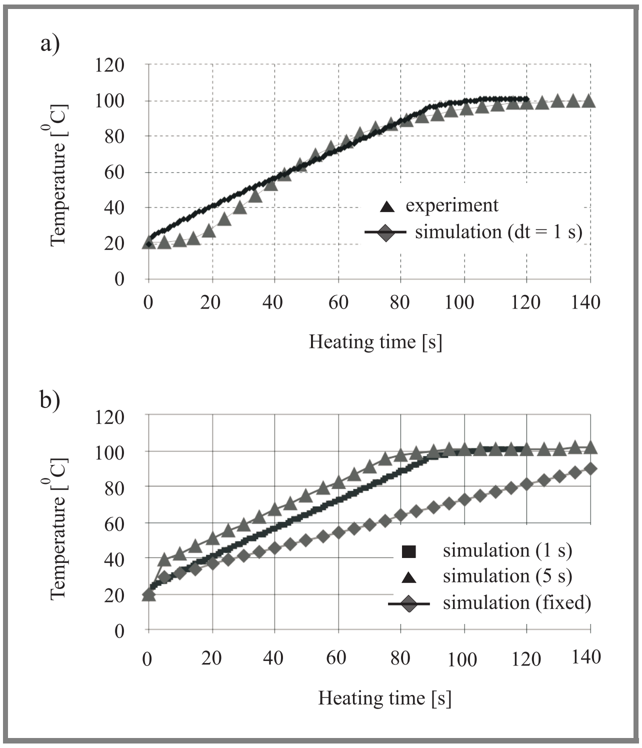


Fig. 7. (a) Comparison of the measurements and the simulation data (in the hot spot); (b) comparison of simulation results (two cases where media parameters where modified with two different heating timesteps – 1 s and 5 s – and one case where media parameters stayed unchanged).

simulation data have been obtained for a heating time step set to 1 s. It has been checked that in this case bigger values can lead to inaccurate data.

The inertia of the thermometers used in the measurements has been checked experimentally. The thermometer has been placed first in a glass filled with water of room temperature (25°C) and then transferred to a glass filled with much hotter water. The output has been registered and we have shown it in Fig. 8. It seems that in case of water the time for the temperature to rise to the correct level is around 4 s. In case of water the contact resistance between the thermometer and the medium is not high. With bread the resistance is much higher so greater inertia can be expected.

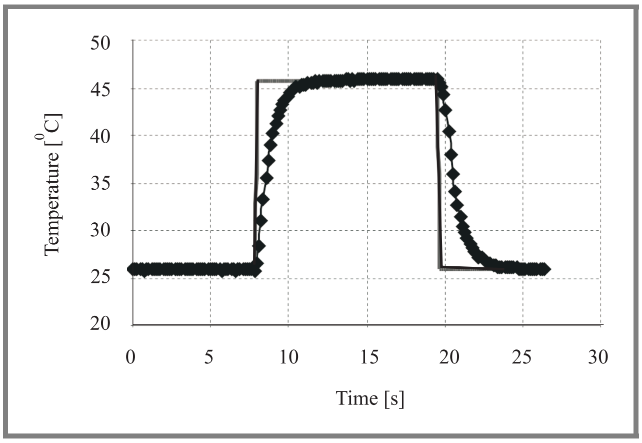


Fig. 8. Response of the thermometer used in experiment (water).

The timestep of the heating is an important factor since choosing too small values will lead to excessively long computation time without any improvement in accuracy. Too big timestep, though, is even more dangerous as it may cause abrupt changes in temperature and, in turn, jumps of the media parameter values which can bring instabilities into the simulations. The data obtained for different timestep values are presented in Fig. 7b. They have been compared with the data obtained without any modification of parameters.

6. Conclusion

We have presented an effective and simple approach to modeling of problems with load rotation, which is an important issue in microwave power applications and has not been previously addressed. It is a post-processing method that is easy to implement with standard mathematical routines found in e.g. Matlab package. We have also considered a couple of examples proving that object rotation may change the dissipated power distribution within the object, and hence optimum geometry of the heating system. This is important in design of microwave ovens and microwaveable food packages. The experimental verification of the simulation tool has been conducted and it has been shown that the accuracy of the computations is high enough to ensure a good agreement with measurements.

Acknowledgement

This work has been performed within EUREKA E!2602 initiative and co-financed by the State Committee for Scientific Research grant no. 134/E-365/SPUB-M/EUREKA/T-10/DZ 206. The authors also kindly acknowledge the helpful discussions with Per Risman, Microtrans AB, on power uniformity criteria and physical parameters of bread as a function of enthalpy.

References

[1] P. O. Risman and M. Celuch-Marcysiak, "Electromagnetic modelling for microwave heating applications", in *13th Int. Conf. MIKON-2000*, Wrocław, Poland, May 2000, pp. 167–182.

[2] M. Sundberg, P. Kildal, and T. Ohlsson, "Moment method analysis of a microwave tunnel oven", *J. Microw. Pow. Electromagn. Ener.*, vol. 33, no. 1, pp. 36–48, 1998.

[3] P. Kopyt, "Optymalizacja układu grzania mikrofalowego pod względem jednorodności rozkładu mocy". Praca magisterska, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska, Warszawa, 2001 (M.Sc. thesis in Polish).

[4] B. Wäppling-Raaholt and P. O. Risman, "FDTD simulation of a microwave heating process: effects of oven parameters on heating uniformity", in *3rd Int. Conf. Predic. Model.*, Leuven, Belgium, Sept. 2000, pp. 271–273.

[5] W. K. Gwarek, "Analysis of an arbitralily-shaped two-dimensional microwave circuits by FD-TD method", *IEEE Trans. Microw. Theory Techn.*, vol. 36, no. 4, pp. 738–744, 1988.

[6] W. K. Gwarek, M. Celuch-Marcysiak, M. Sypniewski, and A. Więckowski, "QuickWave-3D software manual", QWED, Poland, 1999.

[7] V. V. Yakovlev, "Examination of contemporary electromagnetic software capable of modelling problems of microwave heating", in *Proc. 8th Int. Conf. Micro and HF Heating*, Bayreuth, Germany, Sept. 2001, pp. 19–20 (complete paper in *Trends in Microwave and HF Heating*, Springer Verlag, 2001 – to be published).

[8] Matlab 5.1, Mathworks – university licence.

[9] W. K. Gwarek, M. Celuch-Marcysiak, M. Sypniewski, and A. Więckowski, "BHM module", QWED, Poland, 2000.

Paweł Kopyt was born in Warsaw, Poland, on May 15, 1975. He graduated *cum laude* in telecommunications from Warsaw University of Technology, where he received an M.Sc. degree in 2001. At present, he is a Ph.D. student at Institute of Radioelectronics, Warsaw University of Technology. Among his interests are numerical techniques for simulation of electromagnetic and thermal fields and microwave heating.

e-mail: pkopyt@elka.pw.edu.pl
Institute of Radioelectronics
Warsaw University of Technology
Nowowiejska st 15/19
00-655 Warsaw, Poland

Małgorzata Celuch-Marcysiak is an Assistant Professor at the Warsaw University of Technology and Vice-President of QWED company. She received an International Baccalaureate from the United World College of the Atlantic, in 1983, M.Sc. and Ph.D. in electronic engineering from Warsaw University of Technology in 1988 and 1996, respectively. She is one of the main authors of QuickWave electromagnetic simulation software. She contributed to more than 60 technical publications. Her main professional interest is in numerical modeling of electromagnetic problems.

e-mail: m.celuch@ire.pw.edu.pl
Institute of Radioelectronics
Warsaw University of Technology
Nowowiejska st 15/19
00-655 Warsaw, Poland

Assessment of Area Energy Efficiency of LTE Macro Base Stations in Different Environments

Suhail Najm Shahab[1], Ayad Atiyah Abdulkafi[2], and Ayib Rosdi Zainun[1]

[1] *Faculty of Electrical and Electronics Engineering, Universiti Malaysia Pahang, Pahang, Malaysia*
[2] *College of Engineering, Tikrit University, Salahaldin, Iraq*

http://doi.org/10.26636/jtit.2015.1.779

Abstract—Energy efficiency (EE) of wireless telecommunications has become a new challenge for the research community, governments and industries in order to reduce CO$_2$ emission and operational costs. EE of base stations (BSs) in cellular networks is a growing concern for cellular operators to not only maintain profitability, but also to reduce the overall negative impact to the environment and economic issues for wireless network operators. In this paper, a framework focuses on the Area Energy Efficiency (AEE) evaluation of LTE BSs is presented. The parameters affect on the AEE and the coverage area of LTE BS in different scenarios are investigated. AEE analysis has been done using a few key performance indicators including transmit power, bandwidth, load factor with the assumption of different scenarios (urban, suburban and rural). The simulation results show that the LTE BSs have better AEE in an urban environment for cell radius less than 750 m compare with the suburban and rural environments. Furthermore, it is obvious that there is a strong influence of traffic load, BW and transmission power on AEE of LTE network. On the other hand, AEE increases significantly as the BW size increases. Finally, it has been shown that the AEE of LTE macro BS decreases with increasing the percentage of traffic load for all scenarios.

Keywords*—*Area Energy Efficiency, LTE, Macro Base Station.

1. Introduction

Addressing the issue of green communications has benefits to many stakeholders including the industry, academic researchers and government agencies. The cellular industry can realize cost savings and lower their impact to the environment, government agencies realize a fulfillment of administrative goals for energy savings as well as development of standards and metrics, while researchers can push the boundaries of current technologies and theories in material science, distributed computing and system engineering. Telecommunication section and especially cellular networks are parts of Information and Communication Technology (ICT) that is rapidly expanding throughout the globe. With new technologies like Third Generation (3G) and Long Term Evolution (LTE) coming to the market, this section will grow more in a future.

Currently, telecommunication sectors are responsible for about 12% of total energy consumption of the world and generates approximately 1% of CO$_2$ emissions [1] with percentages expected to rise further. In [2] the authors proposed the deployment of LTE macro base station (BS) to study the impact of modulation and coding schemes (MCS), bandwidth (BW) size and transmitted power on the energy efficiency for urban environment. They showed that the higher transmission power results in lower EE. The difference actually diminishes when cell size increases. At its diameter around 1200 m, it was found that the EE is almost equal for all transmission power considered. On the other hand, EE increases significantly as the BW increases. Similar effect on EE is observed when MCS changes from lower order to the higher-order scheme. In cellular networks, the prime energy users are base stations (BSs), backhaul servers and routers. Around 80% energy is consumed by the BSs [3]. Because of this statistic, most of the energy saving research had been focused on the BS.

This paper investigates the area energy efficiency (AEE) issue on LTE networks and more specifically it is based on simulation for the outdoor environments. The environment's scenarios for the simulation were done with three different environments: urban, suburban and rural. Results have conducted and discussed to show the performance of LTE network from the AEE perspective. A comparison analysis is done in terms of energy saving for a specific macro BS deployment between the three different scenarios.

2. Methodology

2.1. Propagation Model

In general, there are many factors that cause the deterioration of signal quality such as distance dependent path losses, shadowing, outdoor/indoor penetration loss and radiation pattern. The received power P_{rx}, from a BS at a distance of d and angle θ from the main lobe of the antenna can be calculated as [4]:

$$P_{rx}(d,\theta,\psi) = P_{tx} - \left[PL(d) + \kappa + A_h(\theta)\right] + \psi_{dB}\,, \qquad (1)$$

where P_{rx}, P_{tx}, PL, κ, A_h, ψ and θ denote to receive power and transmit, path loss, penetration loss, antenna radiation pattern, shadow fading and theta, respectively.

Equation (1) assumes that all the signal gains and losses are expressed in decibels. The random variable ψ is used to model slow fading effects and commonly follows a log normal distribution. The antenna pattern $A_h(\theta)$ depends on the mobile's location relative to the BS which has been adopted from [4]. In addition to path loss and shadowing, another factor which affects the channel quality is penetration loss for users indoors.

In this paper, a 20 dB of attenuation has been assumed to account for outdoor/indoor penetration loss, denoted by κ, which can be found in [5] and [6]. The path loss PL in decibels (dB) for a distance d can be expressed into three different categories, namely urban, suburban and rural areas, which take into account distance, line of sight existence, antenna height, and the average building height with the applicability ranges from 5 to 50 m as proposed in [5] for all environments.

However, the urban scenario usually has a great concentration of BSs due to the demand for capacity. The path loss in urban scenario before the break point d_{BP} can be written in the following form:

$$PL = 22.0\log_{10}(d) + 28.0 + 20\log_{10}(f_c)\,, \tag{2}$$

where d is the distance in meters, and f_c is the carrier frequency in GHz. After d_{BP}, the path loss is founded via:

$$\begin{aligned} PL = 40.0\log_{10}(d) + 7.8 - 18\log_{10}(h'_{BS}) - \\ - 18\log_{10}(h'_{UT}) = 2\log_{10}(f_c)\,, \end{aligned} \tag{3}$$

where h'_{BS} and h'_{UE} are the effective antenna heights at the BS and the User Equipment (UE). The effective antenna heights h'_{BS} and h'_{UE} are computed as follows: $h'_{BS} = h_{BS} - 1.0$ m, $h'_{UE} = h_{UE} - 1.0$ m, where $h_{BS} = 25$ m and $h_{UE} = 1.5$ m are the actual antenna heights proposed in [5] for urban area.

The suburban scenario is modeled to correspond to typical city's periphery with major habitation blocks with several floors. While the remaining territory corresponds to rural low dense populated scenarios that can be crossed by important highways. The path loss in suburban and rural scenarios before the d_{BP} can be written as:

$$\begin{aligned} PL = 20\log_{10}\left(\frac{40\pi d f_c}{3}\right) + \\ + \min(0.03h^{1.72}, 10)\log_{10}(d)\min(0.044^{1.72}, 14.77) + \\ + 0.002\log_{10}(h)d\,. \end{aligned} \tag{4}$$

While after d_{BP}, the path loss for these two scenarios is founded via:

$$\begin{aligned} PL = 20\log_{10}\left(\frac{40\pi d f_c}{3}\right) + \\ + \min(0.03h^{1.72}, 10)\log_{10}(d)\min(0.044^{1.72}, 14.77) + \\ + 0.002\log_{10}(h)d + 40\log_{10}\left(\frac{d}{d_{BP}}\right). \end{aligned} \tag{5}$$

Here h is building height in meters.

2.2. Cell Coverage Area

The cellular system coverage is generally designed for a given minimum received power $P_{\min}$ at the cell boundary. The $P_{\min}$, which is also known as the receiver sensitivity can be written in closed-form for cell coverage area C as [7]:

$$C = Q(a) + \exp\left(\frac{2-2ab}{b^2}\right)Q\left(\frac{2-ab}{b}\right), \tag{6}$$

where:

$$a = \frac{P_{\min} - P_{rx}(R)}{\sigma_{\psi dB}}\,, \quad b = \frac{1-\alpha\log_{10}(e)}{\sigma_{\psi dB}}\,, \tag{7}$$

where α denote to path loss exponents and σ_{dB} is the standard deviation of shadow fading [7].

The reference sensitivity $P_{\min}$ level is the minimum mean received signal strength applied to both antenna ports at which there is sufficient SINR for the specified modulation scheme to meet a minimum throughput requirement of the maximum possible. It is measured with the transmitter operating at full power. $P_{\min}$ is a range of values that can be calculated using the Eq. (8) [8]:

$$P_{\min} = \mathrm{kTBW} + \mathrm{NF} + \mathrm{SINR}_{req} + \mathrm{IM} - \mathrm{G_d}\,, \tag{8}$$

where $kTBW$ is the thermal noise level in units of dBm, in the specified bandwidth (BW), NF is the prescribed maximum noise figure for the receiver where LTE defines an NF requirement of 9 dB for the User Equipment (UE), SINR_{req} is Signal to Interference plus Noise Ratio that required for choosing modulation and coding scheme, IM is the implementation margin and $\mathrm{G_d}$ represents the diversity gain which is equal to 3 dB [8]. $P_{\min}$ is a target minimum received power level below which performance becomes unacceptable [7]. Note that $a = 0$, when the target minimum received power equals the average power at the cell boundary, $P_{\min} = P_{rx}(R)$ and $P_{rx}(R)$ is the received power at the cell boundary due to path loss alone. An extra implementation margin is added to reflect the difference in SINR requirement between theory and practicable implementation [8].

2.3. LTE Data Rate Model

Theoretical peak data rates are difficult to achieve in practical situations only in extremely good channel conditions because of limited by the amount of channel impairments noise and interference from own and other cells. The maximum theoretical data rate for single antenna transmission in static channel can be derived through conventional Shannon's formula which is given in Eq. (9). The data rate R_T in unit of bits per second can be expressed in terms of two parameters which are the bandwidth and the signal to noise ratio SNR.

$$R_T = \mathrm{BW} \times \log_2(1 + \mathrm{SNR})\,. \tag{9}$$

In LTE system, a modified Shannon formula is used to accurately estimate the achieved data rate after taking channel impairments into account.

$$R_T = F \times \mathrm{BW} \times \log_2\left(1 + \frac{\mathrm{SINR}_{req}}{\eta_{\mathrm{SNR}}}\right). \qquad (10)$$

where $F = \eta_{BW} \cdot \eta$ in which the η_{BW} accounts for the system bandwidth efficiency of LTE and η_{SNR} accounts for the SNR implementation efficiency of LTE. It should be noted that LTE is performing less than $1.6 \sim 2$ dB from the Shannon capacity bound because it's not constant and changes with the geometry factor (G-factor), the G-factor distribution is defined as the average own cell power to the other cell power plus noise ratio with OFDMA in a wide system bandwidth this corresponds to the average SINR [8]. It was shown that this impact can be accounted for using the fudge factor (η) multiplying by the parameter (i.e. $\eta = 0.9$ and $\eta_{BW} \cdot \eta = 0.75$). η_{SNR} is a parameter for adjusting SNR efficiency which is almost equal to one [9].

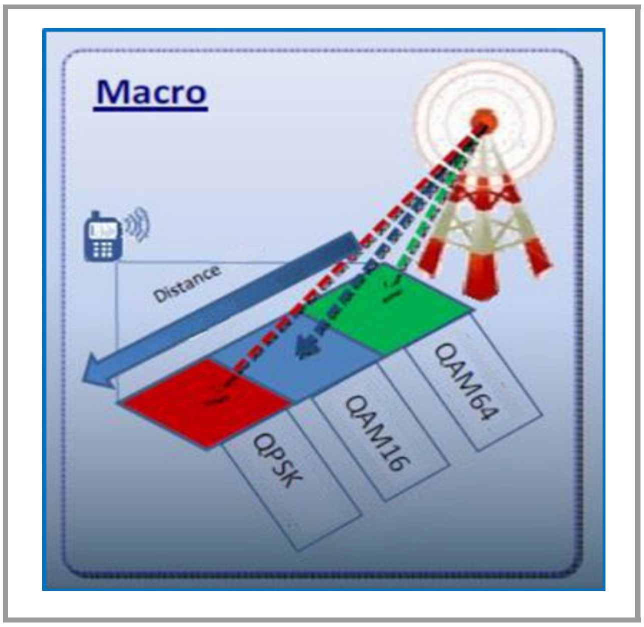


Fig. 1. MCS selected based on user distance.

The MCS selection is depend on the distance between the eNodeB and the UE. The low MCS can be suitable for large distances as the experienced SINR is low while the higher MCS is preferred at short ranges with high data rate demands. Figure 1 shows how the different MCS are selected according to the distance between eNodeB and UE based on the received SINR.

2.4. LTE Power Consumption Model

The main goal of the power consumption model in this paper is to make realistic input parameters available for the simulation. This model also allows fair comparing between different environments and different macrocell BS deployments. The power models have been selected from [10], [11] for different environments cases for LTE deployment.

The power model of macro BS described in [10] has a linear relationship between average radiated power per site and average power consumption. The power consumption calculation is modified to be changed according to the traffic load level and the BS components features. The consumed power P_c by the BS i can be expressed as:

$$P_c = L \cdot N_{sec} N_{ant} \left(\mathrm{A} P_{tx} + \mathrm{B}\right), \qquad (11)$$

where $L \in [0, 1]$ is the load factor and N_{sec} and N_{ant} denote the BS's number of sectors and the number of antennas per sector, respectively. P_c and P_{tx} denote the total power per BS and the power fed to the antenna, respectively. The coefficient A accounts for the part of the power consumption that is proportional to the transmitted power (e.g., radio frequency amplifier power including losses caused by feeders and cooling of sites), while B denotes the power that is consumed independent of the average transmit power and models the power consumed (e.g., signal processing, site cooling, backhaul, and as well as a battery backup) [10], [12]. Both these coefficients are constant for macro BS. The power model is calculating power consumption with respect to transmit power P_{tx} this assumption is valid because currently deployed macro sites power consumption depends upon the traffic load [10]. The parameter L models the activity level of the BS which describes the portion of resources which are allocated for transmission, where zero and full load correspond to no active user in the cell and providing one or more users with all resources available, respectively.

However, it may be unsuitable to observe only power consumption for comparing the networks with different site densities. This is because they may have different coverage's. In order to assess the power consumption of the network relative to its size, the notion of area power consumption APC measured in [W/km^2] is introduced as the total power consumption in a reference cell divided by the corresponding reference area [10], [13]:

$$APC = \frac{P_c}{A_{macro}}, \qquad (12)$$

here A_{macro} is the macro reference area which can be expressed as [10] and [13]:

$$A_{macro} = \frac{3\sqrt{3}}{2} d^2. \qquad (13)$$

It was shown that for a hexagonal deployment the area power consumption metric yields an optimal coverage cell size [10].

2.5. Energy Efficiency

The extrapolation of current trends undertaken by many literatures reveals that for a sustainable growth of wireless communications, an improvement of LTE energy efficiency is required. In this study, energy efficiency assessing a framework is studied via network level simulations.

The total network energy efficiency EE_T which is defined as the ratio of total amount data delivered and the total power consumed measured in bits per Joule [14], is represented by:

$$EE_T = \frac{\sum_{i=1}^{N_{BS}} R_i}{\sum_{i=1}^{N_{BS}} P_{c_i}} . \tag{14}$$

where P_c is the power consumption and R_i is the total data rate with a BS i. N_{BS} is the total number of BSs. As know, cell coverage is a primary concern in the design of wireless data communications systems. Increased inter-site distances (ISDs) generate larger coverage areas. With the same transmission power, different cell size can lead to different individual date rate and accordingly various energy efficiency. Therefore, observing the mere energy efficiency per site is not enough for comparing networks with different cell size. Moreover, another important metric is used through this research to evaluate the energy efficiency of the network relative to its size. The Area Energy Efficiency (AEE) metric which is defined as a bit/Joule/unit area is used as a performance indicator metric. The AEE for LTE network can be expressed as [15]:

$$\text{AEE} = \frac{EE_T}{A_T} , \tag{15}$$

where the aforementioned EE_T and A_T are the total energy efficiency and total area of LTE network respectively.

3. Simulation Procedure and Results

The EE performance of the network corresponding to its size and deployment can be more accurately assessed by comparing the AEE performance under different sector radius and scenarios. In the following subsections, the LTE performance in terms of AEE is presented. Furthermore, the effect of environment type on AEE is demonstrated. Later, by considering different traffic load scenarios, the impact of traffic load on AEE has been explained and discussed. The parameters that are affecting the AEE of LTE macro BS are investigated. The impacts of parameters like different traffic load, BW and P_{tx} on AEE.

3.1. Simulation Procedure

In this section, the simulation procedure and system parameters are discussed. There are three different environments are chosen for study campaigns one is an urban type environment. The second is a suburban site like a small city while the third is with a rural environment. Single LTE macro BS covers a hexagonal shaped area as shown in Fig. 2 in which R is the cell radius and A_{macro} is the coverage area.

The cell size is determined according to the minimum received power level constraints. The receiver sensitivity is calculated based on sufficient SINR for the specified modulation scheme to achieve a minimum requirement of 95% coverage degree. The received SNR is calculated based on the received power level and white noise which are estimated according to the path loss model described in 3GPP TR 36.814 [5]. Then, the achievable data rate within each BS's coverage area is determined based on the SNR distribution in the cell. The power consumption models consist of dynamic power consumption which is fully depended on the traffic load as expressed in Eq. (11). The simulation parameters are based on 3GPP macrocell model with a system bandwidth of 10 MHz with UE height of 1.5 m. The 2.6 GHz spectrum band is used since this is the band allocated to LTE operators in Malaysia [16]. Effective environment height which is subtracted from the actual antenna height for BS and UE to find their effective antenna heights is assumed to be equal to 1 m. IM of 2.5 dB is assumed for all QPSK modes, while 3 dB and 4 dB are generally expected for 16QAM and 64QAM respectively [17]. However, the typical assumptions for the SINR values for different MCS that are used in the simulation assumptions equal the ones in [8]. The proposed simulation model for evaluating the EE in LTE macro BS in different environments is an extension of the work in [18] as shown in Fig. 3.

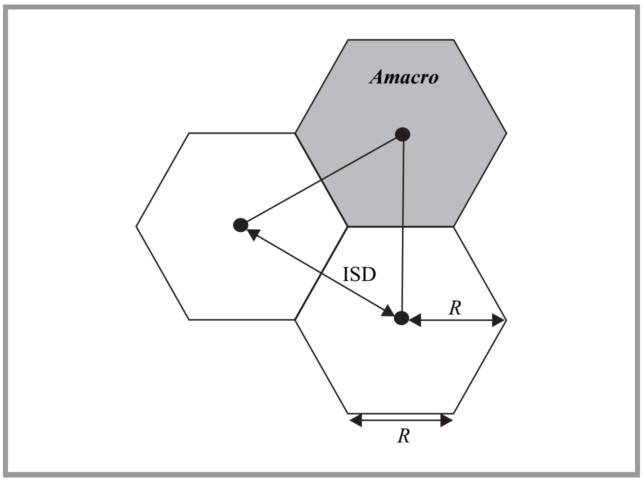


Fig. 2. Corresponding cell geometry.

Table 1
Simulation parameters

Notation	Description	Default
f_c	Carrier Frequency [GHz]	2.6
BW	Bandwidth [MHz]	10
N_{sec}	Number of sectors	3
N_{ant}	Number of antennas	2
MCS	Modulation Coding Scheme	1/3 QPSK [8]
P_{tx}	Transmit Power [dBm]	46
G_d	Diversity gain [dB]	3 [8]
C	Coverage degree	95%
NF	Noise Figure [dB]	9 [8]
A_i B_i	Power consumption parameters [W]	21.45 354.44 [11]

Various parameters have been used in all simulation scenarios to analyze the EE behavior under specific circumstances. Simulation parameters are listed in Table 1 and simulation procedure flow chart shown in Fig. 3.

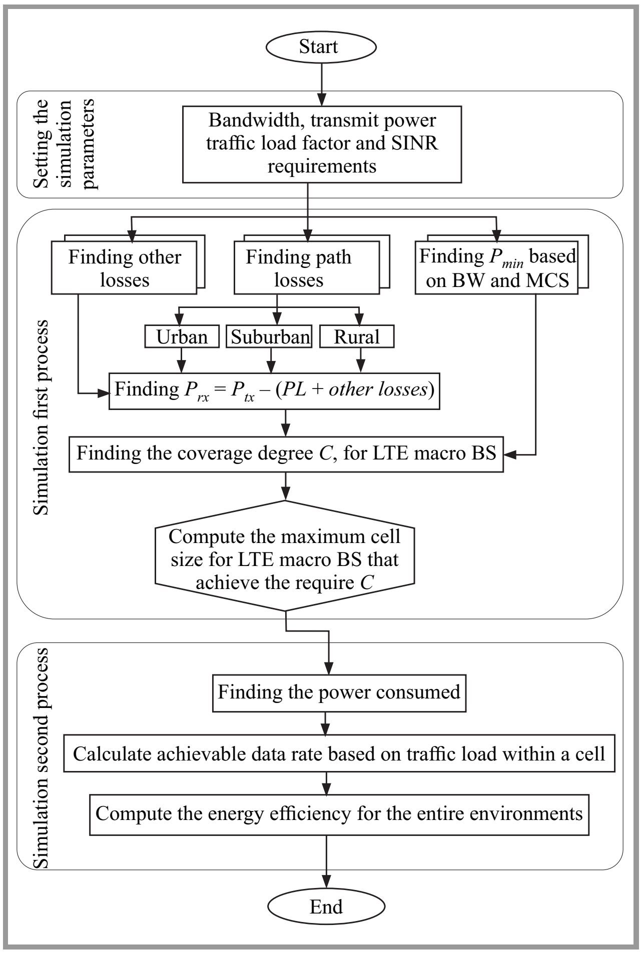

Start
Setting the simulation parameters
Bandwidth, transmit power traffic load factor and SINR requirements
Simulation first process
Finding other losses
Finding path losses
Finding P_{min} based on BW and MCS
Urban
Suburban
Rural
Finding $P_{rx} = P_{tx} - (PL + other\ losses)$
Finding the coverage degree C, for LTE macro BS
Compute the maximum cell size for LTE macro BS that achieve the require C
Simulation second process
Finding the power consumed
Calculate achievable data rate based on traffic load within a cell
Compute the energy efficiency for the entire environments
End

Fig. 3. Simulation model flow chart.

3.2. Simulation Results

AEE for three scenarios at full load. There are different coverage area sizes of LTE BSs due to the deployment environments, there are different data rates for each BS in each environment according to its size and therefore different EE's. Thus, the AEE is used to evaluate the EE of LTE network relative to its size. The AEE has been calculated based on a Eq. (15). In Fig. 4, the AEE versus cell radius for three environments with full load (100%) is plotted. It is obvious that AEE decreases as the macrocell BS's radius increases. Moreover, it can be shown that the LTE BSs have better AEE in urban environment with cell size less than 750 m. For cell radius more than 750 and 1500 m, the LTE performance becomes better in suburban and rural environments respectively. More specifically, at the first 700 m the better AEE is can achieve in urban area but at 710 m the suburban area becomes better than urban and rural, also at 1055 m the rural area became better than urban areas as shown in Fig. 4. This is because the impact of shadowing, path losses as well as the penetration losses has become more significant in the urban area at long distances as compared with the rest environments.

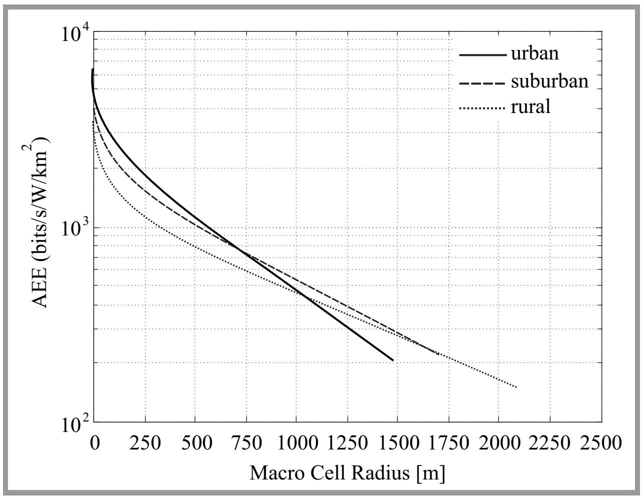


Fig. 4. AEE versus cell radius for three environments.

AEE for three scenarios with different loads. The traffic load is another important factor that affects the network performance. It has a stronger impact on the data rate and the power consumption of LTE network and subsequently on its EE and AEE. The AEE versus cell radius for urban area under different loads shown in Fig. 5. It is clear that the AEE decreases as the traffic load increases. In fact, the AEE's become almost equals as the traffic loads increased as shown in Fig. 5 the curve with traffic load 90% are very closed to the curve with a full traffic load scenarios. Moreover, it can be shown for all environments that the AEE decreases as the traffic load increases due to increasing in power consumption. The same AEE performance can be concluded for suburban and rural areas when varying the traffic load as shown in Figs. 6 and 7 respectively.

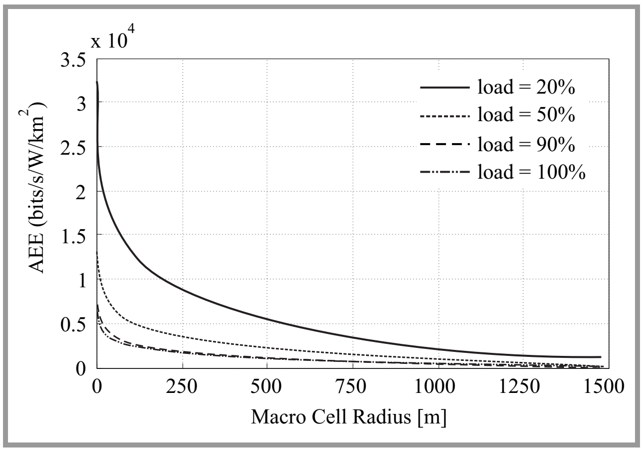


Fig. 5. AEE versus cell radius for urban environment with different loads.

Table 2 summarizes the AEE performance for the three types of environments with different traffic load conditions at 100, 1000 and the cell edge for each environment. As

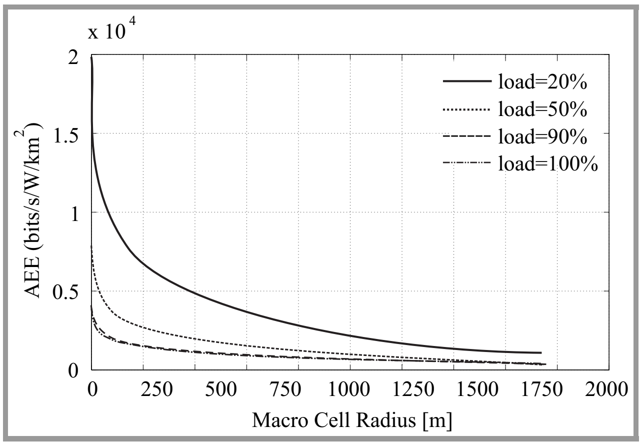


Fig. 6. AEE versus cell radius for suburban environment with different loads.

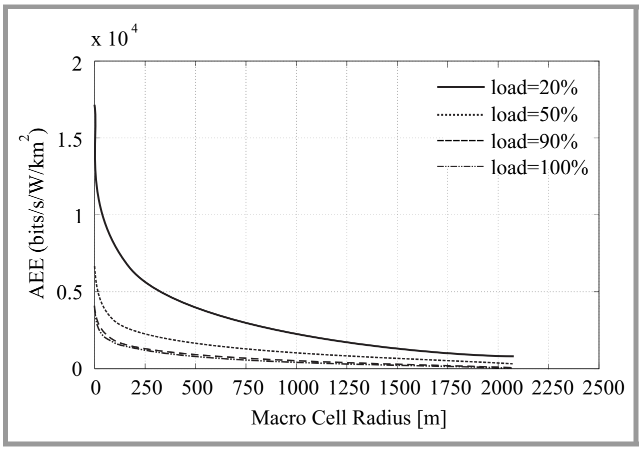


Fig. 7. AEE versus cell radius for rural environment with different loads.

mentioned before, the AEE of LTE macro BS at short distances is better for urban area than suburban and rural for all load conditions. As shown in Table 2 for cell radius more than 1000, the LTE performance becomes better in suburban and rural environments respectively.

Table 3 shows the LTE performance (BW = 10 MHz, 1/3QPSK, full load) in terms of AEE for different transmission powers. However, the AEE decreases as the transmission power increases for the same environment. In addition, it can be concluded that the suburban area achieved better AEE performance due to its suitable cell size compare to urban and rural areas.

Table 2
Training and classification times

Environ-ments	Distance [m]	AEE $\left[\text{bits/s/W/km}^2\right]$			
		20% load	50% load	90% load	Full load
Urban	100	13180	5264	2929	2637
	1000	2416	966.5	536.9	483.2
	1475.7	1041	416.4	231.3	208.2
Suburban	100	10780	4311	2395	2155
	1000	2751	1100	611.4	550.2
	1718.1	1107	442.9	246.1	221.4
Rural	100	7841	3136	1742	1568
	1000	2324	929.6	516.5	464.8
	2074.9	758.1	303.2	168.5	151.6

Table 3
AEE at cell edge for different P_{tx}

Environment	P_{tx} [dBm]		
	43	46	49
Urban	453.7886	208.1916	86.3716
Suburban	482.2614	221.4458	92.0362
Rural	330.7610	151.6243	63.1749

Table 4
AEE at cell edge for different bandwidth

Environ-ment	BW [MHz]					
	1.4	3	5	10	15	20
Urban	8.642	34.101	73.337	208.191	382.168	586.726
Suburban	9.192	36.316	78.206	221.445	407.545	626.886
Rural	6.308	24.962	53.715	151.624	279.118	429.570

Increasing the BW for any type of environment will increase the AEE of LTE macro BS. In fact, the better outcomes can be predicted for suburban area while the urban area comes in the next order and finally the rural area as demonstrated in Table 4 ($P_{tx} = 46$ dBm, 1/3QPSK, full load).

4. Conclusion

One of the most important requirements for wireless communication technologies is to be applicable and universally desirable. AEE for LTE macro BS analysis is the main target for this paper. It is considered as the most important process to achieve mobility within wireless networks. Work evaluation has been done by simulating AEE assessing with different scenarios. Three different environments were chosen for this study including urban, suburban and rural. A framework for evaluating the AEE of LTE network in different environments has been proposed. Using few key performance indicators such as coverage size, area power consumption, energy efficiency and area energy efficiency, the network performance from EE perspective for all the three urban, suburban and rural terrains are compared and evaluated. Although, the LTE BSs have large cell size and good coverage degree in rural areas, the simulation results show that they have better AEE in urban environment with small cell sizes while the AEE becomes better in suburban and rural environments for larger cell radius. Also, it can be concluded that there is a strongly impact of traffic load, bandwidth and transmission power on APC and AEE of LTE macrocell networks. For all the three environments, it has been shown that the AEE of LTE macro BS decreases with increasing the traffic load and this effect becomes the same at high loads. Using the proposed framework, the EE of different deployment scenarios can be evaluated and insights on how to deploy a greener LTE network are provided. The results presented in this work consider only one LTE BS and therefore the impact of the handovers and interference in the LTE network may bring substantial impact on the AEE. These issues have been left for author's future works.

Acknowledgment

This work is supported by University Malaysia Pahang.

References

[1] P. Misar, "Wireless LTE deployment: How it is changing cell site energy and infrastructure design", in *Proc. 32nd IEEE Ann. International Telecommun. Energy Conf. INTELEC 2010*, Orlando, FL, USA, 2010, pp. 510–514.

[2] A. A. Abdulkafi, T. S. Kiong, J. Koh, D. Chieng, A. Ting, and A. M. Ghaleb, "Energy efficiency of LTE macro base station", in *Proc. 1st Int. Symp. Telecommun. Technol. ISTT 2012*, Kuala Lumpur, Malaysia, 2012, pp. 259–264.

[3] M. Pickavet *et al.*, "Worldwide energy needs for ICT: The rise of power-aware networking", in *Proc. 2nd Int. Symp. Adv. Netw. Telecommun. Syst. ANTS 2008*, Bombay, India, 2008, pp. 1–3.

[4] T. T. Tesfay, R. Khalili, J.-Y. L. Boudec, F. Richter, and A. Fehske, "Energy saving and capacity gain of micro sites in regular LTE networks: downlink traffic layer analysis", in *Proc. 6th ACM Workshop on Perform. Monitor. Measur. Heterogen. Wirel. Wired Netw.*, Miami, FL, USA, 2011, pp. 83–92.

[5] 3GPP TR 36.814 V9.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9)", 3rd Generation Partnership Project, Tech. Rep., 2010 [Online]. Available: http://www.3gpp.org

[6] A. A. Abdulkafi, T. S. Kiong, D. Chieng, A. Ting, and J. Koh, "Energy efficiency improvements in heterogeneous network through traffic load balancing and sleep mode mechanisms", *Wirel. Personal Commun.*, vol. 75, no.4, pp. 2151–2164, 2014.

[7] A. Goldsmith, *Wireless Communications*. New York: Cambridge University Press, 2005.

[8] S. Sesia, I. Toufik, and M. Baker, *LTE – The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. United Kingdom: Wiley, 2011.

[9] P. Mogensen *et al.*, "LTE capacity compared to the shannon bound", in *Proc. IEEE 65th Veh. Technol. Conf. VTC2007-Spring*, Dublin, Ireland, 2007, pp. 1234–1238.

[10] F. Richter, A. J. Fehske, and G. P. Fettweis, "Energy efficiency aspects of base station deployment strategies for cellular networks", in *Proc. IEEE 70th Veh. Technol. Conf. Fall VTC-Fall 2009*, Anchorage, AK, USA, 2009, pp. 1–5.

[11] S. Tombaz, M. Usman, and J. Zander, "Energy efficiency improvements through heterogeneous networks in diverse traffic distribution scenarios", in *Proc. 6th Int. ICST Conf. Commun. Netw. in China CHINACOM 2011*, Harbin, China, 2011, pp. 708–713.

[12] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks", in *Future Network and Mobile Summit*, Florence, Italy, 2010, pp. 1–8.

[13] A. J. Fehske, F. Richter, and G. P. Fettweis, "Energy efficiency improvements through micro sites in cellular mobile radio networks", in *Proc. 2nd Int. Worksh. Green Commun.*, in conjunction with *GLOBECOM Workshops 2009*, Honolulu, HI, USA, 2009, pp. 1–5.

[14] A. Chockalingam and M. Zorzi, "Energy efficiency of media access protocols for mobile data networks", *IEEE Trans. Commun.*, vol. 46, pp. 1418–1421, 1998.

[15] W. Wang and G. Shen, "Energy efficiency of heterogeneous cellular network", in *Proc. IEEE 72nd Veh. Technol. Conf. Fall VTC-Fall 2010*, Ottawa, Ontariao, Canada, 2010, pp. 1–5.

[16] Malaysian Communications and Multimedia Commission Annual Report, SKMM-MCMC, 2011 [Online]. Available: http://www.skmm.gov.my/skmmgovmy/media/General/pdf/SKMM_2011.pdf

[17] A. A. Abdulkafi *et al.*, "Energy-aware load adaptive framework for LTE heterogeneous network", *Trans Emerging Tel Tech*, vol. 25, no. 9, pp. 943–953, 2014.

[18] S. N. Shahab, T. S. Kiong, and A. A. Abdulkafi, "A framework for energy efficiency evaluation of LTE network in urban, suburban and rural areas", *Australian J. Basic Appl. Sci.*, vol. 7, no. 7, pp. 404–413, 2013.

Suhail Najm Shahab was graduated in 2010 with B.Eng. in Computer Technology Engineering from Al-Hadba'a University College, Mosel, Iraq. In 2013 he received his M.Sc. in Electrical Engineering with specialization in Wireless Communication from University Tenaga Nasional, Malaysia. Currently, he is enrolled as a Ph.D. candidate in Faculty of Electrical & Electronics Engineering, University Malaysia Pahang, Malaysia. He has authored and co-authored numerous publications in international conferences and journals. His research interest includes LTE, energy-efficient cellular networks, adaptive beamforming, smart antenna.

E-mail: 68suhel@gmail.com
Faculty of Electrical and Electronics Engineering
Universiti Malaysia Pahang
26600 Pekan, Pahang, Malaysia

Ayad Atiyah Abdulkafi received the B.Sc. and M.Sc. degrees in Electrical Engineering (major in telecommunications) from Al-Mustansiriya University, Baghdad, Iraq, in 2001 and 2004, respectively. He received his Ph.D. in Wireless Communication Engineering from University Tenaga Nasional (UNITEN), Malaysia in 2014. He is a staff member in College of Engineering, Tikrit University, Iraq. He is currently a postdoc fellow at Multimedia University, Malaysia. His research interests are in wireless communications, including, LTE, radio resource management and optimization, heterogeneous networks, energy-efficient wireless network design and Green Cellular Networks, OFDM, optical wireless communications, Visible Light Communication.

E-mail: al.ayad@yahoo.com
College of Engineering
Tikrit University
Salahaldin, Iraq

Ayib Rosdi Bin Zainun obtained his B.Eng. in Electrical and Electronics from University Technology Malaysia (UTM), Skudai, Johor, Malaysia in 2000. He received his M.Sc. in Engineering (Adaptive Array Antenna) from Nagoya Institute of Technology, Nagoya, Japan (2005). He has completed Ph.D. (Dye-Sensitize Solar Cell) from University Technology MARA (UiTM), Shah Alam, Selangor, Malaysia (2012). He is currently working as a lecturer at Faculty of Electrical and Electronics Engineering, University Malaysia Pahang. He has been a member of various committees for projects of national interest in Malaysia, and he is referee of various scientific journals. His research is mainly centered on the field of adaptive array antenna and applied science (materials for solar cells applications).

Email: ayib@ump.edu.my
Faculty of Electrical and Electronics Engineering
Universiti Malaysia Pahang
26600 Pekan, Pahang, Malaysia

Military route planning in battlefield simulation: effectiveness problems and potential solutions

Zbigniew Tarapata

http://doi.org/10.26636/jtit.2003.4.201

Abstract — Path searching is challenging problem in many domains such as simulation war games, robotics, military mission planning, computer generated forces (CGF), etc. Effectiveness problems in military route planning are related both with terrain modelling and path planning algorithms. These problems may be considered from the point of view of many criterions. It seems that two criterions are the most important: quality of terrain reflection in the terrain model and computational complexity of the on(off)-line path planning algorithm. The paper deals with two above indicated problems of route planning effectiveness. Comparison of approaches used in route planning is presented. The hybrid, terrain merging-based and partial path planning, approach for route planning in dynamically changed environment during simulation is described. It significantly increase effectiveness of route planning process. The computational complexity of the method is given and some discussion for using the method in the battlefield simulation is conducted. In order to estimate how many times faster we can compute problem for finding shortest path in network with *n* big squares (b-nodes) with relation to problem for finding shortest path in the network with *V* small squares (*s*-nodes) acceleration function is defined and optimized.

Keywords — battlefield simulation, route planning, shortest paths, effectiveness problems, computational complexity.

1. Introduction

For many years in military applications a simulated battlefield is used for training military personnel. There are at least three ways to provide the simulated opponent:

- two groups of trainees in simulators may oppose each other (often used);
- human instructors who are trained to behave in a way that mimics the desired enemy doctrine (seldom used);
- computer system that generates and controls multiple simulation entities using software and possibly a human operator.

The last approach is known as a semi-automated force (SAF or SAFOR) or a computer generated force (CGF). CGF is used in military distributed interactive simulation (DIS) systems to control large numbers of autonomous battlefield entities using computer equipment and software rather than humans in simulators.

The advantages of CGF are well-known [17]:

1) they lower the cost of a DIS system by reducing the number of standard simulators that must be purchased and maintained;

2) CGF can be programmed, in theory, to behave according to the tactical doctrine of any desired opposing force, and so eliminate the need to train and retrain human operators to behave like the current enemy;

3) CGF can be easier to control by a single person than an opposing force made up of many human operators and it may give the training instructor greater control over the training experience.

As an inseparable part of CGF, modules for route planning based on the real-terrain models are used. For example in modular semi-automated forces (ModSAF) in module "SAFsim", which simulates the entities, units, and environmental processes the route planning component is located [14]. Moreover, automated route planning will be a key element of almost any automated terrain analysis system that is a component of a military command and control system. In the work [1] authors describe a combined on-road/off-road planning system that was closely integrated with a geographic information system and a simulation system. Routes can be planned for either single columns or multiple columns. For multiple columns, the planner keeps track of the temporal location of each column and insures they will not occupy the same space at the same time. In the same paper the hierarchic route planner as integrate part of predictive intelligence military tactical analysis system (PIMTAS) is discussed. In the paper [8] authors presented an on-going efforts to develop a prototype for ground operations planning, the route planning uncertainty manager (RPLUM) tool kit. They are applying uncertainty management to terrain analysis and route planning since this activity supports the commander's scheme of maneuver from the highest command level down to the level of each combat vehicle in every subordinate command. They extend the PIMTAS [1] route planning software to accomodate results of reasoning about multiple categories of uncertainty. Authors of the paper [3] presented route planning in the close combat tactical trainer (CCTT).

Kreitzberg [11] has developed the tactical movement analyzer (TMA). The system uses a combination of digitized maps, satellite images, vehicle type and weather data to compute the traversal time across a grid cell. TMA can compute optimum paths that combine both on-road and off-road mobility, and with weather conditions used to modify the grid cost factors. The smallest grid size used is approximately 0.5 km. Author uses the concept of a signal propagating from the starting point and uses the traversal time at each cell in the array to determine the time at which the signal arrives at neighboring cells. Other researchers have chosen to decompose the map into regions that are defined by having a constant traversability across the region [1, 9, 16, 19, 20, 27]. The advantage of this approach is that the number of regions will, in general, be far fewer than the number of grid cells. The disadvantages include difficulty in defining the center of the region and the computation difficulties in determining the optimum paths between two adjacent cells. The optimum region-to-region path can be obtained by using either Dijkstra's continuous algorithm (DCA) developed by Mitchell [15]. In many cases, a multiresolution simulation modelling is used to simplify complex battlefield processes [4, 6, 16, 17].

As integrated part of route planning modules the terrain database-based model is being used. Terrain data can be as simple as an array of elevations (which provides only a limited means to estimate mobility) or as a complex as an elevation array combined with digital map overlays of slope, soil, vegetation, drainage, obstacles, transportation (roads, etc.) and the quantity of recent weather. For example in [1] authors describe heterogeneous reasoning and mediator environment system (HERMES) will allow the answering of queries that require the interrogation of multiple databases in order to determine the start and destination parameters for the route planner.

There are a few approaches in which the map (representing a terrain area) is decomposed into a graph [1, 9, 19, 20]. All of them first convert the map into regions of *go* (open) and *no-go* (closed). The *no-go* areas may be considered as obstacles and are represented as polygons. A few ways for consider the map can be used, for example: visibility diagram, Voronoi diagram, straight-line dual of the Voronoi diagram, edge-dual graph, line-thinned skeleton, regular grid of squares, grid of homogeneous squares coded in quadtree system, etc.

Effectiveness problems in military route planning are related both with terrain modelling and path planning algorithms. These problems may be considered from the point of view of many criterions. It seems that two criterions are the most important: quality of terrain reflection in the terrain model (visibility diagram, Voronoi diagram, regular grid of terrain squares, etc.) and computational complexity of the on(off)-line path planning algorithm. The paper deals with above indicated problems of route planning effectiveness.

In the next section we will discuss in details route planning approaches.

2. Comparison approaches used in route planning

It was said in the previous section that we will deal with effectiveness of two problems of battlefield simulation:

- terrain reflection in the terrain model used in battlefield simulation;
- military route planning using one of terrain models.

If terrain models are concerned a few ways for considering the map were listed in the previous section: Voronoi diagram, straight-line dual of the Voronoi diagram (the Delaunay triangulation), visibility diagram, edge-dual graph, line-thinned skeleton, regular grid of squares, grid of homogeneous squares coded in quadtree system.

The polygonal representations of the terrain are often created in database generated systems (DBGS) through a combination of automated and manual processes [19]. It is important to say that these processes are computationally complicated but are conducted before simulation (during preparation process). Typically, an initial polygonal representation is created from the digital terrain elevation data through the use of an automated triangulation algorithm, resulting in what is commonly referred to as a triangulated irregular network (TIN). A commonly used triangulation algorithm is the Delaunay triangulation. Definition of the Delaunay triangulation may be done via its direct relation to the Voronoi diagram of a set, S, of N 2D points: the straight-line dual of the Voronoi diagram is a triangulation of S.

The **Voronoi diagram** is the solution to the following problem: given a set S of N points in the plane, for each point p_i in S what is the locus of points (x, y) in the plane that are closer to p_i than to any other point of S?

The **straight-line dual** is defined as the graph embedded in the plane obtained by adding a straight-line segment between each pair of points of S whose Voronoi polygons share an edge. Figure 1 depicts an irregularly spaced set of points S, its Voronoi diagram, and its straight-line dual (i.e. its Delaunay triangulation).

The **edge-dual graph** is essentially an adjacency list representing the spatial structure of the map. To create this graph, we assign a node to the midpoint of each map edge which does not bound an obstacle (or the border). Special nodes are assigned to the start and goal points. In each non-obstacle region, we add arcs to connect all nodes at the midpoints of the edges which bound the same region. The fact that all regions are convex guarantees that all such arcs cannot intersect obstacles or other regions. Example of the edge-dual graph is presented in Fig. 2.

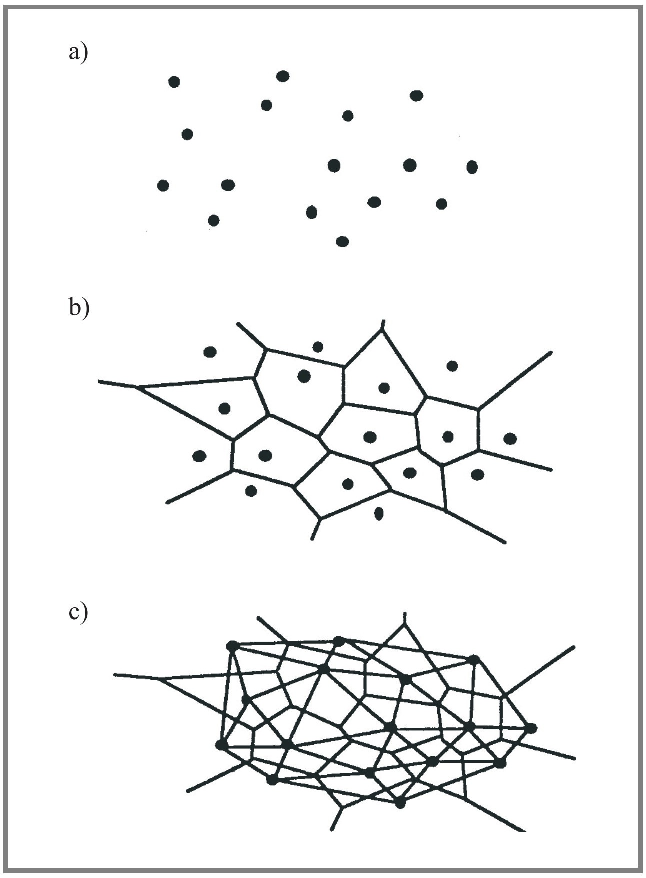


Fig. 1. Voronoi diagram and its Delaunay triangulation [19]: (a) a set *S* of *N* points in the plane; (b) the Voronoi diagram of *S*; (c) the straight-line dual of the Voronoi diagram (the Delaunay triangulation).

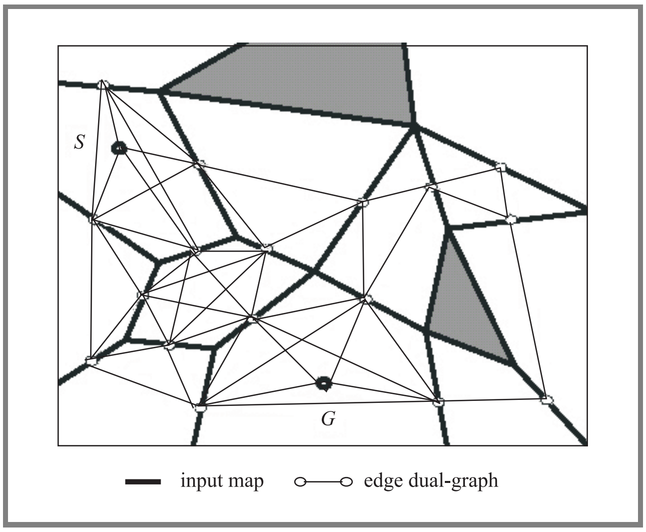


Fig. 2. Edge-dual graph. Obstacles are represented by filled polygons.

The **visibility graph**, is a graph whose nodes are the vertices of terrain polygons and whose edges joint pairs of nodes for which the corresponding segment lies inside polygon. An example is shown in Fig. 3.

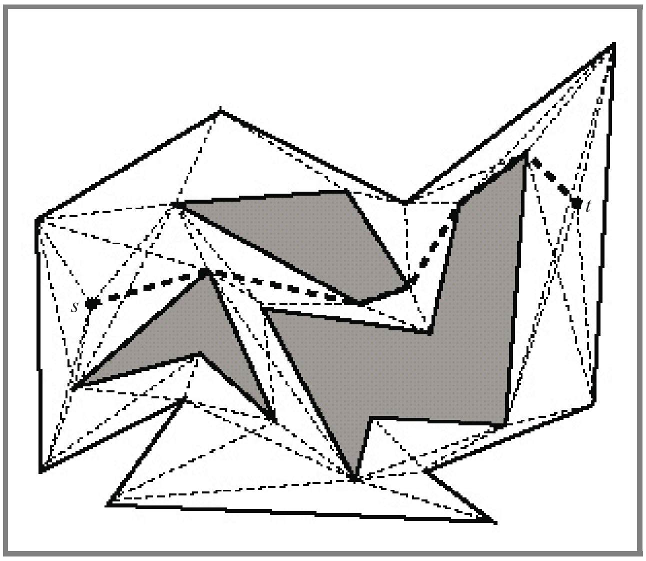


Fig. 3. Visibility graph [15]. There is marked shortest geometric path from source node *s* to destination *t*. Obstacles are represented by filled polygons.

The **regular grid of squares** divides terrain space on the squares with the same size and each square is treated as having homogeneity from the point of view of terrain characteristics. An example of this approach will present in the next sections (see Fig. 6 and Fig. 7).

The **grid of homogeneous squares coded in quadtree system** divides terrain space on the squares with heterogeneous size. The size of square results from its homogeneity according to terrain characteristics. Example of this approach was presented, e.g. in [29].

If paths planning approaches used in battlefield simulation are concerned, there are four main approaches [10]: free space analysis, vertex graph analysis, potential fields, grid based algorithms.

In the **free space approach**, only the space not blocked and occupied by obstacles is represented. For example, representing the center of movement corridors with Voronoi diagrams [19] is a free space approach (see Fig. 1).

Advantage of Voronoi diagrams is that they have efficient representation.

Disadvantages of Voronoi diagrams:

- they tend to generate unrealistic paths (paths derived from Voronoi diagrams follow the center of corridors while paths derived from visibility graphs clip the edges of obstacles);
- the width and trafficability of corridors are typically ignored;
- distance is generally the only factor considered in choosing the optimal path.

In the **vertex graph approach**, only the endpoints (vertices) of possible path segments are represented [15].

Advantages:

- this approach is suitable for spaces that have sufficient obstacles to determine the endpoints.

Disadvantages:

- determining the vertices in "open" terrain is difficult;
- trafficability over the path segment is not represented;
- factors other than distance cannot be included in evaluating possible routes.

In the **potential field approach**, the goal (destination) is represented as an "attractor", obstacles are represented by "repellors", and the vehicles are pulled toward the goal while being repelled from the obstacles.
Disadvantages:

- the vehicles can be attracted into box canyons from which they cannot escape;
- some elements of the terrain may simultaneously attract and repel.

In the **regular grid approach**, a grid overlays the terrain, terrain features are abstracted into the grid, and the grid rather than the terrain is analyzed.
Advantages:

- simplification of the analysis.

Disadvantages:

- "jagged" paths are produced because movement out of a grid cell is restricted to four (or eight) directions corresponding to the four (or eight) neighboring cells;
- granularity (size of the grid cells) determines the accuracy of terrain representation.

A many of route planners in the literature are based on the Dijkstra's shortest path algorithm, A* algorithm [7], geometric path planning algorithms [15] or its variants [12, 13, 18, 26, 31, 32]. For example, A* has been used in a number of computer generated forces systems as the basis of their planning component, to plan road routes [3], avoid moving obstacles [10], avoid static obstacles [18] and to plan concealed routes [14]. Very extensive discussion related to geometric shortest path planning algorithms was presented by Mitchell in [15] (references consist of 393 papers and handbooks). Geometric shortest paths problem is defined as follows: given a collection of obstacles, find an Euclidean shortest obstacle-avoiding path between two given points. Mitchell considers following problems:

- geodesic paths in a simple polygon;
- paths in a polygonal domain (searching the visibility graph, continuous Dijkstra algorithm);
- shortest paths in other metrics (L_p metric, link distance, weighted region metric, minimum-time paths, curvature-constrained shortest paths, optimal motion of non-point robots, multiple criteria optimal paths, sailor's problem, maximum concealment path problem, minimum total turn problem, fuel-consuming problem, shortest paths problem in an arrangement);
- on-line algorithms and navigation without map;
- shortest paths in higher dimensions.

3. Effectiveness problems in route planning

We focus one's attention on path planning algorithms and its effectiveness. Path planning algorithms used in battlefield simulation can be off-line or on-line. Off-line path planning algorithms like A* or Dijkstra's algorithm (listed in the previous section) find the whole solution before starting execution (simulation). They plan paths in advance and usually find optimal solutions. Their efficiency is not considered to be crucial and the moved object just follows the generated path. Although this is a good solution for a static environment, it is rather infeasible for dynamic environments, because if the environment or the cost functions are changed, the remaining path may need to be replanned, which is not efficient for real-time applications (e.g. real-time simulation). Let's recall that standard Dijkstra's algorithm has time complexity $O(V^2)$, where V denotes number of nodes in the graph. This complexity may be improved (if the graph is thin) implementing priority queue as binary heap, obtaining $O(E \cdot \lg V)$, or implementing priority queue as Fibonacci heap, obtaining $O(E + V \cdot \lg V)$, where E describes number of graph's edges.
In Fig. 4 we have graph of calculations time for single shortest path problem using standard Dijkstra algorithm in regular grid network with V nodes[1] (each path was calculated for the left-lower and the right-upper pair of cells (nodes) in grid network (similar to one from Fig. 6).
In Fig. 5 we have graph of calculations time for the same problem but defined as linear programming problem and solved using GAMS solver. From Fig. 4 results that when we must compute shortest path in grid network with e.g. $V = 400$ nodes (grid with size 20×20) then computational time is about 100 ms (for average case) using 1 GFLOPS processor and Dijkstra's algorithm. Let's suppose that we simulate battlefield for two-sided company level on the terrain area with size 16 km^2 (terrain square with 4×4 km size, so 4 km/20 = 200 m is side length for each of 400 cells). If we assume, that each company has 3 platoons then in the same simulation time we must plan movement, in the worst case, for $2 \times 3 = 6$ platoons (as non-divided

[1] Using computer with 1 GFLOPS processor (like PENTIUM III 800 MHz).

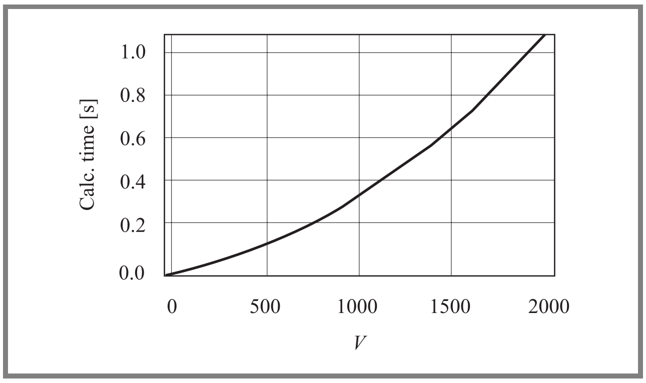


Fig. 4. Calculations time for single shortest path problem using Dijkstra algorithm in regular grid network with V nodes.

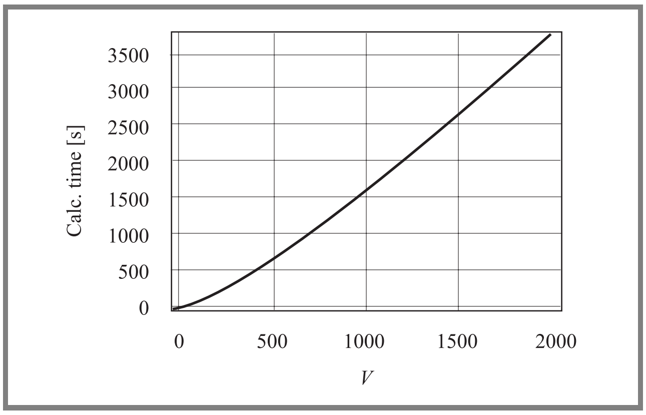


Fig. 5. Calculations time for single shortest path problem defined as linear programming problem and solved using GAMS solver in regular grid network with V nodes.

objects). Because these calculations must be done sequentially (having single processor), so estimation of computational time for all objects is about 6×100 ms = 600 ms. In this case we assumed that all processor power is used for path planning algorithm but it is some simplification, of course. Having, i.e. two-sided battalion fighting ($2 \times 3 \times 3 = 18$ platoons to plan movement in the same simulation time, in the worst case) we need 18×100 ms ≈ 2 s. This delay has significant effect on smoothness of simulation and its visualisation. And we should take into considerations that the network with 20×20 cells is small from among needed in battlefield simulation process.

There are three ways to increase effectiveness of considered problems:

- decreasing the size of terrain-based graph to decrease the computational time of paths planning algorithms [1];
- using specific on-line paths planning algorithms [10, 12, 13, 16, 32];
- using some partial path update approaches [23, 26].

Each of mentioned above ways has some advantages and disadvantages.

Advantage of the first way (**decreasing the size of graph**) is that the number of merged cells into regions will, in general, be far fewer than the number of grid cells. The disadvantages include difficulty in defining the center of the region and the computation difficulties in determining the optimum paths between two adjacent cells.

Some partial path planning algorithms (the second way) plan an off-line path, let the object follow the path, and if any new environment information is gathered, they partially re-plan the existing solution. Similar approach for multi-convoy redeployment in stochastic, dynamically changed environment, was presented in [26, 27]. Disadvantage of this approach is that some times, a small change in the environment may cause re-plan almost a complete path, which may take a long process time (when the network size is big).

The basic idea of **on-line path planning approach** (the third way), in generally, is that the object is moved step-by-step from cell to cell using some heuristic method. This approach is borrowed from movement robots path planning [13, 23, 32]. The decision about the next move (its direction, speed, etc.) depends on the current location of the object and environment status. For example, the idea of RTEF (real-time edge follow) algorithm [32] is to let the object eliminate closed directions (the directions that cannot reach the target point) in order to decide on which way to go (open directions). For instance, if the object has a chance to realize that moving to north and east will not let him reach the goal state, then it will prefer going to south or west. RTEF find out these open and closed directions, so decreasing the number of choices the object has. However, this approach has one basic disadvantage. Namely, in this approach using a few criterions simultaneously to find optimal (or acceptable) path is difficult and it is rather not possible to estimate, in advance, moment of achievement the destination. Moreover, it does not guarantee finding optimal solutions and even suboptimal ones may significanly differ from acceptable.

From this cause, we present in the next section hybrid, cells-merging-based and partial path planning approach for route planning in dynamically changed environment.

Considering route planning in the battlefield simulation we must mention multi-convoy (or multi-object) redeployment and, in consequence, multi-paths planning. Complexity of this process depends on the following conditions [29]:

- count of objects in each convoy (the convoy longer the scheduling of redeployment more complicated);
- have convoys be redeployed simultaneously?
- can convoys be destroyed during redeployment?
- can terrain-based network be destroyed during redeployment?
- have convoys be redeployed through disjoint routes?
- have convoys achieve selected places (nodes) at fixed time?
- do convoys have to start at the same time?

- have convoys determine any action strips for moving?
- can convoys be joined and separated during redeployment?
- have convoys cross through fixed nodes?, etc.

The most often problem related to multi-convoy redeployment is to move a few convoys through disjoint paths simultaneously [25, 30]. Disjoint paths condition results from safety ensuring for moved convoys. In the battlefield simulation finding disjoint paths for moved objects (e.g. tanks inside tank platoon) simplifies its movement because route for each tank do not cross route for each other and we avoid potential collisions. Disjoint paths optimization problem is NP-hard, so some heuristic or other suboptimal approaches are used [2, 21, 25, 31]. Description of some prototype module for maneuvre planning using disjoint paths approach was presented in [28].

4. A new multiresolution approach for increasing route planning effectiveness

Discussed, in the previous section, region approach for terrain included difficulty in defining the center of the region and the computational difficulties in determining the optimum paths between two adjacent cells. In this section we propose some multiresolution-based approach for finding shortest paths in the big grid networks. We assume that we have grid graph $G = \langle \boldsymbol{V},\boldsymbol{A}\rangle$ (see Fig. 7) as representation of terrain squares (see Fig. 6), where $\boldsymbol{V}$ describes set of nodes (squares of terrain), $V = |\boldsymbol{V}|$, $\boldsymbol{A}$ describes set of arcs, $\boldsymbol{A} = \{\langle x,y\rangle \subset \boldsymbol{V}\times\boldsymbol{V} : \text{square } x \text{ is adjacent to square } y\}$.

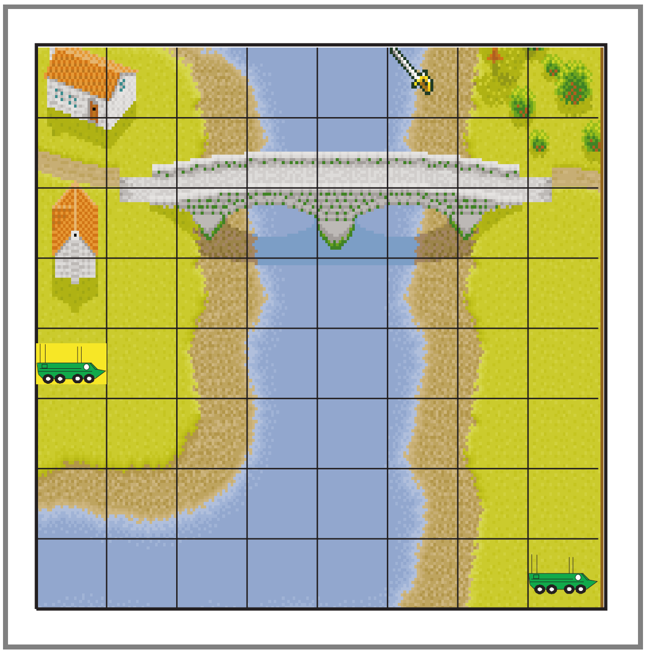


Fig. 6. Terrain space with division on regular grid squares. We want to move object from the right-lower corner to the left side.

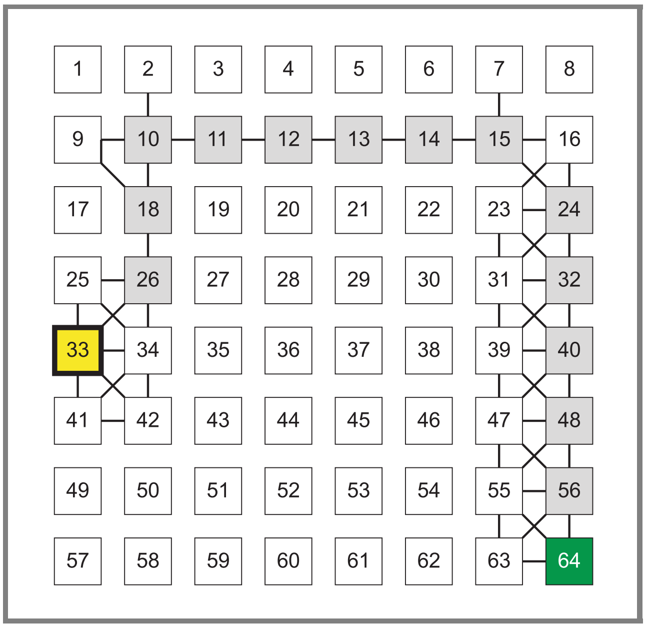


Fig. 7. Grid graph as representation of terrain squares from Fig. 6. There is marked shortest path from node 64 to node 33.

In this graph we may describe some functions (as traversability, visibility, crossing time, crossing probability, detecting probability, etc.) obtaining network as model of movement environment. We assume that for each arc $\langle x,y\rangle \in \boldsymbol{A}$ we have cost $c(x,y)$. The idea of the approach is to merge geographically adjacent small squares (nodes belonging to $\boldsymbol{V}$) into bigger squares (called b-nodes, see Fig. 8) and build b-graph $\overline{G}$ (graph based on the b-nodes, see Fig. 9) using specific transformation. This transformation is based on the assumption that we set arc (b-arc) between two b-nodes $\overline{x} \subset \boldsymbol{V}$, $\overline{y} \subset \boldsymbol{V}$ when exist such two nodes $x \in \overline{x}$, $y \in \overline{y}$ that $\langle x, y\rangle \in \boldsymbol{A}$. In practice, as nodes of $\overline{G}$ graph we will consider strongly connected components of b-nodes. Cost $\overline{c}(\overline{x}, \overline{y})$ of the b-arc $\langle \overline{x}, \overline{y}\rangle \in \overline{\boldsymbol{A}}$ is set on the basis of the biggest cost of some shortest paths calculated inside the subgraph built on the nodes of $\overline{x}$. Next, in the b-graph we find shortest paths between such pairs $\overline{x_s}, \overline{y_t}$ of the b-nodes that source node s and target node t belong to sets $\overline{x_s}, \overline{y_t}$, respectively.

Formal definition of the graph $\overline{G}$ is as follows:

$$\overline{G} = \langle \overline{\boldsymbol{V}},\overline{\boldsymbol{A}}\rangle\,, \tag{1}$$

where:

$\overline{\boldsymbol{V}} = \{\overline{x}_1,\overline{x}_2,\ldots,\overline{x}_n\}$-set of b-nodes, $|\overline{\boldsymbol{V}}| = n$,

$\overline{x}_i = \{x_{i1},x_{i2},\ldots,x_{im}\} \subset \boldsymbol{V}$, $i = \overline{1,n}$,

$\underset{\substack{i,j\\ i\neq j}}{\forall} \overline{x}_i \cap \overline{x}_j = \emptyset$, $i = \overline{1,n}$, $j = \overline{1,n}$, $\bigcup\limits_{i=1}^{n} \overline{x}_i = \boldsymbol{V}$,

$\overline{\boldsymbol{A}} = \left\{ \langle \overline{x},\overline{y}\rangle \subset \overline{\boldsymbol{V}}\times\overline{\boldsymbol{V}} : \underset{x\in\overline{x},y\in\overline{y}}{\exists} \langle x,y\rangle \in \boldsymbol{A} \right\}$.

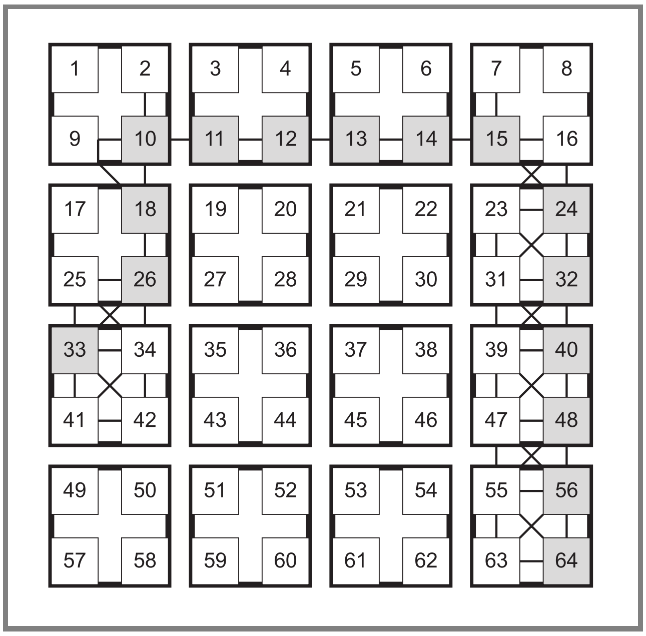


Fig. 8. Merging geographically adjacent small squares from Fig. 7 into $n = 16$ bigger squares (b-nodes).

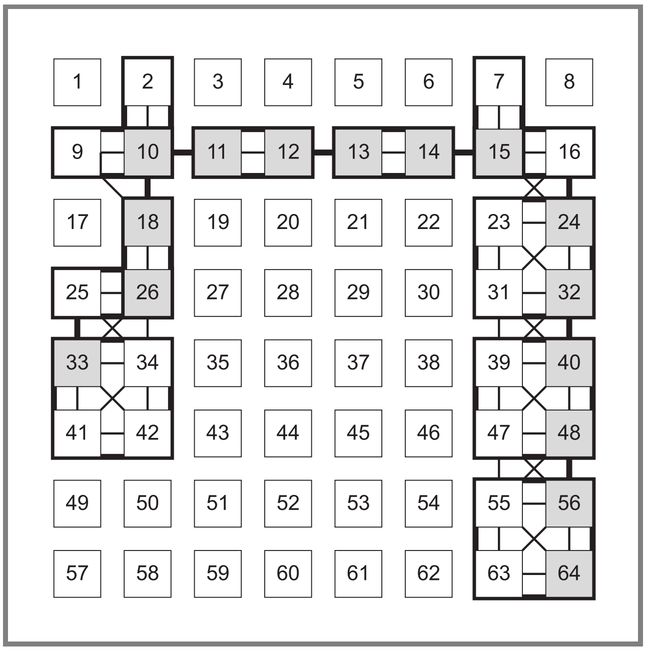


Fig. 9. b-graph for squares merging from Fig. 8. As b-nodes we use strongly connected components of b-nodes from Fig. 8.

Cost function $\overline{c}(\overline{x},\overline{y})$ for b-arc $\langle\overline{x},\overline{y}\rangle$ we determine as:

$$\overline{c}(\overline{x},\overline{y}) = \max_{\{x\in\overline{x}\}} F(x,\overline{y}), \qquad (2)$$

where:

$$F(x,\overline{y}) = \min_{\left\{y\in\overline{y}:\ \underset{z\in\overline{x}}{\exists} \langle z,y\rangle\in\boldsymbol{A}\right\}} L\big(P(x,y)\big),$$

$$L\big(P(x,y)\big) = \sum_{i=0}^{l(P(x,y))-1} c(x_i,x_{i+1}),$$

$$P(x,y) = (x_0 = x, x_1, x_2, \ldots, x_{l(P(x,y))} = y),$$
$$\underset{i=0,l(P(x,y))-1}{\forall} \langle x_i,x_{i+1}\rangle \in \boldsymbol{A}.$$

The merging algorithm for b-graph-based shortest paths planning (MSP-algorithm) is following:

1. merge small squares from graph G (Fig. 7) into n bigger squares (Fig. 8) (n is parameter of the algorithm; we show in further discussion how we can set the optimal value of the n);

2. inside each of the n big squares (b-nodes) determine strongly connected components obtaining at least n subgraphs;

3. set each of subgraphs obtained from the Step 2 as b-nodes and arcs as described by (1) obtaining graph $\overline{G}$ (Fig. 9);

4. find shortest paths between each pair of nodes inside each b-node (subgraph) of $\overline{G}$ to calculate cost $\overline{c}(\overline{x},\overline{y})$ for each arc of $\overline{G}$ using Eq. (2);

5. find shortest path in $\overline{G}$ with cost function $\overline{c}(\cdot,\cdot)$ between such pairs $\overline{x_s},\overline{y_t}$ of b-nodes that source node s and target node t belong to sets $\overline{x_s},\overline{y_t}$, respectively.

It's important to explain that setting in the Step 3 strongly connected components as b-nodes assure that each node inside such component is attainable from each other, so if b-node $\overline{x}$ is connected (through b-arc) with b-node $\overline{y}$ then exist path from each node of $\overline{x}$ to each node of $\overline{y}$.
Let's estimate time complexity of MSP algorithm. We will estimate complexity of each step of the algorithm as follows (we assume that each b-node is strongly connected):

2. determination of strongly connected components in graph G: we have n b-nodes creating n merged subgraphs of G; each subgraph of G has no more than $\lceil\frac{V}{n}\rceil$ nodes, so we have complexity $O\big(n\cdot\lceil\frac{V}{n}\rceil\big) = = O(V)$;
3. we have n b-nodes so we obtain $O(n)$;
4. shortest path problem between each pair nodes in N-nodes graph has complexity $O(N^3)$; if each subgraph of G is strongly connected component of G, then has n b-nodes (creating subgraphs), so each subgraph has $\lceil\frac{V}{n}\rceil$ nodes, hence finding all-pairs shortest paths in single subgraph has complexity $O\Big(\big(\frac{V}{n}\big)^3\Big)$; because we must calculate it n times, so we have $O\Big(n\cdot\big(\frac{V}{n}\big)^3\Big)$;
5. finding shortest paths in graph $\overline{G}$: because $\overline{G}$ has n b-nodes, so using standard Dijkstra's shortest path algorithm we have $O(n^2)$.

We omit the merging Step 1 because, having n, we can prepare this step before simulation. Taking into considerations above estimations we obtain total complexity of the algorithm as $O\left(\frac{V^3}{n^2}+n^2+V\right)$ (we have also omitted $O(n)$ because $n \ll V$).

There is very interesting and important question from the point of view of proposed approach effectiveness: how should we set n to obtain the better effectiveness than for V?

Let's notice that computational complexity of the algorithm based on the network with small squares[2] is $O(V^2)$ and for the algorithm based on the bigger squares is $O\left(\frac{V^3}{n^2}+n^2+V\right)$. It means that, in sense of complexity symbol $O(\cdot)$, the bigger squares approach is better if the following formula is satisfied:

$$\frac{V^3}{n^2}+n^2+V<V^2 \tag{3}$$

or equivalently, when

$$n^4-(V^2+V)n^2+V^3<0\,. \tag{4}$$

Solving this inequality we obtain, that $n \in [n_1, n_2]$, where

$$n_1=\sqrt{\frac{V^2+V-\sqrt{(V^2+V)^2-4V^3}}{2}}\,, \tag{5}$$

$$n_2=\sqrt{\frac{V^2+V+\sqrt{(V^2+V)^2-4V^3}}{2}}\,. \tag{6}$$

For example, for the graph from Fig. 7 $(V=64)$ we obtain $n_1 \approx 8$, $n_2 \approx 64$.

In order to estimate how many times faster we compute problem for finding shortest path in the network with n big squares $\left(\left(\frac{V^3}{n^2}+n^2+V\right)\right)$ with relation to problem for finding shortest path in the network with V small squares $(O(V^2))$ we may formulate acceleration function as follows[3]:

$$A(V,n)=\frac{V^2}{\frac{V^3}{n^2}+n^2+V}\,. \tag{7}$$

Exemplified graphs of $A(V,n)$ are shown in Figs. 10 and 11.

Having grid network with V squares (nodes) we can formulate following optimization problem: to find such cardinal n^*, for which

$$A(V,n^*)=\max_{n\in[n_1,n_2]} A(V,n)\,, \tag{8}$$

where n_1, n_2 are described by formulas (5) and (6).

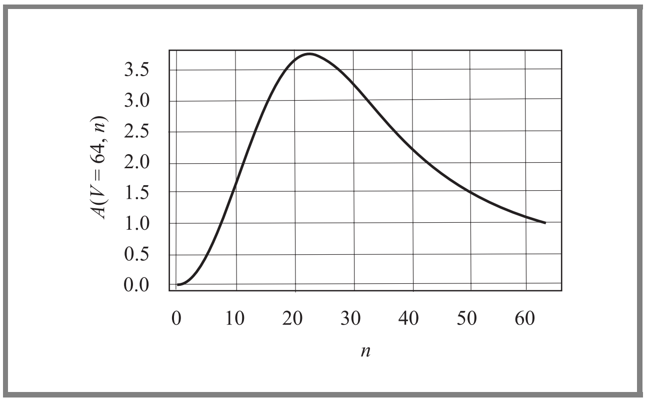


Fig. 10. Graph of $A(V,\ n)$ function for the network with $V=64$ nodes.

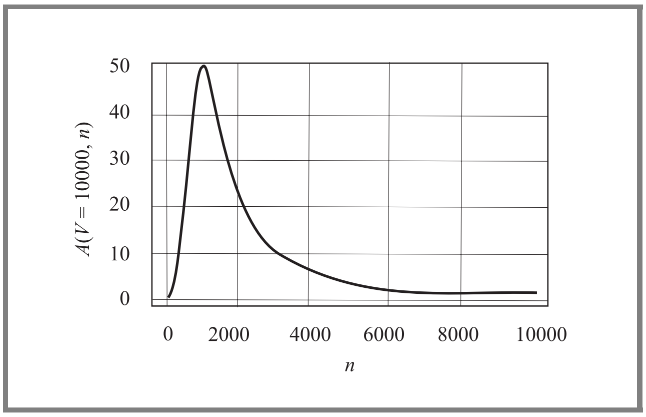


Fig. 11. Graph of $A(V,\ n)$ function for the network with $V=10000$ nodes.

Let's notice, that function (7), omitting constraint for n integer, has real nonnegative maximum for the value $n^*=\sqrt[4]{V^3}$. It may be easily shown that for each $V>0$, $n^* \in [n_1,\ n_2]$. In practice we are interested in such value $n^{**} \approx n^*$, that square root of $\frac{V}{n^{**}}$ is cardinal number (it results from the fact that each of n^{**} big squares consist of $\frac{V}{n^{**}}$ small squares and in grid structure of the network a big square has $\sqrt{\frac{V}{n^{**}}} \times \sqrt{\frac{V}{n^{**}}}$ small squares).

In Table 1 the influence of V on n^*, n^{**} and $A(V,n^{**})$ is shown. It is easy to observe the best acceleration of shortest path algorithm using presented approach in regular grid network with V nodes may be approximated by value $A(V,n^{**}) \approx \frac{1}{2}\sqrt{V}$.

Let's notice that from presented estimations and Table 1 result that for 400-nodes grid graph considered at the beginning of the previous section movement planning for two-sided battalion fighting (for 18 platoons) will be done in time 18 × 100/9.5 ms ≈ 200 ms.

[2] Using standard Dijkstra's shortest paths algorithm (without modifications increasing its effectiveness).

[3] Exact to complexity estimation symbol $O(\cdot)$.

Table 1
Influence of V on n^*, n^{**} and $A(V,n^{**})$

V	n^*	n^{**}	$\lceil \frac{V}{n^{**}} \rceil$	$A(V,n^{**})$
100	32	25	4	4.3
400	90	100	4	9.5
900	164	225	4	12.3
1600	253	169	9	14.7
2500	354	256	9	20.4
10000	1000	1089	9	49.0
40000	2828	2500	16	96.8
90000	5196	5625	16	147.9
160000	8000	6400	25	181.4
250000	11180	10000	25	243.7

5. Conclusions

The approach presented in the paper gives possibilities to significantly decrease computational time in terrain-based route planning when the terrain environment is represented by regular grid of squares. This approach may be applied, i.e. for route planning in the simulated battlefield.

The estimations of presented algorithm effectiveness may be improved through a few ways. The first way is to use in time complexity estimations the best known shortest-path algorithm estimation $(O(E+V\cdot\lg V))$ instead complexity of standard Dijkstra's algorithm $(O(V^2))$ because the regular grid graph is thin (maximal number of direct successors for any node is 8), so $O(8V+V\lg V)<O(V^2)$ nearly for all V (exactly for $V>11$). The second way is to improve Step 4 of the algorithm because it seems to be unnecessary determinations all-pairs shortest paths in each b-nodes (subgraphs). It's seems that is enough to determine shortest paths between "outside" nodes of b-nodes because only these nodes are used to link b-node with another. Moreover, to confirm presented estimations it is essential to conduct calculations in real grid graphs.

Presented suggestions may be contribution for further works.

References

[1] J. R. Benton, S. S Iyengar, W. Deng, N. Brener, and V. S. Subrahmanian, "Tactical route planning: new algorithms for decomposing the map", in *Proc. IEEE Int. Conf. Tools for AI*, Herndon, 1995, pp. 268–277.

[2] A. Bley, "On the complexity of vertex-disjoint length-restricted path problems", Konrad-Zuse-Zentrum fur Informationstechnik, Berlin, 1998 (see also: http://www.zib.de/PaperWeb/abstracts/SC-98-20/).

[3] C. Campbell, R. Hull, E. Root, and L. Jackson, "Route planning in CCTT", in *Proc. 5th Conf. Comput. Gener. Forc. Behav. Repres.*, Tech. Rep., Institute for Simulation and Training, 1995, pp. 233–244.

[4] C. G. Cassandras, C. G. Panayiotou, G. Diehl, W.-B. Gong, Z. Liu, and C. Zou, "Clustering methods for multi-resolution simulation modeling", in *Proc. Conf. Enabl. Technol. Simul. Sci., Int. Soc. Opt. Eng.*, Orlando, USA, 2000, pp. 37–48.

[5] C. Cooper, A. Frieze, K. Melhorn, and V. Priebe, "Average-case complexity of shortest-paths problems in the vertex-potential model", *Rand. Struct. Algor.*, vol. 16, pp. 33–46, 2000.

[6] P. K. Davis, J. H. Bigelow, and J. McEver, "Informing and calibrating a multiresolution exploratory analysis model with high resolution simulation: the interdiction problem as a case history", in *Proc. 2000 Winter Simul. Conf.*, 2000, pp. 316–325.

[7] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths", *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, no. 2, pp. 100–107, 1968.

[8] J. James, B. Sayrs, J. Benton, and V. S. Subrahmanian, "Uncertainty management: keeping battlespace visualization honest", http://citeseer.nj.nec.com/ 386770.html

[9] L. Joe and P. M. Feldman, "Fundamental research policy for the digital battlefield", Res. Rep. DB-245-A, RAND Co., Santa Monica, USA, 1998.

[10] C. R. Karr, M. A. Craft, and J. E. Cisneros, "Dynamic obstacle avoidance", in *Proc. Conf. Distrib. Interact. Simul. Syst. Simul. Train. Aerosp. Envir., Int. Soc. Opt. Eng.*, Orlando, USA, 1995, pp. 195–219.

[11] T. Kreitzberg, T. Barragy, and B. Nevin, "Tactical movement analyzer: a battlefield mobility tool", in *Proc. 4th Join Tactic. Fus. Symp.*, Laurel, 1990.

[12] B. Logan, "Route planning with ordered constraints", in *Proc. 16th Works. UK Plann. Schedul. Spec. Int. Group*, Durham, UK, 1997.

[13] B. Logan and A. Sloman, "Agent route planning in complex terrains", Tech. Rep. CSRP-97-30, University of Birmingham, School of Computer Science, Birmingham, 1997.

[14] M. Longtin and D. Megherbi, "Concealed routes in ModSAF", in *Proc. 5th Conf. Comput. Gener. Forc. Behav. Repres.*, Tech. Rep., Institute for Simulation and Training, 1995, pp. 305–314.

[15] J. S. B. Mitchell, "Geometric shortest paths and network optimization", in *Handbook of Computational Geometry*, J. R. Sack and J. Urrutia. Elsevier Science Publ., B.V. North-Holland, Amsterdam, 1999.

[16] D. K. Pai and L. M. Reissell, "Multiresolution rough terrain motion planning", Department of Computer Sciences, University of British Columbia, Tech. Rep. TR 94-33, Vancouver, 1994.

[17] M. D. Petty, "Computer generated forces in distributed interactive simulation", in *Proc. Conf. Distrib. Interact. Simul. Syst. Simul. Train. Aerosp. Envir., Int. Soc. Opt. Eng.*, Orlando, USA, 1995, pp. 251–280.

[18] S. Rajput and C. Karr, "Unit route planning", Tech. Rep. IST-TR-94-42, Institute for Simulation and Training, Orlando, USA, 1994.

[19] G. A. Schiavone, R. S. Nelson, and K. C. Hardis, "Interoperability issues for terrain databases in distributed interactive simulation", in *Proc. Conf. Distrib. Interact. Simul. Syst. Simul. Train. Aerosp. Envir., Int. Soc. Opt. Eng.*, Orlando, USA, 1995, pp. 89–120.

[20] G. A. Schiavone, R. S. Nelson, and K. C. Hardis, "Two surface simplification algorithms for polygonal terrain with integrated road features", in *Proc. Conf. Enabl. Technol. Simul. Sci., Int. Soc. Opt. Eng.*, Orlando, USA, 2000, pp. 221–229.

[21] A. Schrijver and P. Seymour, "Disjoint paths in a planar graph – a general theorem", *SIAM J. Discr. Math.*, no. 5, pp. 112–116, 1992.

[22] H. Sherali, K. Ozbay, and S. Subrahmanian, "The time-dependent shortest pair of disjoint paths problem: complexity, models and algorithms", *Networks*, no. 31, pp. 259–272, 1998.

[23] A. Stentz, "Optimal and efficient path planning for partially-known environments", in *Proc. IEEE Int. Conf. Robot. Automat., ICRA'94*, vol. 4, pp. 3310–3317.

[24] P. D. Stroud and R. C. Gordon, "Automated military unit identification in battlefield simulation", LAUR-97-849, *SPIE Proc.*, vol. 3069, Los Alamos National Laboratory, Los Alamos, 1997.

[25] Z. Tarapata, "Algorithm for simultaneous finding a few independent shortest paths", in *Proc. 9th Eur. Simul. Symp., ESS'97, Soc. Comput. Simul.*, Passau, Germany, 1997, pp. 89–93.

[26] Z. Tarapata, "Simulation method of aiding and estimation of transportation columns movement planning in stochastic environment", in *Proc. 13th Eur. Simul. Multiconf., Soc. Comput. Simul. Int.*, Warsaw, Poland, 1999, pp. 613–619.

[27] Z. Tarapata, "Computer simulation of individual and grouped military objects redeployment", *Bull. Milit. Univ. Technol.*, no. 1, pp. 147–162, 2000.

[28] Z. Tarapata, "Computer tool for supporting and evaluating convoys redeployment planning", *Oper. Res. Decis.*, no. 1, pp. 91–107, 2000.

[29] Z. Tarapata, "Some aspects of multi-convoy redeployment modelling and simulation", in *Proc. 21st AFCEA Eur. Symp. & Exposit.*, Prague, 2000 (compact disk publication).

[30] Z. Tarapata, "Modelling, optimisation and simulation of groups movement according to group pattern in multiresolution terrain-based grid network", in *Proc. Reg. Conf. Milit. Commun. Inform. Syst.*, Zegrze, Poland, 2001, vol. I, pp. 241–251.

[31] Z. Tarapata, "Fast method for redeploying multi-convoy in multiresolution grid network", *Bull. Milit. Univ. Technol.*, 2003 (in press).

[32] C. Undeger, F. Polat, and Z. Ipekkan, "Real-time edge follow: a new paradigm to real-time path search", SCS Publications, 2001 (see also: http://citeseer.nj.nec.com/489498.html).

Zbigniew Tarapata has graduated from Cybernetics Faculty at Military University of Technology (MUT) in Warsaw. He received his Master's degree in computer science in 1995 and Doctor's degree in the same field, in 1998. From 1995 to July 1999 he worked as an assistant and since July 1999 – as a senior lecturer at Operations Research Division of Institute of Mathematics and Operations Research of Cybernetics Faculty of MUT. His scientific interests and work are related to the following subjects: mathematical and simulation modelling of systems; transport optimization; combat modelling, simulation, optimization and prediction; graph and network optimization; algorithms effectiveness; methods and tools of multicriteria decision making and supporting.

e-mail: ztarap@isi.wat.waw.pl
Institute of Mathematics and Operations Research
Faculty of Cybernetics
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Traffic Engineering in Software Defined Networks: A Survey

Mohammad R. Abbasi[1], Ajay Guleria[2], and Mandalika S. Devi[1]

[1] *Department of Computer Science and Application, Panjab University, Chandigarh, India*
[2] *Computer Center, Panjab University, Chandigarh, India*

http://doi.org/10.26636/jtit.2016.4.757

Abstract—An important technique to optimize a network and improve network robustness is traffic engineering. As traffic demand increases, traffic engineering can reduce service degradation and failure in the network. To allow a network to adapt to changes in the traffic pattern, the research community proposed several traffic engineering techniques for the traditional networking architecture. However, the traditional network architecture is difficult to manage. Software Defined Networking (SDN) is a new networking model, which decouples the control plane and data plane of the networking devices. It promises to simplify network management, introduces network programmability, and provides a global view of network state. To exploit the potential of SDN, new traffic engineering methods are required. This paper surveys the state of the art in traffic engineering techniques with an emphasis on traffic engineering for SDN. It focuses on some of the traffic engineering methods for the traditional network architecture and the lessons that can be learned from them for better traffic engineering methods for SDN-based networks. This paper also explores the research challenges and future directions for SDN traffic engineering solutions.

Keywords*—*application awareness, Software Defined Networking, traffic engineering.

1. Introduction

A major problem with underlying communication network is the dynamic nature of the network applications and their environment. This means that the performance requirements of the transferred data flows, like Quality of Service (QoS), can vary over time. The applications operate in a wide range of environments, i.e. wired and wireless with a variety of networking devices. For the applications to perform effectively, the underlying network should be flexible enough to dynamically change in response to any changes in the application requirements and their environment. The current approaches are either based on static or overprovisioned overlay networks, or require the applications to change in accordance with the network performance.

An important way to address this problem is through traffic engineering (TE). It is the process of analyzing the network state, predicting and balancing the transmitted data load over the network resources. It is a technique used to adapt the traffic routing to the changes in the network condition. The aim of traffic engineering is to improve network performance, QoS and user experience, by efficient use of resources, which can reduce operation cost too. The QoS techniques assign the available resources to the prioritized traffic to avoid congestion for this traffic. However, these techniques do not provide additional resources to the traffic that requires QoS. The traditional routing techniques do not provide any mechanism to allocate network resources in an optimal way.

To address this problem the research community started working on traffic engineering and proposed new ways to improve network robustness in response to the growth of traffic demands. Traffic engineering reduces the service degradation due to congestion and failure, e.g. link failure. Fault tolerance is an important property of any network. It is to ensure that if a failure exists in the network, still the requested data can be delivered to the destination.

Computer networks consist of numerous networking devices, such as switches, middle boxes (e.g. firewalls) and routers. Traditional network architecture is distributed, as shown in Fig. 1, where each networking device has both the control plane and the data plane. The control plane is the intelligent part of networking devices. It makes decision about forwarding and routing of data-flow. The data plane is the part of a networking device that carries user traffic. It executes the control plane's commands and forwards the data.

Network operators have to manually configure these multi-vendor devices to respond to a variety of applications and event in the network. Often they have to use limited tools such as command line interface (CLI) and sometimes scripting tools to convert these high-level configuration policies into low-level policies. This makes the management and optimization of a network difficult, which can introduce errors in the network. Other problems with this architecture can cause oscillations in the network, since control planes of the devices are distributed, innovation is difficult because the vendors prohibit modification of the underlying software in the devices.

To overcome these problems, the idea of network programmability was introduced, particularly with the introduction of Software Defined Networking (SDN) [1]. SDN allows a network to be programmed so that its behavior can

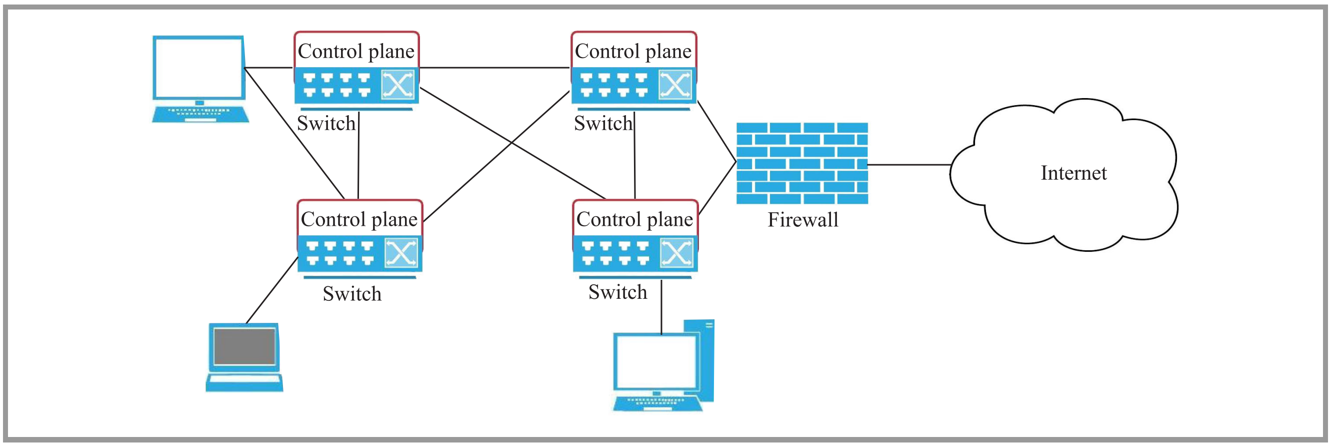


Fig. 1. Traditional network architecture.

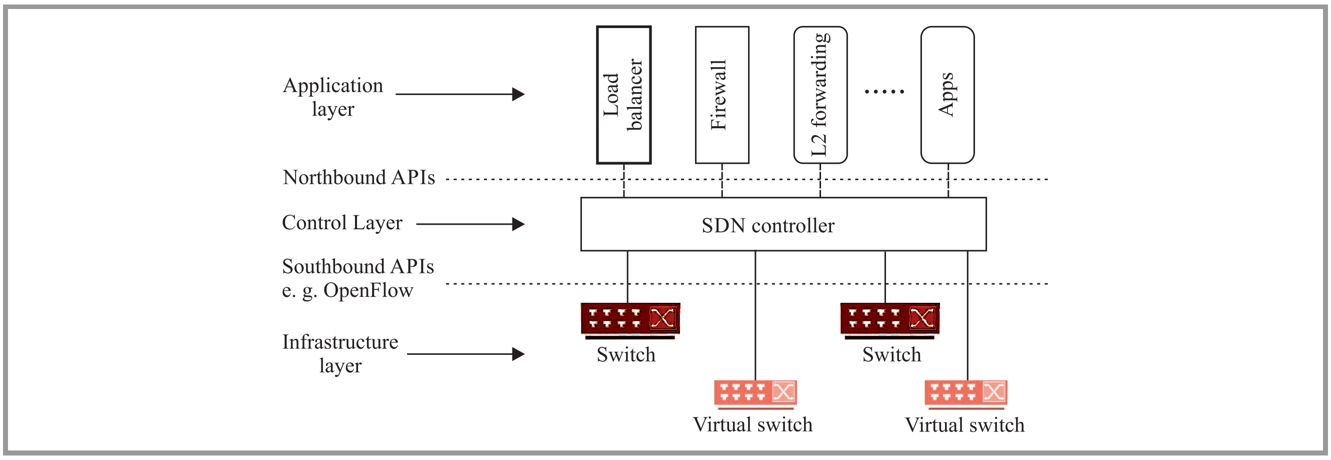


Fig. 2. An example of SDN architecture.

be changed actively on demand and in a fine-grained manner. It is a new networking model, where the control plane and the data plane are separated. The idea behind SDN is to simplify network management and enable innovation, i.e. to develop and deploy new network applications and services with ease, also to manage and optimize network performance through high-level policy enforcement.

To optimize these heterogeneous networks, both classic networks and SDN-based networks, a number of TE techniques have been introduced. Most are based on tweaking wide area TE and routing mechanism, such as Equal Cost Multi-Path routing (ECMP), Intermediate System to Intermediate System (IS-IS), and Multi protocol Label Switching (MPLS) [2], [3].

From traffic engineering point of view, even though these techniques perform well, they suffer from several limitations such as, they take routing decision locally, and it is difficult to change the link weights dynamically. In addition, while sending traffic these techniques consider few criteria, such as link capacity.

SDN separates the control plane and data plane of networking devices and introduces a well-defined interface, the OpenFlow protocol [4], between the two planes. The SDN architecture (Fig. 2) and the OpenFlow takes the intelligence, control functions, out of networking devices and place them in a centralized servers called controller, and provides centralized control over a network. The SDN/OpenFlow controller acts as an operating system for the network. It executes the control applications and services, such as routing protocols and L2 forwarding. This configuration abstracts the underlying network infrastructure. Therefore, it enables the applications and network services to treat the network as a logical entity.

One of the most widely used SDN enabler is the OpenFlow v.1.3 protocol. It allows the controller to manage the OpenFlow switches. The OpenFlow switches contain one or more flow tables, a group table, and a secure OpenFlow channel (Fig. 3). The flow tables and the group table are used for packet lookup and then to forward the packets. The OpenFlow channel is an abstraction layer. It establishes a secure link between each of the switches and the controller via the OpenFlow protocol. This channel abstracts the underlying switch hardware. As of OpenFlow version 1.5, a switch can have one or more OpenFlow channels that are connected to multiple controllers.

SDN is, generally, a flow-based control strategy. Through the OpenFlow a controller can define how the switches should treat the flows. In a SDN when a source node sends

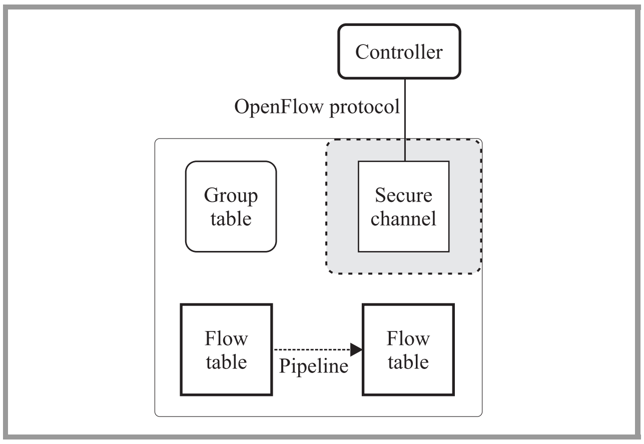


Fig. 3. Main OpenFlow switch components.

data to the destination, the switch sends the first packet to the controller, since it doesn't know how to treat this packet. The controller calculates the path for this packet and installs the appropriate rules in the switches on the packet's path.

The new networking paradigm, SDN, has introduced new characteristics such as:

- separation of the control plane functionality, and the data plane functionality;
- centralized architecture allows the controller to have a central view of the deployed network. The controller has the global view of the network devices, servers, and virtual machines;
- network programmability, SDN provides an open standard, which allows external applications to program the network;
- facilitates innovation, new protocols and control applications can be introduced because OpenFlow provides the required abstractions, so we do not need to know the switch internals and configuration;
- flow management, through the OpenFlow a controller can define flows in different granularity, and how the switches should treat the flows.

The rest of the paper surveys some of the TE techniques, and it is organized as follows: Section 2 provides some of the TE mechanisms available for the classic network architecture and the assimilation from them. Section 3 describes an overview of SDN TE solutions. In Section 4, research challenges and future directions are discussed. Sections 5 and 6 conclude the paper.

2. Review of Classic Traffic Engineering Techniques

Classic traffic engineering techniques are based on tweaking wide area TE and routing mechanism such as ECMP or existing routing protocols such as IS-IS or MPLS [2], [3], [5], [6]. The Open Shortest Path First (OSPF) and IS-IS routing protocols do not adapt to the changes in the network condition because the link weights are static and these protocols lack any performance objectives while selecting the paths. The traffic engineering extensions to IS-IS and OSPF standard, extends these protocols by incorporating the traffic load while selecting a path. In these approaches during link state advertisements, routers advertise the traffic load along with link costs. After routers exchange link costs and traffic loads, then they calculate the shortest path for each destination. These standards require the routers to be modified to collect and exchange traffic statistics [5], [6].

Fortz *et al*. [7] propose a traffic engineering mechanism that monitors network wide view of the traffic pattern and network topology, then changes the link weights accordingly. This mechanism is based on the interior gateway protocols, like IS-IS. The authors says that classic inter-domain gateway protocols are effective traffic engineering tools in a network, and ensure robustness in terms of scalability and failure recovery. The introduced mechanism keeps the router and routing protocols unchanged. The mechanism is a centralized approach where it monitors the network topology and traffic, then optimizes the link costs to provide the best path possible to address the network goals. Routing protocols, like OSPF, select the path with minimum cost. If multiple paths with the same minimum cost are available then the traffic can be equally distributed among these paths. This is the concept behind ECMP. As depicted in Fig. 4, ECMP is a routing technique which balances the load over multiple paths by routing the packets to multiple-paths with equal cost. Various routing protocols such as OSPF and IS-IS explicitly support ECMP routing [8].

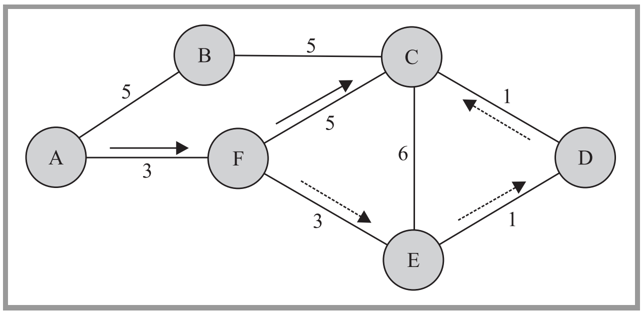


Fig. 4. An example of ECMP – there are two paths with equal cost to the destication node C, i.e. (A, F, C) and (A, F, E, D, C).

Multi-protocol Label Switching, MPLS, provides a tunneling mechanism. It creates end-to-end connections between the nodes. MPLS can integrate short path labels with IP routing mechanism, where the ingress routers assign short fixed labels to the packets, instead of long network addresses. The networking devices use this label to forward the packets to the destination through label-switched path (LSP). This reduces the routing table lookup overhead. The MPLS based traffic engineering, MPLS-TE, first reserve the resources for end-to-end path and then transfer the data. It establishes a labeled switched path over links with sufficient bandwidth. This technique assures that enough resources are available for a flow. Since MPLS-TE works on available bandwidth in one aggregated class, it does not support QoS [9]. To provide QoS capability DiffServe-aware MPLS-TE techniques have been introduced, which combine both the Differentiated Services (DiffServ) and MPLS traffic engineering techniques to provide QoS [10]. Dongli *et al.* [9] analyze the QoS performance of DiffServe-aware MPLS traffic engineering techniques. The experimental results show that DiffServ-aware MPLS-TE can provide good QoS for traffics such as VoIP and other data, but due to the variable bit-rate property of the video data, these techniques cannot guarantee QoS for video data. As compared with conventional routing protocol MPLS is more flexible in selecting paths, since it sets up virtual circuit paths to send the traffic. The disadvantage of MPLS is that network operators need to manage the resource allocation to each path, and change the network configuration to adjust the path to the traffic condition. Because MPLS-TE transfers the aggregated traffics along allocated LSPs, it suffers from scalability and robustness [11]. In MPLS-TE it is necessary to use backup links so that if any link fails the traffic can be transferred through different paths.

An important way to balance the traffic over network resources is to disseminate the traffic over multiple paths. Gojmerac *et al.* [12] introduce an adaptive multi-path routing, which allows dynamic traffic engineering. Unlike other solutions, using global network information, the proposed technique focuses on local information in each node. This means the routers exchange information about links only to their immediate neighbors. So the nodes only have the information regarding their neighbors. During multi-path routing any neighboring node which is closer to the destination has a smaller cost than the current node. This neighboring node is considered as a viable candidate for the next hop. The advantage of taking routing decision based on local information is that it can reduce the signaling and memory overhead. The downside to the approach is, since the nodes do not have the global knowledge of the network state, it may not result in optimum routing of the traffic. Also due to the inherent limitation of the traditional network architecture it cannot adapt to the rapid changes in the traffic pattern and it can cause oscillation in the network.

Frank *et al.* in a [13] propose a content-aware traffic engineering technique for content distribution/delivery networks. The content providers duplicate the contents over distributed server infrastructures to provide better services to the users in different locations. The authors argue that it is essential for the content providers to know network topology and measure network state before mapping user request to the servers, which can introduce new challenges such as assigning users to the servers and performing traffic engineering. ISPs have the knowledge of the individual links status and network topology. This information can separate the server selection task from content delivery task, and help the content providers to focus on mapping the user to a server that provides better user experience. The introduced traffic engineering uses the information provided by ISPs along with the user's location to dynamically adapt to the traffic demand for the contents on the servers. This framework focuses on the traffic demand rather than routing, and uses the knowledge of the content providers (e.g. server status), and ISPs' knowledge (e.g. the network state and the user's location). For this reason this framework can complement the existing routing protocols and traffic engineering because it emphasizes on traffic demand rather than on traffic routing. Routing protocol such as OSPF and IS-IS are used to produce a routing matrix. With this matrix it tries to adjust a set of flow demands to reduce the maximum link utilization. The results of the experiments show that this framework has improved the user experience while reducing maximum link utilization and traffic delay.

Several energy-aware traffic engineering solutions have been proposed in [14]–[16]. These solutions incorporate traffic engineering to reduce the energy cost while trying to keep the network performance unaffected.

Vasic *et al.* [16] introduce an online traffic engineering technique. It spreads the load among multiple paths to reduce the energy consumption without affecting traffic rate. It presumes that energy-aware hardware is used in the network. These devices are capable to adjust its operating rate to its utilization, also they can sleep whenever it is possible to save energy. To enhance energy saving and keep the transfer rate steady, it transfers the data over multiple paths. The authors propose a number of techniques where they shift data to the links with low energy consumption, or they try to remove the traffic from as many possible links to allow the links and routers to sleep.

Most of the discussed approaches agree on the point that to engineer traffic in an efficient way a network-wide approach is required. When short-term changes happen in traffic volume the traffic engineering solution should quickly decide on how to route the traffic to different paths to balance link utilization. Under such circumstances where traffic pattern changes frequently, it is important for the traffic engineering solution to be stable. Otherwise, it can cause oscillation. Traffic oscillation can have a number of undesirable effects on the network, for example, switch-buffer overflow, out-of-order packets, poor allocation of network resources to the users, traffic delay and service degradation [17]. The solutions that have the above characteristics are difficult to implement in the traditional network architecture since we need to have access to global information in real-time, which is a tedious work in this paradigm. To find an optimal solution, most of the proposed solutions are based on local measurements, i.e. require the networking devices to decide independently on how to send the packets. In the traditional networks, generally, the link costs are kept static for a long period. Since the link cost is fixed, the traffic is transferred through the same path for a long period, until the link costs are changed.

For a traffic engineering technique to have an optimum effect on the network, it should have the following characteristics:

- it should utilize multi-path diversity in the network,
- it should make routing decisions based on the global view of the network,
- it should consider the flow values.

3. Review of Traffic Engineering Techniques in SDN

In SDN-based networks the controller can dynamically change the network state, for example, in traditional networks the link cost for routing protocols such as IS-IS are kept static for a long period. If congestion happens in the network it may lead to poor delivery of data till the link costs are changed or the problem is resolved. However, in SDN these values can be changed more dynamically to adapt to the changes. More innovative routing mechanism can be implemented, or the existing routing protocols can be modified, so that they can change dynamically as per network state to enhance resource utilization, avoid failure and congestion, and improve QoS. With the advances in SDN several traffic engineering techniques have been introduced by the research community. Table 1 summarizes some of the TE techniques in SDN.

To connect their Data Centers across the world and meet their performance requirements, Google introduced a Software Defined WAN architecture called B4 [18]. B4 is designed to resolve the problems in Wide Area Network (WAN) such as reliability, failures, and performance. It assigns bandwidth to the competing services, dynamically shifts traffic pattern, and overcomes network failure. B4 is designed to allow rapid deployment of new or standard protocols and control functions. One of such introduced functionalities is a traffic engineering mechanism, which allows applications to dynamically adapt in response to changes in the network behavior or failure. This architecture employs the routing and traffic engineering as separate services. The TE is layered on top of the routing protocols. This enables the network with a fallback strategy. If the TE service faced with a serious problem, it would be stopped so that the packets are forwarded using short path forwarding mechanism. This architecture consists of 3 logical layers:

- global layer, allows centralized control of the entire WAN through logically centralized applications such as the Central Traffic Engineering server (CTE) and SDN gateway (it allows centralized control of the network);
- site controller layer which includes the OpenFlow controller and network control applications such as routing services;
- switch hardware layer includes the switches, and performs traffic forwarding.

CTE is responsible for tasks such as measuring the unoccupied network bandwidth for multi-path forwarding, assigning and adjusting resource demands among the services, and actively relocating traffic from failed links and switches. SDN gateway provides the network topology graph for CTE. CTE uses this graph to compute the aggregated traffic at site to site edges. Then, an abstract of the computed result is fed to TE optimization algorithm to fairly allocate resources among the competing application groups/services. To achieve fairness it allocates resources using Min-Max fairness technique. Based on the applications' priority it allocates bandwidth to the applications. It uses hashed-based ECMP to balance the load among multiple links.

Hedera [19] is introduced to make an effective use of the bandwidth in a data center. Hedera detects the elephant-flows at the edge switches. The Hedera implementation uses periodic pulling, where it collects statistics every five seconds to detect large flows. At first switches send a new flow using its default flow matching rules on one of its equal-cost paths, until the flow size grows and meet the threshold. Then, the flow is marked as elephant-flow. The default threshold is 10% of network interface controller (NIC). At this point Hedera's central scheduler uses its global view of the network and calculates a better path for the flow and route the traffic. To effectively use the bandwidth the scheduler calculates the path in a way that it is non-conflicting, and it can accommodate the flow. This method can improve the bandwidth utilization, but because it uses periodic pulling, it can cause high resource utilization and overhead.

The main design goal of DevoFlow [20] is to improve network scalability and performance by keeping the flows in the data plane as much as possible without losing the centralized view of the network. This reduces the interaction between control plane and data plane. DevoFlow uses aggressive use of wildcards to reduce the controller and switches interactions. Therefore, switches take routing decision locally, while controller manages the overall control of the network and routes the significant flows, i.e. elephant-flows. It uses techniques such as packets sampling to collect switch statistic and detect the elephant-flows. The flows that have transferred a certain number of bytes is marked as large flow. The suggested threshold is 1–10 MB. In the beginning DevoFlow forwards the traffic using DevoFlow's multi-path wildcard rules. When an elephant flow is detected the controller will calculate the path that is least congested, and re-route the traffic to this path.

The flow detection mechanisms used in Hedera and Devoflow have high resource overhead. To overcome this problem Mahout [21] modifies the end-hosts to detect elephant-flows. It uses a shim layer in the operating system to mark the significant flows. The shim layer monitors the TCP socket buffer and marks the flows when in a given period the buffer exceeds the rate threshold. It

Table 1
Overview of traffic engineering techniques in SDN

Technique	Description	Routing	Comments
B4 [18]	• it uses a centralized TE, layered on top of the routing protocols, • to achieve fairness it allocates resources using Min-Max fairness technique.	• it uses hashed-based ECMP to balance the load among multiple links.	• if TE service can be stopped so that the packets are forwarded using short path forwarding mechanism.
Hedera [19]	• detects the elephant-flows at the edge switches, • if threshold is met, i.e. 10% of NIC bandwidth, the flow is marked as elephant flow, • uses periodic pulling, every 5 s.	• uses the global view of network and calculate the better paths, which are non-conflicting, for the elephant flows.	• it achieves 15.4 b/s throughput, • achieves better optimal bisection of bandwidth of network, in comparison to ECMP, • periodic pulling can cause high resource utilization in switches.
DevoFlow [20]	• detects the elephant-flows at the edge switches, • if threshold is met, i.e. 1–10 MB, it marks the flow as elephant-flow.	• it uses aggressive use of wild carded OpenFlow rules, and a static multi-path routing algorithm to forward the traffic.	• it can improve throughput up to 32% in CLOS network.
Mahout [21]	• detects the elephant-flows at end-host using a shim layer, the default threshold is 100 k, and then the flow is marked as elephant-flow, • it uses in-band signaling to inform the controller about the elephant-flows.	• it computes the best path for elephant-flow; otherwise it forwards other flows using ECMP, • it calculates the path that is least congested by pulling the elephant-flow statistics and link utilization from switches.	• it can detect elephant flow, if threshold is 100 k, in 1.53 ms, • it has 16% better bisection than ECMP.
MicroTE [22]	• detect the elephant flows at end-host, • it calculates the mean of traffic matrix between ToR-ToR, if the mean and traffic is between δ of each other, default is 20%, then it is predictable.	• uses short term predictability to route the traffic on multiple paths, • the remaining flows are routed by the EMCP scheme with heuristic threshold.	• if traffic is predictable it perform close to optimal performance otherwise it performs like ECMP.
MiceTrap [23]	• it addresses the mice-flows, • uses end-host elephant-flow detection to distinguish between mice-flows, and elephant-flows.	• it aggregates the mice-flows to improve scalability, • it route the mice-flows using a weighted multi.	• N/A.
Rethinking Flow Classification in SDN [26]	• it is a tag-based classification, • source-edge switch tags the packets based on the application classes.	• the tag is also an identifier for matching & forwarding the packets	• it is 3 ms faster than the OpenFlow field matching, • it requires introduction of new API's to the data plane.
Atlas [25]	• it classifies each application uniquely. • it uses C5.0 machine learning tool to classify the applications, • it requires user to install agents on their mobile devices to collect information to train ML trainer.	• it routes the flow based on applications, and network requirements.	• it has about 94% accuracy, • requires extension to OpenFlow.
MSDN-TE [32]	• it gathers network state information and considers the actual path's load to forward the flows on multiple paths.	• it dynamically selects the best shortest path among the available paths.	• it has better performance over other forwarding mechanisms such as Shortest Path First, • it reduced download time by 48%.

uses an in-band signaling mechanism to mark the flows as elephant-flows and inform the controller about the significant flows. Mahout uses ECMP to route normal traffic. When an elephant-flow is detected the controller calculates the best path for this flow. To calculate the best paths the controller pulls the elephant-flow statistics and link utilization from the switches to select the least congested path. This method can detect the elephant-flows faster, with lower processing overhead than other method. But, it requires modification of the end-hosts.

In [22] Benson *et al*. present a traffic engineering mechanism for data center network called MicroTE, which uses an end-host elephant flow detection to detect the elephant flows. It exploits short-term prediction, and quickly adapts to the changes in traffic pattern. To efficiently handle the network load, it takes advantage of multiple paths in the network and coordinates traffic scheduling by using global view of traffic across the available network paths. The authors argue that the traffic nature of data center networks is bursty, and during long-run time the traffic is unpredictable, above 150 s, but it is predictable in short-time scale of 1–5 s. The TE methods for ISPs do not perform well in data center environments because they work on the granularity of hours, but TE for Data Centers should work on granularity of seconds.

Unlike MicroTE, MiceTrap [23] incorporates the end-host flow detection to handle mice-flows and uses OpenFlow group table (multi-path group type) to aggregate the incoming mice-flows for each destination. The authors believe that TE mechanism, which handles elephant-flows, can cause congestion to mice-flows, i.e. short-lived flows. Also the resources should be distributed according to flow values. Managing mice-flows using ECMP and giving preference to elephant flows can degrade QoS. MiceTrap architecture consists of end-host elephant flow detection module, multi-path aggregates implemented in OpenFlow switch, and a controller. It uses the kernel-level shim layer approach to mark the elephant flow detection. The shim layer method monitors the TCP socket buffer and marks the flows when in a given period the buffer exceeds the specified threshold. Multi-path Mice-flow Aggregator, aggregates the incoming mice-flows for each destination. This reduces the rules for traffic management because if each mice-flow is managed by an exact flow rule, it will cause a bottleneck and limit the scalability. The advantage of using group table is that it saves bandwidth since one single group message can update a set of flows when the traffic distribution is changed. It uses a weighted routing algorithm which forwards aggregated traffic into multiple paths by considering the current network load while calculating the paths.

These are effective solutions for data center networks, but they share the network resources based on the flow size and do not consider the flow-value. An important way to consider the flow-values is to classify the applications. Two promising techniques for application classification are Deep Packet Inspection (DPI) and Machine Learning (ML) classification method. In comparison to techniques such as port-based classification, these techniques have a high classification accuracy. DPI methods inspect the payload of packets and search for known patterns, keywords or regular expressions that are characteristic of a given application. These methods are more accurate, but with higher overhead (in terms of memory and processing). ML methods exploit several flow-level features to classify the traffic. To classify the flows these methods look for known flow behavior such as packet counts, data bytes, TCP flags, etc. [24].

In the work [25] Qazi *et al*. try to investigate how to integrate application awareness in SDN-based networks and how to classify traffics with high accuracy. A framework called Atlas is introduced, which is capable of classifying the traffic in the network and enforcing higher layer policies. The presented framework uses a ML tool called C5.0 to classify the flows based on the application types. It shows 94% accuracy. The Atlas framework can classify each specific application. It can classify each VoIP application uniquely rather than classifying them as a common VoIP flows. Such framework should be scalable so the application detection and enforcing application-aware policy is done in a smooth and uninterrupted manner. They have deployed the framework in the HP Lab wireless network. It requires the users to install software agents on their mobile devices. These agents collect information such as active network sockets, Netstat logs for each application. The agents send this information to the controller, where the controller runs machine learning trainer. The OpenFlow switch statistics are extended to store first n packet size of each flow and announce it to the controller. The controller collects such flow features along with the information sent by the agents to train the ML tool by using the ML trainer.

Hamid *et al*. in [26] introduce a tag-based classification architecture, where the source-edge switches tag the packets based on the application classes. This way the network operator can apply different policies for each of the application classes. The tag is also used as an identifier for matching the packets which reduce the matching overhead. After a tagged packet is delivered to the destination edge switch, the switch removes the tag and performs the required actions, if there is any action, and sends the packet to the destination host. The experimental result shows this tag-based approach is 3 ms faster than the hash-based field matching methods like OpenFlow field matching, and reduces processing overhead. To solve the backward compatibility, unlike MPLS, the tag is added to the end of packet instead of its middle. This way, if the variable length packet is supported, there is no need for whole packet parsing. Otherwise, the whole packet should be parsed. The downside of this approach is that it requires changes in the switch internal by introducing a new API to the switch data plane. This API is an application layer processor in the data plane.

A promising way to address challenges and problems in distributed environments, such as a computer network, is with the help of intelligent agents, i.e. Multi-Agent System (MAS) and mobile agents. Bieszczad *et al*. [27] describes how intelligent agents can be used to facilitate network management. It explains the potential of Intelligent Agents to tackle various difficulties in different network management areas such as fault management, security, performance management, accounting, etc. SDN provides a good platform for the agents to tackle such difficulties.

In [28] Skobelev *et al.* propose a task-scheduling system for SDN-based networks. This system incorporates MAS to overcome task-scheduling problems in the distributed systems, i.e. where the servers and computational resources are distributed. The MAS task scheduler associates the basic system entities with an agent. It consists of three main agents:

- task agent represents the task that should be processed with minimum cost by a server in the network,
- resource agent provides the system with a server to process tasks,
- commutator agent is responsible for providing information about network state and task allocation to the nodes.

This system is developed in C#, .NET framework, as a Windows application.

The research [29]–[31] show that by providing application-awareness and feedback from clients' machines to the network, the user Quality of Experience (QoE), and resource utilization can be improved. These works use agents on user side to collect information (like audio/video quality, waiting time, etc.) and send this information to the controller to adjust the network state accordingly to improve users' QoE.

To address traffic forwarding and traffic engineering in SDN, Dinh et al. [32] introduced a multipath-based forwarding traffic engineering mechanism called MSDN-TE. The goal of this mechanism is to forward the traffic in such a way that it avoids congestion on any link in the network. MSDN-TE dynamically selects the best available shortest paths and forwards the incoming traffic. This TE mechanism gathers network state information and considers the actual path's load to forward the flows on multiple paths. The MSDN-TE is a module which extends OpenDayLight controller. It consists of three components:

- a monitoring function which is used for gathering information about network states and flows in the network; for example, flow's static, link utilization, network topology, etc. The path matrices are refreshed every 10–15 s;
- the TE algorithm calculates n number of paths, which have the lowest traffic load, between the source and destination node. To select the shortest paths Eppstein [33] algorithm is used;
- the actuating function supports TE algorithm module. It takes certain actions and dynamically allots flows to the selected paths. Compared with the Shortest Path First and spanning tree, MSDN-TE shows better performance in regard to download time, delay and packet drops. For example, it reduced packet drops by 72.9% in AGIS [34] simulated topology and more than 90% in Abilene [34] simulated topology.

4. Traffic Engineering Research Challenges

As SDN becomes widely used, the research community and industry introduce new protocols and control applications like TE mechanisms. However, like any new technology the potential of SDN to simplify and improve network management comes with new challenges that need to be addressed. In this section, some of the challenges and future directions for traffic engineering in SDN are discussed, namely, fault-tolerance, energy-awareness, flow-update scheduling and consistency, and data-flow scheduling and dissemination.

4.1. High-availability

Fault tolerance is an important feature of any computer network. It means if an unexpected error or problem happens, like the failure of a link or a switch, the services in the network will continue to be accessible. The faults in a network can cause congestion and packet loss. These conditions can last for seconds due to the time it takes for TE mechanism to respond to the faults and update the network, i.e. update the topology and switches. In a SDN-based network two types of failure are control plane failure and data plane failure, like failure of links and switches. Besides physical/logical failure of control plane, control plane failure can also refer to a situation when the controller fails to update switches in the right time, so the switches continue to forward the traffic with the old rules. This can lead to congestion because the link capacity is not considered. There are two approaches to address the faults in the network: proactive where the paths are calculated and reserved beforehand, and reactive where the resources are not reserved until failure happens. The paths are calculated dynamically or decided in advance. Proactive approach has faster fault recovery since the paths are already calculated. When a fault occurs there is minimum interaction between the controller and the switch. This approach is about 5 times faster than reactive approach [35], [36]. But, reactive approach has a lower cost because the link capacity is not reserved, so it requires less memory in the control plane.

In [37] Liu *et al.* have introduced a proactive fault management TE mechanism, known as forward fault correction (FFC), which handles both data plane and control plane faults. In this approach if the number of faults is smaller than a configurable bound f, it can ensure protection against failure and congestion in the network. Depending on the value of f, and the traffic distribution flowing in the network, FFC provisions a certain amount of link capacity to avoid failure in the network. However, since the link capacity is pre-provisioned it can have lower throughput. Kim *et al.* in [38] introduced a fault-tolerance framework called CORONET, which uses a centralized controller to forward packets and can work with any network topology. CORONET recovers from switch or link failures in a short period. It uses multi-path routing methods. Its architecture consists of modules to discover network topology, route planning, shortest-route path calculation, and traffic assignment. To simplify packet forwarding, and minimize the number of flow-rules CORONET uses VLANs mechanism in the switches. Therefore, it is also scalable.

A traffic engineering framework should detect faults in the network and re-route the sensitive applications' traffic by avoiding the failed areas to allow these applications to work seamlessly and avoid service degradation. SDN characteristics such as failover mechanism introduced in OpenFlow protocol, global view of network, and its capability to dynamically change network state facilitate failure recovery. However, it is still challenging since the controller needs to calculate the new paths and install the flow-rules in the switches. The TE should achieve low communication overhead with a trade-off between the latency and memory usage.

4.2. Energy-awareness

To guarantee QoS in a network, high-end networking devices that have a high power consumption are used. To reduce delay and increase reliability, these resources are usually over-provisioned to increase network capacity. However, this leads to concerns about greenhouse gas emission and power wastage. A number of researches show that non-negligible percentage of world power consumption and CO_2 emission is due to information and communication technologies [39]. This motivated the researchers to propose new algorithms and devices to address these difficulties [14]–[16]. These techniques adapt the network elements' active time according to the traffic load.

The techniques proposed for classical network architecture are not as effective as they can be. There are few studies on energy-aware techniques for SDN-based networks. In [40] Giroire *et al.* have introduced an energy-aware routing technique for SDN that gathers traffic matrix, calculate the routing paths to guarantee QoS, and put the idle links and nodes into sleeping mode. This technique considers both memory limitation of routers and link capacity. SDN features such as centralized view of network and network programmability can help to introduce new efficient centralized energy-aware TE techniques that allow a network to dynamically adapt to the traffic load and network condition with the goal of achieving good performance and use network resources effectively to reduce power consumption. The centralized TE mechanism can shut down a set of switches and links, when the traffic demand is low to reduce power utilization while satisfying user experience.

4.3. Data-flow Scheduling

After the rules are installed in the switches, a switch will match and send the incoming packets to the destination. An important way to ensure QoS in a network is flow scheduling, where the packets that require better QoS are scheduled and transferred first. Flow priority is the main scheduling method [41], [42]. In this method, the flows with higher priorities are sent first. SDN provides a good platform to introduce new software-controlled flow schedulers that are capable of flow-oriented multi-policy scheduling. This abstraction can help to introduce advanced network configurability.

Bo-Yu *et al.* [43] have introduced an Iterative Parallel Grouping Scheduling Algorithm, IPGA-scheduling. It is designed to address the prioritized flow scheduling problem, which is required for QoS differentiation among different prioritized flows, and also energy saving in data center networks. This system is a Compute Unified Device Architecture (CUDA) system within SDN controller. CUDA is a parallel computing architecture developed by Nvidia for graphics processing.

In [42] Rifai *et al.* proposes coarse-grained scheduling. The authors argue that the data center networks and the Internet traffic nature mostly consist of short flows and most of the flows are carried by TCP. Therefore, the emphasis of this scheduling system is on flow-size scheduling. This system uses switch's OpenFlow flow statistics to identify the flow-size along with multiple queues per port to implement 802.1p QoS. The 802.1p standard delegates 8 queues per port. Two size-based schedulers are introduced. Both of these schedulers have two queues per port, and it is assumed that they are managed by strict priority scheduler.

Using a scheduler can improve the network performance. The majority of the schedulers are developed around the idea of "one size fits all", or consider only the flow size, and the flow value. The type of the application is ignored. These approaches examine, mostly, packet's priority and port workload while assigning the flows to a port. For example, in a VoIP network, the VoIP applications need to have the highest priority to ensure QoS. Even though priority-based solutions can address these requirements, they require precise configuration of the network which is time-consuming and error-prone.

4.4. Flow-update

In an operating network, the controller may change the configuration of the switches several times through flow-updates. Flow-update refers to updating the current switch configurations, forwarding rules, with new configurations. Flow-updates are important for various tasks such as fault management, adapting to changes in traffic pattern, etc. Flow-update is a challenging task since improper update of multiple flows can cause problems such as congestion, service degradation, and inconsistency in the network. Hence, flow-update scheduling is an important issue to be addressed. If a new rule is assigned for each flow it can increase the resource cost (e.g. processing and memory) in both the data plane and control plane. Also, the time that it takes for a flow to be added in a switch adds to the latency. There should be a tradeoff between load-balancing and latency.

The most common approaches for flow-update scheduling problem is the two-phase update mechanism, where controller first installs the new forwarding-rules into the switches, if all packets that require the old rules are transferred, then the new installed rules will be used and the old forwarding-rules are removed from the switches. Compared to the one-phase approach, the advantage of this approach is that the chance of the controller to fail in updating the switches is lower. However, Li *et al.* [44] argue that the two-phase update mechanism is not effective, since it does not consider the switch's flow-table size. Thus, to address the multi-flow update problem, a step-by-step approach is introduced. This problem is formulated as an optimization problem to minimize the maximum link utilization, which is an important network performance metric. In this approach the controller schedules the flow updates and then updates each flow step by step, i.e. the path of a flow is changed to the new one in a step, so if there are *n* flow updates then the process is completed in *n* steps. This method considers both the link capacity and the flow-table size.

In [45] Mahajan believe that flow-update, to ensure consistency, has a number of properties such as loop free, packet drop free, switch memory limitation, load balancing, etc. Depending on the type of a network, different consistency, or combination of the consistency properties are needed, for example load-balancing or loop free network.

5. Conclusion

In this paper we have reviewed literature in the field of traffic engineering for both traditional network architecture and SDN, and examined some of the TE challenges and future directions. SDN is a new networking paradigm which simplifies the network management and enables innovation. It tries to address many problems in the traditional network architecture by simplifying network management through centralized management of a network, introducing network programmability, and providing a global view of a network and its state. New traffic engineering techniques are required to exploit these features for better control and management of traffic. Different TE mechanism should be included in SDN to control congestion and manage traffic for different applications in various QoS-sensitive scenarios such as video or business data, and to provide required QoS while balancing the load among the available resources in a network. To improve the network load handling, a traffic engineering mechanism should enable a network to react in real-time and classify a variety of traffic types from different applications. Routing optimization is one of the main techniques in TE. It should take advantage of multiple paths in the network and coordinate traffic scheduling by using global view of traffic across the available network paths. Beside load-balancing and optimization of resources, other components of TE are QoS and resilience form failure. SDN is currently capable of enforcing policy for lower layers, i.e. Layer 2-4, but not many studies have explored the higher layer policy enforcement. By identifying the packets sent by the applications to the network, it is possible to enforce higher layer, application layer, policies. Higher layer policy enforcement can help to engineer resilient and flexible network. Such networks can be optimized for each application to provide a good QoS and improve user experience. The authors described how an end-host flow detection mechanism and Multi-Agent System can improve network performance and scalability while reducing complexity.

6. Future Work

In terms of future work, authors plan to propose an efficient traffic engineer framework, which makes the SDN-based networks more application-aware. In this work a multi-agent based software framework consisting of a number of algorithms for application classification, and data scheduling and dissemination will be developed. The agents are responsible for application classification of user's traffic. Then, the data scheduling and dissemination algorithms will calculate the best path and order to process and forward the flow to the destination. All these modules will work together to ensure high-availability, load-balancing and optimizing resource utilization, and also to ensure high-QoS rating for QoS sensitive applications. This framework can help to automate the network configuration to achieve high QoS for the desired applications. By combining techniques such as scheduling, application classification, and MAS, a network can deliver better services.

References

[1] B. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks", *IEEE Commun. Surv. & Tutor.*, vol. 16, no. 3, pp. 1617–1634, 2014.

[2] B. Fortz and M. Thorup, "Internet traffic engineering by optimizing OSPF weights", in *Proc. 19th IEEE Ann. Joint Conf. of the IEEE Comp. & Commun. Soc. INFOCOM 2000*, Tel Aviv, Israel, 2000, vol. 2, pp. 519–528.

[3] X. Xiao, A. Hannan, B. Bailey, and L. M. Ni, "Traffic engineering with MPLS in the Internet", *Network*, vol. 14, no. 2, pp. 28–33, 2000.

[4] O. N. Foundation, "OpenFlow – open networking foundation" [Online]. Available: https://www.opennetworking.org/sdn-resources/openflow (accessed Aug. 23, 2016).

[5] K. Ishiguro, A. Lindem, A. Davey, and V. Manral, "Traffic engineering extensions to OSPF Version 3", RFC 5329, IETF Trust, 2008 [Online]. Available: https://tools.ietf.org/html/rfc5329

[6] T. Li and H. Smit, "IS-IS extensions for Traffic Engineering", RFC 5305, IETF Trust, 2008 [Online]. Available: https://tools.ietf.org/html/rfc5305

[7] B. Fortz, J. Rexford, and M. Thorup, "Traffic engineering with traditional IP routing protocols", *Commun. Mag.*, vol. 40, no. 10, pp. 118–124, 2002.

[8] D. Thale and C. Hopps, "Multipath issues in unicast and multicast next-hop selection", RFC 2991, IETF Trust, 2000 [Online]. Available: https://tools.ietf.org/html/rfc2991

[9] D. Zhang and D. Ionescu, "QoS performance analysis in deployment of DiffServ-aware MPLS Traffic Engineering", in *Proc. 8th ACIS Int. Conf. on Software Engin., Artif. Intell., Netw., & Parallel/Distrib. Comput. SNPD 2007*, Qingdao, China, 2007, vol. 3, pp. 963–967.

[10] F. Le Faucheur *et al.*, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, IEFT Trust, 2002 [Online]. Available: https://tools.ietf.org/rfc/rfc3270.txt

[11] I. F. Akyildiz *et al.*, "A new traffic engineering manager for DiffServ/MPLS networks: design and implementation on an IP QoS Testbed", *Computer Commun.*, vol. 26, no. 4, pp. 388–403, 2003.

[12] I. Gojmerac, T. Ziegler, F. Ricciato, and P. Reichl, "Adaptive multipath routing for dynamic traffic engineering", in Proc. Global Telecommun. Conf. GLOBECOM'03, San Francisco, CA, USA, 2003, vol. 6, pp. 3058–3062.

[13] I. Poese, B. Frank, G. Smaragdakis, S. Uhlig, A. Feldmann, and B. Maggs, "Enabling content-aware traffic engineering", *ACM SIGCOMM Comp. Commun. Rev.*, vol. 42, no. 5, pp. 21–28, 2012.

[14] M. Zhang, C. Yi, B. Liu, and B. Zhang, "GreenTE: Power-aware traffic engineering", in *Proc. 18th IEEE Int. Conf. on Netw. Protocols ICNP 2010*, Kyoto, Japan, 2010, pp. 21–30.

[15] E. Amaldi, A. Capone, L. G. Gianoli, and L. Mascetti, "A MILP-based heuristic for energy-aware traffic engineering with shortest path routing", in *Network Optimization*, J. Pahl, T. Reiners, and S. Voß , Eds. *LNCS*, vol. 6701, pp. 464–477. Springer, 2011.

[16] N. Vasić and Dejan Kostić, "Energy-aware traffic engineering", in *Proc. of the 1st Int. Conf. on Energy-Effic. Comput. & Netw. e-Energy'10*, Passau, Germany, 2010, pp. 169–178.

[17] L. Zhang and D. Clark, "Oscillating behavior of network traffic: A case study simulation", *Internetworking: Res. and Exper.*, vol. 1, no. 2, pp. 101–112, 1990.

[18] S. Jain *et al.*, "B4: Experience with a globally-deployed software defined WAN", *ACM SIGCOMM Comp. Commun. Rev.*, vol. 43, no. 4, pp. 3–14, 2013.

[19] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic Flow Scheduling for Data Center Networks", in *Proc. 7th USENIX Symp. on Netw. Syst. Design & Implemen. NSDI'10*, San Jose, CA, USA, 2010, vol. 10, pp. 19–19.

[20] A. R. Curtis, J. C. Mogul, J. Tourrilhes, P. Yalagandula, P. Sharma, and S. Banerjee, "DevoFlow: scaling flow management for high-performance networks", *ACM SIGCOMM Comp. Commun. Rev.*, vol. 41, no. 4, pp. 254–265, 2011.

[21] A. R. Curtis, W. Kim, and P. Yalagandula, "Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection", in *Proc. 30th IEEE Int. Conf. Comp. Commun. IEEE INFOCOM 2011*, Shanghai, China, 2011, pp. 1629–163.

[22] T. Benson, A. Anand, A. Akella, and M. Zhang, "MicroTE: Fine grained traffic engineering for data centers", in *Proc. 7th Conf. on Emerg. Networking Experim. & Technol. Co-NEXT'11*, Tokyo, Japan, 2011, p. 8.

[23] R. Trestian, G.-M. Muntean, and K. Katrinis, "MiceTrap: Scalable traffic engineering of datacenter mice flows using OpenFlow", in *IFIP/IEEE Int. Symp. on Integr. Netw. Managem. IM 2013*, Ghent, Belgium, 2013, pp. 904–907.

[24] S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, and M. Mellia, "Reviewing traffic classification", in *Data Traffic Monitoring and Analysis*, E. Biersack, C. Callegari, and M. Matijasevic, Eds. *LNCS*, vol. 7754, pp. 123–147. Springer, 2013.

[25] Z. A. Qazi *et al.*, "Application-awareness in SDN", *ACM SIGCOMM Comp. Commun. Rev.*, vol. 43, no. 4, pp. 487–488, 2013.

[26] H. Farhadi and A. Nakao, "Rethinking flow classification in SDN", in *Proc. IEEE Int. Conf. on Cloud Engin. IC2E 2014*, Boston, MA, USA, 2014, pp. 598–603.

[27] A. Bieszczad, B. Pagurek, and T. White, "Mobile agents for network management", *Commun. Surveys*, vol. 1, no. 1, pp. 2–9, 1998.

[28] P. O. Skobelev, O. N. Granichin, D. S. Budaev, V. B. Laryukhin, and I. V. Mayorov, "Multi-agent tasks scheduling system in software defined networks", *J. of Physics: Conf. Series*, vol. 510, no. 1, p. 012006, 2014 (doi: 10.1088/1742-6596/510/1/012006).

[29] M. Jarschel, F. Wamser, T. Hohn, T. Zinner, and P. Tran-Gia, "SDN-based application-aware networking on the example of YouTube video streaming", in *Proc. 2nd Eur. Worksh. on Softw. Defined Netw. EWSDN 2013*, Berlin, Germany, 2013, pp. 87–92.

[30] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, "Towards network-wide QoE fairness using openflow-assisted adaptive video streaming", in *Proc. ACM SIGCOMM Worksh. on Future Human-Centric Multim. Netw. FhMN 2013*, Hong Kong, China, 2013, pp. 15–20.

[31] H. Nam, K.-H. Kim, J. Y. Kim, and H. Schulzrinne, "Towards QoE-aware video streaming using SDN", in *Proc. Global Commun. Conf. GLOBECOM 2014*, Austin, TX, USA, 2014, pp. 1317–1322.

[32] K. T. Dinh, S. Kukliński, W. Kujawa, and M. Ulaski, "MSDN-TE: Multipath Based Traffic Engineering for SDN", in *Intelligent Information and Database Systems. Asian Conference on Intelligent Information and Database Systems*, N. T. Nguyen, B. Trawiński, and R. Kosala, Eds. Springer, 2016, pp. 630–639.

[33] D. Eppstein, "Finding the *k*-shortest paths", *SIAM J. Comput.*, vol. 28, pp. 652–673. 1999.

[34] S. Knight, H. X. Nguyen, N. Falkner, R. Bowden, and M. Roughan, "The Internet topology zoo", *IEEE J. on Selec. Areas in Commun.*, vol. 29, no. 9, pp. 1765–1775, 2011.

[35] S. Sharma, D. Staessens, D. Colle, M. Pickavet, and P. Demeester, "Enabling fast failure recovery in OpenFlow networks", in *Proc. 8th Int. Worksh. on the Des. of Reliable Commun. Netw. DRCN 2011*, Kraków, Poland, 2011, pp. 164–171.

[36] D. Staessens, S. Sharma, D. Colle, M. Pickavet, and P. Demeester, "Software defined networking: Meeting carrier grade requirements", in *Proc. 18th IEEE Worksh. on Local & Metropolitan Area Networks LANMAN 2011*, Chapel Hill, NC, USA, 2011, pp. 1–6.

[37] H. H. Liu, S. Kandula, R. Mahajan, M. Zhang, and D. Gelernter, "Traffic engineering with forward fault correction", *ACM SIGCOMM Comp. Commun. Rev.*, vol. 44, no. 4, pp. 527–538, 2014.

[38] H. Kim, J. R. Santos, Y. Turner, M. Schlansker, J. Tourrilhes, and N. Feamster, "Coronet: Fault tolerance for software defined networks", in *Proc. 20th IEEE Int. Conf. on Network Prot. ICNP 2012*, Austin, TX, USA, 2012, pp. 1–2.

[39] R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, "Energy efficiency in the future Internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures", *Commun. Surveys & Tutor.*, vol. 13, no. 2, pp. 223–244, 2011.

[40] F. Giroire, J. Moulierac, and T. K. Phan, "Optimizing rule placement in software-defined networks for energy-aware routing", in *Proc. Global Commun. Conf. GLOBECOM 2014*, IEEE, Austin, TX, USA, 2014, pp. 2523–2529.

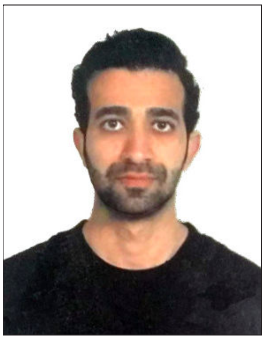
[41] F. Pop, C. Dobre, D. Comaneci, and J. Kołodziej, "Adaptive scheduling algorithm for media-optimized traffic management in software defined networks", *Computing*, vol. 98, no. 1-2, pp. 147–168, 2016 (doi: 10.1007/s00607-014-0406-9).

[42] M. Rifai, D. Lopez-Pacheco, and G. Urvoy-Keller, "Coarse-grained scheduling with software-defined networking switches", in *Proc. 2015 ACM Conf. on Spec. Interest Group on Data Commun. SIGCOMM'15*, London, UK, 2015, pp. 95–96. 2015.

[43] B. Y. Ke, P.-. Tien, and Y.-L. Hsiao, "Parallel prioritized flow scheduling for software defined data center network", in *Proc. 14th Int. Conf. on High Perform. Switch. & Rout. IEEE HPSR 2013*, Taipei, Taiwan, 2013, pp. 217–218.

[44] Y. Liu, Y. Li, Y. Wang, and J. Yuan, "Optimal scheduling for multi-flow update in Software-Defined Networks", *J. of Network & Computer Applications*, vol. 54, no. C, pp. 11–19, 215 (doi: 10.1016/j.jnca.2015.04.009).

[45] R. Mahajan and R. Wattenhofer, "On consistent updates in software defined networks", in *Proc. 12th ACM Worksh. on Hot Topics in Netw. HotNets-XII*, College Park, MD, USA, 2013, p. 20.



Mohammad Reza Abbasi received his MCA degree in Computer Science and Applications from Panjab University, Chandigarh, India, in 2013. He is currently pursuing his Ph.D. in Panjab University. His research interests include software defined networking, network management, and network virtualization.

E-mail: mabbasi@pu.ac.in
Department of Computer Science & Application
Panjab University
160014 Chandigarh, India

Ajay Guleria received his Ph.D. degree in Computer Science and Engineering from National Institute of Technology Hamirpur. Presently he is working as Senior System Manager in Panjab University Chandigarh. His current research areas of interest include software defined networking, information centric networking, network security and vehicular ad hoc networks. He is a member of IEEE, ISTE.

E-mail: ag@pu.ac.in
Computer Center
Panjab University
160014 Chandigarh, India

Mandalika S. Devi is a Professor in the Department of Computer Science and Applications, Panjab University, Chandigarh. She received her Ph.D. degree in Computer Science and Systems Engineering from Andhra University, Visakhapatnam and M.E. in Computer Science and Engineering, from NIT, Allahabad. She has completed M.Sc. in Applied Mathematics from Andhra University, Visakhapatnam. Before joining Panjab University, she served Indian Space Research Organization, Sriharikota, and National Institute of Technical Teachers' Training and Research, Chandigarh. Her areas of expertise include algorithms, image processing, distributed artificial intelligence and educational computing.

E-mail: syamala@pu.ac.in
Department of Computer Science & Application
Panjab University
160014 Chandigarh, India

Information for Authors

Journal of Telecommunications and Information Technology (JTIT) is published quarterly since 2000. It comprises original contributions, dealing with a wide range of topics related to telecommunications and information technology. **All papers are subject to peer review.** Topics presented in the JTIT report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

JTIT 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, voice communications devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology.

We encourage submissions from a diverse range of authors from across all countries and backgrounds.

Manuscript

Latex files are preferred and Editorial Office provides a style to prepare the material along with the documentation. We also accept Microsoft Word and PDF files. A typical article is 10 pages long (approximately 6,000 words) and must include the following contents:

- Authors' names and affiliations in the following format:
  First name and surname (last name), academic title,
  Position held,
  ORCID number,
  E-mail address from the University's domain,
  Faculty and name of the University,
  Link to University website.
- Abstract (150-200 words). The abstract should contain statement of the problem, assumptions and methodology, results and conclusion or discussion on the importance of the results. Abstracts must not include mathematical expressions or bibliographic references.
- Keywords related to the content of the article. About four keywords or phrases in alphabetical order should be used, separated by commas.
- The content of the article in a typical structure, i.e.: introduction, related work, conducted research, conclusions, references.

Figures, Tables and Photos

Together with the article, please send files with graphics with the highest resolution available, 150 dpi or more in bitmap resolution (jpg, png) and vector (cdr, svg, ps, pdf) formats are welcomed.

References

We use four main citation styles for a journal article, for an Internet article, for a conference paper, and for a book. Below are examples of citations. In each item, the DOI number or link to the PDF of the cited article should be provided.

[1] R.K. Meyers and A.H. Desoky, "An implementation of the blowfish cryptosystem", *2008 IEEE International Symposium on Signal Processing and Information Technology*, 2008 ((https://doi.org/10.1109/IS-SPIT.2008.4775664).

[2] K. Nowicki and T. Uhl, *Ethernet End-to-End*, 1st ed. Germany, Shaker-Publisher, 2008 (ISBN: 978383832271404).

[3] C. Shorten and T.M. Khoshgoftaar, "A survey on image data augment-tation for deep learning", *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019 (https://doi.org/10.1186/s40537-019-0197-0).

[4] S. Wong *et al.*, "Traffic forecasting using vehicle-to-vehicle communication", *3rd Annual Conference on Learning for Dynamics and Control*, pp. 917–929, 2021 (https://arxiv.org/pdf/2104.05528).

Submission

The paper with full PDF version and anonymous PDF version for the blind review process should be submitted on the JTIT website https://www.jtit.pl/jtit/about/submissions.

Reviewing Process

The article is initially approved by the Editor-In-Chief and if the decision is positive, is then sent to the reviewers. Depending on the subject of the article, it takes few weeks. In the next step, reviews are showed to authors who have 2 weeks to correct the article. Finally, the corrected text can be re-presented to the reviewer for reevaluation, which will take another 2 weeks.

As a result, after about 3 months, we are able to send the text for publication in the upcoming issue of JTIT.

When the reviews are inconsistent, additional corrections are necessary, or the reviewer expects additional verification because the corrections ordered by the author are insufficient or additional problems arise, the review of the article may be extended by another month or more.

Editorial Work

Positively reviewed and corrected article is next prepared by the editorial office for publication. At the end of this process the author receives an copyedited version for approval.

Licensing

Manuscript submitted to JTIT should not be published or simultaneously submitted for publication elsewhere. By submitting a manuscript the author grants license to the National Institute of Telecommunications, for the use of the paper in the fields of exploitation: reproducing and fixing the paper, distributing the paper by means of introduction to trade, letting for use or rental of the original or copies, and distributing the paper by means of public exhibition, screening, presentation and broadcast as well as rebroadcast, and making the paper publicly available in such a manner that anyone could access it at a place and time selected thereby, or by making it available in a way not allowing selection of time or place, including by means of Internet or other networks.

Ghostwriting Declaration

We require formal declaration that the process of writing the paper was not influenced by any third party. In the article, all the contributions of other people are clearly indicated. The theories presented, methods used, analysis and research, as well as the copyrights to the drawings, photographs and other figures belong to the authors or are clearly credited in the text. The author must also indicate whether his work has received financial support and if the realization of the whole project was possible thanks to the permission and cooperation with scientific institutions, associations and others.

Other Information

- The JTIT being an Open Access Journal (OAJ) has no article processing charges (APCs). The published articles can be downloaded freely without payment.
- JTIT supports open access and using continuous publishing "publish-as-you-go" scheme. This means that we no longer wait to accumulate several articles into a quarterly issue before publication. Rather, articles are continuously added to current issues after acceptation. Publish-as-you-go reduces publication lag for our authors, and make the newest research available quickly. After completing the review process, an article is published online in the current issue with DOI registration. When the issue period ends, a new issue is activated. So accepted articles are published without waiting for the quarterly issue end.

National Institute
of Telecommunications

Editorial Office

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

phone +48 22 512 81 83
fax +48 22 512 84 00
e-mail: journal@jtit.pl
editorial.office@il-pib.pl
https://www.jtit.pl

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY **SPECIAL ISSUE 2024**