Preface/Editorial

Special issue of Acta Polytechnica Hungarica linked to ISCAMI 2016

This issue of Acta Polytechnica Hungarica contains papers that were invited from the participants of ISCAMI 2016 (International Student Conference on Applied Mathematics and Informatics 2016, Malenovice, Czech Republic). Among the 16 submissions, after a thorough reviewing process, 11 of them have been accepted. The accepted papers have not only stayed true to the theme of the conference, but have also given a fresh twist - they can be easily categorised as papers dealing with applications of Maths to Informatics. On the one hand, they cover a variety of techniques from Soft Computing (Fuzzy Set Theory and Neural Networks) to Optimization and Differential Equations. On the other hand, they cover a myriad of application areas from Medical Diagnostics to Inventory Management to Acoustics to Information Processing to Nutritional Gastronomy! We briefly sketch the topic of each paper published in this special issue.

In the paper "Application of Evaluation Criteria to Cartographic Projections" written by Daniel Szatmári, Margita Vajsáblová and Denisa Mojšová the map projections designed by minimax type criteria, Airy-Kavraiskii's variational criterion and map projections with a minimal RMS distortion in the category of conic, azimuthal and cylindrical projections have been discussed. This article aims to compare the mentioned criteria based on the achieved values of scale distortion in the selected European countries.

In the following paper "The Significance of the Integrated Multicriteria ABC-XYZ Method for the Inventory Management Process", Milan Stojanović, Dušan Regodić analyze a methodology based on the periodical review and assessment of product inventories and the anticipation of demand. The main goal is to present the activities and pace of the fulfillment of inventories derived on the basis of the ABC-XYZ classification.

Dušan Marček, in the paper "The Category Proliferation Problem in ART Neural Networks", concentrates on problems of category proliferation and methods of minimizing of their occurrence and he proposes a new model for the optimized algorithm KMART (Kondadadi & Kozma Modified ART), called IKMART (Improved KMART), which enables to optimize the dilemma of stability/plasticity, increase the precision of categorization and influence the speed of categorization. Some results from the categorization of real text documents, which contextually overlap, are also discussed.

The next paper "Parallelization and validation of algorithms for zebrafish cell lineage tree reconstruction from big 4D image data" is a work of Robert Špir, Karol Mikula, Nadine Peyrieras. The authors present numerical algorithms,

postprocessing and validation steps for an automated cell tracking and cell lineage tree reconstruction from large-scale 3D+time two-photon laser scanning microscopy images of early stages of zebrafish (Danio rerio) embryo development. They also compare the results with ground truth data obtained by manual checking of cell links by biologists and measure the accuracy of their algorithm.

In their paper "New Approach to Fuzzy Decision Matrices", Pavla Rotterová and Ondřej Pavlačka introduce a new approach in which a fuzzy decision matrix does not describe discrete random variables but fuzzy rule bases, when the states of the world are modeled by fuzzy sets defined on the universal set on which the probability distribution is given, and the evaluations of the alternatives are expressed by fuzzy numbers. The proposed solution is illustrated with an example.

The next article "Directional monotonicity of fuzzy implications" of Katarzyna Miś concentrates on some properties of fuzzy implication functions, which are key operation in fuzzy logic. Firstly, the known notion of special implications are discussed, next the notion of inversely special implications as directional decreasing functions is introduced. The author presents several results connected with inversely special R-implications. She also discusses this new property for other families of fuzzy implications like (S,N)-implications, f-implications and g-implications.

In the paper titled "A Nutrition Adviser's Menu Planning for a Client Using a Linear Optimization Model", Lucie Schaynová, presents a new linear optimization model which improves a nutritional adviser's steps and prevents mistakes when preparing a diet plan for a client manually. The model, among other factors, takes into account the client's favourite or the adviser's recommended recipes, prevents unbalanced nutrition, respects the client's eating habits and ensures wastage of food. The model involves linear constraints which also ensures that two incompatible recipes are not used in the same meal and that a recipe is not used in an incompatible meal.

In the following paper "Acoustical Simulations based on FVM Solution of the Helmholtz Equation", Izabela Riečanová and Angela Handlovičová numerically simulate the data obtained by acoustic measurements. These measurements were performed in specialized acoustic laboratory. Their main idea was to study the reflection of different frequencies from boards with openings of various size and shape. The Finite volume method has been used to make the simulations, where the Helmholtz equation is solved using the impedance boundary conditions. The results of simulations are presented.

Roksana Brodnicka and Henryk Gacki in their article "Asymptotic stability of an evolutionary nonlinear Boltzmann-type equation", present a sufficient condition for the asymptotic stability with respect to total variation norm of semigroup generated by an abstract evolutionary non-linear Boltzmann-type equation in the space of signed measures with the right-hand side being a collision operator.

Petr Bujok in the paper "On Modification of Population-Based Approach Used in Adaptive Differential Evolution Algorithm" introduces new approach for the mutation operation in the differential evolution (DE) algorithm. The aim of this technique is to enhance the mutation strategy to avoid the local minimum area. The proposed method is applied in several well-known DE or adaptive DE variants. Selected DE variants and the corresponding counterparts are used to solve the problems of CEC 2015 test suite.

In the contribution "Computer-Aided Diagnostics of Schizophrenia: Comparison of Different Feature Extraction Methods", its authors Radomír Kůs and Daniel Schwarz present an analysis of two brain morphometry techniques and various feature extraction methods utilized in computer-aided diagnostics of schizophrenia. The methodology was incorporated into a classification pipeline and applied to distinguish between first-episode patients and healthy controls on the basis of magnetic resonance images of their brains.

We thank the submitting authors, since in all cases the submitted works were not a mere extension of their presentations at the conference but a thorough revamp that made their contributions both wholesome and substantial. We thank the reviewers who have been kind enough to give off their time, effort and knowledge that has further enriched the accepted works. Our jobs were made much more easy and palatable because of them. Thanks again to all of them.

Our gratitude goes also to the editorial team of Acta Polytechnica Hungarica for the excellent support they provided us throughout this process. Finally, we thank the organizers of the ISCAMI 2016, in particular doc. RNDr. Martin Štěpnička, for more reasons than one. Firstly, for the picturesque setting in which the conference was conducted - a setting that is so heavenly that the mind feels immediately uncluttered - the perfect state for uninhibited thinking. Secondly, for their thoughtfulness and method in the organization and logistical support. A special mention should be made for their empathetic understanding of the financial constraints that people in the academia have to negotiate, especially at the level of students.

Michał Baczyński
University of Silesia in Katowice, Institute of Mathematics
Bankowa 14, 40-007 Katowice, Poland
E-mail address: michal.baczynski@us.edu.pl

Balasubramaniam Jayaram
Indian Institute of Technology Hyderabad, Department of Mathematics
Kandi (V), Sangareddy (M), Hyderabad - 502 285, Telangana, India
E-mail address: jbala@iith.ac.in

Radko Mesiar
Slovak University of Technology in Bratislava
Radlinského 11, 810 05 Bratislava, Slovak Republic
E-mail address: radko.mesiar@stuba.sk

Application of Evaluation Criteria to Cartographic Projections

Daniel Szatmári, Margita Vajsáblová, Denisa Mojšová

Slovak University of Technology in Bratislava
Faculty of Civil Engineering
Radlinského 11, 810 05 Bratislava, Slovak Republic
daniel.szatmari@stuba.sk, margita.vajsablova@stuba.sk, xmojsova@stuba.sk

Abstract: The choice of the optimal cartographic projection, especially for large-scale maps, is an actual problem affected by the precision of positioning geodetic points using the new GNSS technologies in the coordinate systems. In this contribution we describe the map projections designed by minimax type criteria, Airy-Kavraiskii's variational criterion and map projections with a minimal RMS distortion in the category of conic, azimuthal and cylindrical projections. The aim of this paper is to compare the mentioned criteria based on the achieved values of scale distortion in the selected European countries.

Keywords: cartographic projection; conformal projection; scale distortion

1 Introduction

Design of the most suitable map projection, an actual problem affected by the precision of positioning geodetic points using the new GNSS technologies in coordinate systems, involves two classic options:

i. We choose a group of projections according to a purpose of the future map and calculate parameters of the projection according to distortion requirement (requirement of one standard parallel, requirement of two standard parallels, etc.),

ii. We calculate parameters of the ideal projection without any restrictions using minimax and variational criteria described in the following sections.

The aim of this paper is to present a combination of the above described options: we calculate the parameters of the optimized map projection by minimization of the root mean square (RMS) distortion for the chosen group of projections of the reference ellipsoid because it has not been given attention for this issue – especially for the reference ellipsoid.

2 Cartographic Projections

Cartographic projections can be defined as a mathematical transformation of a surface of an ellipsoid (defined by its semimajor axis a and eccentricity e) or a sphere (defined by its radius R) onto a plane [4]. A point on the surface of an ellipsoid is referenced by its latitude φ and longitude λ or by isometric coordinates q, λ:

$$q = \ln\left[\tan\left(\frac{\varphi}{2} + 45° \right) \sqrt{\left(\frac{1 - e\sin\varphi}{1 + e\sin\varphi} \right)^{e}} \right] \tag{1}$$

The mathematical transformation between its ellipsoidical coordinates φ, λ and planar coordinates x, y is given by map equations:

$$x = f_1(\varphi, \lambda), \quad y = f_2(\varphi, \lambda). \tag{2}$$

In case of conic and azimuthal projection, the polar coordinates ρ, ε are used. The map equations of these projections in general:

$$\rho = g_1(\varphi, \lambda), \quad \varepsilon = g_2(\varphi, \lambda). \tag{3}$$

A point on the surface of a sphere is referenced by its latitude U and longitude V or by isometric coordinates Q, V:

$$Q = \ln \tan\left(\frac{U}{2} + 45° \right) \tag{4}$$

Different projections cause different types of distortions. The scale distortion of a projection is characterized by the scale distortion factor m defined as the ratio of a differentially small distance, dS, on a mapping plane and the corresponding differential element ds on the reference surface. The angular distortion $\Delta\omega$ of a projection is defined as the difference of an angle ω' measured on the mapping plane and the corresponding angle ω on the reference surface. The distortion of the area of a projection is characterized by the area distortion factor m_{area} defined as the ratio of a differential area element dP on the mapping plane and the corresponding differential area element dp on the reference surface. Conformal projections (projections with zero angular distortion) are the most frequently applied map projections in geodetic coordinate systems. The map equations for isometric coordinates in a conformal projection have to satisfy the following conditions:

$$x + iy = f(q + i\lambda), \quad x - iy = f(q - i\lambda). \tag{5}$$

Cartographic projections can be evaluated (with respect to extremal and minimax criteria) by the maximal value of the scale distortion $|m - 1|_{\max}$ or using the RMS

value of scale distortion throughout the territory according to Airy's, Jordan's and Kavraiskii's variational criteria [1], [5], [9]. The most popularized variational criterion for the valuation of map projections is Airy-Kavraiskii's criterion, where the characteristic value of the cartographic projection of the domain Δ with area $\underline{p}_\Delta$ on reference surface is:

$$I^2 = \frac{1}{p_\Delta} \iint_\Delta \ln^2 m \, \mathrm{d}\, p. \tag{6}$$

The characteristic value of the cartographic projection for the $\underline{n}$ chosen points is:

$$I^2 = \frac{1}{n} \sum_{i=1}^{n} \ln^2 m_i. \tag{7}$$

A minimax or variational projection can be derived for the reference sphere using the following procedure described in [13]. In conformal projections it holds:

$$\frac{\partial^2 \ln \nu}{\partial Q^2} + \frac{\partial^2 \ln \nu}{\partial V^2} = 0 \tag{8}$$

where

$$\nu = m \cos U. \tag{9}$$

The solution of (8) has the shape [6], [7]:

$$\ln \nu = \sum_{j=0}^{n} \left(a_j \psi_j + b_j \tau_j \right) \tag{10}$$

where $\underline{a_j}$ and $\underline{b_j}$ are the coefficients of the conformal projection and $\underline{\psi_j}$ and $\underline{\tau_j}$ are determined by:

$$\psi_j + i\,\tau_j = (Q + iV)^j \,. \tag{11}$$

After the separation of the real and imaginary components of the complex variables (if $\underline{n} = 4$), we obtain for the scale distortion factor [8]:

$$\begin{aligned} \ln m = {} & a_0 + a_1 Q + a_2 \left(Q^2 - V^2 \right) + a_3 \left(Q^3 - 3QV^2 \right) + \\ & + a_4 \left(Q^4 - 6Q^2V^2 + V^4 \right) + b_1 V + 2b_2 QV + b_3 \left(3Q^2V - V^3 \right) + \\ & + b_4 \left(4Q^3V - 4QV^3 \right) - \ln \cos U. \end{aligned} \tag{12}$$

2.1 Minimax Type Projections

Minimax projections [2], where:

$$\left| m_{\max} - 1 \right| = \left| m_{\min} - 1 \right| \tag{13}$$

can be derived by two consecutive steps:

1) need to minimize the natural logarithm of the scale distortion of the closed boundary points:

$$I^2 = \sum_{i=1}^{n_1} \ln^2 m_i = \min \tag{14}$$

where the scale distortion factor is calculated by (12).

After minimizing the condition (14):

$$\frac{\partial I^2}{\partial a_j} = 0, \quad \frac{\partial I^2}{\partial b_j} = 0 \tag{15}$$

we get a system of nine equations in nine variables, and the coefficients $\underline{a}_0$-$\underline{b}_4$ of the projection can be calculated.

2) If the natural logarithm of the scale distortion factor of the boundary points is equal to zero, the extremal values of the projection´s scale distortion will be in the middle of the projected area. This can be reduced by the scale factor $\underline{m}_s = 2/(\underline{m}_{\max} + \underline{m}_{\min})$.

2.2 Variational Type Projections

A variational type of projection can be derived after the application of Airy-Kavraiskii's criterion (7) and minimizing the natural logarithm of the scale distortion for $\underline{n}_2$ points inside the given area where the scale distortion factor is calculated by (12).

After minimizing the condition (7):

$$\frac{\partial I^2}{\partial a_j} = 0, \quad \frac{\partial I^2}{\partial b_j} = 0 \tag{16}$$

we get a system of nine equations in nine variables, and the coefficients $\underline{a}_0$-$\underline{b}_4$ of the projection can be calculated.

The main disadvantage of minimax and variational projections is the lack of geometric interpretation. However the process of deriving of these projections is applied for the reference sphere. Before the calculus we have to transform the ellipsoidical coordinates $\underline{\varphi}$, $\underline{\lambda}$ on the spherical coordinates $\underline{U}$, $\underline{V}$ by Gauss' conformal projection:

$$\tan\left(\frac{U}{2} + 45°\right) = k\left[\tan\left(\frac{\varphi}{2} + 45°\right)\sqrt{\left(\frac{1 - e\sin\varphi}{1 + e\sin\varphi}\right)^e}\right]^\alpha, \quad V = \alpha\,\lambda \tag{17}$$

with the parameters α and k calculated for the central parallel of the given territory. The radius of the sphere is determined so that the reference sphere and the reference ellipsoid have the same Gaussian curvature [10].

3 Methods of Distortion Optimization in Conformal Cartographic Projections on Developable Surfaces

In 1933 Kavraiskii formulated the method of calculation of the parameters for conformal conic projection of the reference sphere [3] with minimal RMS value of distortion – the scale distortion between two parallels was minimized by Airy's criterion (the procedure was also published in the Baltic Geodetic Commission Report in 1936). This method for the reference sphere, is inadequate, therefore, we formulate this process especially for the reference ellipsoid for conformal conic, conformal azimuthal and conformal cylindrical projections with the requirement of minimal RMS value of scale distortion in the projected area.

3.1 Conformal Conic Projection of the Reference Ellipsoid with Minimal RMS Value of Scale Distortion

A conformal conic projection was first introduced by Johannes Heinrich Lambert (1728-1777). Conic projections are appropriate for oblong territories along geographic parallels.

The map equations of Lambert's conformal conic projection are:

$$\rho = \rho_0 \left[\frac{\tan\left(\frac{\varphi_0}{2} + 45°\right)}{\tan\left(\frac{\varphi}{2} + 45°\right)} \sqrt{\left(\frac{(1 - e\sin\varphi_0)(1 + e\sin\varphi)}{(1 + e\sin\varphi_0)(1 - e\sin\varphi)} \right)^e} \right]^n, \quad \varepsilon = n\lambda \tag{18}$$

where φ_0 is the ellipsoidal latitude of the standard parallel and ρ_0 is its polar radius. The parameters ρ_0, n and φ_0 are also the three constants of the conic projection affecting its accuracy.

The scale distortion m of a conformal conic projection is calculated by:

$$m = \frac{n\rho}{N\cos\varphi} \tag{19}$$

where N is a radius of curvature in the prime vertical:

$$N = \frac{a}{\sqrt{1 - e^2\sin^2\varphi}}. \tag{20}$$

The process of minimization of the RMS value of scale distortion throughout the territory is more effective using only two parameters, therefore we have defined the following substitution in [15]:

$$k = \rho_0 \left(\tan\left(\frac{\varphi_0}{2} + 45° \right) \sqrt{\left(\frac{1 - e\sin\varphi_0}{1 + e\sin\varphi_0} \right)^e} \right)^n \tag{21}$$

then the conformal conic projection of the reference ellipsoid has only two constants $\underline{n}$ and $\underline{k}$, then its map equations are:

$$\rho = \frac{k}{\tan^n\left(\frac{\varphi}{2} + 45° \right)} \sqrt{\left(\frac{1 + e\sin\varphi}{1 - e\sin\varphi} \right)^{e\,n}}, \quad \varepsilon = n\lambda \tag{22}$$

The scale distortion factor on the projected area is optimized according to Airy-Kavraiskii's variational criterion (7) by minimizing the value of $\underline{I}$. The projected territory is divided by ellipsoidal latitude $\underline{\varphi}$ to $\underline{j}$ segments $\Delta\underline{p_i}$ with area $\underline{p_i}$, for $\underline{i} = 1, \ldots, \underline{j}$:

$$I^2 = \frac{\sum_{i=1}^{j} p_i \ln^2 m_i}{\sum_{i=1}^{j} p_i}, \text{ where } \sum_{i=1}^{j} p_i = p \tag{23}$$

The scale distortion factor $\underline{m_i}$ of the conformal conic projection of the ellipsoid for the determined area $\Delta\underline{p_i}$ is evaluated after the substitution (22) in the equation (19):

$$m_i = \frac{n\,k}{N_i \cos\varphi_i \tan^n\left(\frac{\varphi_i}{2} + 45° \right)} \sqrt{\left(\frac{1 + e\sin\varphi_i}{1 - e\sin\varphi_i} \right)^{e\,n}} \tag{24}$$

where $\underline{\varphi_i}$ is the ellipsoidal latitude of medial parallel of the $\underline{i}^{\text{th}}$ band and $\underline{N_i}$ is its radius of curvature in the prime vertical.

Now, let us introduce a term $\underline{h_i} = \ln \underline{m_i}$ which can be expressed using (24):

$$\begin{aligned} h_i &= \ln m_i = \ln(n\,k) - \ln(N_i \cos\varphi_i) + \\ &+ n\left(-\ln \tan\left(\frac{\varphi_i}{2} + 45° \right) + \frac{e}{2}\ln(1 + e\sin\varphi_i) - \frac{e}{2}\ln(1 - e\sin\varphi_i) \right) \end{aligned} \tag{25}$$

and after the following substitutions:

$$b = \ln(n\,k),\ \ \gamma_i = -\ln(N_i \cos\varphi_i),$$
$$\alpha_i = -\ln\,\tan\left(\frac{\varphi_i}{2} + 45°\right) + \frac{e}{2}\ln(1 + e\sin\varphi_i) - \frac{e}{2}\ln(1 - e\sin\varphi_i) \tag{26}$$

we can evaluate the coefficients α_i and γ_i for each of the bands and formulate j equations, whereby the equation for i^{th} band of the projected territory is:

$$h_i = \alpha_i n + b + \gamma_i \tag{27}$$

The characteristic I, in (23) is a function of two variables n and b; $I^2 = f(n, b)$ after the substitution (26). We obtain the minimal value of I, if the partial derivative of this function is equal to zero:

$$\frac{\partial\sum_{i=1}^{j} p_i h_i^2}{\partial n} = 0,\ \ \frac{\partial\sum_{i=1}^{j} p_i h_i^2}{\partial b} = 0 \tag{28}$$

Therefore the normal equations are:

$$\begin{aligned} n\sum_{i=1}^{j} p_i\alpha_i^2 + b\sum_{i=1}^{j} p_i\alpha_i + \sum_{i=1}^{j} p_i\alpha_i\gamma_i = 0 \\ n\sum_{i=1}^{j} p_i\alpha_i + b\sum_{i=1}^{j} p_i + \sum_{i=1}^{j} p_i\gamma_i = 0 \end{aligned} \tag{29}$$

The parameter n and the coefficient b are the solution of this system of equations. The parameter k is evaluated from (26).

The ellipsoidal latitudes φ_1 and φ_2 of the preserved parallels can be calculated for example by Newton's method from the conditions for their scale distortion factor:

$$\begin{aligned} m_1 &= \frac{n\,k\sqrt{1 - e^2\sin^2\varphi_1}}{a\cos\varphi_1\,\tan^n\left(\frac{\varphi_1}{2} + 45°\right)}\sqrt{\left(\frac{1 + e\sin\varphi_1}{1 - e\sin\varphi_1}\right)^{e\,n}} = 1 \\ m_2 &= \frac{n\,k\sqrt{1 - e^2\sin^2\varphi_2}}{a\cos\varphi_2\,\tan^n\left(\frac{\varphi_2}{2} + 45°\right)}\sqrt{\left(\frac{1 + e\sin\varphi_2}{1 - e\sin\varphi_2}\right)^{e\,n}} = 1 \end{aligned} \tag{30}$$

3.2 Conformal Azimuthal Projection of the Reference Ellipsoid with Minimal RMS Value of Scale Distortion

The author of the conformal azimuthal projection of the sphere (also known as stereographic projection) is Hipparchus. Azimuthal projections are appropriate for circle-shaped territories. In this chapter, we derive the formulas to calculate parameters of the conformal azimuthal projection of the reference ellipsoid with the requirement of minimal RMS value of scale distortion in the projected area.

The map equations for the conformal azimuthal projection of the ellipsoid are:

$$\rho = c \tan\left(45° - \frac{\varphi}{2}\right)\sqrt{\left(\frac{1+e\sin\varphi}{1-e\sin\varphi}\right)^{e}}, \quad \varepsilon = \lambda \tag{31}$$

where $\underline{c}$ is a constant of the azimuthal projection, its value affects the accuracy of projection.

The scale distortion of the conformal azimuthal projection is calculated by:

$$m = \frac{\rho}{N\cos\varphi} \tag{32}$$

We have realized the process of minimization of the RMS value of scale distortion throughout the territory by minimizing the value $\underline{I}$ of Airy-Kavraiskii's variational criterion (7) after dividing the projected territory by ellipsoidal latitude $\underline{\varphi}$ to $\underline{i}$ segments $\Delta\underline{p_i}$.

We obtain the formula for the scale distortion factor $\underline{m_i}$ of the conformal azimuthal projection of the ellipsoid for the determined area $\Delta\underline{p_i}$ after substitution (31) in the equation (32):

$$m_i = \frac{c}{2N_i \cos^2\left(45° - \frac{\varphi_i}{2}\right)}\sqrt{\left(\frac{1+e\sin\varphi_i}{1-e\sin\varphi_i}\right)^{e}} \tag{33}$$

where $\underline{\varphi_i}$ is the ellipsoidal latitude of the medial parallel of the $\underline{i}^{\text{th}}$ band and $\underline{N_i}$ is its radius of curvature in the prime vertical.

Now, as before, we can express $\underline{h_i} = \ln \underline{m_i}$ from (33):

$$h_i = \ln m_i = \ln c - \ln 2 - \ln N_i - 2\ln\cos\left(45° - \frac{\varphi_i}{2}\right) + \\ + \frac{e}{2}\ln(1+e\sin\varphi_i) - \frac{e}{2}\ln(1-e\sin\varphi_i) \tag{34}$$

and apply the following substitutions:

$$b = \ln c,\ \gamma_i = -\ln 2 - \ln N_i - 2\ln\cos\left(45° - \frac{\varphi_i}{2}\right) + \\ + \frac{e}{2}\ln(1 + e\sin\varphi_i) - \frac{e}{2}\ln(1 - e\sin\varphi_i) \tag{35}$$

We can evaluate the coefficients γ_i for each of the bands and formulate j equations, whereby the equation for i^{th} band of the projected territory is:

$$h_i = b + \gamma_i \tag{36}$$

The characteristic I is a function of parameter b; $I^2 = f(b)$. We obtain the minimal value of I, if the partial derivative of this function is equal to zero:

$$\frac{\partial \sum_{i=1}^{j} p_i h_i^2}{\partial b} = 0 \tag{37}$$

From there we obtain the normal equation:

$$b\sum_{i=1}^{j} p_i + \sum_{i=1}^{j} p_i \gamma_i = 0 \tag{38}$$

We can evaluate the coefficient b from (38), then the parameter c from (35).

The ellipsoidal latitude φ_0 of the preserved parallel can be calculated for example by Newton's method after substitution (20) instead of N_0 into the condition for its scale distortion (33):

$$m_0 = \frac{c\sqrt{1 - e^2\sin^2\varphi_0}}{2a\cos^2\left(45° - \frac{\varphi_0}{2}\right)}\sqrt{\left(\frac{1 + e\sin\varphi_0}{1 - e\sin\varphi_0}\right)^e} = 1 \tag{39}$$

3.3 Conformal Cylindrical Projection of the Reference Ellipsoid with Minimal RMS Value of Scale Distortion

A conformal cylindrical projection of a sphere designed by Mercator is one of the most common map projections. Cylindrical projections are appropriate for oblong territories along the equator or an orthodrome (e.g. geographic meridian). In this chapter, we derive the formulas to calculate parameters of the conformal cylindrical projection of the reference ellipsoid with the requirement of minimal RMS value of scale distortion in the projected area.

The map equations for the conformal cylindrical projection of the ellipsoid are:

$$x = n \ln \left[\tan\left(45° - \frac{\varphi}{2} \right) \sqrt{\left(\frac{1 + e \sin \varphi}{1 - e \sin \varphi} \right)^{e}} \right], \; y = n\lambda \tag{40}$$

where n is the constant of the cylindrical projection (geometric characteristic – radius of the cylinder), its value affects the accuracy of the projection.

The scale distortion of conformal cylindrical projection is calculated by:

$$m = \frac{n}{N \cos \varphi} \tag{41}$$

We have realized the process of minimization of the RMS value of scale distortion factor throughout the territory, as before, by minimizing the value of I of Airy-Kavraiskii's variational criterion (7) after dividing the projected territory by ellipsoidal latitude φ to j segments Δp_i.

We express $h_i = \ln m_i$ from m_i of conformal cylindrical projection of the ellipsoid for the determined area Δp_i from equation (41):

$$h_i = \ln m_i = \ln n - \ln N_i - \ln \cos \varphi_i \tag{42}$$

where φ_i is the ellipsoidal latitude of the medial parallel of the i^{th} band and N_i is its radius of curvature in the prime vertical. After the following substitutions:

$$b = \ln n, \;\; \gamma_i = -\ln N_i - \ln \cos \varphi_i \tag{43}$$

we can evaluate the coefficients γ_i for each of the area segments and formulate j equations, whereby equation for i^{th} band of the projected territory is (36).

The characteristic I is a function of the parameter b; $I^2 = f(b)$. Its minimal value is obtained, if the condition (37) is satisfied. Therefore we evaluate the coefficient b from the normal equation (38) and obtain the radius of the cylinder n from (43).

The ellipsoidal latitude φ_0 of the preserved parallel can be calculated for example by Newton's method after substitution (20) instead of N_0 into equation (41), therefore the condition for its scale distortion is:

$$m_0 = \frac{n \sqrt{1 - e^2 \sin^2 \varphi_0}}{a \cos \varphi_0} = 1 \tag{44}$$

4 Applications of the Optimized Conformal Cartographic Projections

4.1 Conformal Projections of Slovakia

The problem of a new map projection in Slovakia is very real. The currently used cartographic projection in Slovakia is the Křovák's projection, which was designed in 1922 solely for Czechoslovakia. It is an oblique case of a conformal conic projection based on two preserved parallels. Bessel's reference ellipsoid is transformed into a sphere (17), which is transformed to a secant cone in oblique position. The scale distortion of the projection is from –10 to +11 cm/km, the RMS value of scale distortion in Slovakia according to Airy-Kavraiskii's variational criterion (7) is 7.1 cm/km.

After the dissolution of the former Czechoslovak Republic the shape of Slovak country is not optimal for the mentioned Křovák's projection anymore. This calls for a new design of cartographic projection based on the requirements of the Geodesy, Cartography and Cadaster Authority of Slovakia. In 2010 a new cartographic projection was proposed: Lambert's conformal conic projection in normal position with scale distortion from –6.7 to +6.7 cm/km [14] and RMS value of scale distortion in Slovakia according to Airy-Kavraiskii's variational criterion (7) equal to 5.0 cm/km.

The parameters of the aforementioned conic projections (Křovák, Lambert) $\underline{\rho}_0$ and $\underline{n}$ were calculated by criteria of scale distortion of selected parallels. Alternative method is to calculate the parameters of a conformal conic projection with the requirement of a minimal RMS value of scale distortion for the whole projected territory described in chapter 3.1.

Within the latter method Slovakia is projected onto the reference ellipsoid GRS80 between parallels with latitudes $\underline{\varphi}_S = 47° 43′ 09.6235″$ and $\underline{\varphi}_N = 49° 36′ 04.6826″$. The parameters $\underline{n}$ and $\underline{k}$ of a conformal conic projection optimized by minimal RMS value of scale distortion after dividing the territory of Slovakia to 20 segments are:

$\underline{n} = 0.750\,955\,513\,8$

$\underline{k} = 11\,642\,467.97$ m

Ellipsoidal latitudes $\underline{\varphi}_1$ and $\underline{\varphi}_2$ of the standard parallels calculated by (30) are:

$\underline{\varphi}_1 = 48° 07′ 45.6717″$

$\underline{\varphi}_2 = 49° 12′ 54.3553″$

Then the scale distortion of the projection is from –4.4 to +9.0 cm/km and the RMS value of scale distortion in Slovakia according to Airy-Kavraiskii's variational criterion (7) is 3.4 cm/km.

Figure 1 illustrates the percentage distribution of scale distortions of conformal conic projections in Slovakia. The dark bar represents Lambert's conformal conic projection (for example, 45 % of the territory has scale distortion from –5 to +5 cm/km), the light bar represents the optimized conformal conic projection (96 % of the territory has scale distortion from the same interval, by comparison with Lambert's projection, it is more than double). Although the maximal scale distortion of the optimized conic projection (+9.0 cm/km) is bigger than the maximal scale distortion of the non-optimized conic projection (+6.7 cm/km), this value is exceeded only on 1.4 % of the projected area. On the other side **the optimized conic projection has smaller distortion over a larger area**.

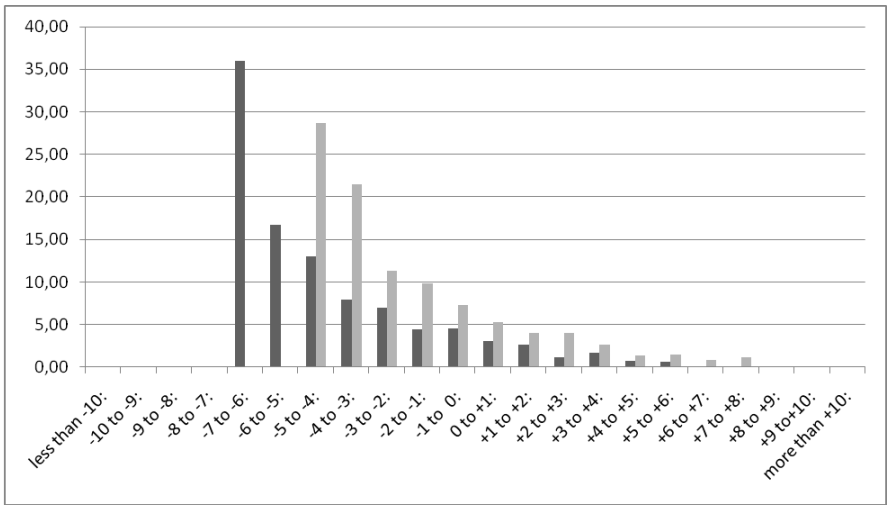

Figure 1
Percentage distribution of scale distortions of conformal conic projections in Slovakia

Table 1
Comparison of conformal projections of Slovakia

Cartographic projection	Scale distortion [cm/km]		RMS value of scale distortion [cm/km]
	from	to	
Křovák	–10.0	+11.0	7.1
Minimax	–4.9	+4.9	2.6
Variational	–2.6	+7.0	1.8
Conformal conic – Lambert	–6.7	+6.7	**5.0**
Optimized conformal conic	–4.4	+9.0	**3.4**

Minimax projection for the territory of Slovakia was designed in [12]. The scale distortion of the projection is from –4.9 cm/km to +4.9 cm/km. The RMS value of scale distortion in Slovakia according to Airy-Kavraiskii's variational criterion (7) is 2.6 cm/km.

Variational projection for Slovakia was designed in [11]. The scale distortion of the projection is from –2.6 cm/km to +7.0 cm/km. The RMS value of scale distortion in Slovakia according to Airy-Kavraiskii's variational criterion (7) is 1.8 cm/km.

The comparison of the aforementioned conformal projections of the territory of Slovakia is shown in Table 1.

4.2 Conformal Projections of the Netherlands

The currently used cartographic projection in the Netherlands is a conformal azimuthal projection in oblique position called Stereographic projection. Bessel's reference ellipsoid is transformed into a sphere (17), which is transformed to a secant plane in oblique position. Ellipsoidal coordinates of the cartographic pole situated in a town of Amersfoort are:

$\underline{\varphi} = 52° 09′ 22.178″$

$\underline{\lambda} = 5° 23′ 15.500″$

The scale distortion of the projection is from –9 to +10 cm/km, the RMS value of scale distortion according to Airy-Kavraiskii's variational criterion (7) is 5.6 cm/km.

The Netherlands are situated on the reference sphere between the cartographic parallel with latitude $\underline{S_S} = 88° 25′ 26.6818″$ and the cartographic pole $\underline{S} = 90°$. We have designed a conformal azimuthal projection optimized by minimal RMS value of scale distortion using the method derived in chapter 3.2. The parameter $\underline{c}$ of this projection after dividing the country's territory to 20 segments is:

$\underline{c} = 12\,734\,816.084$ m.

Spherical cartographic latitude $\underline{S_0}$ of a standard parallel calculated by (39) if $\underline{e} = 0$:

$$m_0 = \frac{c}{2R\cos^2\left(45° - \frac{S_0}{2}\right)} = 1 \tag{45}$$

is $\underline{S_0} = 89° 03′ 10.824″$. (Spherical cartographic latitude of the cartographic pole is $\underline{S} = 90°$.)

Then the scale distortion of the projection is from –6.8 to +12.1 cm/km and the RMS value of scale distortion of the optimized azimuthal projection according to Airy-Kavraiskii's variational criterion (7) is 4.5 cm/km.

Figure 2 illustrates the percentage distribution of scale distortions of conformal azimuthal projections in the Netherlands. The dark bar represents the currently used conformal azimuthal projection (e.g. 73% of the territory has scale distortion from –7 to +7 cm/km), the light bar represents the optimized conformal azimuthal projection (e.g. 95% of the territory has scale distortion from the same interval). Although the maximal scale distortion of the optimized azimuthal projection (+12.1 cm/km) is bigger than the maximal scale distortion of the non-optimized azimuthal projection (+10 cm/km), on the other side **the optimized azimuthal projection** designed by us in this chapter **has smaller distortion over a larger area**.

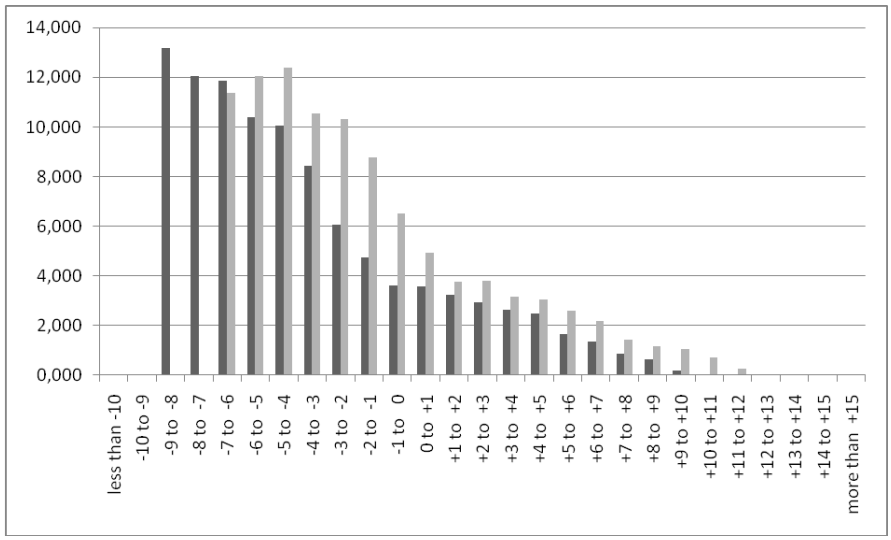

Figure 2

Percentage distribution of scale distortions of conformal azimuthal projections in the Netherlands

The scale distortion of the minimax projection (for the process of calculation see chapter 2.1) is from –5.1 cm/km to +5.1 cm/km. The RMS value of scale distortion according to Airy-Kavraiskii's variational criterion (7) is 2.9 cm/km.

The scale distortion of the variational projection (for the process of calculation see chapter 2.2) is from –2.8 cm/km to +7.4 cm/km. The RMS value of scale distortion in the Netherlands according to Airy-Kavraiskii's variational criterion (7) is 2.0 cm/km.

The comparison of the aforementioned conformal projections of the territory of the Netherlands is shown in Table 2.

Table 2
Comparison of conformal projections of the Netherlands

Cartographic projection	Scale distortion [cm/km]		RMS value of scale distortion [cm/km]
	from	to	
Minimax	–5.1	+5.1	2.9
Variational	–2.8	+7.4	2.0
Conformal azimuthal	–9.0	+10.0	**5.6**
Optimized conformal azimuthal	–6.8	+12.1	**4.5**

4.3 Conformal Projections of Hungary

For the Hungarian civilian base maps the Uniform National Projection system (EOV) is currently used which is a conformal cylindrical projection in oblique position. The GRS 1967 reference ellipsoid is transformed into a sphere (17), which is transformed to a secant cylinder in oblique position. The scale distortion of the projection is from –7 to +26 cm/km [16] and the RMS value of scale distortion according to Airy-Kavraiskii's variational criterion (7) is 6.8 cm/km.

Hungary is situated on the reference sphere between cartographic parallels with latitudes $\underline{S_S} = -1° 23' 47.6528''$ and $\underline{S_N} = 1° 27' 46.2515''$. We have designed a conformal cylindrical projection optimized by minimal RMS value of scale distortion using the method derived in chapter 3.3. The parameter $\underline{n}$ of this projection after dividing the country's territory to 20 segments is:

$\underline{n} = 6\,379\,314.331$ m

Spherical cartographic latitude $\underline{S_0}$ of a preserved parallel can be calculated by (44) if $\underline{e} = 0$:

$$m_0 = \frac{n}{R \cos S_0} = 1 \tag{46}$$

Then the scale distortion of the optimized projection is from –6.8 to +25.1 cm/km and the RMS value of scale distortion of the optimized cylindrical projection according to Airy-Kavraiskii's variational criterion (15) is 6.7 cm/km.

Figure 3 illustrates the percentage distribution of scale distortions of conformal cylindrical projections in Hungary. The dark bar represents the currently used conformal cylindrical projection (EOV), the light bar represents the optimized conformal cylindrical projection. The comparison showed that map projection used in Hungary (EOV) is **the only currently used map projection with distortions nearby optimal**.

The scale distortion of the minimax projection (for the process of calculation see chapter 2.1) is from –10.3 cm/km to +10.3 cm/km. The RMS value of scale distortion according to Airy-Kavraiskii's variational criterion (7) is 5.4 cm/km.

The scale distortion of the variational projection (for the process of calculation see chapter 2.2) is from –6.2 cm/km to +16.5 cm/km. The RMS value of scale distortion in Hungary according to Airy-Kavraiskii's variational criterion (7) is 4.1 cm/km.

The comparison of the aforementioned conformal projections of the territory of Hungary is shown in Table 3.

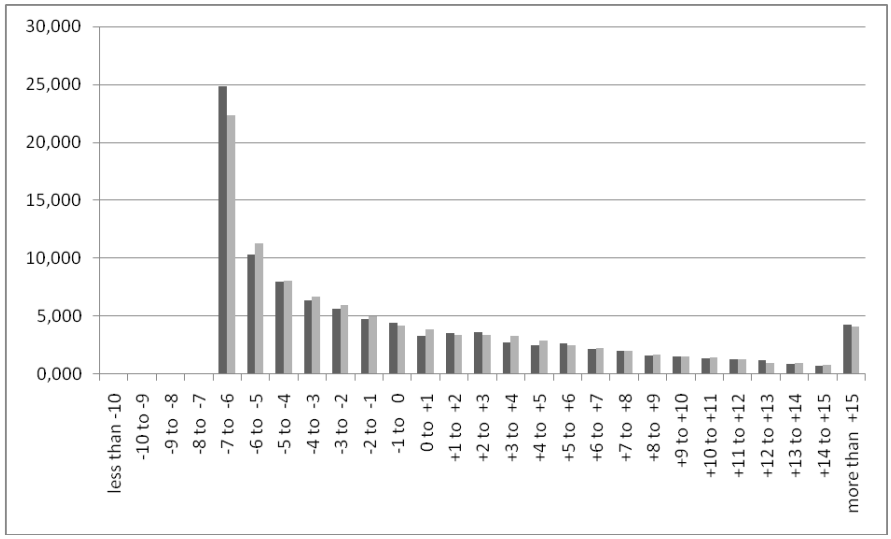

Figure 3
Percentage distribution of scale distortions of conformal cylindrical projections in Hungary

Table 3
Comparison of conformal projections of Hungary

Cartographic projection	Scale distortion [cm/km]		RMS value of scale distortion [cm/km]
	from	to	
Minimax	–10.3	+10.3	5.4
Variational	–6.2	+16.5	4.1
Conformal cylindrical (EOV)	–7.0	+26.0	**6.8**
Optimized conformal cylindrical	–6.8	+25.1	**6.7**

Conclusions

The final statement, which is the optimal map projection, significantly depends on the given criteria. In terms of extremal distortions the application of the most appropriate group of projections (conic, azimuthal, cylindrical) according the geometrical characteristics of the territory is suitable. A non-standard approach, that minimizes the RMS distortion throughout the territory, optimizes the distribution of distortions of the projected territory. These claims were confirmed by the results demonstrated in tables 1-3. Using the RMS minimization is a good alternative especially for conic, azimuthal and cylindrical projections because these projections are more understandable for cartographic practice and the projections with optimized RMS distortion have smaller distortions over a larger area.

Acknowledgement

This work was supported by the grant VEGA 1/0682/16.

References

[1] AIRY, G. B. Explanation of a projection by balance of errors for maps applying to a very large extent of the Earth's surface and comparison with other projections. *Philosophical Magazine and Journal of Science*, Vol. 22, pp. 409-421, 1861

[2] CHEBYSHEV, P. L. Sur la construction des cartes géographiques. Oeuvres de P.L. Chebyshev. Chelsea, 1962

[3] FIALA, F. *Mathematical Cartography.* ČSAV, Praha, 1955

[4] GRAFAREND, W. E. – KRUMM W. F. *Map Projections. Cartographic Information systems*. Springer – Verlag Berlin Heidelberg. ISBN 978-36-420-7178-2, 2006

[5] HOJOVEC, V. – ŠMEHIL, J. Criterion for the valuation of cartographic projections in terms of distortion. *Geodetic and Cartographic Review*, Vol. 14(56), No. 1, pp. 9-12 [in Czech] 1968

[6] HOJOVEC, V. – KLÁŠTERKA, J. – RENDLOVÁ, H. Using of the Chebyshev's and variational criteria in conformal projections. *Geodetic and Cartographic Review*, Vol. 21(63), No. 1, pp. 3-6 [in Czech] 1975

[7] HOJOVEC, V. Application of optimization criteria in conformal projections. *Geodetic and Cartographic Review*, ISSN 0016-7096. Vol. 42/84, No. 7, pp. 133-138 [in Czech] 1996

[8] HOJOVEC, V. – BOŘÍK, M. – MIKUTA, V. – MINÁŘ, P. Results of optimization of the conform cartographic projection for the Czech Republic. *Geodetic and Cartographic Review*, ISSN 0016-7096. Vol. 43(85), No. 12, pp. 253-256 [in Czech] 1997

[9] KAVRAJSKIJ, V. V. *Collected works 2*. Moscow, 1959

[10] PRESSLEY, A. *Elementary Differential Geometry*. Springer – Verlag London, ISBN 1-85233-152-6, 2001

[11] SZATMÁRI, D. Optimization of the conformal cartographic projection. In: *Advances in Architectural, Civil and Environmental Engineering: 24th Annual PhD Student Conference*. Bratislava, pp. 54-60. ISBN 978-80-227-4301-3 [in Slovak] 2014

[12] SZATMÁRI, D. Optimization of conformal cartographic projections for the Slovak Republic according to Chebyshev's theorem. *Slovak Journal of Civil Engineering*, ISSN 1210-3896, Vol. 23, No. 4, pp. 19-24, 2015

[13] URMAJEV, N. A. *Methods for finding new cartographic projections*. Moscow, 1947

[14] VAJSÁBLOVÁ, M. Proposal of the New Cartographic Projection of the Slovak Republic Territory. *Geodetic and Cartographic Review*. ISSN 0016-7096. Vol. 57 (99), No. 8, pp. 185-190 [in Slovak] 2011

[15] VAJSÁBLOVÁ, M. *Aspects of the design of a cartographic projection for the territory of Slovakia*. *Edition of scientific works*. Slovak University of Technology, Bratislava, ISBN 978-80-227-4393-8. [in Slovak] 2015

[16] www.geod.bme.hu/staff_h/varga/vetulet.html

The Significance of the Integrated Multicriteria ABC-XYZ Method for the Inventory Management Process

Milan Stojanović, Dušan Regodić

University of Singidunum, 32 Danijelova St., 11000 Belgrade, Serbia
milan.stojanovic.13@singimail.rs, dregodic@singidunum.ac.rs

Abstract: Inventory optimization in the supply chain is one of the most important goals in logistical business operations given the fact that optimized inventories directly impact the efficiency and profitability of the business. In the contemporary conditions of business processes, the goal of an enterprise's business operations reflects in the maximal reduction in the level of inventories, simultaneously retaining a certain level of services provided, in order for them to become and remain competitive in the market. Understanding the significance of inventories enables optimal uninterrupted business doing, for which reason exactly the ABC-XYZ method, as one of the ways to efficiently manage inventories, is used in this paper. Given the fact that there are limitations to the ABC classification, the limitation to one single criterion and the non-existence of a demand analysis at determining the needed inventories, the problem is overcome by the introduction of the XYZ classification. The merging of the mentioned classifications results in the integrated ABC-XYZ classification model, which can be used, on the basis of a multi-criteria and multi-dimensional approach, to classify inventories and make a proposal for their optimization.

Keywords: inventory management; management; ABC-XYZ analysis; analytic hierarchical process

1 Introduction

The Globalization process sets new rules of the game for enterprises daily. The most important task of the management of an enterprise is to maximize a profit, and their achievement of such a goal in the contemporary conditions of business operations, is often faced with a large number of limitations. The key question that every enterprise poses and requires an answer to, is the question of how to become better and more successful than the competition. To be competitive in the market today is not a success, but rather question of survival, which is to a large extent is dependent on the enterprise's dilemma: Which inventory level is optimal, to enable profit maximization while keeping captured capital at as low a level as possible?

Inventory management represents a very important segment of the business conducted by modern enterprises and as such, it is crucial for the success of an enterprise's business operations. For that exact reason, it requires close monitoring, as well as, constant improvement in compliance with contemporary standards. The inventory level of the products that an enterprise has should be in accordance with the market needs, i.e. with the estimation of demand for a particular period. The goal of inventory management is to find the quantity of inventories of products that is sufficient to uninterruptedly meet the market requirements and reduce the costs incurred through inventories keeping.

The first step in the determination of the expected demand is the collection and organizing of pieces of information about the previous sale and goods movement through the supply chain. Sales enterprises have in their assortments a great number of articles, so there is a need for such articles to be classified and for establishing a system in which the movement of inventories will be recorded and monitored. Then, an inventory management methodology needs to be defined.

In this work, a methodology based on the periodical review and assessment of product inventories and the anticipation of demand are presented, and a proposal is made for the activities and pace of the fulfillment of inventories derived on the basis of the ABC-XYZ classification. The paper consists of the literature overview, the methodological part and a practical example as well.

The research subject is the analysis of the inventories of the Win Win Shop d.o.o. (limited liability Company) enterprise in the retail chain in the Serbian territory. The company does business in the territories of Bosnia and Herzegovina, Montenegro and Serbia. Currently, there are 101 retail shops in Serbia. The company's assortment offers a large selection of IT equipment, AV equipment, domestic appliances, small household gadgets, video surveillance devices, watches, jewelry, kids' toys, healthcare devices and many other devices. Win Win Shop's vision is to preserve and advance its position in the market. It aspires to improve its core activity, which implies the IT sector, as well as to take the position of a leader in other business niches for the other products in the offer.

Its fundamental goal is to give customers as much as possible for the invested money, while appreciating their time and loyalty. Win Win is a partner with the most famous world brands and producers, such as: HP (laptops, computers, servers, printers, monitors, toner cartridges…), Asus (laptops, netbooks, monitors, motherboard, graphical cards…), Acer (laptops, netbooks, tablets and monitors), Del (laptops and monitors), Toshiba (laptops and monitors), Lenovo (laptops), MSI (laptops, motherboards, graphical cards…), Fujitsu (laptops), Samsung HP (laptops, tablets, mobile phones, monitors, printers, toners…), ViewSonic (monitors and tablets), and many other.

The goal is to make the most favorable offer in all the segments of business doing, excellent prices, deferred payment in installments, a big choice, a wide offer of technical goods.

In accordance with the company's goals, each retail shop needs to keep a certain quantity of inventory, depending on the situation in the market. In order to achieve high profitability, quick adaptations to changes in the market and monitoring changes in the IT sector is what the company needs, reasoning it is necessary that the quantities of inventories should be as small as possible, simultaneously retaining a wide assortment. The considerations related to a reduction in costs are reflected in a reduction in the costs of the distribution and control of the products that make weak contributions to the sales results.

2 The Methodology

The analysis of inventories by means of the ABC classification is a widely used approach. The conventional ABC classification was developed at General Electric in the 1950s. The ABC analysis has been used in inventories management since the 1950s [1]. The traditional ABC ranking of products is conducted on the basis of only one criterion, and it is most frequently the annual usage (AU). In real time business operations, there is quite often a need to take into consideration some other criteria, too, for the defining of the importance and the quantities of the needed inventories of a product, so in a very short time the classical ABC classification was replaced with the multi-criteria ABC classification (multiple criteria inventory classification (MCIC)). The methods combining the known multi-criteria decision-making techniques with the ABC ranking were developed.

One of the known approaches is the application of the analytic hierarchical process (AHP) [2], and the application of the AHP in the inventories management integrated approaches [3]. Certain authors have developed the weighted linear optimization of inventories [4, 5]. Ng suggested a model, named after the NG-model, according to which, all criteria are translated into the scalar result of each element undergoing classification [6]. The extension of the NG-model presupposes that the effects of weights should be maintained until the final solution [7].

There is also an approach to the ABC classification that uses artificial neural networks (ANN) [8] and artificial-intelligence-based classification techniques [9]. The newer models that deal with the problem issues of the ABC classification introduce fuzzy logic so as to include criteria that are nominal, those depending on the preferences and experiences of the management, those whose implementation can be simple [10]. The two-dimensional ABC-XYZ [11, 12] and the multidimensional approaches ABC-XYZ-VED [13] have also been developed.

2.1 The ABC Classification

The significance of the ABC analysis is reflected in the fact that it enables the monitoring of inventories as well as the determination of potentially useful inventories and those that do not contribute to the goals but rather are costs and are a burden for the enterprise. The ABC classification enables inventories management at several levels, in compliance with their importance. Inventories are categorized into groups according to the Pareto principle, which is based on the observation that there are a small number of elements that dominate in the achievement of the achieved results in different situations.

The Pareto principle represents the rule of 80:20, which means that 20% of the sold articles contribute 80% to the sales results. The ABC analysis in combination with the Pareto rule enables the forming of three groups of products: usually, around 20% of the products that contribute 80% to the total value belong to Group A; into Group B, the products that contribute around 15% are classified; ultimately, Group C consists of the products whose contribution is about 5%. This distribution is arbitrary, and the groups are defined in accordance with the enterprise's needs.

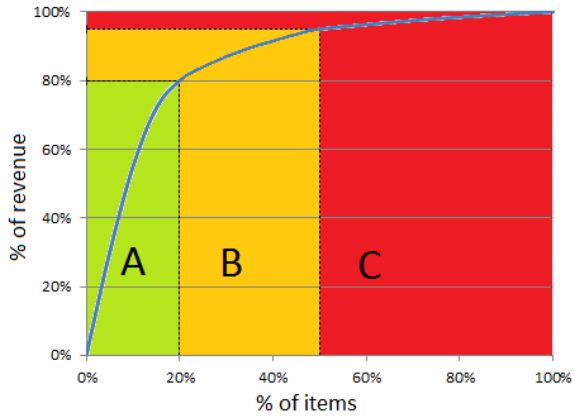

Figure 1
The example of the ABC curve

There are a few steps to follow in conducting an ABC analysis:

1) The selection of an eligible criterion. The criterion of choice usually depends on the purpose of the analysis. For example, the scrap rate is often used for quality control; the percentage (%) of a market share is used for marketing research; the annual usage is to a large extent used for inventories management.

2) The collection and checking of required data. All collected data must be accurate and the units of measure must be consistent.

3) Making the necessary calculations. When inventories management is concerned, this includes:

The calculation of the annual usage, where:

$$GV_i = c_i * x_i, \quad (1)$$

(c_i – the unit price and x_i – the volume of demand). The ranking of the elements is carried out in descending order according to the AU value. A calculation is also made of the cumulative value according to the AU, and their value in percentages.

4) The determination of the number of groups and the breakpoint for each of the groups, i.e. the rule of the classification for each group.

5) The classification of the elements into the groups on the basis of the set rule.

6) Adaptation in accordance with some other conditions [14].

The ABC method is very well-known for its simplicity, but the same is criticized for the fact that it uses only one criterion for classification. Ever since Flores and Whybark [15] suggested that more than one criterion should be perceived, this field has actively been researched in [16]. That the ABC analysis should encompass several criteria has been widely accepted.

The methodology used in this paper includes the three main steps after the identification of the relevant criteria. First, the weights of certain criteria should be determined; second, each element per each criterion should be assigned a value. If elements are measured by different units, the second step includes the repeated scaling on the scale from 0 to 1, or 0 to 100. The final step is the combining of the weight coefficients and the values of the elements per certain criteria and obtaining the total values of the weights as per each element.

This approach reveals each element of inventories per each criterion, after which different results are combined by using the weighted additive function. Many analysts use the framework provided by the Analytic Hierarchy Process (the AHP method) [17]. The AHP is used so as to compare the criteria with the aim to determine the weight coefficient of each criterion. A comparison of the pairs of a thousand elements by adhering to each criterion is an impossible task to do. Instead of that, alternatives are assessed according to each criterion by using weights. These weight coefficients are determined once and the same can be used as long as the criteria themselves or the treatment of the same by the management do not or does not, respectively, change. During the decision-making process, it will express the joint conclusion of multiple experts as to the optimal solution [18].

The result can be used in order to rank the elements according to different categories. First, the decision maker identifies all the criteria important for the given problem. Second, the criteria are arranged following a certain hierarchy.

Third, a series of the comparison of pairs transforms subjective estimations into a set of weight coefficients [2].

In the process of comparing in pairs, the value from an appropriate comparison scale from 1 to 9 is assigned as a result of the comparison of the two alternatives (or two criteria) with each other. After the matrix of the comparison in pairs is formed, the weights of the alternatives (or criteria) are calculated.

Table 1
Saaty's Evaluation Scale [19]

Degree of preference (aij)	Verbal Judgment	Description
1	Of the same significance	Two elements have identical significance with respect to the goal.
3	Weak dominance	An experience or a judgment is slightly more in favor of one element in comparison to another.
5	Strong dominance	An experience or a judgment is substantially more in favor of one element in comparison to another.
7	Demonstrated dominance	The dominance of one element is confirmed in practice.
9	Absolute dominance	The dominance of the highest degree.
2,4,6,8	Intermediate value	A compromise or a further classification is needed.

The matrix A of the dimensions nxn is formed at the level of the criteria, in which there are the elements of $a_{ii} = 1$ (the elements of the matrix on the main diagonal are units), and the elements of a_{ji} are the reciprocal values of a_{ij}, i≠j, i, j=1, 2, ..., n, Equation (2).

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 1/a_{12} & \cdots & \cdots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 1/a_{1n} & \cdots & \cdots & a_{nn} \end{bmatrix} \quad (2)$$

The coefficient weights for the given matrix of comparison are calculated according to the formula: $A\omega = \lambda_{max}\omega$, where λ_{max} is the biggest eigenvalue of A and ω is the eigenvector corresponding to λ_{max}. Because of the features of the *A* Matrix, it follows that $\lambda max = n$, and the difference $\lambda max - n$ is used in measuring the consistency of estimations. In the case of inconsistency, the closer λmax is to *n*, the more consistent the estimation is.

The Consistency Index (CI) represents the measure of the deterioration of n from λ_{max}, and can be represented by the following formula:

$$CI = \frac{\lambda_{max} - n}{n-1} \quad (3)$$

By means of the Consistency Index, it is also possible to calculate the consistency ratio

$$CR = CI/RI \quad (4)$$

Where, RI is the random consistency index. The CR value should be less than 0.1, or otherwise the evaluation of the criteria is considered as inconsistent and the same should be repeated [20].

It is presumed that there are N elements, and that they should be classified into the A, B or C groups, depending on the classification according to the J criterion. Any one of the elements according to any one of the criteria is labeled with x_{ij}. There is a presumption that all the criteria are positively linked to the level of importance, i.e. the higher the value of the element per certain criterion, the bigger a chance for that element to be classified into Class A.

The proposed approach with weight coefficients is used in order to ensure that each element, as per several criteria, generates one result, called the optimal result of the element (5). The weight coefficients used for optimization are calculated as a group of coefficients whose total must equal 1 and which satisfy the conditions, Equations (6), (7) and (8) [4].

$$\max \qquad S_i = \sum_{j=1}^{J} w_{ij} * x_{ij} \quad (5)$$

$$\sum_{j=1}^{J} w_{ij} = 1, \quad (6)$$

$$w_{ij} - w_{i(j-1)} \geq 0, \quad j = 1, \dots, (J-1) \quad (7)$$

$$w_{mj} \geq 0, \quad j = 1, \dots, J \quad (8)$$

2.2 The XYZ Classification

The quantity of products in inventories in one sales enterprise should be in compliance with demand. The XYZ is used in those sales enterprises in which demand can dramatically vary from one to another of certain products. The XYZ analysis distributes the elements into the three groups, according to the characteristics of consumption. Group X consists of the products for which there is continuous demand, characterized by very slight oscillations, for which reason it is possible to forecast demand for this group with great accuracy; into Group Y, the products sold discontinuously, with fluctuations in demand, are classified, and forecasts for this group of products are of middle-degree accuracy; Group Z encompasses the products sold from time to time, and with big differences in the volume of demand, so the forecasting of demand is very difficult and with little accuracy [21].

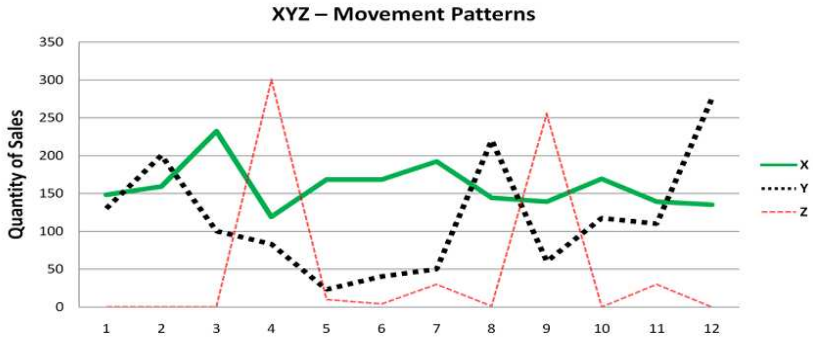

Figure 2
The demand patterns of XYZ products [21]

In the XYZ analysis, ranking is conducted according to the demand variability criterion viewed against average demand. It is needed to determine the variation coefficient, which is calculated as the ratio of the standard deviation and average sales. The variation coefficient is a relative measure of the dispersion of the probability distribution.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{9}$$

$$CV = \frac{\sigma}{\bar{x}} \tag{10}$$

The next step is the defining of the product groups and their forming on the basis of the obtained calculations. The proposed division is: Group X from 0% to 10% for the products whose demand can accurately be estimated; Group Y from 10% to 25% for the products whose demand can relatively accurately be predicted; Group Z from 25% to ∞ for the products whose demand can be predicted with very little precision [22]. As in the case of the ABC classification, ranks are arbitrary.

2.3 The Integrated ABC-XYZ Approach

The integrated ABC-XYZ approach is used to determine the activities for each of the defined groups of articles in a paired comparison matrix.

Group A/X consists of those elements with a big share in the total value, continuous consumption and the great accuracy of the demand forecast. These products make it possible to precisely plan and order, so there is no need to keep large safety quantities of inventories.

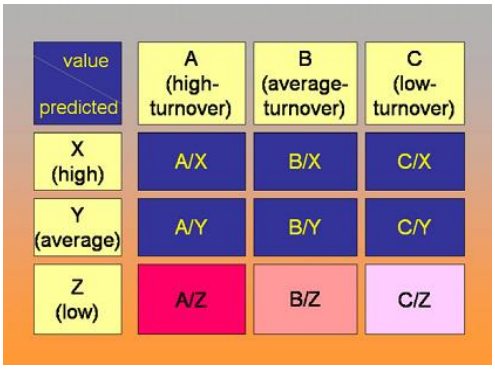

Figure 3
The combined ABC and XYZ analyses [23]

Group A/Y includes the products with a big share in the total value, but their consumption is discontinuous and the precision of their forecasting is lower. This group of products should be dedicated adequate attention when planning is concerned, so as to achieve purchase prices at the lowest cost possible.

Group A/Z consists of those products with a high share in the total value, but they are sold from time to time and demand for them can be forecasted with little accuracy. Inventories management is the most complicated within this group.

Group B/X consists of the products with a middle share in the total value, continuous consumption, demand for which is forecasted with great accuracy. When this group of products is concerned, the dynamics of purchase should be determined, simultaneously with determining the smallest inventory levels.

Group B/Y consists of the products with a middle share in the total value, discontinuous consumption and a middle-degree of accuracy for their forecasting demand.

Group C/X consists of the products with a small share in the total value, continuous consumption and the great accuracy of the forecasting of needs. These products should be ordered in accordance with the needs.

The products belonging to the groups B/Z, C/Y and C/Z have negligible impacts on an enterprise's business operations, so, they are purchased rarely and their planning is frequently neglected or left to suppliers in combination with some other product.

In general, the categories AX, BX and AY can be said to qualify for just-in-time approaches, whereas efforts must be minimized for the items of low value with bad demand predictability, which are located in the CZ category. All the remaining material groups in between must be individually investigated.

The following table summarizes the characteristics of the nine different material classes after combining the ABC-Analysis with the XYZ-Analysis.

Table 2
Part Characteristics in the Combined ABC-XYZ-Matrix [24]

	A	B	C
X	high value, high predictability continuous demand	medium value, high predictability continuous demand	low value, high predictability continuous demand
Y	high value, medium predictability fluctuating demand	medium value, medium predictability fluctuating demand	low value, medium predictability fluctuating demand
Z	high value, low predictability irregular demand	medium value, low predictability irregular demand	low value, low predictability irregular demand

In addition to that, different inventory strategies are also possible. The matching target inventory levels are shown in Table 3.

Table 3
Part Characteristics in the Combined ABC-XYZ-Matrix [25]

	A	B	C
X	low inventory	low inventory	low inventory
Y	low inventory	medium inventory	high inventory
Z	medium inventory	medium inventory	high inventory

In the continuation of the paper, we are going to demonstrate the application of the integrated ABC-XYZ approach on a practical example.

3 Empirical Study

The research was being conducted during the period of 12 months in the course of the year 2015, and the results were being collected on a monthly basis. In order to determine the optimal quantities of the products, the analysis was carried out as per product groups. The data analyzed in the paper relate to 44 articles from within the group of IT products consisting of Laptop/Notebook computers, and they consist of the price per unit of product and the realized monthly demand.

In order to include additional criteria in the classification as well, the data were collected about the delivery time (LT) from the supplier and the criticality (C) of certain articles. The criticality criterion is qualitative and is determined on the 3-value scale: 0.1, 0.5 and 1, where 0.1 represents the article which is not critical for the total offer of the enterprise and 1 is the article which is the key one for the enterprise's offer. The collected data and their transformed values are accounted for in Table 6 below.

For the ABC classification, the following parameters are set:

- Group A consists of the elements encompassing 0-80% of the obtained total value of the elements.

- Group B consists of the elements encompassing 80-95% of the obtained total value of the elements.
- Group C consists of the elements encompassing 95-100% of the obtained total value of the elements.

For the XYZ classification, the following parameters are set [25]:

- Group X consists of the elements whose variation coefficient is less than 0.5.
- Group Y consists of the elements whose variation coefficient is between 0.5 and 1.
- Group Z consists of the elements whose variation coefficient is bigger than 1.

Table 4

The initial data and their transformed values

Item No.	Total	Price (€)	Annual usage (€)	Lead time (day)	Criticality	AU transformed	LT transformed	C transformed
1	8	810.18 €	6,481.40 €	15	0.5	0.01	1	0.44
2	42	416.58 €	17,496.50 €	15	1	0.05	1	1.00
3	53	441.58 €	23,403.92 €	15	0.5	0.07	1	0.44
4	134	333.25 €	44,655.50 €	7	1	0.14	0.2	1.00
5	18	391.58 €	7,048.50 €	7	0.5	0.01	0.2	0.44
6	7	433.25 €	3,032.75 €	7	0.5	0.00	0.2	0.44
7	69	365.75 €	25,236.75 €	5	1	0.08	0	1.00
8	28	407.58 €	11,412.10 €	5	1	0.03	0	1.00
9	59	332.32 €	19,606.68 €	5	1	0.06	0	1.00
10	33	366.65 €	12,099.45 €	5	1	0.03	0	1.00
11	68	333.25 €	22,661.00 €	5	1	0.07	0	1.00
12	25	333.25 €	8,331.25 €	5	0.1	0.02	0	0.00
13	9	879.53 €	7,915.80 €	7	1	0.02	0.2	1.00
14	5	842.50 €	4,212.50 €	7	0.5	0.00	0.2	0.44
15	4	861.02 €	3,444.07 €	7	0.5	0.00	0.2	0.44
16	67	349.92 €	23,444.42 €	5	0.1	0.07	0	0.00
17	21	346.20 €	7,270.20 €	5	0.1	0.02	0	0.00
18	37	313.80 €	11,610.60 €	5	0.1	0.03	0	0.00
19	409	566.58 €	231,732.58 €	5	1	0.78	0	1.00
20	30	374.92 €	11,247.50 €	5	0.1	0.03	0	0.00
21	23	458.25 €	10,539.75 €	5	1	0.03	0	1.00
22	12	578.61 €	6,943.30 €	7	0.5	0.01	0.2	0.44
23	7	1,203.61 €	8,425.26 €	15	0.5	0.02	1	0.44
24	6	1,374.92 €	8,249.50 €	15	0.1	0.02	1	0.00
25	6	849.92 €	5,099.50 €	15	0.1	0.01	1	0.00
26	5	1,268.43 €	6,342.13 €	15	0.1	0.01	1	0.00
27	3	1,833.24 €	5,499.73 €	15	0.5	0.01	1	0.44

28	3	2,040.82 €	6,122.45 €	15	0.5	0.01	1	0.44
29	10	541.58 €	5,415.83 €	7	0.1	0.01	0.2	0.00
30	5	722.13 €	3,610.67 €	7	0.1	0.00	0.2	0.00
31	4	1,879.62 €	7,518.47 €	7	0.1	0.02	0.2	0.00
32	12	1,374.92 €	16,499.00 €	7	0.5	0.05	0.2	0.44
33	8	1,933.25 €	15,466.00 €	7	0.5	0.04	0.2	0.44
34	5	1,366.58 €	6,832.92 €	7	0.5	0.01	0.2	0.44
35	2	1,412.42 €	2,824.83 €	7	0.1	0.00	0.2	0.00
36	155	308.25 €	47,778.75 €	5	1	0.15	0	1.00
37	186	308.25 €	57,334.50 €	5	1	0.19	0	1.00
38	823	308.25 €	253,689.75 €	5	1	0.86	0	1.00
39	144	333.25 €	47,988.00 €	5	1	0.15	0	1.00
40	118	341.58 €	40,306.83 €	5	0.5	0.13	0	0.44
41	1181	249.91 €	295,141.74 €	5	0.5	1.00	0	0.44
42	196	366.58 €	71,850.33 €	5	1	0.24	0	1.00
43	103	366.58 €	37,758.08 €	5	1	0.12	0	1.00
44	115	366.58 €	42,157.08 €	5	1	0.13	0	1.00

In order to determine the weights of the criteria, the AHP procedure was conducted, and according to it, the weight value of 0.387 was obtained for the AU criterion; the weight value of 0.169 was obtained for the Lead Time criterion, and for the Criticality criterion, that value was 0.443. The score of each article and the group it belongs to according to the carried out ABC classification are presented in Table 5 below.

Table 5
The results of the ABC classification

Item No.	Total	Price (€)	AU	LT	C	Score	Group
2	42	416.58 €	17,496.50 €	15	1	0.446	A
4	134	333.25 €	44,655.50 €	7	1	0.445	A
13	9	879.53 €	7,915.80 €	7	1	0.443	A
7	69	365.75 €	25,236.75 €	5	1	0.443	A
8	28	407.58 €	11,412.10 €	5	1	0.443	A
9	59	332.32 €	19,606.68 €	5	1	0.443	A
10	33	366.65 €	12,099.45 €	5	1	0.443	A
11	68	333.25 €	22,661.00 €	5	1	0.443	A
19	409	566.58 €	231,732.58 €	5	1	0.443	A
21	23	458.25 €	10,539.75 €	5	1	0.443	A
36	155	308.25 €	47,778.75 €	5	1	0.443	A
37	186	308.25 €	57,334.50 €	5	1	0.443	A
38	823	308.25 €	253,689.75 €	5	1	0.443	A
39	144	333.25 €	47,988.00 €	5	1	0.443	A

42	196	366.58 €	71,850.33 €	5	1	0.443	A
43	103	366.58 €	37,758.08 €	5	1	0.443	A
44	115	366.58 €	42,157.08 €	5	1	0.443	A
3	53	441.58 €	23,403.92 €	15	0.5	0.201	A
23	7	1,203.61 €	8,425.26 €	15	0.5	0.198	A
1	8	810.18 €	6,481.40 €	15	0.5	0.198	A
28	3	2,040.82 €	6,122.45 €	15	0.5	0.198	A
32	12	1,374.92 €	16,499.00 €	7	0.5	0.198	B
27	3	1,833.24 €	5,499.73 €	15	0.5	0.197	B
33	8	1,933.25 €	15,466.00 €	7	0.5	0.197	B
5	18	391.58 €	7,048.50 €	7	0.5	0.197	B
22	12	578.61 €	6,943.30 €	7	0.5	0.197	B
34	5	1,366.58 €	6,832.92 €	7	0.5	0.197	B
14	5	842.50 €	4,212.50 €	7	0.5	0.197	B
15	4	861.02 €	3,444.07 €	7	0.5	0.197	B
6	7	433.25 €	3,032.75 €	7	0.5	0.197	C
40	118	341.58 €	40,306.83 €	5	0.5	0.197	C
41	1181	249.91 €	295,141.74 €	5	0.5	0.197	C
24	6	1,374.92 €	8,249.50 €	15	0.1	0.001	C
26	5	1,268.43 €	6,342.13 €	15	0.1	0.001	C
25	6	849.92 €	5,099.50 €	15	0.1	0.001	C
31	4	1,879.62 €	7,518.47 €	7	0.1	0.000	C
29	10	541.58 €	5,415.83 €	7	0.1	0.000	C
30	5	722.13 €	3,610.67 €	7	0.1	0.000	C
12	25	333.25 €	8,331.25 €	5	0.1	0.000	C
16	67	349.92 €	23,444.42 €	5	0.1	0.000	C
17	21	346.20 €	7,270.20 €	5	0.1	0.000	C
18	37	313.80 €	11,610.60 €	5	0.1	0.000	C
20	30	374.92 €	11,247.50 €	5	0.1	0.000	C
35	2	1,412.42 €	2,824.83 €	7	0.1	0.000	C

After the multi-criteria ABC classification, the groups were formed, in which Group A contains 21 articles, which accounts for 47.73% of the total number of the analyzed articles; Group B contains 8 articles and 18.18% of the total number of the articles; and Group C consists of 15 articles and accounts for 34.1%.

While performing the XYZ analysis, the data needed are those about the monthly sales in the observed period. The arithmetic mean and the standard deviation for each one of the determined articles are calculated. Then, the variation coefficient is calculated, on the basis of which coefficient products undergo the classification into the groups X, Y and Z, according to the set parameters and the results of the classification are displayed in Table 6 below.

Table 6
The results of the XYZ classification

Item No.	Arithmetic mean	Standard deviation	Variation coefficient	Group
17	1.75	0.829	0.47	X
19	34.08	11.594	0.34	X
36	12.92	5.560	0.43	X
38	68.58	17.217	0.25	X
39	12.00	5.788	0.48	X
41	98.42	19.543	0.20	X
42	16.33	7.930	0.49	X
43	8.58	3.523	0.41	X
2	3.50	2.693	0.77	Y
4	11.17	7.548	0.68	Y
5	1.50	1.118	0.75	Y
7	5.75	5.182	0.90	Y
8	2.33	2.211	0.95	Y
9	4.92	4.406	0.90	Y
10	2.75	2.005	0.73	Y
11	5.67	4.230	0.75	Y
12	2.08	1.706	0.82	Y
21	1.92	1.498	0.78	Y
32	1.00	0.816	0.82	Y
37	15.50	8.865	0.57	Y
40	9.83	5.505	0.56	Y
44	9.58	6.251	0.65	Y
1	0.67	0.943	1.41	Z
3	4.42	4.591	1.04	Z
6	0.58	0.759	1.30	Z
13	0.75	1.090	1.45	Z
14	0.42	0.640	1.54	Z
15	0.33	0.624	1.87	Z
16	5.58	6.304	1.13	Z
18	3.08	3.328	1.08	Z
20	2.50	2.843	1.14	Z
22	1.00	1.080	1.08	Z
23	0.58	0.759	1.30	Z
24	0.50	0.645	1.29	Z
25	0.50	0.957	1.91	Z
26	0.42	0.640	1.54	Z
27	0.25	0.595	2.38	Z
28	0.25	0.433	1.73	Z

29	0.83	1.213	1.46	Z
30	0.42	0.493	1.18	Z
31	0.33	0.850	2.55	Z
33	0.67	1.929	2.89	Z
34	0.42	0.862	2.07	Z
35	0.17	0.373	2.24	Z

After the XYZ analysis, which also introduces the level of demand in the consideration of the inventories, the three groups of articles are formed: the X group – to which the articles with high and relatively stable demand belong, and in which, in this case, there are 8 articles, which accounts for 18.18% of the total number of the analyzed articles; group Y – which is characterized by the articles following a particular trend of demand, namely the 14 such articles, accounting for 31.82% of the total number of the articles; and the C group, in which demand is irregular and unpredictable, with 22 articles, accounting for 50% of the total number of the analyzed articles.

Table 7

The result of the integrated ABC-XYZ analysis is given in the table

Item No.	The score obtained through the ABC classification	Variation coefficient	ABC classification	XYZ classification
19	0.443	0.340	A	X
36	0.443	0.430	A	X
38	0.443	0.251	A	X
39	0.443	0.482	A	X
42	0.443	0.486	A	X
43	0.443	0.410	A	X
2	0.446	0.769	A	Y
4	0.445	0.676	A	Y
7	0.443	0.901	A	Y
8	0.443	0.948	A	Y
9	0.443	0.896	A	Y
10	0.443	0.729	A	Y
11	0.443	0.746	A	Y
21	0.443	0.781	A	Y
37	0.443	0.572	A	Y
44	0.443	0.652	A	Y
1	0.198	1.414	A	Z
3	0.201	1.039	A	Z
13	0.443	1.453	A	Z
23	0.198	1.301	A	Z
28	0.198	1.732	A	Z

5	0.197	0.745	B	Y
32	0.198	0.816	B	Y
14	0.197	1.536	B	Z
15	0.197	1.871	B	Z
22	0.197	1.080	B	Z
27	0.197	2.380	B	Z
33	0.197	2.894	B	Z
34	0.197	2.069	B	Z
17	0.000	0.474	C	X
41	0.197	0.199	C	X
12	0.000	0.819	C	Y
40	0.197	0.560	C	Y
6	0.197	1.301	C	Z
16	0.000	1.129	C	Z
18	0.000	1.079	C	Z
20	0.000	1.137	C	Z
24	0.001	1.291	C	Z
25	0.001	1.915	C	Z
26	0.001	1.536	C	Z
29	0.000	1.456	C	Z
30	0.000	1.183	C	Z
31	0.000	2.550	C	Z
35	0.000	2.236	C	Z

The results of the integration of the ABC and XYZ classifications accounted for in Table 6 enable the formation of the 9 groups of products, where it is possible to suggest a special inventories management strategy with respect to each group.

Table 8

The division of the articles from the aspect of the multi-criteria ABC-XYZ analysis

	A	B	C
X	19,36,38,39,42,43		41, 17
Y	2,4,7,8,9,10,11,21,37,44	32,5	40, 12
Z	13,3,23,1,28	27,33,22,34,14,15	6,24,26,25,31,29,30,16,18,20,35

The results presented in the table show that Group AX contains 6 products, which account for 13.63% of the total number of the analyzed articles; Group AY consists of 10 products, accounting for 22.72%; these two groups of products represent those products that are dedicated greatest attention to from the logistical point of view. There is constant and predictable demand for these products, and they have a high share in the total financial result of the enterprise. There is a need for the constant monitoring of these articles and for the establishment of such a system of purchase that will be continual, monitoring demand according to quantities.

The articles from the AZ group, namely the 5 articles found in the group, are those with a high yield because their unit price is high but demand for them appears from time to time. A proposal is made with respect to these articles that they should have minimal inventories depending on demand. The BY articles require the keeping for safety inventories. The BZ, CY and CZ groups require the least attention, so they can also be analyzed. Some articles can be declared as unneeded and they can be exempt from making further orders, whereas when the other articles are concerned, it is possible to form group orders so as to reduce the costs of purchase, simultaneously forming certain inventories in order to fulfill the requirements of demand.

Conclusion

Our contemporary, competitive environment calls for efficiency in the circulation of goods from the supplier to the consumer, for all the segments of the supply chain. In order to achieve the targeted level of service towards consumers, it is necessary that inventories should be managed in a satisfactory and effective manner. In the search for a balance between these two contradictory goals, managers draw on various techniques, which, unfortunately, are often experiential.

Supply chain management is a process of an efficient integration of producer and supplier; and storeroom and buyer, in such a manner that produced goods are distributed in optimal quantities to reduce the costs of business operations, while simultaneously satisfying the buyer.

Inventories play an exceptionally big role in retail business enterprises. Losses from inventories, accounting for up to 1% of retail sales, are assessed as good, while in numerous retail shops the same can account for over 3% of sales. According to some research studies, leading enterprises in the retail field lose from 10% to 25% of their profits due to the inappropriate management of their inventories.

Today, inventory management is one of the most important tasks an enterprise is faced with, on a daily basis. The main goal of inventory management is to minimize the volume and the time of the engagement of working capital in inventories. Consequently, if inventories are treated in a poor manner, interruptions in production are possible, as well as, the loss of inventories due to being stored for too long. In order to avoid that loss, there are numerous systems and methods for inventory management, the ABC analysis being one of the most popular.

There are, however, limitations to the ABC analysis, which are overcome by introducing the XYZ analysis. The XYZ can be said to be a secondary analysis of inventories, which enables the following step in the inventories analysis – the application of the demand variability criterion in comparison with the average level of demand. A symbiosis of the two analyses results in the integrated model for the ABC-XYZ for the classification and optimization of inventories. The

purpose of the application of this method is the establishment of the optimal inventory level, which is one of the key conditions for cost reductions within an enterprise.

The presented inventory analysis system, focused on Laptop/Notebook computers, is indicative of the practical application of the ABC-XYZ analysis. As we can see, the products that should be paid greater attention to, as well as, those that should be paid lesser attention to, in the purchase operation, have been identified. Moreover, we have also determined which products are not necessary in the product mix.

Given the fact that there are few significant possibilities for reducing costs in an Enterprise, the optimization of inventories represents one of the key ways for an Enterprise to be more profitable. The application of the ABC-XYZ analysis would improve the decision-making process in an enterprise, with respect to its inventories domain, which consequently contributes to a reduction in costs and assures, a better competitive position for an Enterprise.

References

[1] Dickie, H. F., ABC Inventory Analysis Shoots for Dollars, Not Pennies, Factory Management and Maintenance (1951) pp. 92-94

[2] Flores, B. E., Olson, D. L., & Dorai, V. K., Management of multicriteria inventory classification, Mathematical and Computer Modelling, Vol. 16, No. 12 (1992) pp.71-82

[3] Cedillo-Campos, M., & Cedillo-Campos, H., w@reRISK method: Security risk level classification of stock keeping units in a warehouse, Safety Science 79 (2015) pp. 358-368

[4] Ramanathan, R., ABC inventory classification with multiple-criteria using weighted linear optimization, Computers and Operations Research, Vol. 33, No. 3 (2006) pp. 695-700

[5] Hatefi, S., & Torabi, S., A Common Weight Linear Optimization Approach for Multicriteria ABC Inventory Classification, Hindawi Publishing Corporation, Advances in Decision Sciences (2015)

[6] Ng, W., A simple classifier for multiple criteria ABC analysis, European Journal of Operational Research 177 (2007) pp. 344-353

[7] Hadi-Vencheh, A., An improvement to multiple criteria ABC inventory classification, European Journal of Operational Research 201, (2010) pp. 962-965

[8] Partovi, F. Y., & Anandarajan, M., Classifying inventory using an artificial neural network approach, Computers & Industrial Engineering, 41(4) (2002) pp. 389-404

[9] Yu, M.-C., Multi-criteria ABC analysis using artificial-intelligence-based, Expert Systems with Applications 38 (2011) pp. 3416-3421

[10] Chu, C., Liang, G., & Liao, C., Controlling inventory by combining ABC analysis and fuzzy classification, Computers & Industrial Engineering 55 (2008) pp. 841-851

[11] Buliński, J., Waszkiewicz, C., & Buraczewski, P., Utilization of ABC/XYZ analysis in stock planning in the enterprise, Annals of Warsaw University of Life Sciences – SGGW Agriculture No 61 (Agricultural and Forest Engineering) (2013) pp. 89-96

[12] Choudary, Y., & Balaji, N., "Inventory Planning Optimization" The Challenges with Segmenting and Extrapolating Demand, International Journal of Scientific & Engineering Research, Volume 6, Issue 3 (2015) pp. 159-163

[13] Stoll, J., Kopf, R., Schneider, J., & Lanza, G., Criticality analysis of spare parts management: a multi-criteria classification regarding a cross-plant central warehouse strategy, Production Engineering. Research and Development (2015)

[14] Chu, C., & Chu, Y., Computerized ABC Analysis: The Basis for Inventory Management, Computers & Industrial Engineering, Vol. 13 (1987) pp. 66-70

[15] Flores, B., & Whybark, D., Multiple Criteria ABC Analysis, International Journal of Operations & Production Management Vol. 6, No. 3 (1986) pp. 38-45

[16] Ye Chen, K. W., A case-based distance model for multiple criteria ABC analysis, Computers & Operations Research 35 (2008) pp. 776-796

[17] Partovi, F., & Burton, J., Using the Analytic Hierarchy Process for ABC Analysis, International Journal of Operations & Production Management, Vol. 13, No. 9 (1993) pp. 29-44

[18] Saaty, T. The analytic hierarchy process. New York: McGraw-Hill (1980)

[19] Szűts, A., Krómer, I., Developing a Fuzzy Analytic Hierarchy Process for Choosing the Energetically Optimal Solution at the Early Design Phase of a Building, Acta Polytechnica Hungarica Vol. 12, No. 3 (2015) pp. 25-39

[20] Balaji, K., & Senthil Kumar, V., Multicriteria Inventory ABC Classification in an Automobile Rubber Components Manufacturing Industry, Procedia CIRP 17 (2014) pp. 463-468

[21] Nowotyńska, I., An Application of XYZ Analysis in Company Stock Management, Modern Management Review, Vol. XVIII, 20 (2013) pp. 77-86

[22] Dhoka, D., & Choudary, Y., "XYZ" Inventory Classification & Challenges, IOSR Journal of Economics and Finance (IOSR-JEF) Volume 2, Issue 2 (2013) pp. 23-26

[23] Clevert, D.-A., & all., Cost analysis in interventional radiology – A tool to optimize management costs, European Journal of Radiology 61 (2007) pp. 144-149

[24] Wannenwetsch, H. Integrierte Materialwirtschaft und Logistik. Beschaffung, Logistik, Materialwirtschaft und Produktion. 4th ed. Berlin: Heidelberg (2010)

[25] Sommerer, G. Unternehmenslogistik. Ausgewählte Instrumentarien zur Planung und Organisation logistischer Prozesse. München, Hanser (1998)

[26] Scholz-Reiter, B., Heger, J., Meinecke, C., & Bergmann, J., Integration of demand forecasts in ABC-XYZ analysis: practical investigation at an industrial company, International Journal of Productivity and Performance Management, Vol. 61, Iss: 4 (2012) pp. 445-451

The Category Proliferation Problem in ART Neural Networks

Dušan Marček

Department of Applied Informatics, Faculty of Economic,
VŠB-Technical University of Ostrava,
Sokolská 33, 702 00 Ostrava, Czech Republic
Dusan.Marcek@vsb.cz

Michal Rojček

Department of Informatics, Faculty of Education,
Catholic University in Ružomberok,
Hrabovská cesta 1, 034 01 Ružomberok, Slovak Republic
Michal.Rojcek@ku.sk

Abstract: This article describes the design of a new model IKMART, for classification of documents and their incorporation into categories based on the KMART architecture. The architecture consists of two networks that mutually cooperate through the interconnection of weights and the output matrix of the coded documents. The architecture retains required network features such as incremental learning without the need of descriptive and input/output fuzzy data, learning acceleration and classification of documents and a minimal number of user-defined parameters. The conducted experiments with real documents showed a more precise categorization of documents and higher classification performance in comparison to the classic KMART algorithm.

Keywords: Improved KMART; Category Proliferation Problem; Fuzzy Clustering; Fuzzy Categorization

1 Introduction

The number of various electronic documents grows enormously every day. It appears that it is necessary to search for new algorithms for their fast and reliable classification [1] [2]. New document classification algorithms contribute to this objective; however, descriptive data for classifiers are mostly not available. Therefore, fully controlled classification approaches are not entirely appropriate for broader deployment, for example on the web. Categorization approaches contained in algorithms of non-controlled learning appear to be more suitable for

broader deployment [3]. A wide application of neural networks based on the theory of adaptive resonance (ART) was found in document clustering and classification tasks. Some of the applications are briefly described in Section 2, more may be found, for example, in the works [4]–[9].

This work is organized into 6 sections. Section 2 generally deals with problems of category proliferation and methods of minimizing of their occurrence. In Section 3, we present the types of ART networks based on fuzzy clustering, by which, it is possible to categorize overlapping data into more categories with various membership degrees. In Section 4, we present the learning algorithm of a KMART (Kondadadi & Kozma Modified ART) network with the cluster creating principle. The core of the contribution is created by Sections 5 and 6. In Section 5, we propose a new model for the optimized algorithm KMART, called IKMART (Improved KMART), which enables to optimize the dilemma of stability/plasticity, increase the precision of categorization and influence the speed of categorization. In Section 6, we present results of experiments of the categorization of real text documents, which contextually overlap. The conclusion provides a brief summary.

2 The Category Proliferation Problem in ART Networks

The category proliferation problem, which was described in the works [10]–[14], often occurs in the categorization of documents using ART or ARTMAP networks. Category proliferation leads to the creation of a large number of categories, which mostly decrease the precision of categorization [10].

Category proliferation may occur due to various reasons such as noise [15], training with large datasets (overtraining) [16], or due to unsuitable setting of network parameters [17]. Another reason of the category proliferation occurrence, as stated in literature, may be a state when a network is trained with data of related content [16], [18]. For contextually related input documents there are various categories as well as their mutual intersections created by a network, thus, it is not easy to correctly generalize an input area of documents.

Various methods on how to deal with the category proliferation problem in Fuzzy ART and Fuzzy ARTMAP networks have been devised. More broadly, there are basically two kinds of methods for the minimization of category proliferation: (1) post-process methods, which are realized in networks after the completion of a training process. These methods are based on the cut rule [19], which removes redundant categories based on their frequency of use and precision, or (2) adjustment methods in construction of a learning algorithm in order to avoid a large number of categories even before they are created [20]. This method

includes modifications in the way of learning [21] and actualization of the network weight system [22], with the objective of decreasing category proliferation resulting from noisy inputs, as well as, Fuzzy ARTMAP variants, Distributed ARTMAP [23], Gaussian ARTMAP [23] and boosted ARTMAP [23].

Isawa [20] designed the improvement of a Fuzzy ART algorithm, called C-FART, based on the connection of overlapping categories in order to remove the category proliferation problem. An important feature of this approach is control of the threshold parameter *AT* for individual categories and its change in the learning process. The parameter determines if categories merge or if they stay unmerged. In the work [10] there were suggested changes in the learning algorithm of a Fuzzy ARTMAP network, which enable a network to predict more than one class during classification. There was introduced a threshold value of activation, which enabled a network to create more than one prediction of a class when it was necessary, especially for patterns of overlapping areas between classes. A part of this algorithm is also the suppression of formation of small categories, which improved the categorization and predictive precision. Other features dealing with the category proliferation problem in ART networks can be found for example in literature [11], [21], [24], [25], which focus mostly on removing of the category proliferation problem in ARTMAP networks caused by noisy data. In the works [10], [14] authors deal with the creation of proliferation from the perspective of overlapping input data. In these works, data are categorized only into one winning category, which is unsuitable for text document processing applications, because in output categories there is removed the possible content context of documents with different categories.

For the correction of creation of new categories, there is a vigilance parameter ρ used in most ART networks, however, its change has only little effect. This is notable especially on a set of synthetic documents. The greatest progress in this direction has been reached by Isawa [14], who introduced a threshold parameter *AT* within a Fuzzy ART algorithm for similar categories and its change during the learning process. However, this approach does not guarantee complete stability (immutability) of categories; it only reduces several similar categories by connecting them.

3 Fuzzy Clustering and the Categorization by a Fuzzy ART Network

The literature overview stated in the previous section showed that none of the published works in the area of category proliferation problems solves fuzzy approaches enabling to categorize overlapping data into more categories with a varying membership degree. The stated works categorize data only into one winning category. For example, if there exists, a document that belongs to the

category of atheism as well as to the category of Christianity, it is expected to be classified into both categories with a certain membership degree, not only into a winning category. Therefore, there were further developed ART networks based on fuzzy clustering, which are suitable for binary and analogous input data. There have been methods published, which suggest various ways of fuzzy clustering such as a system of concept duplication [17] for an ART1 network, an IFART (Improved Fuzzy ART) system for a Fuzzy ART network [26], and a KMART system also for a fuzzy ART network [5]. Based on the stated methods, the KMART method appears to be the most suitable for the concept of fuzzy clustering in ART networks, because the method of concept duplication is demanding on computing memory and moreover, it implements an evidence parameter, which has large memory requirements at low values and at higher values a network starts to behave unstably [17]. An IFART network is based on the post-process method, which calculates the membership of data in clusters after their formation by a very difficult calculating process, because after the clustering process it has to go through all data (e.g. documents) in all clusters and calculate membership degrees of every data instance in all clusters based on cluster centers [26]. In a KMART network, a membership of documents in individual clusters is calculated directly in the learning algorithm. This approach to fuzzification is simple from the calculating and implementation perspective and it brings also further advantages such as reduction of user-defined input parameters [5]. Its learning algorithm with the description of cluster formation is stated in the following section.

4 Algorithm and the Description of Cluster Formation in a KMART Network

In the work [5], there was suggested a variation of the existing Fuzzy ART algorithm [27], so that it is possible to apply Fuzzy clustering. This system is called KMART according to its authors Kondadadi & Kozma [3] and its steps are stated in Table 1.

The learning algorithm KAMART is based on a modified version of a fuzzy art network. Instead of choosing maximal similarity of a category and using the vigilance test for verification if a category is close enough to an input pattern, there can be controlled every category in the recognition layer by application of the vigilance test. If a category passes the vigilance test, then an input document is inserted into this particular category.

Measurement of similarity lies within the vigilance test that defines the membership degree of a given input sample, in an actual cluster. It enables a document to be in more clusters with a different membership degree. All prototypes that pass the vigilance test are actualized according to the learning rule

(4). This modification has two other advantages compared to a fuzzy ART network. Firstly, a fuzzy ART network is time consuming because it requires iterative browsing during searching for a winning category that satisfies the vigilance test. In the described modification, this searching is not necessary because every node in the recognition layer has already been controlled. This makes the model less difficult for calculation. Another advantage is that by eliminating the category choice step, we are avoiding the use of a choice parameter α. This will reduce a number of user-defined parameters in the system. This modification does not violate the underlying principal of an ART network, i.e. to avoid the dilemma of stability and plasticity. KMART is still an incremental clustering algorithm and before learning a new input it controls the input and it learns an output pattern only if it corresponds to any of the stored patterns with a certain tolerance.

Table 1
Learning algorithm of a KMART (Kondadadi & Kozma Modified ART) network

1. Load a new input vector (document) I containing binary or analogous components.

   Let I:=[subsequent input vector]

2. Calculate membership degrees for all outputs $y(j)$ (it is a membership degree of a document in *j* category) based on the relationship:

   $$y(j) := \frac{|I \wedge w_j|}{|I|}, \quad (1)$$

   Where, $\wedge$ is fuzzy AND operator, defined as: $(x \wedge y) = \min(x_i, y_i)$.

3. Match the calculated value $y(j)$ to the matrix map, on a place of actually processed category j $(j > 1)$ and document doc $(doc > 1)$:

   $$map(j, doc) := y(j) \quad (2)$$

4. Vigilance test:

   If $y(j) \geq \rho$, then go to the step 5, otherwise go to the step 6. (3)

5. Actualize the winning neuron (learning rule):

   $$w_j^{(new)} := \beta\left(I \wedge w_j^{(old)}\right) + (1 - \beta) w_j^{(old)} \quad (4)$$

6. Return: go to the step 2, while ≤ max number of categories, otherwise go to the step 1. If there is no other vector (document) in order or $w^{(new)} = w^{(old)}$, then finish.

5 Proposal of a Modified Model of KMART Network for Fuzzy Clustering and the Categorization of Contextually-related Documents

It has been shown that by modification of the original Fuzzy ART neural network there can be reached the excellent results in the area of clustering and categorization of text documents [5], [7], [28]–[30]. One of the above described modifications, which enables fuzzy clustering is a KMART network [5]. There are also newer approaches to fuzzy clustering for ART networks [17], [26], however, these have serious deficiencies described in section 3. Therefore, our proposed modified model of a KMART network is based on the KMART network stated in the work [5]. The objective of the proposed modification is to remove the category proliferation problem caused by the influence of text documents overlapping in content, apply a fuzzy approach in the categorization of these documents and optimize features of the model – especially stability and plasticity of categories, the precision of categorization and computing speed – on real text documents. The model consists of two separate parts (see Figure 1): the fuzzy clustering part (KMART) and the fuzzy categorization part (modified KMART). These parts are interconnected by a mutual layer, which is created by matrixes of fuzzy categories and documents *map* and network weights w_{ij}.

The function of the fuzzy clustering part of the model, based on the KMART network, is designed to keep plasticity of categories. It means that a training set of text documents chosen by a user will suitably create or expand a number of categories. In the second run, one representative document is sufficient to add a new category. A representative document should ideally contain as much as possible common keywords with the categorized documents from the fuzzy categorization part, which should belong to this category. As both parts work with network weights w_{ij} in both directions, i.e. for writing and reading, both arrows in Figure 1 are double-headed. Only output values of documents' membership degree in individual categories are recorded in the matrix of documents and categories *map*, thus the communication direction is single.

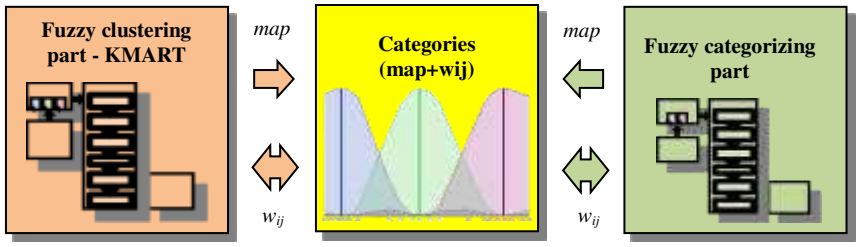

Figure 1
General view of the model architecture after connection of both parts

The function of the fuzzy categorization part is designed to maintain stability of the categories. As this part of the model is prevented from the possibility to create new categories, the absolute stability of categories even in case of contextually overlapping documents is assured, which contributes to solve the category proliferation problem. The fuzzy categorization part is based on the learning algorithm of a KMART network, and it is based on the following three adjustments of the original algorithm from Table 1.

After the calculation of membership degrees for all outputs $y(j) := \frac{|I \wedge w_j|}{|I|}$ and their integration into the output matrix *map*, there is omitted the vigilance test $y(j) \geq \rho$, based on which it is decided if a new category will or will not be created. This step (step no. 4 from the algorithm in Table 1) was completely removed together with the difficult set up of the vigilance parameter ρ. The membership degree y is calculated for all documents and categories based on the equation (5) (step no. 2 from the algorithm in Table 2). The creation of new categories was prevented by this adjustment. At the same time, there was cancelled the burden of creation of new categories (by omitting the increment of category calculation and adding new rows to the matrix *map* and weights w_{ij}).

The second adjustment lies in a partial removal of the step for the weight adaptation (learning rule) $w_j^{(new)} := \beta\left(I \wedge w_j^{(old)}\right) + (1-\beta)w_j^{(old)}$. Removing of this step in the algorithm in Table 1 will not violate the precision of a set of synthetic documents or in a training set of real documents. In case of testing of a real document set, it is necessary to return this step back because the precision of categorization would be decreased. In case of removing of the weight adaptation there will also be removed the last user-defined parameter, which is the learning speed β.

The third adjustment of the algorithm assures its stability and resistance against its cycling. The KMART algorithm can reach a stable state in case of satisfying of the condition: $w^{(new)} = w^{(old)}$. It means that in the previous and current state there is no change of weights ($\Delta w = 0$).

It often happens in practice, that e.g. in case of wrong set up of parameters weighs will oscillate and the stability condition is not fulfilled ($\Delta w > 0$). The adjustment consists of removal of this condition. The algorithm ends when membership degrees for all incoming documents to all exiting categories are calculated.

Regarding the categorization part in Figure 1, the matrix of documents and categories *map* as well as the network weights w_{ij} are shared also for the second categorization part of the model. Thus, the categorization part of the model is connected to a learned network through these two matrixes and it uses it for its processes. After the description of performed adjustments in the algorithm KMART, there is the new fuzzy categorization algorithm IKMART stated in Table 2.

Table 2
Steps of the new algorithm IKMART

1. Load a new input vector (document) I containing binary or analogous parts.

   Let I:=[subsequent input vector].
   If there is no document in order, go to the step 6.

2. Calculate membership degrees for all outputs $y(j)$ (it is a membership degree of a document to *j* category) based on the relationship:

   $$y(j) := \frac{|I \wedge w_j|}{|I|}, \tag{5}$$

   where $\wedge$ is fuzzy operator AND, defined as: $(x \wedge y) = \min(x_i, y_i)$.

3. Match the calculated value $y(j)$ to the matrix map, on a place of the actually processed category j $(j > 1)$ and document doc $(doc > 1)$:

   $$map(j, doc) := -y(j) \tag{6}$$

   Negative value $-y$ is a distinguishing feature in order to identify which algorithm calculated the given value in the mutual matrix *map*. Algorithm KMART uses positive values.

4. Weight adaptation w_j:

   $$w_j^{(new)} := \beta\left(I \wedge w_j^{(old)}\right) + (1 - \beta) w_j^{(old)} \tag{7}$$

5. Return to the step 2, until $j \leq max$, where *max* stands for the maximum number of categories, otherwise go to the step 1.
6. The end of algorithm.

In the following, we present the behavior of the algorithm IKMART and results of the testing on a real situation with real text documents.

6 Experiments – The Categorization of Real Text Documents

The objective of the experiment is to verify if the proposed model reaches the required stability of categories and if there occurs an improvement of quality and

speed in comparison to the original KMART model also on real text documents, which are contextually overlapping.

Figure 2 schematically shows the overlapping of sets of individual documents. Based on Figure 2, we define two basic characteristics for the evaluation of categorization quality: Precision and Recall [31].

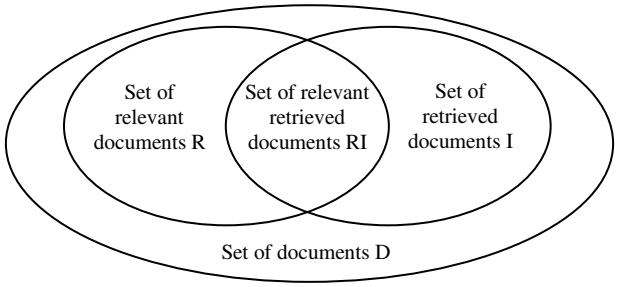

Figure 2
Relationship between the document sets

Precision P can be defined based on the relationship:

$$P = \frac{|RI|}{|I|}, \tag{8}$$

where $|RI|$ is a number of retrieved relevant documents and $|I|$ is a number of all retrieved documents. Recall R can be defined as a ratio of a number of retrieved relevant documents ($|RI|$) the number of relevant documents ($|R|$):

$$R = \frac{|RI|}{|R|} \tag{9}$$

For the calculation of categorization quality there is usually used the so-called F-measure (or also F1 score). The F-measure is a value, which is a compromise between the precision P and recall R and it serves to overall evaluation of quality of the information processing model. It is expressed by the following relationship:

$$F - measure = 2 \cdot \frac{P \,.\, R}{P+R} \tag{10}$$

Text documents are selected from the corpus 20 Newsgroups[1]. It is a corpus consisting of English texts from email discussion groups. The corpus in total contains 20 topics (categories) such as: sport, computers, religion, politics, science, electronics, medicine and so on.

The *training matrix* contains 500 selected pre-processed text documents from the corpus 20 Newsgroups, each with 118 terms. The documents are divided into five categories, in each of them there are 2 x 50 = 100 documents. In order to create more precise clusters, documents are duplicated (2 x repeated in every category).

[1] Available at: http://qwone.com/~jason/20Newsgroups/

Since the documents are for an algorithm without learning, this set does not contain information (description) to which categories should a given document belong. Therefore, it was necessary to repeat 50 documents for each category. Thus, there was reached more precise clustering of documents into categories. Otherwise the KMART network created an incorrect structure of categories. *The training matrix* of documents and terms is built by the method Term Frequency - Inverse Document Frequency (TF-IDF). *The input matrix* contains the following categories: 1. Hockey, 2. Christianity, 3. PC hardware, 4. Atheism, 5. MAC hardware. *The testing matrix* contains 100 pre-processed documents from the same corpus as the training matrix, each with 118 terms. Documents are divided into two categories with 50 documents, while every document belongs to two categories at the same time. The testing matrix is again set up by the method TF-IDF and it contains the following two different double combinations. The first combination is labeled as Windows (expected context with 3^{rd} and 5^{th} category from the training matrix) and the second one is the combination with the label Religion (expected context with 2^{nd} and 4^{th} category from the training matrix).

In the process of the experiment, the KMART network was firstly provided with the training set. The network created the structure of five categories within the clustering process (hockey, Christianity, PC hardware, Atheism, MAC hardware). The process was subsequently repeated in order to prove that the network had learned correctly. At the most optimal value of parameters $\rho = 0.61$ and $\beta = 1$ (determined experimentally), there was the maximum membership degree 0.927 reached for the training set (see Table. 3).

Table 3
Results of algorithms with the training set – real documents

Algorithm and input set	β	ρ	F-measure	CPU time [s]	Number of iterations	Number of created categories
KMART $_{TRAIN}$	1	0.61	0.927	1.547	3	5
KMART $_{TRAIN}$	1	0.61	0.927	0.567	1	0
Fuzzy Kat $_{TRAIN}$ without weight adaptation	-	-	0.927	0.524	1	0
Fuzzy Kat $_{TRAIN}$ with weight adaptation	1	-	0.335	0.551	1	0

The Fuzzy categorization algorithm IKMART was further modified in this experiment so that for reaching of a more precise categorization we applied also the step of weight adaptation (learning rule) according to the expression (7). Thus, there were created two versions of the fuzzy categorization algorithm IKMART: without the weight adaptation and with the weight adaptation. Experiments showed that in case of the training set there were reached significantly higher

values of a membership degree of documents in categories with the original version of the algorithm without the weight adaptation (see Table 3).

Table 4 shows the reached values of the quality of document categorization into categories and values of algorithm performance given by CPU times on the testing set of real documents. At the fuzzy categorization, there is monitored if a document reached the first highest membership degree (1st HMD) in its category, i.e. if there is a document about hockey at the input of the KMART network it should reach 1st HMD in the cluster identified as the hockey category. If there is a document at the input of the network that has context with e.g. two already created categories, then there is the calculation of membership degrees monitored in both categories, i.e. 1st HMD and 2nd HMD. Then there are calculated F-measures for these two categories (1st HMD and 2nd HMD). Until now, the behavior of individual algorithms of the training set was monitored. The first part of the experiment finished here. Results are stated in Table 3. Subsequently, it was necessary to test the algorithm, with the testing set.

The experiment in the second part started from the beginning by repeated training of the KMART network by the training set and then all the algorithms were tested by the testing set with documents from new categories: Windows and Religion. **It was proven that the use of the Fuzzy categorization algorithm with the weight adaptation reaches better values in all monitored parameters than the repeated use of the KMART algorithm.** Unlike experiments with synthetic documents [32], this did not create any new category (which is correct) but both F-measures were lower than in the Fuzzy categorization algorithm with the weight adaptation and equal to or lower than in the Fuzzy categorization algorithm without the weight adaptation. It is caused by the incorrect document categorization, what is also shown in Figure 3. The CPU time was lower for both versions of the Fuzzy categorization algorithm, because the KMART network needed 2 iterations for stabilization. If there was the parameter $\beta<1$ in the KMART network, then saving of the CPU time in case of the Fuzzy categorization algorithm would be significantly higher.

Table 4

Results of algorithms with the training and testing set – real documents

Algorithm and input set	β	ρ	F-measure 1st HMD	F-measure 2nd HMD	Average of F-measures	CPU time [s]	Number of iterations	Number of created categories
KMART $_{TRAIN}$	1	0.61	0.927	-	-	1.475	3	5
KMART $_{TEST}$	1	0.4	0.667	0.667	0.667	0.260	2	0
Fuzzy Kat $_{TEST}$ without weight adaptation	-	-	0.830	0.667	0.749	0.110	1	0
Fuzzy Kat $_{TEST}$ with weight adaptation	0.4	-	**0.953**	**0.758**	**0.856**	**0.110**	1	0

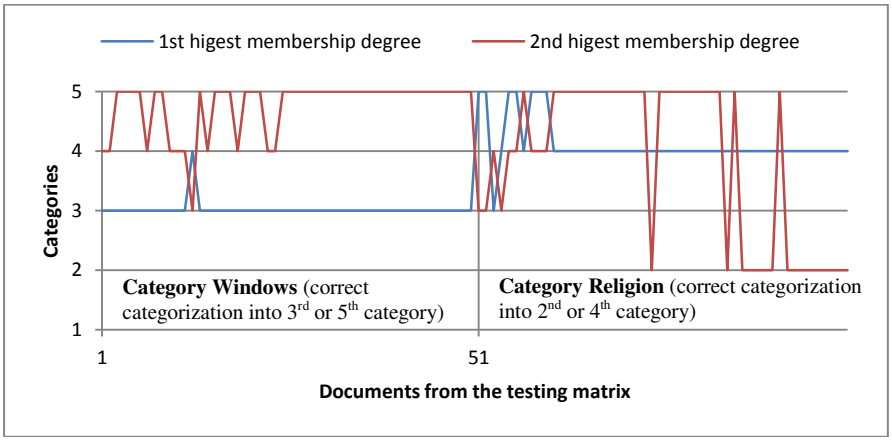

Figure 3
Graf of the document categorization from the testing matrix into categories by the Fuzzy categorizing algorithm with the weight adaptation

Figure 3 shows the behavior of the fuzzy categorization algorithm with the weight adaptation of the testing matrix. The graph illustrates the more concrete and precise progress of both membership degrees than it was in case of the KMART network. In case of documents belonging to both categories (Windows and Religion) from the testing matrix, there were correctly recognized both expected categories from the training matrix.

Conclusions

This work is devoted to the issue of decreasing category proliferation, as an adverse effect occurring in a network of the ART type, which in the end decreases the precision of document categorization. The core of the contribution lies in the proposal of the new architecture of the ART network type, with the aim to minimize category proliferation and at the same time increase the category performance. In the article, there was proposed a model for an Improved KMART (IKMART) network consisting of a block of clustering, operated by the fuzzy clustering algorithm KMART and a block of fuzzy categorization operated by the developed categorization algorithm IKMART. These are interconnected with the matrix of documents and the matrix of fuzzy categories. The IKMART model was verified for the categorization of real overlapping contextually similar text documents. Results of the verification showed that the proposed model provides stability of categories and a better qualitative, as well as, performance values on a domain of real text documents belonging to more categories than the separate basic model KMART. It can be concluded that the proposed model contributed to solving of the category proliferation problem in ART networks, caused by content related documents, with more existing categories. Next proposed, is the model IKMART compared to the conventional fuzzy clustering model Fuzzy C-Means, alternatively, with further variations, such as, Gustafson-Kessel Fuzzy C-Means, or Kernel-based Fuzzy C-Means.

Acknowledgement

The article was prepared within the project TA04031376 „Research / development training methodology aerospace specialists L410UVP-E20“. This project is supported by Technology Agency Czech Republic. This article was also supported within Operational Programme Education for Competitiveness – Project No. CZ.1.07/2.3.00/20.0296.

References

[1] I. Černák and A. Kelemenová, Artificial life on selected models, methods and means. Ružomberok: VERBUM-Editorial Center Faculty of Education Ružomberok, 2010

[2] I. Černák and M. Lehotský, „Some possibilities for implementing neural networks to the teleinformatic practice", in Kognice a umělý život VI, 2006, pp. 125-128

[3] E. K. Jacob, „Classification and Categorization : A Difference that makes a Difference", Libr. Trends, Vol. 52, 2004, pp. 515-540

[4] N. Ngamwitthayanon and N. Wattanapongsakorn, „Fuzzy-ART in network anomaly detection with feature-reduction dataset", 7th Int. Conf. Networked Comput., 2011, pp. 116-121

[5] R. Kondadadi and R. Kozma, „A modified fuzzy ART for soft document clustering", Proc. 2002 Int. Jt. Conf. Neural Networks. IJCNN'02, Vol. 3, 2002

[6] L. Massey, „On the quality of ART1 text clustering", in Neural Networks, 2003, Vol. 16, pp. 771-778

[7] G.-B. V P., L.-P. C., de-Moya-Anegon F., and H.-S. V., „Comparison of neural models for document clustering", Int. J. Approx. Reason., Vol. 34, 2003, pp. 287-305

[8] S. Kim and D. C. Wunsch, „A GPU based Parallel Hierarchical Fuzzy ART clustering", 2011 Int. Jt. Conf. Neural Networks, 2011, pp. 2778-2782

[9] I. Dagher, „Fuzzy ART-based prototype classifier", Computing, Vol. 92, No. 1, 2010, pp. 49-63

[10] W. Y. Sit, L. O. Mak, and G. W. Ng, „Managing category proliferation in fuzzy ARTMAP caused by overlapping classes", IEEE Trans. Neural Networks, Vol. 20, 2009, pp. 1244-1253

[11] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, „Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps", IEEE Trans. Neural Networks, Vol. 3, No. 5, 1992, pp. 698-713

[12] A. Al-Daraiseh, A. Kaylani, M. Georgiopoulos, M. Mollaghasemi, A. S. Wu, and G. Anagnostopoulos, „GFAM: Evolving Fuzzy ARTMAP neural networks", Neural Networks, Vol. 20, 2007, pp. 874-892

[13] G. A. Carpenter and B. L. Milenova, „Distributed ARTMAP", IJCNN'99. Int. Jt. Conf. Neural Networks. Proc. (Cat. No.99CH36339), Vol. 3, 1999

[14] H. Isawa, H. Matsushita, and Y. Nishio, „Fuzzy Adaptive Resonance Theory Combining Overlapped Category in consideration of connections", 2008 IEEE Int. Jt. Conf. Neural Networks (IEEE World Congr. Comput. Intell., 2008

[15] E. P. Hernandez, E. G. Sanchez, Y. A. Dimitriadis, and J. L. Coronado, „A neuro-fuzzy system that uses distributed learning for compact rule set generation", in IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics, 1999, Vol. 3

[16] P. Henniges, E. Granger, and R. Sabourin, „Factors of overtraining with fuzzy ARTMAP neural networks", in Proceedings of the International Joint Conference on Neural Networks, 2005, Vol. 2, pp. 1075-1080

[17] L. Massey, „Conceptual duplication", Soft Comput., Vol. 12, No. 7, pp. 657-665, 2007

[18] M. Georgiopoulos, A. Koufakou, G. C. Anagnostopoulos, and T. Kasparis, „Overtraining in fuzzy ARTMAP: Myth or reality?", in IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), 2001, Vol. 2

[19] G. A. Carpenter and A.-H. Tan, „Rule extraction: From neural architecture to symbolic representation", Conn. Sci., Vol. 7, No. 1, 1995, pp. 3-27

[20] E. Parrado-Hernández, E. Gómez-Sánchez, and Y. A. Dimitriadis, „Study of distributed learning as a solution to category proliferation in Fuzzy ARTMAP based neural systems", Neural Networks, Vol. 16, 2003, pp. 1039-1057

[21] C. J. L. C. J. Lee, C. G. Y. C. G. Yoon, and C. W. L. C. W. Lee, „A new learning method to improve the category proliferation problem in fuzzy ART", in Proceedings of ICNN'95 - International Conference on Neural Networks, 1995, Vol. 3

[22] J. S. Lee, C. G. Yoon, and C. W. Lee, „Learning method for fuzzy ARTMAP in a noisy environment", Electron. Lett., Vol. 34, No. 1, 1998, pp. 95-97

[23] G. A. Carpenter, B. L. Milenova, and B. W. Noeske, „Distributed ARTMAP: A neural network for fast distributed supervised learning", Neural Networks, Vol. 11, No. 5, 1998, pp. 793-813

[24] A. Kaylani, M. Georgiopoulos, M. Mollaghasemi, and G. C. Anagnostopoulos, „AG-ART: An adaptive approach to evolving ART architectures", Neurocomputing, Vol. 72, No. 10-12, 2009, pp. 2079-2092

[25] R. Alves, C. Padilha, J. Melo, and A. D. Neto, „ARTMAP with modified internal category geometry to reduce the category proliferation", in IJCCI 2012 - Proceedings of the 4th International Joint Conference on Computational Intelligence, 2012, pp. 653-658

[26] S. Ilhan and N. Duru, „An improved method for fuzzy clustering", 2009 Fifth Int. Conf. Soft Comput. Comput. with Words Perceptions Syst. Anal. Decis. Control, 2009

[27] G. A. Carpenter, S. Grossberg, and D. B. Rosen, „Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system", Neural Networks, Vol. 4, No. 6, 1991, pp. 759-771

[28] S. Hsieh, C.-L. Su, and J. Liaw, „Fuzzy ART for the document clustering by using evolutionary computation", WSEAS Trans. Comput., Vol. 9, 2010, pp. 1032-1041

[29] N. Hoa and T. Bui, „A New Effective Learning Rule of Fuzzy ART", in Conference on Technologies and Applications of Artificial Intelligence, 2012, pp. 224-231

[30] C. Djellali, „Enhancing text clustering model based on truncated singular value decomposition, fuzzy art and cross validation", in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013, pp. 1078-1083

[31] S. Büttcher, C. L. A. Clarke, and G. V. Cormack, Information Retrieval: Implementing and Evaluating Search Engines. Cambridge: The MIT Press, 2010

[32] M. Rojček, „System for Fuzzy Document Clustering and Fast Fuzzy Classification", in 15th IEEE International Symposium on Computational Intelligence and Informatics, CINTI 2014, 2014, pp. 39-42

Parallelization and validation of algorithms for Zebrafish cell lineage tree reconstruction from big 4D image data

Robert Spir, Karol Mikula, Nadine Peyrieras

Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, Slovak University of Technology, Radlinskeho 11, 810 05 Bratislava, Slovakia, spir@math.sk, mikula@math.sk, nadine.peyrieras@inaf.cnrs-gif.fr.

Abstract: The paper presents numerical algorithms, post processing and validation steps for an automated cell tracking and cell lineage tree reconstruction from large-scale 3D+time two-photon laser scanning microscopy images of early stages of Zebrafish (Danio rerio) embryo development. The cell trajectories are extracted as centered paths inside segmented spatio-temporal tree structures representing cell movements and divisions. Such paths are found by using a suitably designed and computed constrained distance functions and by a backtracking in the steepest descent direction of a potential field based on a combination of these distance functions combination. Since the calculations are performed on big data, parallelization is required to speed up the processing. By careful choice and tuning of algorithm parameters we can adapt the calculations to the microscope images of vertebrae species. Then we can compare the results with ground truth data obtained by manual checking of cell links by biologists and measure the accuracy of our algorithm. Using an automatic validation process and visualization tool that can display ground truth data and our result simultaneously, along with the original 3D data, we can easily verify the correctness of the tracking.

Keywords: cell tracking; validation; big data; parallel computation

1 Introduction

With the recent research in biology and medicine the in vivo imaging of various organisms at cell level at very early stages of development without corrupting the cell integrity and normal evolution of the embryo is becoming possible thanks to the advancement of the modern microscopy techniques. Using two-photon laser scanning microscopy we can obtain long periods of the 3D+time images of an embryonic development with relatively short time step beginning just a few hours after the fertilization. By expression of the fluorescence protein trough the RNA injection at the one-cell stage we can obtain the labeling of cell nuclei and membranes

(Fig. 1). Thanks to the transparency for the laser scanning and thanks to the similarity with human cells in many aspects, this can be used to perform various analyses of the embryogenesis and the growth of organisms and the Zebrafish (Danio reio) embryogenesis is studied extensively. The results are used both in basic and applied biology and medicine research, e.g. in the anticancer drug design. To process such a large amount of data there is a need for efficient and parallel algorithms that will allow us to quickly process this data and obtain satisfying results.

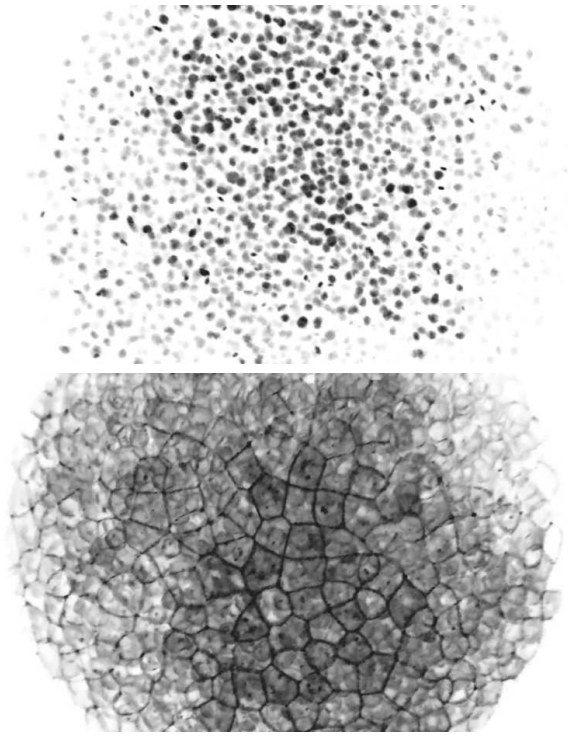

Figure 1
Volume rendering of cell nuclei (top) and cell membrane (bottom) data from confocal microscope.

In this paper we focused on the cell tracking part of our algorithm, mainly on the efficient OpenMP parallel implementation of 4D Rouy-Tourin scheme with fixing [11] and on the detailed description of various post processing steps and result validation by comparison with ground truth data. These topics are not discussed in [7, 9].

The paper is organized as follows. In the section 2 we briefly describe the individual steps of our image processing workflow. In the next section, section 3, we present our approach to cell trajectories and lineage tree extraction and discuss the numerical approaches used in the tracking method and parameters that can be used to tune and improve the tracking results. In section 4 we discuss the parallelization of the distance function calculation. Then we discuss extraction of the cell trajectories and post processing of the results in section 5. In the last section, section 6, we discuss

the numerical experiments on real 3D+time image sequences of the early zebrafish embryogenesis and perform the automatic validation of the results by comparison with the ground truth data containing correct cell links in time verified manually by biologists and manual visual checking using our visualization tool.

2 Image processing workflow

Developing algorithms for the analysis of the embryogenesis images, we first start with the image filtering to reduce the noise in the original 3D images obtained from a two-photon laser microscope. For the filtering we are using a geodesic mean curvature flow method (GMCF) based on the nonlinear diffusion equation that was developed by Chen et al. in [1] and efficiently discretized by Kriva et al. in [2, 3]. We apply the filtering to the cell nuclei and membrane images and in this step we eliminate various small structures and noise which do not represent the cell nuclei or membrane structures. For the computation we are using a parallel MPI implementation of the algorithm and Red-Black SOR parallel solver.

In the next step we detect the cell nuclei identifiers using a level-set center detection method (LSCD) developed by Frolkovic et al. in [3, 4]. This model is based on the nonlinear advection-diffusion equation which is applied to the result of GMCF filtering. The basic assumption is that cell nuclei are represented by humps of relatively higher image intensity where the diameter of cell nuclei is rather large compared to the diameter of spurious inner cell structures. Thus applying a sufficient number of evolutionary steps of LSCD, the spurious structures are shrunk and smoothed but larger humps are still significant and we can look for the remaining local maxima in the volume representing the coordinates of the cell nuclei. The method is parallelized with MPI using Red-Black SOR parallel solver to solve the linear system. These cell nuclei coordinates will represent the basic input data for our segmentation and tracking algorithms.

Using the cell identifiers, we can move to the segmentations of inner cell structures, the cell nuclei. Here we use the generalized subjective surface equation (GSUBSURF) of advection-diffusion type developed by Sarti et al. in [3, 5, 6]. The cell identifiers are used as seeds for the segmentations.

With the cell nuclei coordinates from the LSCD and nuclei segmentations from GSUBSURF we came up with the idea of the construction of 4D segmentation that will sufficiently approximate the real cell shapes. The first approach was to use 3D ellipsoids with constant half-axes for all cells in all time steps to obtain some very rough approximation of cell nuclei and combine them to the 4D segmentation. Here we found that thanks to the imperfection of the input data a non negligible number of cell identifiers that should represent the cell nuclei were missing and the result of the tracking was many short disjointed trajectories. To improve the results, the next step was considering 4D ellipsoids constructed also in the time dimension. Using such time overlapping we were able to achieve connected trajectories (close gaps) even in cases when the cell nuclei are missing in single time step [7, 8]. We improved the approximation of cell nuclei by using the diameters of 4D ellipsoid

from performed segmentations of each single cell in the whole dataset so each 4D ellipsoid is unique and is approximating the corresponding cell. By shrinking or enlarging this diameter one can further refine the quality of the tracking results. Calculating suitable designed distance functions inside the 4D segmentation and using their combination to construct a potential field, we can extract cell trajectories representing their spatio-temporal movement during the embryogenesis. Here we can use a special weighted combination of the distance functions during the potential field construction to further improve the tracking results [9].

There are also developed other approaches to the problem of cell tracking. In [10] Amat et al. are using the methods based on sequential Bayesian approach with Gaussian mixture models are used. In [3] there has been presented a method to build the cell lineage tree for the complex stages of Zebrafish embryo development based on stochastic simulated annealing minimization of a heuristic energy functional.

3 Cell tracking

Our method is composed of five basic steps:

- Construction of a 4D segmentation yielding the 4D spatio-temporal tree structure
- Computation of constrained 4D distance functions inside this 4D segmentation and designing of a proper combination of them in order to built a potential field for tracking
- Extraction of cell trajectories using the steepest descent direction of the potential field
- Centering the extracted trajectories inside the 4D spatio-temporal trees in order to eliminate duplicates
- Construction of the cell lineage tree by detecting trajectories which merge together when going backward in time indicating mitosis and thus a branching node of the lineage tree

3.1 4D segmentations

4D segmentation is a spatio-temporal structure which approximates the space-time movement of cell nuclei. According to Zanella et al. in [6, 12] the shape of cell nuclei during zebrafish embryogenesis is reasonably approximated by spheres or ellipsoids. Thus, in order to construct 4D segmentation we use cell identifiers detected in all time steps, $s_m^l, m = 1,\ldots,n_C^l, l = 1,\ldots,N_\theta$ (m denotes cell identifier index at time step l and N_θ is the number of time steps) by LSCD method from [4] and create 4D ellipsoids around all these points. To determine halfaxes of the ellipsoids we use real cell nuclei segmentations obtained using the generalized subjective surface (GSUBSURF) method [6, 12] paired with cell coordinates from the

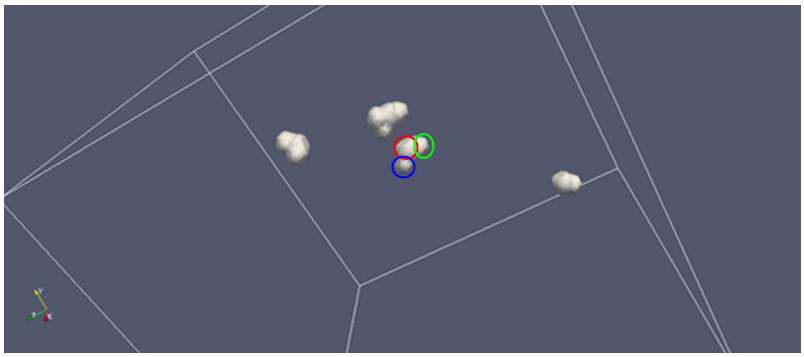

Figure 2
Projection of the 4D segmentation to one time step. For each cell identifier we can see the central sphere (inside the red circle) surrounded by the overlaps of the sphere from the previous time step (inside the green circle) and from the next time step (inside the blue circle). Three further parts of 4D segmentation are shown as well.

cell detection step. We calculate the volume of the real segmented nucleus and compute the radius of a sphere with the same volume. This radius is then used as spatial half-axes for the constructed ellipsoid. Here, we also introduce a parameter S representing shrinking of the half-axes (if $S < 0$) or expanding of them (if $S > 0$). A slight shrinking of the real radius is used later in the tracking algorithm since it helps to have a spatially non-overlapping tubular structure representing the cell movement. This parameter is tuned by comparison of the tracking with ground truth data and its optimal choice can improve the quality of tracking results. In temporal direction we are using halfaxis equal to $d\theta$ corresponding to the image acquisition interval (Fig. 2). The nonzero temporal halfaxis is important due to the time overlap which we create and thus we improve connectivity of 4D spatio-temporal tree structures. Thanks to the time overlap we interconnect branches of the 4D spatio-temporal tree where a cell center was not detected in one frame but it was detected in two neighboring frames and thus we correct false negative errors of the cell center detection algorithm.

3.2 Computing 4D distance functions and building the potential field

Our 4D segmentation containing 4D spatio-temporal tree structures can be represented by a 4D piecewise constant function, with some BIG, a sufficiently large number greater than maximum distance in the data (e.g. $BIG = 10^4$), value outside of the segmentation and with zero value inside it. Using this information, we compute two types of distance functions inside the 4D spatio-temporal tree structures. We call them constrained because all the calculations are constrained by the boundaries of the 4D segmentation. Due to that fact, the computed distances between 4D points of the 3D+time image sequence are not the standard Euclidean distances in R^4 but they represent minimal Euclidean paths between the points inside the 4D segmentation.

The first type of distance function will be denoted by $D_L(x_1,x_2,x_3,\theta)$. It is calculated gradually inside all simply connected regions, starting from cell centers in the lowest possible time step θ. After the calculation is completed in all regions reachable from these cell centers, we fix the computed values and continue the calculation from centers in the next time step but only in regions where the values are not yet fixed. Using this approach we calculate the distance function $D_L(x_1,x_2,x_3,\theta)$ inside the whole 4D segmentation. At the end all doxels inside the 4D segmentation contain the value of the distance to the farthest (backwardly in time) cell identifier to which it is continuously connected. The second type of distance function, $D_B(x_1,x_2,x_3,\theta)$, represents the distance of any inner point of the 4D segmentation to the boundary of the 4D segmentation. Again, $D_B(x_1,x_2,x_3,\theta)=0$ for all (x_1,x_2,x_3,θ) outside of the 4D segmentation.

Based on these facts we build the new potential field

$$V(x_1,x_2,x_3,\theta)=D_L(x_1,x_2,x_3,\theta)-\alpha D_B(x_1,x_2,x_3,\theta), \tag{1}$$

which is used for the extraction of cell trajectories. The parameter $\alpha>0$ is used to adjust the weight of D_B function to tune and improve tracking results.

Both distance functions $D_L(x_1,x_2,x_3,\theta)$ and $D_B(x_1,x_2,x_3,\theta)$ are computed numerically on the doxel structure of the 3D+time image sequence using the Rouy-Tourin scheme [13]. We numerically solve the time relaxed Eikonal equation, which has the following form

$$d_t+|\nabla d|=1 \tag{2}$$

for the unknown function $d(x_1,x_2,x_3,\theta,t)$. Here we solve a spatially 4D problem, so ∇d is the 4D gradient of the function d, i.e. the vector of partial derivatives with respect to x_1,x_2,x_3 and θ variables. For discretization of the equation (2) we use the spatially 4D Rouy-Tourin scheme.

We identify here the 4D doxels with finite volumes V_{ijkl} having four indices. Without losing generality, we rescale the time step $d\theta$ to be equal to $dx_1=dx_2=dx_3$ and denote their common value by h_D (standardly we set $h_D=1$). Let d^n_{ijkl} denote the approximate value of solution d in the barycenter of V_{ijkl} in a discrete step $t^n=n\tau_D$ where τ_D is the length of step discretizing t variable. Then, for every V_{ijkl} we define the index set N_{ijkl} of all (p,q,r,s) such that $p,q,r,s\in\{-1,0,1\}$, $|p|+|q|+|r|+|s|=1$. In order to build the scheme for any $(p,q,r,s)\in N_{ijkl}$, we define

$$D^{pqrs}_{ijkl}=\left(\min\left(d^{n-1}_{i+p,j+q,k+r,l+s}-d^{n-1}_{ijkl},0\right)\right)^2 \tag{3}$$

and then also

$$M^{pqrs}_{ijkl}=\max\left(D^{-p,-q,-r,-s}_{ijkl},D^{p,q,r,s}_{ijkl}\right). \tag{4}$$

Using this notations, the 4D Rouy-Tourin scheme for solving equation (2) has the following form

$$d^n_{ijkl}=d^{n-1}_{ijkl}+\tau_D-\frac{\tau_D}{h_D}\sqrt{M^{1000}_{ijkl}+M^{0100}_{ijkl}+M^{0010}_{ijkl}+M^{0001}_{ijkl}}\,. \tag{5}$$

Since the scheme produces monotonically increasing updates that are gradually approaching a steady state, we can implement (5) in a computationally efficient way [11]. Let us consider the index set I of all indices (i,j,k,l) and the set $\mathscr{F}^n$ that contains the indices $(i,j,k,l) \in I$ of the finite volumes where the steady state has been already approximately reached, i.e. $|d^n_{ijkl} - d^{n-1}_{ijkl}| < \varepsilon$, where ε is some chosen small threshold value. The basic principle is that we perform all computations only in the finite volumes that have not yet reached the steady state. The number of these finite volumes and the computational time needed to complete one time step of the procedure gradually decrease until the values in all cells are fixed. The method is given by *Algorithm 1*:

- ✓ Do while $\mathscr{F}^n \neq I$
  - ✓ Do for all $(i,j,k,l) \in I$
    - ✓ if $(i,j,k,l) \in \mathscr{F}^n$ then continue
    - ✓ else
      - ✓ update d^n_{ijkl} using (5)
      - ✓ if $|d^n_{ijkl} - d^{n-1}_{ijkl}| < \varepsilon$ then $\mathscr{F}^n = \mathscr{F}^n \cup \{(i,j,k,l)\}$
- ✓ $n = n+1$

4 OpenMP parallel implementation of 4D Rouy-Tourin scheme

The two main standards for parallel programing are MPI - Message Passing Interface [14] and OpenMP - Open Multi-Processing [15]. With the MPI, the parallel calculation runs in multiple independent processes, so this standard is suitable for systems with distributed memory. Here the processes communicate using MPI subroutines to exchange data using some communication media (e.g. local network or high speed interconnect). The disadvantage is that this communication always includes some overhead. On the other hand, OpenMP runs the calculation in single process with multiple parallel threads. Each thread has access to shared portion of the memory, so the communication tends to be fast with minimal overhead. The disadvantage is that this approach cannot be used on systems with distributed memory. When dealing with 4D data and 4D computations by using MPI parallelization, we would need to send large amount of data during interprocess communication. E.g., we would need around 120MB of data to send and receive between each neighboring parallel processes during each time step of distance function calculation. So, we decided to use OpenMP for parallelization, which is faster in this case. The drawback of this approach is that we need to run the calculation on single shared memory computer and we are limited by the amount of the available memory of such a server. For example, for a 320 time step dataset with $512 \times 512 \times 120$ voxels 3D volumes we need roughly 128GB of shared memory. Since current generation HPC servers have 256GB or more of shared memory at disposal, OpenMP approach is applicable. In comparison, with MPI approach in standard cluster configuration,

where all internode communication is going through 1Gbit network, it spent 2 seconds on the communication part, when running on two servers. The total time of calculation of a single Rouy-Tourin time step took 5-7 seconds, depending on the number of already fixed points. We can see that the communication part is significant, contributing 30% of the single time step execution. When running on a single server, so the communication does not need to go through network, it took 1 second, which is still a measurable difference. In OpenMP version there is no communication needed.

Using OpenMP directives we were able to parallelize most parts of the whole calculation by dividing the data in the time dimension and doing the calculations on the data in parallel. Since the distance function is calculated in an explicit manner using only the data form previous time step, no communication is needed. The only parts that remain serial are reading of the input data and writing of the results, since these parts are limited by the speed of disk drives and there is no sense on parallelizing them.

On standard multiprocessor servers, the global memory space is divided into separate units, so-called NUMA nodes, where each processor has fast access to a local NUMA node, but slow access to other NUMA nodes. E.g. each NUMA node could contain 32GB of memory, which equals 256GB of total memory with 8 processors. Since in our calculations we need 128GB of memory, even a serial code can use the memory from multiple nodes. In this case, it is more efficient to use *numactl* utility to set node memory interleaving policy, when the memory is allocated from all nodes in a round-robin fashion. Since the other CPUs are not under load and the memory access is evenly distributed between nodes, the performance hit is lower than if the program uses memory only from certain nodes. In Fig. 3 we can see memory distribution when running serial calculation using default settings and when using node memory interleaving set with *numactl* utility. Using the memory interleaving, the memory usage is equal across all nodes, while with default settings, only memory from nodes 1,2,3,5 and 7 is used. In this test, the second distance function D_B calculation took 752 seconds with default settings and 558 seconds with memory interleaving, which is more than 25% faster.

Using the parallel processing, it is important to bind an individual parallel process or in our case, thread, to a single NUMA node, or processor core. Thus, when the thread allocates some local memory, it does not happen that the process scheduler switches the thread to a different CPU where the thread does not have any data in local memory. When the thread is bound to a certain CPU, it will run on this CPU for its whole lifetime. For CPU binding with OpenMP programming we can use the environment variable OMP_PROC_BIND. If it is set to TRUE, then the individual threads will be bound to single CPU cores in a sequential manner. This is not optimal, because when we want to run e.g. only 4 parallel threads, they will be bound to the first four cores of the first CPU, so they will all run on a single NUMA node and have to share a single memory access. The optimal would be a binding to NUMA nodes in round-robin fashion. To use this type of binding, we can use GOMP_CPU_AFFINITY environment variable that is not in the OpenMP standard, but is an extension of GCC compiler. Setting this variable to ”0 4 8 12” will bind the

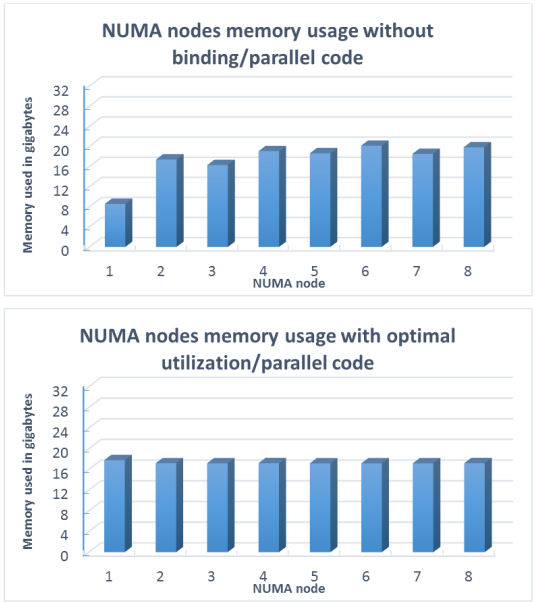

Figure 3
on top, node memory usage with default settings, memory is used only on certain nodes, on bottom, node memory is used evenly with memory interleaving.

first thread to CPU core 0, second to 4 etc. and each thread will run on single NUMA node with dedicated memory controller and the combined memory bandwidth will be 4-times wider than in the previous case. In our test, with calculation running in parallel on the first four CPU cores on a single NUMA node, the calculation of D_B took 362 seconds, when running on first cores of four NUMA nodes the calculation took only 143 seconds, which is more than twice as fast.

When running the calculation in 32 parallel threads without any binding, it took 57 seconds and with binding it took 32 seconds which is again nearly twice as fast. Without the binding, the parallel threads are switched between CPUs and NUMA nodes by the operating system process scheduler. The node memory usage of parallel calculation with and without binding can be seen in Fig. 4.

Since the OpenMP program runs as a single process with multiple threads in a single shared memory space, we can use a special programming technique which utilizes the so-called lazy allocation feature of Linux kernel where the physical memory is actually allocated only after writing in it. In the program for computation of the distance function we allocate one large array using single call of the malloc() function which allocates only virtual memory in Linux and no physical memory is used. Then we write zeroes into this array in a parallel `for` cycle and only then the real physical memory is allocated. Since we use static OpenMP scheduling, each parallel thread will allocate memory on node on which it is actually running and then in all subsequent parallel cycles, the thread will always get the same part of the calculation and will use only local memory allocated in the first parallel cycle.

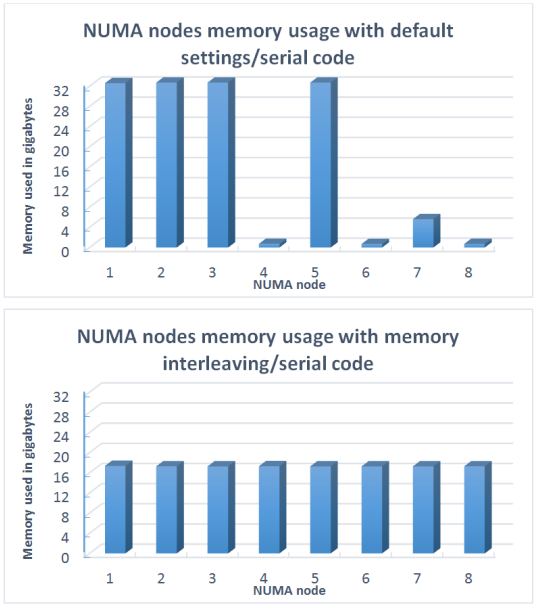

Figure 4
on top, node memory usage without thread binding is uneven, on bottom memory usage with thread binding and optimal allocation, each thread is accessing only the local memory and the calculation is faster.

With this approach we achieve optimal uniform NUMA node memory usage even without memory interleaving and combining with thread binding the threads are mainly using the local node memory and accesses to memory on other nodes are minimized.

The speed-up of the parallel calculation when running multiple threads can be seen in Table 1. Because with more threads we can use local memory optimally, the speed-up is more than double in some cases.

Table 1
Parallel speed-up of the calculation with increasing number of threads.

Threads	1	2	4	8	16	32
Time (s)	752	324	143	73	43	32
Speedup	–	2.32	5.26	10.30	17.49	23.5

The total speed-up of parallel computation running in 32 parallel threads compared to serial code is 23.5 times.

5 Extraction of the cell trajectories

The cell trajectory is represented by a series of points in space-time (discrete spatio-temporal curve) for which we prescribe the condition that there exists exactly one point in every time step $l = N_b, \ldots, N_e$, $1 \leq N_b < N_e \leq N_\theta$. The extraction of cell trajectories is realized in two steps

- at first, we use backtracking in time by the steepest descent direction of the potential V built in (1) starting from all cell identifiers $s_m^l, m = 1, \ldots, n_C^l$, detected in all time steps $l = 2, \ldots, N_\theta$, recursively moving to the nearest points with strictly lower value of the potential,
- then we center all the extracted trajectories inside the 4D spatio-temporal trees by using the constrained distance function D_B in order to eliminate duplicates.

For the trajectories extraction we are again using the OpenMP parallelization and extracting each trajectory in parallel. In this case, we are using the *numactl* utility to set the memory interleaving policy, because during the backtracking we can move through the whole dataset and we cannot predict the parts of memory through which the trajectory will move. It is noteworthy that the longest part of this process is the writing of the results to disk which cannot be parallelized. The parallel backtracking and centering part lasts around 5 minutes and the writing part lasts 10 minutes.

The trajectories are stored in a text file with the following format:

- Each trajectory is defined by 1 line header containing a unique ID and the trajectory length
  ```
  --trajectory:  3632944 length:  28
  ```
- Then there are exactly *length* lines with coordinates in time and space representing the trajectory points
  ```
  184 393 90 452
  184 393 91 453
  184 393 90 454
  184 392 90 455
  184 392 90 456
  ...
  ```

5.1 Post processing of the extracted trajectories and cell lineage tree reconstruction

To fully reconstruct the cell lineage tree, we need to do further post processing on the trajectory extraction result. We have to eliminate duplicate points when two trajectories merge and continue backwards in time together and indicate mitosis on the merged trajectory. In the post processing step we also improve the results by reconnecting still disjointed trajectories and fix mitoses of more than two cells joining in the same time. Our post processing software is written in C# language and it can be easily run in multiplatform environment using .NET Framework in the Windows family of operating systems or mono in unix systems without recompilation.

Let us consider a mother cell which is going to divide at time step l, i.e. at time step $l+1$ it has two descendants and in later times maybe more due to further divisions. Without losing generality, let the number of descendant of this cell in the whole image sequence be $N_d = 2^m$. Up to time l the life of the cell is represented by exactly N_d trajectories which are, however, until time l composed by the same spatio-temporal points. From the time $l+1$, the half of trajectories differs from the second half, but every half is again composed by the same points until the next division. The representation of cell life by the multiple trajectories which are partially the same does not cause any problem in visualization and/or reconstruction of a single cell or cell population movements and divisions. However, the reduction of the equal multiple parts of trajectories is necessary for the reconstruction of the cell lineage tree and is explained below.

The cell lineage tree is stored in a structure with a specific format. For each node of the cell lineage tree we store the spatio-temporal coordinates of the corresponding point, its unique global ID and the global ID of its mother in the extracted trajectory. In case of the first point of the trajectory, the mother ID is set to zero. To fill in the lineage tree structure, we subsequently take all points of the extracted trajectories, starting with the longest trajectories, and add subsequently the node representation of those points to the structure. For every trajectory we start by the first point and check whether the node corresponding to this point already exists in the structure. If it exists, it means that we have already added a trajectory which has some part equal to the current one. And also that there exists a later time after which the trajectories differ. We skip all equal points which are already represented as nodes in the lineage tree structure. Only the first different point of the current trajectory is added to the structure, together with the mother ID of the last equal point from the previously added trajectory. Using such approach we obtain the whole cell lineage tree where each node exists only once and the nodes are logically linked together in the same manner as it is in the real cell mother-daughter relation.

The lineage tree is then stored in a file with the following format

```
...
100007;8;44;-72;28;0;0;-1;-65536
100008;9;-83;19;32;0;0;-1;-1
100009;9;-36;2;39;1;0;-1;-16776961
100010;25;-79;18;33;1;9;-1;-1
...
```

where each point has global ID, local ID, three coordinates in space and one coordinate in time, local ID of the mother point and two optional parameters. In our case, the first -1 represents that the point was not validated and the last number represents the color of the point in ARGB format for visualization purposes. In this example we can see that the point with the global ID 100008 in time 0 is the mother point of the point with the global ID 100010 in time 1 since they are related using their local ID's. When the local ID of the mother equals 0, then the point has no mother and represents the starting point of the trajectory.

Now if we had perfect data, no further post processing would be needed. Since the real data are noisy and further errors can be introduced by wrong center detection, segmentation and near trajectory overlapping in space, we need to do additional steps to improve the results and try to fix errors. The most common errors are points where there are more than two trajectories merging and where the trajectories are disjointed because the corresponding connecting center in single time-step is missing and the time overlap in segmentation is insufficient.

For each point we calculate its movement direction using the central difference between the next and previous point of the trajectory, the average movement direction by averaging the movement direction using three points forward and backward and the average direction with respect to the points from the surrounding trajectories by averaging the direction with points in the near vicinity given by a fixed threshold. Now we can move to individual post processing steps.

- In the first step we disconnect all points where there are more than two trajectories merging and leave only the two nearest trajectories to merge.

- Next we try to reconnect the ending trajectories with the beginning trajectories in the next step searching for the nearest trajectory in the direction of the average cell movement gradually increasing the search radius from one to ten doxels.

- In the last step we again try to reconnect the ending trajectories with the beginning trajectories, but in the distance of two time steps, gradually increasing the search radius up to ten doxels.

Since the reconnection of trajectories in each time step is independent, we can use parallelization. For the paralelization we are using Task Parallel Library [16] included in Microsoft .NET Framework 4.0 and higher and also supported by the mono framework. We are using the *Parallel.For* construct to process multiple time steps in parallel. Using

```
Parallel.For(0, nodes.Length - 1, i =>
                                  {
                                  ...
                                  }
```

the runtime will automatically split the calculation to all available processors. Various parallelization options can be set using the *ParallelOptions* class, which we are not using in this case. Unfortunately currently there is no way to control the NUMA node binding so the best we can do is to use the NUMA memory interleaving policy when running the post processing on a NUMA server. We tested the parallel speed-up on standard quad-core desktop CPU intel core i7-4770k. The serial code took 413 seconds and the parallel version took 144 seconds, which represents the speed-up of 2.87. After this post processing we have the tracking results ready for a validation.

6 Validation of the results of numerical experiments on real Zebrafish embryogenesis data with ground truth data

For the testing of the method and for the reconstruction of the cell lineage tree we use one of the dataset produced by BioEmergences platform [17] with the acquisition step $d\theta = 50$ seconds, N_θ, number of time steps, is equal to 480 and dimension of every 3D image is 512x512x104 voxels. The real voxel size is $dx_1 = dx_2 = dx_3 =$ 1.33 micrometer in every spatial direction. In the last time step $N_\theta = 480$ the biologist marked cell populations forming various organs of the embryo and we can use this information to visualize the formation of these organs in time, Fig. 5. Using developed approach we can track those cell populations. Since the size of a single 3D image is about 27MB, for 480 time steps we deal with about 13 GB of raw data which requires usage of high-performance parallel computing (HPC) facilities especially when designing spatially 4D numerical algorithms.

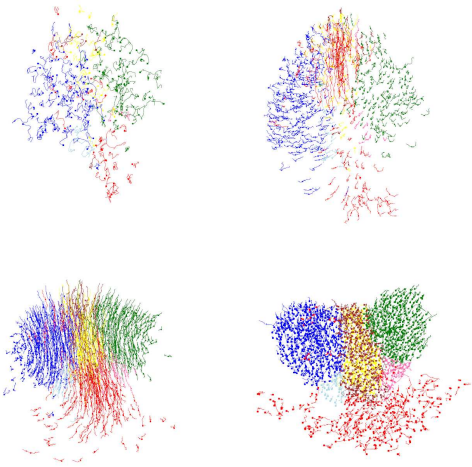

Figure 5
Visualization of the organ formation in time steps 1, 160, 320 and 480.

Before tracking, all 3D images of the processed data were filtered by 10 steps of geodesic mean curvature flow (GMCF) model [2, 3] and the cell nuclei identifiers were detected by 15 steps of level set center detection (LSCD) algorithm [4, 3]. The cell nuclei were segmented using the generalized subjective surface (GSUBSURF) method [3]. From several millions of cell identifiers we built the 4D segmentation and then the cell trajectories were extracted. The correctness of the mother-daughter cell links for the first dataset was validated using the ground truth data and the results are presented in Table 2 and Fig. 6.

The ground truth data (GTD) contains 38797 manually checked links between cells in time during all 480 time steps which we can use to validate the correctness of cell

links in our results. It also contains 71 valid mitoses that can be used for the accuracy of mitosis detection. Since the mitoses in ground truth data and in our tracking can be shifted by a few time steps in time, we are using the following procedure for the validation.

- In the first step, we find the corresponding trajectory points in our results with points in the ground truth data by finding the points with the same coordinates or with the shortest distance to the ground truth data points.
- Then we can easily check for the correctness of the cell links by checking if the link between two points in ground truth data is the same as the link between the corresponding points in our results. If it is the same we have a correct mother-daughter link, if the link is missing or we have the link to a different corresponding cell, then we have a wrong mother-daughter link.
- For the mitosis detection, we have three different possibilities:
  - The mitosis is in the same time step in the ground truth data and our results. This is the most simple case and can be validated easily just by checking the links between points.
  - The mitosis occurs later in our dataset than in the ground truth data, but on the same trajectory. We call this case "forward mitosis" and we move forward by the links between cells, checking the links in daughter cells after mitosis in the ground truth data and corresponding cells in our results looking for mitosis in the next five time steps.
  - In the third case, the mitosis occurs later in the ground truth data than in our dataset, but on the same trajectory. We call this case "backward mitosis" and we move backward through corresponding points and search for mitosis in our results in the previous five time steps.

Using this approach we can validate the results of our tracking algorithm by comparing it with the ground truth data. We are interested in two parameters, correctness of the cell links and the number of correctly detected mitoses.

To tune the tracking results of the 4D segmentation approach (ellipsoids created from diameters of real cell nuclei segmentations) we adjust two parameters mentioned in section 3.1, S and α. First, we can expand or shrink the nuclei radii used for building the 4D segmentation. By comparison with ground truth data we concluded that the best results for these two datasets were obtained when we shrink the radii by $S = -0.5$, cf. Fig. 6. The second parameter is α, used in the construction of the potential V. We try to find the optimal α to obtain the best ratio of correct links and detected mitoses. We tested the tracking for $\alpha \in [0.5, 4]$ and obtained the best results around $\alpha = 1.6$, see Table 2, where we achieved 94% accuracy of cell link detection and 21% of mitosis detection. When comparing the number of detected mitoses (Table 3) we can see that it is decreasing very fast and around $\alpha = 3.2$, where we have 96.5% correct links, we have only 3 correct mitosis, which is only 4% of all 71 mitosis in the ground truth data. The problem is caused by the fact, that the cell nuclei right after the mitosis "jump" away from each other, thus creating large gaps in the 4D segmentation. Improving the mitosis detection is

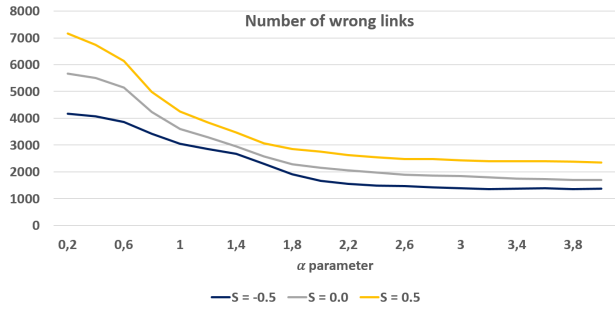

Figure 6
Number of wrong links in tracking compared to ground truth data, depending on α and S. With α increasing towards 2, the number of wrong links is decreasing, then it stabilizes and the best result is obtained for $S = -0.5$, cf. also Table 2.

difficult and still an open problem. Possible direction in this research would be to use membrane images for creating the 4D segmentation where no jumps during the cell split occur. The quality of membrane images by a confocal microscopy is still not sufficient for this purose, but one could use simulated cell membranes by using approximate Voronoi shapes constructed around the cell identifiers, which gives promising results.

Table 2
Comparison of the tracking result obtained with segmentation from real nuclei segmentations with ground truth data depending on α, with $S = -0.5$.

α	**Correct mother links**	**Correct daughter links**	**Wrong mother links**	**Wrong daughter links**
0.4	34856	34719	3935	4072
0.8	35510	35371	3280	3419
1.2	36064	35946	2726	2844
1.6	36551	36493	2240	2298
2.0	37125	37126	1668	1667
2.4	37288	37309	1506	1485
2.8	37349	37386	1448	1411
3.2	37402	37437	1395	1360
3.6	37357	37415	1440	1382
4.0	37377	37434	1420	1363

For manual visual validation we created a tool to visualize the result alongside the ground truth data. One can display the tracking with the trajectories with optional length, highlight trajectory start and end, cell mitosis and the volume rendering of original data along with the 2D slices can also be displayed.

Here we present the Figures 7-9 showing our tool for visual validation. In Fig. 7 one can see the visualization of the volume rendering of the original cell along with the trajectories from the ground truth data (left) and our tracking result. Here the

Table 3
Comparison of the correctly detected mitosis with segmentation from real nuclei segmentations with ground truth data depending on α, with $S = -0.5$.

α	**Correctly detected mitosis**	**Correctly detected forward mitosis**	**Correctly detected backward mitosis**	**Total correctly detected mitosis**
0.4	10	0	6	16
0.8	6	0	10	16
1.2	5	0	14	19
1.6	3	0	12	15
2.0	2	0	4	6
2.4	2	0	1	3
2.8	2	0	1	3
3.2	2	0	1	3
3.6	2	0	1	3
4.0	2	0	2	4

trajectory in our result is the same as the trajectory manually validated by biologists even for a long time interval (20 time steps forward and backward in this case). In Fig. 8 and Fig. 9 the cells before mitosis are displayed. One can see that the mitosis will occur since the trajectory is divided. Correctly detected cell mitosis is displayed in Fig. 8. In Fig. 9 on the right, in our result, the mitosis is not detected and only one branch of the trajectory is registered.

Conclusions

In this paper we presented an algorithm for the cell tracking on large-scale 3D+time microscopy data. We created parallel implementation of the 4D Rouy-Tourin scheme to speed up the data processing and then we applied this method to complex stages of the zebrafish early embryogenesis microscopic images. Using our automatic post-processing and validation workflow we can easily compare the tracking with ground truth data. Using the visualization tool we can visually verify the correctness by displaying the ground truth data along with our result and volume rendering of the original images.

Acknowledgement

This work was supported by Grants APVV-15-0522 and VEGA 1/0608/15.

References

[1] Y. Chen, B. C. Vemuri, L. Wang, Image denoising and segmentation via non-linear diffusion, Comput. Math. Appl. 39 (2000) 131–149. doi:10.1016/S0898-1221(00)00050-X.

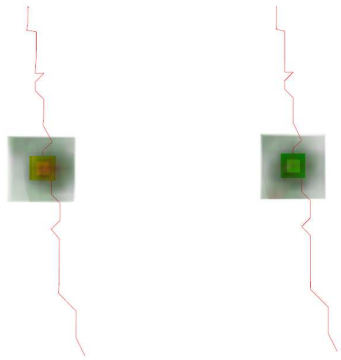

Figure 7
Visualization of a single cell trajectory from our manual validation tool. Data from the ground truth data is displayed on the left and our tracking result is on the right. Even a long trajectory is reconstructed correctly.

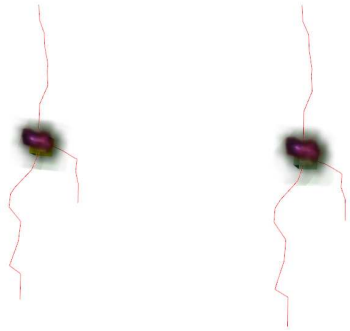

Figure 8
Visualization of a single cell trajectory from our manual validation tool. Data from the ground truth data is displayed on the left and our tracking result is on the right. The cell division is detected correctly.

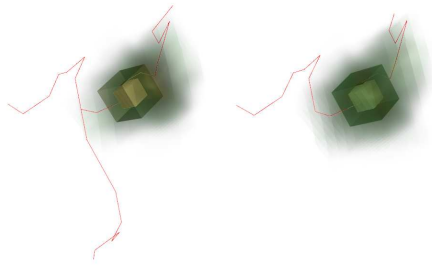

Figure 9
Visualization of a cell with dividing trajectory from our manual validation tool. Data from the ground truth data is displayed on the left and our tracking result is on the right. Only one branch of the divided cell is registered in our tracking result.

[2] Z. Krivá, K. Mikula, N. Peyriéras, B. Rizzi, A. Sarti, O. Stašová, 3d early embryogenesis image filtering by nonlinear partial differential equations, Medical Image Analysis 14 (4) (2010) 510–526. doi:10.1016/j.media.2010.03.003.

[3] E. Faure, T. Savy, B. Rizzi, C. Melani, M. Remešíková, R. Špir, O. Drblíková, R. Čunderlík, G. Recher, B. Lombardot, M. Hammons, D. Fabrèges, L. Duloquin, I. Colin, J. Kollár, S. Desnoulez, P. Affaticati, B. Maury, A. Boyreau, J. Y. Nief, P. Calvat, P. Vernier, M. Frain, G. Lutfalla, Y. Kergosien, P. Suret, R. Doursat, A. Sarti, K. Mikula, N. Peyriéras, P. Bourgine, An algorithmic workflow for the automated processing of 3d+time microscopy images of developing organisms and the reconstruction of their cell lineage, Nature Communications 7, Article number: 8674. doi:10.1038/ncomms9674.

[4] P. Frolkovič, K. Mikula, N. Peyriéras, A. Sarti, A counting number of cells and cell segmentation using advection-diffusion equations, Kybernetika 43 (6) (2007) 817–829.

[5] A. Sarti, R. Malladi, J. A. Sethian, Subjective surfaces: A method for completing missing boundaries, Proceedings of the National Academy of Sciences of the United States of America 97 (12) (2000) 6258—-6263. doi:10.1073/pnas.110135797.

[6] K.Mikula, Peyriéras, M. Remešíková, A.Sarti, 3d embryogenesis image segmentation by the generalized subjective surface method using the finite volume technique, Finite Volumes for Complex Applications V, Problems & Perspectives (Eds. R.Eymard, J.M.Herard), ISTE and WILEY, London (2008) 585–592.

[7] K. Mikula, N. Peyriéras, R. Špir, Numerical algorithm for tracking cell dynamics in 4d biomedical images, Discrete and Continuous Dynamical Systems - Series S 8 (5) (2015) 953–967. doi:10.3934/dcdss.2015.8.953.

[8] K. Mikula, R. Špir, M. Smíšek, E. Faure, N. Peyriéras, Nonlinear pde based numerical methods for cell tracking in zebrafish embryogenesis, Aplied Numerical Mathematics 95 (2015) 250–266. doi:10.1016/j.apnum.2014.09.002.

[9] K. Mikula, R. Spir, N. Peyrieras, Cell lineage tree reconstruction from time series of 3d images of zebrafish embryogenesis, Submitted to MCBMIIA2016.

[10] F. Amat, W. Lemon, D. P. Mossing, K. McDole, Y. Wan, K. Branson, E. W. Myers, P. J. Keller, Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data, Nature Methods 11 (2014) 951–958. doi:10.1038/nmeth.3036.

[11] P. Bourgine, P. Frolkovič, K. Mikula, N. Peyriéras, M. Remešíková, Extraction of the intercellular skeleton from 2d microscope images of early embryogenesis, Lecture Notes in Computer Science 5567 (Proceeding of the 2nd International Conference on Scale Space and Variational Methods

in Computer Vision, Voss, Norway, June 1-5,2009) , Springer (2009) 38–49doi:10.1007/978-3-642-02256-2_4.

[12] C. Zanella, M. Campana, B. Rizzi, C. Melani, G. Sanguinetti, P. Bourgine, K. Mikula, N. Peyrieras, A. Sarti, Cells segmentation from 3-d confocal images of early zebrafish embryogenesis, EEE Transactions on Image Processing 19 (3) (2011) 770–781. doi:10.1109/TIP.2009.2033629.

[13] E. Rouy, A. Tourin, Viscosity solutions approach to shape-from-shading, SIAM Journal on Numerical Analysis 29 (3) (1992) 867—-884. doi:10.1137/0729053.

[14] MPI Forum, http://mpi-forum.org/, accessed: 2017-03-29.

[15] OpenMP Home page, http://www.openmp.org/, accessed: 2017-03-29.

[16] Microsoft, Task parallel library, https://msdn.microsoft.com/en-us/library/dd460717, accessed: 2016-07-25.

[17] BioEmergences platform web site, http://bioemergences.iscpif.fr/bioemergences/index.php, accessed: 2017-03-29.

New Approach to Fuzzy Decision Matrices

Pavla Rotterová, Ondřej Pavlačka

Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, 17. listopadu 1192/12, 771 46 Olomouc, Czech Republic, pavla.rotterova01@upol.cz, ondrej.pavlacka@upol.cz

Abstract: Decision matrices represent a common tool for modeling decision-making problems under risk. They describe how the decision-maker's evaluations of the considered alternatives depend on the fact which of the possible and mutually disjoint states of the world will occur. The probabilities of the states of the world are assumed to be known. The alternatives are usually compared on the basis of the expected values and the variances of their evaluations. However, the states of the world as well as the alternatives evaluations are often described only vaguely. Therefore, we consider the following problem: the states of the world are modeled by fuzzy sets defined on the universal set on which the probability distribution is given, and the evaluations of the alternatives are expressed by fuzzy numbers. We show that the common approach to this problem, based on employing crisp probabilities of the fuzzy states of the world computed by the formula proposed by Zadeh, is not appropriate. Therefore, we introduce a new approach in which a fuzzy decision matrix does not describe discrete random variables but fuzzy rule bases. The problem is illustrated by an example.

Keywords: decision matrices; fuzzy decision matrices; decision making under risk; fuzzy states of the world; fuzzy rule bases system

1 Introduction

A decision matrix is often used as a tool of risk analysis in decision making under risk [3], [4], [7], [14]. It describes how the decision-maker's evaluations of the considered alternatives depend on the fact which of the possible and mutually disjoint states of the world will occur. The probabilities of occurrences of these states of the world are supposed to be known. Thus, the evaluations of the alternatives are discrete random variables. The alternatives are usually compared on the basis of the expected values and the variances of their evaluations. The decision-maker selects the alternative that maximizes his/her expected evaluation or maximizes the expected evaluation and simultaneously minimizes the variance.

In practical applications, the states of the world as well as the evaluations of the alternatives can be determined vaguely. The states of the world are mostly

described verbally, like "the gross domestic product will increase moderately during next year". Sometimes, it can be problematic to express the evaluations of alternatives precisely because we may not have enough information. For instance, the evaluation under a certain state of the world can be described as “about 5%”. In some cases, it is more natural for a decision-maker to express the evaluations by selecting a term from a given linguistic scale.

The vaguely described pieces of information can be mathematically modeled by means of tools of fuzzy sets theory. Different views of uncertainty and fuzzy decisions in a decision matrix are discussed in [7]. Multiple attribute decision making problems, described by a decision matrix with crisp and fuzzy data, are analyzed in [1]. In [2], a fuzzy decision matrix is applied to a group decision making. An application of risk analysis with fuzzy sets employing the decision matrix is presented in [3]. In [4], the authors considered decision matrices with fuzzy targets. In [5], the hesitant fuzzy decision matrix, i.e. a decision matrix containing fuzzy sets with a different definition of membership function then the original one proposed by Zadeh [15], is considered.

A decision matrix with the fuzzy states of the world and the fuzzy evaluations of the alternatives under the particular fuzzy states of the world is called *a fuzzy decision matrix*. In [**Error! Reference source not found.**2], the authors considered a model where the fuzzy states of the world are expressed by fuzzy sets on the universal set on which the probability distribution is given. They proposed to proceed in the same way as in the case of the crisp (i.e. exactly described) states of the world; they set the probabilities of the fuzzy states of the world applying the formula proposed by Zadeh in [17]. Within this approach, the evaluations of the alternatives are understood as discrete random variables taking on fuzzy values with the probabilities of the fuzzy states of the world.

In [10], the authors showed that the Zadeh’s probabilities of fuzzy events lack the common interpretation of a probability measure. Another problem is a precise definition of "the occurrence of the particular fuzzy state of the world" (see the discussion in Section 3.3). Therefore, an alternative to how the information contained in a fuzzy decision matrix can be treated was proposed in [8]. The way is based on the idea that a fuzzy decision matrix does not determine discrete fuzzy random variables, but a system of fuzzy rule bases (a fuzzy rule base was introduced in [16]). However, only the crisp (i.e. not fuzzy) evaluations of alternatives were considered in [8] which makes the problem much simpler. The main aim of the paper is to extend this approach to the case where the evaluations of alternatives are expressed by fuzzy numbers, and to derive the formulas for correct computations of fuzzy expected values and fuzzy variances of evaluations of alternatives. The obtained fuzzy characterstics will be compared with those obtained by the approach considerd in [12].

The paper is organized as follows. A decision matrix tool is briefly recalled in Section 2. In Section 3, the common approach to the fuzzification of a decision

matrix is analysed and the related problems are discussed. Our new approach to this problem is introduced and analysed in Section 4. In Section 5, both approaches are compared by an illustrative example.

2 Decision Matrices

In this section, let us describe a decision matrix as a tool for supporting a decision making under risk.

Let us consider *a probability space* $(\Omega, \mathcal{A}, P)$ where Ω denotes a non-empty universal set of all elementary events, $\mathcal{A}$ is a σ-algebra of subsets of Ω, i.e. $\mathcal{A}$ represents the set of all considered random events, and P: $\mathcal{A} \rightarrow [0,1]$ denotes a probability measure.

Now, let us describe a decision matrix under risk, considered e.g. in [3], [4], [7] and [14]. The decision matrix is shown in Table 1. In the matrix, $x_1, x_2, \ldots, x_n$ represent the alternatives of a decision-maker, $S_1, S_2, \ldots, S_m$, where $S_j \in \mathcal{A}$ for $j = 1, 2, \ldots, m$, denote the mutually disjoint states of the world, i.e. $S_j \cap S_k = \emptyset$ for any $j, k \in \{1, 2, \ldots, m\}$, $j \neq k$, and $\bigcup_{j=1}^{m} S_j = \Omega$, $p_1, p_2, \ldots, p_m$ stand for the probabilities of the states of the world $S_1, S_2, \ldots, S_m$, i.e. $p_j = P(S_j)$, and for any $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, m\}$, $h_{i,j}$ means the decision-maker's evaluation if he/she chooses the alternative x_i and the state of the world S_j occurs. The evaluation of the alternative x_i is commonly understood as a discrete random variable H_i: $\{S_1, S_2, \ldots, S_m\} \rightarrow \mathbb{R}$ which takes on the values $h_{i,j} = H_i(S_j)$ with the probabilities p_j, $j = 1, 2, \ldots, m$.

Table 1
Crisp decision matrix

	S_1 p_1	S_2 p_2	… …	S_m p_m		
x_1	$h_{1,1}$	$h_{1,2}$	…	$h_{1,m}$	EH_1	$var\ H_1$
x_2	$h_{2,1}$	$h_{2,2}$	…	$h_{2,m}$	EH_2	$var\ H_2$
…	…	…	…	…	…	…
x_n	$h_{n,1}$	$h_{n,2}$	…	$h_{n,m}$	EH_n	$var\ H_n$

The alternatives are usually compared on the basis of the expected values and the variances of their evaluations (an overview of decision making rules can be found e.g. in [**Error! Reference source not found.**]). The expected values of the decision-maker's evaluations, denoted by $EH_1, EH_2, \ldots, EH_n$, are given for any $i \in \{1, 2, \ldots, n\}$ by:

$$EH_i = \sum_{j=1}^{m} p_j \cdot h_{i,j}. \tag{1}$$

The variances of the decision-maker's evaluations, denoted by *var* H_1, *var* H_2, …, *var* H_n, are calculated for any $i \in \{1, 2, \ldots, n\}$ as follows:

$$var\ H_i = \sum_{j=1}^{m} p_j \cdot (h_{i,j} - EH_i)^2. \tag{2}$$

The alternative that maximizes the expected evaluation and minimizes the variance of the evaluation is selected as the best one.

3 Fuzzy Decision Matrices

Now, let us describe the common approach to the generalization of a decision matrix to the case where the states of the world and the evaluations of the alternatives are expressed by fuzzy sets, considered e.g. in [**Error! Reference source not found.**2]. Within this approach, the probabilities of the fuzzy states of the world, computed by the formula proposed by Zadeh in [17], are used for computations of the characteristics of the evaluations of the alternatives.

3.1 Fuzzy States of the World

Vaguely defined states of the world can be mathematically expressed by fuzzy sets. *A fuzzy set A* on a non-empty set Ω is determined by its membership function μ_A: $\Omega \rightarrow [0, 1]$. Let us denote the family of all fuzzy sets on Ω by $\mathcal{F}(\Omega)$. *A support* of A and *a core* of A are given as $Supp\ A := \{\omega \in \Omega \mid \mu_A(\omega) > 0\}$ and $Core\ A := \{\omega \in \Omega \mid \mu_A(\omega) = 1\}$, respectively. A_α means *an α-cut* of A, i.e. $A_\alpha := \{\omega \in \Omega \mid \mu_A(\omega) \geq \alpha\}$ for any $\alpha \in (0,1]$.

Remark Any crisp set $A \subseteq \Omega$ can be seen as a fuzzy set $A \in \mathcal{F}(\Omega)$ of a special kind where its characteristic function χ_A coincides with the membership function μ_A of the fuzzy set. In fuzzy models, this convention allows us to consider also precisely described events given by crisp sets.

In fuzzy decision matrices, fuzzy states of the world are described by the fuzzy events. According to Zadeh [17], *a fuzzy event* $A \in \mathcal{F}(\Omega)$ is a fuzzy set whose α-cuts are random events, i.e. $A_\alpha \in \mathcal{A}$ for all $\alpha \in (0,1]$. As an analogy to a decomposition of the universal set Ω by crisp states of the world, the fuzzy states of the world, denoted by S_1, S_2, …, S_m, have to form *a fuzzy partition* of the universal set Ω, i.e. for any $\omega \in \Omega$, it has to hold that

$$\sum_{j=1}^{m} \mu_{S_j}(\omega) = 1. \tag{3}$$

Zadeh [17] extended the crisp probability measure P to the case of fuzzy events. Let us denote this extended measure by P_Z. A probability $P_Z(A)$ of a fuzzy event A is defined as follows:

$$P_Z(A) := E(\mu_A) = \int_{\omega \in \Omega} \mu_A(\omega) dP. \tag{4}$$

3.2 Fuzzy Evaluations of Alternatives under the Particular Fuzzy States of the World

As was mentioned in Introduction, it can be difficult for a decision-maker to evaluate each alternative under each state of the world by a real number. One reason can be a lack of information caused e.g. by inaccuracies of measurements or a lower quality of data transmissions. Another reason can be that it is more natural for the decision-maker to describe the evaluations linguistically rather than by numbers.

Linguistic terms or uncertain quantities can be mathematically modeled by fuzzy numbers. A *fuzzy number A* is a fuzzy set on the set of all real numbers $\mathbb{R}$ such that its core A is non-empty, its α-cuts A_α are closed intervals for any $\alpha \in (0, 1]$, and its support *Supp A* is bounded. The family of all fuzzy numbers on $\mathbb{R}$ will be denoted by $\mathcal{F}_N(\mathbb{R})$. In some models, fuzzy evaluations can be restricted only to a closed interval, mostly [0,1]. A *fuzzy number defined on the interval* $[a, b]$ is a fuzzy number whose α-cuts belong to the interval $[a, b]$ for all $\alpha \in (0,1]$. The family of all fuzzy numbers on the interval $[a, b]$ will be denoted by $\mathcal{F}_N([a, b])$.

Thus, there are two ways of expressing a fuzzy evaluation of an alternative. The first way is to specify the evaluation directly by a fuzzy number. For instance, some expert can evaluate the particular alternative directly by the fuzzy number "about five percent profit", whose membership function is shown in Figure 1.

The second possibility of expressing the fuzzy evaluation of the alternative consists in the fact that the evaluation is modeled by a linguistic variable (linguistic variables were introduced in [16]). A decision-maker evaluates the alternatives under the particular states of the world by appropriate linguistic terms whose mathematical meanings are described by fuzzy numbers. A set of linguistic terms $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_r$ forms *a linguistic scale on* $[a, b]$ if $T_1, T_2, \ldots, T_r \in \mathcal{F}_N([a, b])$ representing their mathematical meanings form a fuzzy partition of $[a, b]$.

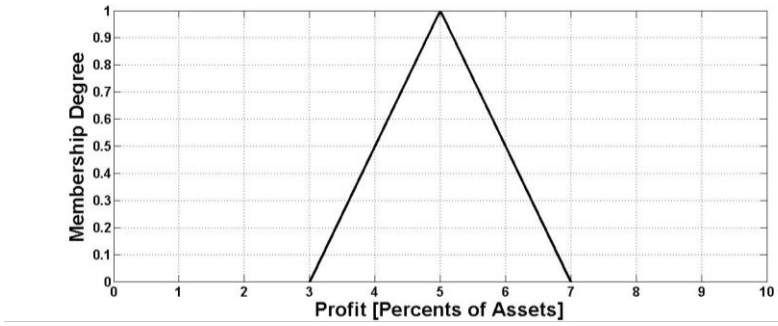

Figure 1
Example of an expertly specified evaluation

Example Let us consider a linguistic scale shown in Figure 2. This scale is formed by the linguistic terms "a big loss" (BL), "a small loss" (SL), "approximately without profit" (AWP), "a small profit" (SP), and "a big profit" (BP). In some cases, a selection of some linguistically described value like "a small profit" from the given linguistic scale can be more convenient for a decision-maker.

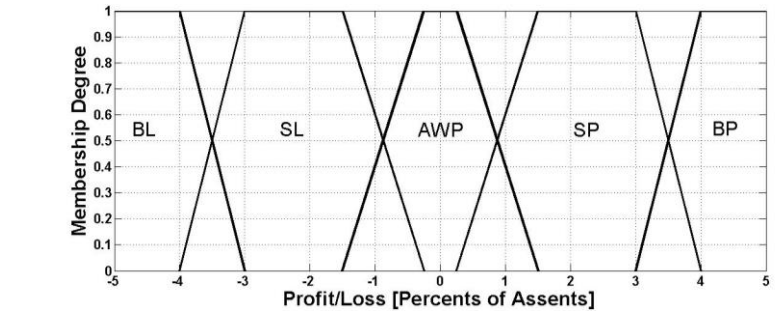

Figure 2
Example of a linguistic scale

3.3 Common Approach to a Fuzzy Decision Matrix

Let us describe a common approach to a fuzzy decision matrix that was considered e.g. in [**Error! Reference source not found.**2].

In the fuzzy decision matrix given in Table 2, x_1, x_2, …, x_n denote the alternatives of the decision-maker and S_1, S_2, …, S_m stand for the fuzzy states of the world. Probabilities of the fuzzy states of the world S_1, S_2, …, S_m, calculated according to (4), are denoted by p_{Z1}, p_{Z2}, …, p_{Zm}, i.e. $p_{Zj} = P_Z(S_j)$. For any $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, m\}$, $H_{i,j}$ expresses the fuzzy evaluation of the alternative x_i under the fuzzy state of the world S_j.

Table 2
Fuzzy decision matrix

	S_1	S_2	…	S_m		
	p_{Z1}	p_{Z2}	…	p_{Zm}		
x_1	$H_{1,1}$	$H_{1,2}$	…	$H_{1,m}$	EH_1^Z	$var\ H_1^Z$
x_2	$H_{2,1}$	$H_{2,2}$	…	$H_{2,m}$	EH_2^Z	$var\ H_2^Z$
…	…	…	…	…	…	…
x_n	$H_{n,1}$	$H_{n,2}$	…	$H_{n,m}$	EH_n^Z	$var\ H_n^Z$

Thus, the evaluation of the alternative x_i is understood as a discrete fuzzy random variable H_i^Z: $\{S_1, S_2, \ldots, S_m\} \rightarrow \mathcal{F}_N(\mathbb{R})$ where $H_i^Z(S_j) = H_{i,j}$ for j=1, 2, …, m. Its fuzzy expected value, denoted by EH_i^Z, is computed according to the generalized formula (1) where the probabilities p_j of the states of the world are replaced by the Zadeh's probabilities p_{Zj} of the fuzzy states of the world and the crisp evaluations $h_{i,j}$ are replaced by the fuzzy evaluations $H_{i,j}$, i.e.

$$EH_i^Z = \sum_{j=1}^{m} p_{Zj} \cdot H_{i,j}. \tag{5}$$

The α-cuts $EH_{i,\alpha}^Z = \left[Eh_{i,\alpha}^{ZL}, Eh_{i,\alpha}^{ZU}\right]$ are obtained for all $\alpha \in (0,1]$ as follows: Let $H_{i,j,\alpha} = \left[h_{i,j,\alpha}^L, h_{i,j,\alpha}^U\right]$, j = 1, 2, …, m. The boundary values of $EH_{i,\alpha}^Z$ are obtained by

$$Eh_{i,\alpha}^{ZL} = \sum_{j=1}^{m} p_{Zj} \cdot h_{i,j,\alpha}^L \tag{6}$$

and

$$Eh_{i,\alpha}^{ZU} = \sum_{j=1}^{m} p_{Zj} \cdot h_{i,j,\alpha}^U. \tag{7}$$

Computation of the fuzzy variance $var\ H_i^Z$ is more complex. It was shown in [9] that the formulas proposed in [**Error! Reference source not found.**2] were not correct because the relationships between the fuzzy evaluations $H_{i,1}$, $H_{i,2}$,…, $H_{i,m}$, and the fuzzy expected evaluation EH_i^Z were not involved in the calculation. This fact causes that the uncertainty of the resulting fuzzy variance is falsely increased. The proper formulas for the computation of the fuzzy variance were proposed in [9]. For any $i \in \{1, 2, ..., n\}$ and any $\alpha \in (0,1]$, the α-cut of the fuzzy variance $var\ H_{i,\alpha}^Z = \left[var\ h_{i,\alpha}^{ZL}, var\ h_{i,\alpha}^{ZU}\right]$ has to be calculated as follows: Let us denote

$$z_i\left(h_{i,1}, h_{i,2}, \ldots, h_{i,m}\right) = \sum_{j=1}^{m} p_{Zj} \cdot \left(h_{i,j} - \sum_{k=1}^{m} p_{Zk} \cdot h_{i,k}\right)^2. \tag{8}$$

Then,

$$var\ h_{i,\alpha}^{ZL} = \min\left\{z_i\left(h_{i,1}, h_{i,2}, ..., h_{i,m}\right) \middle|\ h_{i,j} \in H_{i,j,\alpha},\ j = 1,2,..,m\right\} \tag{9}$$

and

$$var\ h_{i,\alpha}^{ZU} = \max\left\{z_i\left(h_{i,1}, h_{i,2}, ..., h_{i,m}\right) \middle|\ h_{i,j} \in H_{i,j,\alpha},\ j = 1,2,..,m\right\}. \tag{10}$$

As it is written in section 2, the element $h_{i,j}$ of the matrix given in table 1 describes the decision-maker's evaluation of the alternative x_i if the state of the world S_j occurs. If we consider the fuzzy states of the world instead of the crisp ones, a natural question arises: What does it mean to say "if the fuzzy state of the world S_j occurs"? Let us suppose that some $\omega \in \Omega$ has occurred. If $\mu_{S_j}(\omega) = 1$, then it is clear that the evaluation of the alternative x_i is exactly $h_{i,j}$. However, what is the evaluation of x_i if $0 < \mu_{S_j}(\omega) < 1$ (which also means that $0 < \mu_{S_k}(\omega) < 1$ for some $k \neq j$)? Thus, perhaps it is not appropriate in the case of a decision matrix with the fuzzy states of the world to treat the evaluation of x_i as a discrete random variable H_i^Z that takes on the fuzzy values $H_{i,1}$, $H_{i,2}$, ..., $H_{i,m}$.

Moreover, it was pointed out by Rotterová and Pavlačka [10] that the Zadeh's probabilities p_{Z1}, p_{Z2}, …, p_{Zm} of the fuzzy states of the world express the expected membership degrees in which the particular states of the world will occur. Thus, they do not have in general the common probabilistic interpretation - a measure of a chance that a given event will occur in the future, which is desirable in the case of a decision matrix.

Therefore, we cannot say that the values EH_1^Z, EH_2^Z, ..., EH_n^Z, given by (6) and (7), and $var\ H_1^Z$, $var\ H_2^Z$, ..., $var\ H_n^Z$, given by (9) and (10), express the expected values and variances of evaluations of the alternatives, respectively. Ordering of the alternatives based on these characteristics is questionable.

4 Fuzzy Rule Bases System Determined by the Fuzzy Decision Matrix

In this section, let us introduce a different approach to the model of decision making under risk described by the decision matrix with fuzzy states of the world presented in Table 2. Taking the problems discussed in the previous section into account, we suggest not to treat the evaluation of the i^{th} alternative x_i, $i \in \{1, 2, \ldots, n\}$, as a discrete random variable H_i^Z taking on the fuzzy values $H_{i,1}$, $H_{i,2}$, ..., $H_{i,m}$ with the probabilities p_{Z1}, p_{Z2}, …, p_{Zm}. Instead of this, we propose to understand the information about the evaluation of the alternative x_i as the following fuzzy rule base:

If the state of the world is S_1, then the evaluation of x_i is $H_{i,1}$.
If the state of the world is S_2, then the evaluation of x_i is $H_{i,2}$.
⋮ (11)
If the state of the world is S_m, then the evaluation of x_i is $H_{i,m}$.

In [8], it was shown that in the case of the fuzzy decision matrix with crisp evaluations under the particular fuzzy states of the world, it is appropriate to use the *Sugeno's method of fuzzy inference*, introduced in [11]. The obtained output from the fuzzy rule base was expressed by a real number.

In the paper, we deal with the fuzzy evaluations of the alternatives under the fuzzy states of the world. Thus, the so-called *generalised Sugeno's method of fuzzy inference,* introduced in [13], should be applied for obtaining an output from the fuzzy rule base (11). According to this method, the evaluation of an alternative x_i for a given $\omega \in \Omega$ is computed in the following way:

$$H_i^S(\omega) = \frac{\sum_{j=1}^{m} \mu_{S_j}(\omega) \cdot H_{i,j}}{\sum_{j=1}^{m} \mu_{S_j}(\omega)} = \sum_{j=1}^{m} \mu_{S_j}(\omega) \cdot H_{i,j}. \tag{12}$$

For any $\alpha \in (0,1]$, let us denote $H_{i,j,\alpha} = \left[h_{i,j,\alpha}^L, h_{i,j,\alpha}^U\right]$, $j = 1, 2, \ldots, m$, and $H_{i,\alpha}^S(\omega) = \left[h_{i,\alpha}^{SL}(\omega), h_{i,\alpha}^{SU}(\omega)\right]$. Then, the boundary values of $H_{i,\alpha}^S(\omega)$ are computed as follows:

$$h_{i,\alpha}^{SL}(\omega) = \sum_{j=1}^{m} \mu_{S_j}(\omega) \cdot h_{i,j,\alpha}^L$$

and

$$h_{i,\alpha}^{SU}(\omega) = \sum_{j=1}^{m} \mu_{S_j}(\omega) \cdot h_{i,j,\alpha}^U.$$

Remark In the formula (12), the denominator equals to one due to the assumption that the fuzzy states of the world S_1, S_2, …, S_m form a fuzzy partition of Ω. It is worth to note that in our approach, this assumption can be omitted.

Since we operate within the given probability space $(\Omega, \mathcal{A}, P)$, H_i^S is a fuzzy random variable such that $H_i^S: \Omega \to \mathcal{F}_N(\mathbb{R})$.

Remark It can be easily seen from (12) that in the case of the crisp states of the world S_j, $j = 1, 2, \ldots, m$, and the crisp evaluations $h_{i,j}$, $i = 1, 2, \ldots, n$, under the particular fuzzy states of the world, the fuzzy random variables H_i^S coincide with discrete random variables H_i taking on the values $h_{i,j}$ with the probabilities p_j,

$j = 1, 2, \ldots, m$. Thus, this new approach can be seen as an extension of a decision matrix to the case of the fuzzy states of the world and the fuzzy evaluations of alternatives where appropriate.

Analogously, as in the common approach to the fuzzy decision matrix, the ordering of the alternatives $x_1, x_2, \ldots, x_n$ can be based on the fuzzy expected values and the fuzzy variances of the random variables H_i^S, $i = 1, 2, \ldots, n$. Let us introduce the formulas for computations of the α-cuts of EH_i^S and $var\ H_i^S$.

For any $\alpha \in (0,1]$, the α-cut of the fuzzy expected output from the fuzzy rule base given by (11), denoted by $EH_{i,\alpha}^S = \left[Eh_{i,\alpha}^{SL}, Eh_{i,\alpha}^{SU}\right]$, is obtained as follows:

$$Eh_{i,\alpha}^{SL} = \min\left\{\int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}\,dP \,\middle|\, h_{i,j} \in H_{\mathrm{i,j},\alpha}, j = 1,2,..,m\right\}$$
$$= \int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j,\alpha}^{L}\,dP \tag{13}$$

and

$$Eh_{i,\alpha}^{SU} = \max\left\{\int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}\,dP \,\middle|\, h_{i,j} \in H_{\mathrm{i,j},\alpha}, j = 1,2,..,m\right\}$$
$$= \int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j,\alpha}^{U}\,dP. \tag{14}$$

The α-cut $var\ H_{i,\alpha}^S = \left[var\ h_{i,\alpha}^{SL}, var\ h_{i,\alpha}^{SU}\right]$ of the fuzzy variance of the output from the fuzzy rule base is obtained as follows: Let us denote

$$\mathrm{s}_i\left(h_{i,1}, h_{i,2}, \ldots, h_{i,m}\right)$$
$$= \int_{\omega\in\Omega}\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j} - \int_{t\in\Omega}\sum_{k=1}^{m}\mu_{S_k}(t)\cdot h_{i,k}\,dP\right)^2 dP. \tag{15}$$

Then,

$$var\ h_{i,\alpha}^{SL} = \min\left\{\mathrm{s}_i\left(h_{i,1}, h_{i,2}, \ldots, h_{i,m}\right) \middle|\, h_{i,j} \in H_{\mathrm{i,j},\alpha}, j = 1,2,..,m\right\} \tag{16}$$

and

$$var\ h_{i,\alpha}^{SU} = \max\left\{\mathrm{s}_i\left(h_{i,1}, h_{i,2}, \ldots, h_{i,m}\right) \middle|\, h_{i,j} \in H_{\mathrm{i,j},\alpha}, j = 1,2,..,m\right\}. \tag{17}$$

Now, let us compare the fuzzy expected values EH_i^Z and EH_i^S, and the fuzzy variances $var\ H_i^Z$ and $var\ H_i^S$.

Theorem 1 For $i = 1, 2, \ldots, n$, the expected fuzzy evaluation EH_i^Z and the expected output from the fuzzy rule base EH_i^S coincide.

Proof For any $\alpha \in (0,1]$, let $EH_{i,\alpha}^S = \left[Eh_{i,\alpha}^{SL}, Eh_{i,\alpha}^{SU}\right]$ be the α-cut of the expected output from the fuzzy rule base and $EH_{i,\alpha}^Z = \left[Eh_{i,\alpha}^{ZL}, Eh_{i,\alpha}^{ZU}\right]$ be the α-cut of the fuzzy expected evaluation. For the boundary values of $EH_{i,\alpha}^S$, it holds:

$$Eh_{i,\alpha}^{SL} = \int_{\omega\in\Omega} \sum_{j=1}^{m} \mu_{S_j}(\omega) \cdot h_{i,j,\alpha}^L dP = \sum_{j=1}^{m} \int_{\omega\in\Omega} \mu_{S_j}(\omega) dP \cdot h_{i,j,\alpha}^L$$

$$= \sum_{j=1}^{m} p_{Zj} \cdot h_{i,j,\alpha}^L = Eh_{i,\alpha}^{ZL}$$

and

$$Eh_{i,\alpha}^{SU} = \int_{\omega\in\Omega} \sum_{j=1}^{m} \mu_{S_j}(\omega) \cdot h_{i,j,\alpha}^U dP = \sum_{j=1}^{m} \int_{\omega\in\Omega} \mu_{S_j}(\omega) dP \cdot h_{i,j,\alpha}^U$$

$$= \sum_{j=1}^{m} p_{Zj} \cdot h_{i,j,\alpha}^U = Eh_{i,\alpha}^{ZU}.$$

Thus, all the α-cuts are the same. Therefore, $EH_i^S = EH_i^Z$. □

In [8], the authors showed that in the case of a fuzzy decision matrix where the evaluations under the particular fuzzy states of the world are expressed by real numbers, the variances $var\, H_i^Z$ and $var\, H_i^S$ are real numbers as well, and $var\, H_i^Z \geq var\, H_i^S$. Now, let us compare the fuzzy variances $var\, H_i^Z$ and $var\, H_i^S$.

Theorem 2 For $i = 1, 2, \ldots, n$, the fuzzy variance $var\, H_i^Z$ of the fuzzy evaluation is greater or equals to the fuzzy variance $var\, H_i^S$ of the output from the fuzzy rule base (11).

Proof Let $z_i(h_{i,1}, h_{i,2}, \ldots, h_{i,m})$ and $s_i(h_{i,1}, h_{i,2}, \ldots, h_{i,m})$ be the auxiliary functions defined by (8) and (15), respectively. For the sake of simplicity, let us denote for a given $h_{i,j} \in H_{i,j,\alpha}$, $j = 1, 2, \ldots, m$,

$$Eh_i = \sum_{j=1}^{m} p_{Zj} \cdot h_{i,j} = \int_{\omega\in\Omega} \sum_{j=1}^{m} \mu_{S_j}(\omega) \cdot h_{i,j} dP.$$

We can express the difference of $z_i(h_{i,1}, h_{i,2}, \ldots, h_{i,m})$ and $s_i(h_{i,1}, h_{i,2}, \ldots, h_{i,m})$ as follows:

$$\mathrm{d}_i\left(h_{i,1},h_{i,2},\ldots,h_{i,m}\right)= z_i\left(h_{i,1},h_{i,2},\ldots,h_{i,m}\right)- s_i\left(h_{i,1},h_{i,2},\ldots,h_{i,m}\right)$$

$$= \sum_{j=1}^{m} p_{Zj}\cdot\left(h_{i,j}-Eh_i\right)^2 - \int_{\omega\in\Omega}\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}-Eh_i\right)^2 dP$$

$$= \sum_{j=1}^{m}\int_{\omega\in\Omega}\mu_{S_j}(\omega)dP\cdot\left(h_{i,j}^2-2\cdot h_{i,j}\cdot Eh_i+\left(Eh_i\right)^2\right)$$

$$-\int_{\omega\in\Omega}\left(\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}\right)^2-2\cdot\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}\cdot Eh_i+\left(Eh_i\right)^2\right)dP$$

$$= \int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}^2 dP-2\cdot Eh_i\cdot\int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}dP$$

$$+\left(Eh_i\right)^2\cdot\int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)dP-\int_{\omega\in\Omega}\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}\right)^2 dP$$

$$+2\cdot Eh_i\cdot\int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}dP-\left(Eh_i\right)^2\cdot\int_{\omega\in\Omega}dP$$

$$= \int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}^2 dP-\left(Eh_i\right)^2-\int_{\omega\in\Omega}\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}\right)^2 dP$$

$$+\left(Eh_i\right)^2=\int_{\omega\in\Omega}\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}^2 dP-\int_{\omega\in\Omega}\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}\right)^2 dP$$

$$= \int_{\omega\in\Omega}\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}^2-\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}\right)^2\right)dP$$

where relations (3), (13), (14) and the following relation from measure theory:

$$\int_{\omega\in\Omega} dP = P(\Omega)=1,$$

were applied.

The integrand $\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}^2-\left(\sum_{j=1}^{m}\mu_{S_j}(\omega)\cdot h_{i,j}\right)^2\right)$ is clearly non-negative (it represents the variance of a discrete random variable that takes on the values $h_{i,j}$, $j = 1, 2, \ldots, m$, with the "probabilities" $\mu_{S_j}(\omega)$, $j = 1, 2, \ldots, m$). It is equal to

zero if and only if $h_{i,j} = h_{i,k}$ for any $j \neq k$ such that both p_{Zj} and p_{Zk} are positive. Thus, the function $d_i(h_{i,1}, h_{i,2}, ..., h_{i,m})$ is always non-negative.

However, $d_i(h_{i,1}, h_{i,2}, ..., h_{i,m})$ is the auxiliary function for computation of the fuzzy difference D_i between $var\, H_i^Z$ and $var\, H_i^S$. For any $\alpha \in (0,1]$, the α-cut of the fuzzy difference $D_{i,\alpha} = \left[d_{i,\alpha}^L, d_{i,\alpha}^U\right]$ is given as follows:

$$d_{i,\alpha}^L = \min\left\{\mathrm{d}_i(h_{i,1}, h_{i,2}, ..., h_{i,m}) \middle| \, h_{i,j} \in H_{\mathrm{i,j},\alpha}, j = 1,2,..,m\right\}$$

and

$$d_{i,\alpha}^U = \max\left\{d_i(h_{i,1}, h_{i,2}, ..., h_{i,m}) \middle| \, h_{i,j} \in H_{\mathrm{i,j},\alpha}, j = 1,2,..,m\right\}.$$

Due to the non-negativity of the auxiliary function d_i, the α-cut of the fuzzy difference $D_{i,\alpha}$ contains only non-negative values, i.e. $var\, H_{i,\alpha}^Z \geq var\, H_{i,\alpha}^S$. Hence, $var\, H_i^Z \geq var\, H_i^S$. □

Thus, although the fuzzy expected values EH_i^Z and EH_i^S coincide, the fuzzy variances $var\, H_i^Z$ and $var\, H_i^S$ differ in general. This can affect the ranking of the considered alternatives, which is illustrated by the example in Section 5.

Now, let us focus on the interpretation of EH_i^S and $var\, H_i^S$. Both characteristics describe a random variable that explains outputs from the fuzzy rule base (11). There are no such interpretational problems as those discussed in the previous section. So this approach seems to be more appropriate for the practical use.

5 Illustrative Example

Let us illustrate the difference between both described approaches on the similar problem as was considered in [9]. Let us compare two stocks, A and B, with respect to their future yields. We consider the following states of the economy: "great economic drop" (GD), "economic drop" (D), "economic stagnation" (S), "economic growth" (G), and "great economic growth" (GG). Let us assume that the considered states of the economy are given only by the development of the gross domestic product, abbreviated as GDP. Further, we assume that the next year prediction of GDP development [%] shows a normally distributed growth of GDP with parameters $\mu = 1.5$ and $\sigma = 2$.

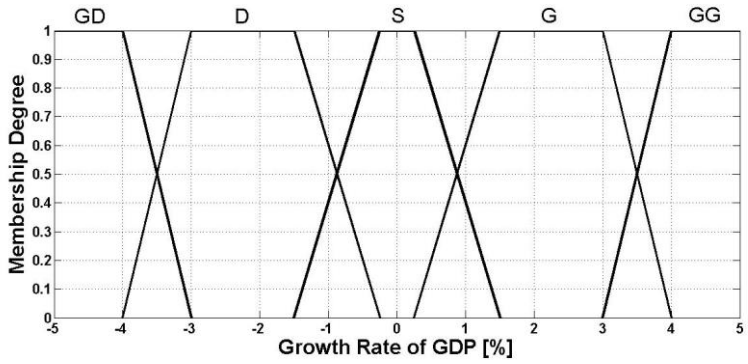

Figure 3
Linguistic scale of the states of the economy

A considered state of the economy can be expressed by a *trapezoidal fuzzy number* which is determined by its *significant values* a_1, a_2, a_3, and a_4 such that $a_1 \leq a_2 \leq a_3 \leq a_4$. The membership function of any trapezoidal fuzzy number $A \in \mathcal{F}_N(\mathbb{R})$ is for any $x \in \mathbb{R}$ in the form as follows:

$$\mu_A(x) = \begin{cases} \dfrac{x - a_1}{a_2 - a_1} & \text{if } x \in [a_1, a_2), \\ 1 & \text{if } x \in [a_2, a_3], \\ \dfrac{a_4 - x}{a_4 - a_3} & \text{if } x \in (a_3, a_4], \\ 0 & \text{otherwise.} \end{cases}$$

The trapezoidal fuzzy number A determined by its significant values is denoted further by (a_1, a_2, a_3, a_4).

Let us assume that the states of the economy are mathematically expressed by trapezoidal fuzzy numbers that form a linguistic scale shown in Figure 3. Moreover, let us consider that the predictions of future stock yields (in %) are set expertly.

Significant values of the fuzzy states of the economy and of the fuzzy stock yields are shown in Table 3. The probabilities of the fuzzy states of the economy were calculated according to the formula (4) and are used only in the calculation of the characteristics of the output with respect to the common approach described in Section 3.

Table 3
Considered fuzzy decision matrix

Economy states	GD = (-∞, -∞, -4, -3)				D = (-4, -3, -1.5, -0.25)			
Probabilities	0.0067				0.1146			
A yield (%)	-36	-34	-31	-16	-20	-17	-10	0
B yield (%)	-45	-40	-32	-25	-22	-17	-11	0
Economy states	S = (-1.5, -0.25, 0.25, 1.5)				G = (0.25, 1.5, 3, 4)			
Probabilities	0.2579				0.4596			
A yield (%)	-5	-3	3	10	6	12	17	24
B yield (%)	-5	-3	3	5	8	12	16	18
Economy states	GG = (3, 4, ∞, ∞)							
Probabilities	0.1612							
A yield (%)	22	27	34	36				
B yield (%)	20	26	33	40				

The resultant fuzzy expected values and the fuzzy variances can be compared, for instance, according to their centers of gravity. The center of gravity of a fuzzy number $A \in \mathcal{F}_N(\mathbb{R})$ is a real number cog_A given as follows:

$$cog_A = \frac{\int_{-\infty}^{\infty} x \cdot \mu_A(x) dx}{\int_{-\infty}^{\infty} \mu_A(x) dx}.$$

The fuzzy expected values *EA* and *EB*, computed by the formulas (6) and (7) (or (13) and (14)) are trapezoidal fuzzy numbers. Their significant values are given in Table 4. The fuzzy variances $var\ A^Z$ and $var\ B^Z$, obtained by the formulas (9) and (10), as well as $var\ A^S$ and $var\ B^S$, computed by (16) and (17), are not trapezoidal fuzzy numbers. Their membership functions are shown in Figures 4 and 5. The significant values of the fuzzy variances are also given in Table 4 (by these significant values we understand end points of the core and of the closure of the support). We can see that the fuzzy variances of the outputs from the fuzzy rule bases reach lower values than the variances obtained by the common approach.

From the results given in Table 4, it is obvious that the center of gravity of the fuzzy expected value *EA* is greater than the center of gravity of the fuzzy expected value *EB*. Therefore, without considering the variances the decision-maker should prefer the stock A.

Table 4
Resultant stocks characteristics

Stock Characteristic	Significant Values (%)				Centre of Gravity
EA	2.48	6.92	12.71	19.30	10.44
EB	2.79	6.72	11.97	15.84	9.33
var A^Z	38.48	117.61	255.30	365.28	195.21
var B^Z	40.12	117.06	245.53	369.68	194.83
var A^S	33.44	105.75	229.30	324.36	173.98
var B^S	35.41	105.23	220.70	332.41	174.98

In this example, we can also see that the change in the fuzzy variance computation can cause a change in the decision-maker's preferences. Based on *var* A^Z and *var* B^Z, the decision-maker is not able to make a decision on the basis of the rule of the expected value and the variance described in Section 2, while based on *var* A^S and *var* B^S, the decision-maker should prefer the stock A (the higher expected value and the lower variance than the stock B compared on the basis of centers of gravity of variances approximated by trapezoidal fuzzy numbers).

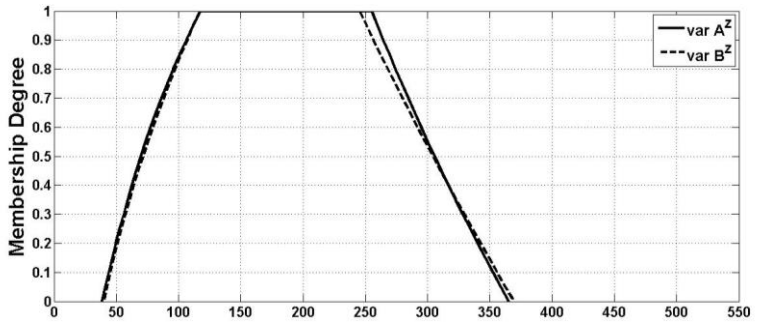

Figure 4
Membership functions of *var* A^Z and *var* B^Z

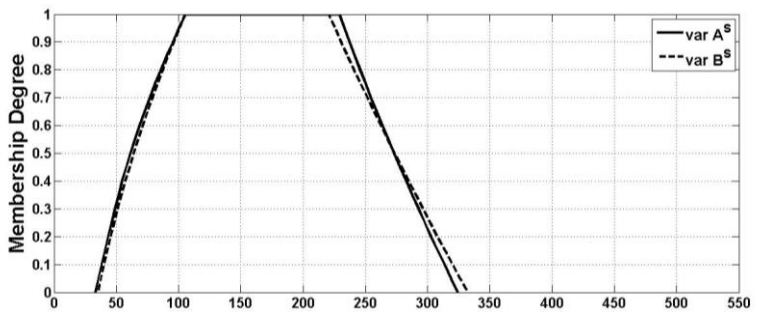

Figure 5
Membership functions of *var* A^S and *var* B^S

Conclusions

We have dealt with the problem of the extension of a decision matrix for the case of the fuzzy states of the world and the fuzzy evaluations of the alternatives. We have analyzed the common approach to this problem proposed in [**Error! Reference source not found.**2] that is based on applying the Zadeh's probabilities of the fuzzy states of the world. We have found out that the meaning of the obtained characteristics of the evaluations of the alternatives, namely the fuzzy expected values and the fuzzy variances, is questionable. Therefore, we have introduced a new approach that is based on the idea that a fuzzy decision matrix does not determine discrete fuzzy random variables, but fuzzy rule bases. In such a case, the obtained characteristics of the evaluations, based on which the alternatives are compared, are clearly interpretable. We have proved that the resulting expected values of the evaluations are for both approaches the same, whereas the variances generally differ. In the numerical example, we have shown that the final ordering of the alternatives, according to both approaches, can be different.

Future work in this field will be focused on the case, where the underlying probability measure is fuzzy. For instance, the parameters of the underlying probability distribution, like μ and σ in the case of the normal distribution considered in the numerical example in Section 5, could be expertly set with fuzzy numbers.

Acknowledgement

This work was supported by the project No. GA 14-02424S of the Grant Agency of the Czech Republic *Methods of Operations Research for Decision Support under Uncertainty* and by the grant IGA_PrF_2016_025 *Mathematical Models* of the *Internal Grant Agency of Palacký University Olomouc*.

References

[1] Chen, S., Hwang, Ch.: *Fuzzy Multiple Attribute Decision Making Methods: Methods and Applications*. Springer Berlin Heidelberg, 1992

[2] Cheng, Ch.: A simple fuzzy group decision making method. In: *Fuzzy Systems Conference Proceedings 1999*. IEEE International, Seoul, South Korea, 1999, 910-915

[3] Ganoulis, J.: *Engineering Risk Analysis of Water Pollution: Probabilities and Fuzzy Sets*. VCH, Weinheim, 2008

[4] Huynh, V. et. al.: A Fuzzy Target Based Model for Decision Making Under Uncertainty. *Fuzzy Optimization and Decision Making* **6**, 3 (2007), 255-278

[5] Liao, H., Xu, Z.: *Fuzzy Decision Making Methodologies and Applications*. Springer Singapore, 2017

[6] Matos, M. A.: Decision under risk as a multicriteria problem. *European Journal of Operational Research* **181** (2007) 1516-1529

[7] Özkan, I., Türken, I. B.: Uncertainty and fuzzy decisions. In: *Chaos Theory in Politics. Understanding Complex Systems*. (S. Banerjeeet al. eds.), Springer Netherlands, Dordrecht, 2014, 17-27

[8] Pavlačka, O., Rotterová, P.: Fuzzy Decision Matrices Viewed as Fuzzy Rule-Based Systems. In: *Proceedings of the 34th International Conference Mathematical Methods in Economics*. (A. Kocourek, M. Vavroušek, eds.), Technical University of Liberec, Liberec, 2016, 641-646

[9] Rotterová, P., Pavlačka, O.: Computing Fuzzy Variances of Evaluations of Alternatives in Fuzzy Decision Matrices. In: *Proceedings of the 34th International Conference Mathematical Methods in Economics*. (A. Kocourek, M. Vavroušek, eds.), Technical University of Liberec, Liberec, 2016, 735-740

[10] Rotterová, P., Pavlačka, O.: Probabilities of Fuzzy Events and Their Use in Decision Matrices. *International Journal of Mathematics in Operational Research* **9**, 4 (2016) 423-435

[11] Sugeno, M.: *Industrial applications of fuzzy control*. Elsevier Science Pub. Co., New York, 1985

[12] Talašová, J., Pavlačka, O.: Fuzzy Probability Spaces and Their Applications in Decision Making. *Austrian Journal of Statistics* **35**, 2&3 (2006) 347-356

[13] Talašová, J.: *Fuzzy methods of multiple criteria evaluation and decision making* (in Czech) Publishing House of Palacký University, Olomouc, 2003

[14] Yoon, K. P., Hwang, Ch.: *Multiple Attribute Decision Making: An Introduction*. SAGE Publications, California, 1995

[15] Zadeh, L. A.: Fuzzy sets, *Information and Control* **8**, 3 (1965) 338-353

[16] Zadeh, L. A.: The Concept of a Linguistic Variable and its Application to Approximate Reasoning - I. *Information Sciences* **8** (1975) 199-245

[17] Zadeh, L. A.: Probability Measures of Fuzzy Events. *Journal of Mathematical Analysis and Applications* **23**, 2 (1968) 421-427

Directional Monotonicity of Fuzzy Implications

Katarzyna Miś

Institute of Mathematics, University of Silesia in Katowice
Bankowa 14, 40-007 Katowice, Poland, kmis@us.edu.pl

Abstract: In this paper we consider special fuzzy implications as directional increasing functions and we introduce the notion of inversely special fuzzy implications as directional decreasing functions. We recall some results connected with special R-implications shown by Sainio et al. [A characterization of fuzzy implications generated by generalized quantifiers, Fuzzy Sets and Systems 159, 2008, pp. 491-499] and we present several new results connected with inversely special R-implications. Also, we discuss this new property for other families of fuzzy implications like (S,N)-implications, f-implications and g-implications.

Keywords: fuzzy implications; special implications; inversely special implications; directional monotonicity

1 Introduction

Standard monotonicity is one of the key properties of any function. Some functions used in fuzzy logic like t-norms, t-conorms, copulas are increasing in each variable. However, a very important fuzzy connective, a fuzzy implication, is hybrid monotonic – it is decreasing in the first variable and increasing in the second one. For such functions, among others, a notion of directional monotonicity was introduced. Our motivation is the article *Directional monotonicity of fusion functions* (Bustince et al. [3]) in which the authors investigated it deeper for different families of functions. In our paper we refer it to fuzzy implications. It turns out that there are some directional increasing and decreasing implications among which we support with examples.

In this paper we consider special fuzzy implications as directional increasing functions and inversely special fuzzy implications as directional decreasing functions (see Section 3). The first notion was introduced in 1996 by Hájek and Kohout [5]. Later, Sainio et al. [11] and Jayaram and Mesiar [6] showed some results concerning R-implications, which we cite here in Section 4. In Section 5 we formulate main new results for inversely special R-implications, while in Section 6 we consider other classes of inversely special fuzzy implications: (S,N)-implications, f-implications and g-implications.

2. Basic Definitions

This section contains definitions, properties and characterizations of directional monotonicity, fuzzy connectives and convex functions that will be used in the main part of this paper.

2.1 A Directional Monotonicity

The notion of the directional monotonicty was introduced in 2015 for functions which are not monotonic in each variable. Such functions are for example weighted arithmetic means, OWA operators, the Choquet and Sugeno integrals. As we mentioned before, fuzzy implications are monotonic in each variable separately, but not together. However, all these types of functions can be monotonic in a way described below.

Definition 1 *(Bustince et al. [3, Definition 2]).* Let $n \in \mathbb{N}, n \geq 2, \mathbb{I}$ be the unit interval $[0,1]$ and $r \in \mathbb{R}^n$ such that $r = (r_1, \dots, r_n) \neq (0, \dots, 0)$. A function $F: \mathbb{I}^n \longrightarrow \mathbb{I}$ is:

i. r-increasing, if for all $x \in \mathbb{I}^n$ and $c > 0$ such that $x + cr \in \mathbb{I}^n$, it holds that
$$F(x + cr) \geq F(x)$$
ii. r-decreasing, if for all $x \in \mathbb{I}^n$ and $c > 0$ such that $x + cr \in \mathbb{I}^n$, it holds that
$F(x + cr) \leq F(x)$.

Lemma 2 Let $r = (r_1, \dots, r_1), r_1 > 0$ and $n \in \mathbb{N}, n \geq 2$. A function $F: \mathbb{I}^n \longrightarrow \mathbb{I}$ is:

1) r-increasing if and only if it is $\mathbb{1}$-increasing,
2) r-decreasing if and only if it is $\mathbb{1}$-decreasing,

where $\mathbb{1} = \underbrace{(1, \dots, 1)}_{n}$

Proof. We show it only for r-increasing functions. The proof for r-decreasing functions is parallel. Let $x = (x_1, \dots, x_n)$ for $x_1, \dots, x_n \in \mathbb{I}$. If a function F is $(r_1, \dots, r_1)$-increasing, then for $c > 0$ such that $(x_1 + cr_1, \dots, x_n + cr_1) \in \mathbb{I}^n$ we have

$$F(x_1 + cr_1, \dots, x_n + cr_1) \geq F(x_1, \dots, x_n)$$
$$F(x_1 + d \cdot 1, \dots, x_n + d \cdot 1) \geq F(x_1, \dots, x_n)$$

Hence, F is $\mathbb{1}$-increasing, for $d > 0$ and $d = cr_1$.

If F is $\mathbb{1}$-increasing, then for applicable $c > 0$ we have

$$F(x_1 + c, \dots, x_n + c) \geq F(x_1, \dots, x_n)$$
$$F(x_1 + d \cdot r_1, \dots, x_n + d \cdot r_1) \geq F(x_1, \dots, x_n)$$

where $d = \frac{c}{r_1}$ and $d > 0$. Therefore, F is r-increasing.

Let us consider a notion of directional monotonicity for two different types of functions.

Example 3

1. The Fodor implication is given by the formula
   $I_{FD}(x,y) = \begin{cases} 1, & x \le y \\ \max\{1-x, y\}, & x > y \end{cases}$ for $x, y \in [0,1]$
   For $x = 0.2,\ y = 0.1,\ c = 0.4$ we have
   $$I_{FD}(x,y) = 0.8 > I_{FD}(x+c, y+c) = 0.5$$
   For the same x, y and $c = 0.71$ we have
   $$I_{FD}(x,y) = 0.8 < I_{FD}(x+c, y+c) = 0.81$$
   Therefore, I_{FD} is not $(1,1)$-increasing neither $(1,1)$-decreasing.
2. The Goguen implication, given by the formula
   $I_{GG}(x,y) = \begin{cases} 1, x \le y \\ \frac{y}{x}, x > y \end{cases}$ for $x, y \in [0,1]$, is (r_1, r_2)-increasing for $r_1, r_2 \ge 0$ such that $r_2 \ge r_1$. Indeed, for $x \le y$ and $c > 0$ such that $x + cr_1$, $y + cr_2 \in [0,1]$ we have $x + cr_1 \le y + cr_2$ when $r_1 \le r_2$ and then $I(x,y) = 1 \le 1 = I(x + cr_1, y + cr_2)$. For $x > y$ and applicable $c > 0$ we have $\frac{y}{x} \le \frac{y+cr_2}{x+cr_1} \iff r_1 y - r_2 x \le 0$, which is true if $r_1 \le r_2$.
3. Let $F: [0,1]^2 \to [0,1]$ be a function given by the formula
   $F(x,y) = (1-\lambda) \cdot \max\{x,y\} + \lambda \cdot \min\{x,y\}$, $\lambda \in [0,1]$ (see [2]).
   Then it is r-decreasing for all $r \in [0,1]^2$ such that $r = (r_1, r_2)$ and $r_1 + \frac{\lambda}{1-\lambda} r_2 \le 0,\ r_1 + \frac{1-\lambda}{\lambda} r_2 \le 0$. Indeed, for all $r_1, r_2, x, y \in [0,1]$ and $c > 0$ such that $x + cr_1, y + cr_2 \in [0,1]$ we have
   $(1-\lambda) \cdot \max\{x,y\} + \lambda \cdot \min\{x,y\} \ge$
   $$(1-\lambda) \cdot \max\{x + cr_1, y + cr_2\} + \lambda \cdot \min\{x + cr_1, y + cr_2\}$$
   this leads us to the following inequalities:
   $r_1 \le -\frac{\lambda}{1-\lambda} r_2$, when $x \ge y$ and $r_1 \le -\frac{1-\lambda}{\lambda} r_2$ for $x < y$

The notion of the directional monotonicity is a generalization of another one, i.e., weak monotonicity (see [12]). Thanks to Lemma 2 we can say that weak monotonic function is a directional one in the direction of the vector $\mathbb{1}$.

More general facts and properties of directional monotonic functions can be found in Bustince et al. [3].

2.2 Fuzzy Connectives

We assume that the reader is familiar with the classical results concerning basic fuzzy logic connectives, but to make this work more self-contained, we place some of them here.

Definition 4 *(Fodor and Roubens [4]).* A function $N:[0,1] \longrightarrow [0,1]$ is called a fuzzy negation if

- $N(0) = 1$ and $N(1) = 0$
- N is decreasing

The basic example of a fuzzy negation is the classical strong negation N_C, i.e.,

$N_C(x) = 1 - x, x \in [0,1]$.

2.2.1 T- norms, t-conorms and Copulas

This part contains basic definitions and theorems, which are necessary to define some families of fuzzy implications.

Definition 5 *(Fodor and Roubens [4]).* A function $T:[0,1]^2 \longrightarrow [0,1]$ is called a triangular norm (t-norm) if it satisfies the following conditions:

- $T(1,x) = x$ for all $x \in [0,1]$
- $T(x,y) = T(y,x)$ for all $x, y \in [0,1]$
- $T(x,y) \leq T(u,v)$ for all $0 \leq x \leq u \leq 1, 0 \leq y \leq v \leq 1$
- $T\big(x,T(y,z)\big) = T(T(x,y),z)$ for all $x,y,z \in [0,1]$

Definition 6 *(Fodor and Roubens [4]).* A function $S:[0,1]^2 \longrightarrow [0,1]$ is called a triangular conorm (t-conorm) if it satisfies the following conditions:

- $S(0,x) = x$ for all $x \in [0,1]$
- $S(x,y) = S(y,x)$ for all $x, y \in [0,1]$
- $S(x,y) \leq S(u,v)$ for all $0 \leq x \leq u \leq 1, 0 \leq y \leq v \leq 1$
- $S\big(x,S(y,z)\big) = S(S(x,y),z)$ for all $x,y,z \in [0,1]$

Definition 7 *(Klement et al. [7, Definitions 2.9, 2.13]).* A t-norm T is said to be:

- Archimedean, if for each $(x,y) \in (0,1)^2$ there is an $n \in \mathbb{N}$ such that $x_T^{[n]} < y$, where by the notation $x_T^{[n]}$ we understand $x_T^{[n]} = \begin{cases} 1, & \text{if } n = 0 \\ x, & \text{if } n = 1 \\ T\left(x, x_T^{[n-1]}\right), & \text{if } n > 1 \end{cases}$
- Nilpotent, if it is continuous and for each $x \in (0,1)$ there is an $n \in \mathbb{N}$ such that $x_T^{[n]} = 0$.
- Strict, if it is continuous and strictly monotonic, i.e., $T(x,y) < T(x,z)$ whenever $x > 0$ and $y < z$.

The following theorem is usually used to characterize continuous Archimedean t-norms, its first proof can be found in the article written by Ling [9].

Theorem 8 *(Klement et al. [7, Theorem 5.1]).* For a function $T:[0,1]^2 \longrightarrow [0,1]$ the following statements are equivalent:

i. T is a continuous Archimedean t-norm

ii. T has a continuous additive generator, i.e., there exists a continuous, strictly decreasing function $f: [0,1] \longrightarrow [0,\infty]$ with $f(1) = 0$ such that $T(x,y) = f^{-1}(\min\{f(x) + f(y), f(0)\})$, for $x, y \in [0,1]$. Moreover, such representation is unique up to a positive multiplicative constant.

The following theorem tells about a method of constructing new t-norms from some family of given t-norms.

Theorem 9 *(Klement et al. [7, Theorem 3.43])*. Let $(T_\alpha)_{\alpha \in A}$ be a family of t-norms and $((a_\alpha, e_\alpha))_{\alpha \in A}$ be a family of non-empty, pairwise disjoint open subintervals of $[0,1]$. Then the following function $T: [0,1]^2 \longrightarrow [0,1]$ is a t-norm:

$$T(x,y) = \begin{cases} a_\alpha + (e_\alpha - a_\alpha) \cdot T_\alpha\left(\frac{x-a_\alpha}{e_\alpha - a_\alpha}, \frac{y-a_\alpha}{e_\alpha - a_\alpha}\right), & \text{if } (x,y) \in [a_\alpha, e_\alpha]^2 \\ \min\{x,y\}, & \text{otherwise.} \end{cases} \quad (1)$$

This theorem allows us to formulate the following definition.

Definition 10 *(Klement et al. [7, Definition 3.44])*. Let $(T_\alpha)_{\alpha \in A}$ be a family of t-norms and $((a_\alpha, e_\alpha))_{\alpha \in A}$ be a family of non-empty, pairwise disjoint open subintervals of $[0,1]$. The t-norm T defined by (1) is called the ordinal sum of the summands $< a_\alpha, e_\alpha, T_\alpha >, \alpha \in A$, and we shall write $T = (< a_\alpha, e_\alpha, T_\alpha >)_{\alpha \in A}$.

In the following theorem, we recall a very important characterization of continuous t-norms.

Theorem 11 *(Klement et al. [7, Theorem 5.11])*. For a function $T: [0,1]^2 \to [0,1]$ the following statements are equivalent:

i. T is a continuous t-norm.
ii. T is uniquely representable as an ordinal sum of continuous Archimedean t-norms, i.e., T is defined by a formula (1).

We present a definition of a copula below. This notion is necessary to show its relationship with t-norms.

Definition 12 *(Klement et al. [7, Definition 9.4])*. A function $C: [0,1]^2 \longrightarrow [0,1]$ is a copula if, for all $x, y, u, v \in [0,1]$ with $x \leq u$ and $y \leq v$, it satisfies the following conditions:

- $C(x,y) + C(u,v) \geq C(x,v) + C(u,y)$
- $C(x,0) = C(0,x) = 0$
- $C(x,1) = C(1,x) = x$

Definition 13. A function $f: [0,1]^2 \longrightarrow [0,1]$ is said to be 1-Lipschitz if it satisfies the Lipschitz property with constant 1 i.e.,

$|f(x_1, y_1) - f(x_2, y_2)| \leq |x_1 - x_2| + |y_1 - y_2|$ for all $x_1, x_2, y_1, y_2 \in [0,1]$.

The next theorem is the full characterization of t-norms which are copulas.

Theorem 14 *(Moynihan [10, Theorem 3.1], Klement et al [7, Theorem 9.10])*. For a t-norm T the following statements are equivalent:

i. T is a copula.
ii. T is 1-Lipschitz.

2.2.2 Convex Functions

This section contains known theorems, which describe continuous and convex functions. Properties presented here are needed in the next part of the work for additive generators of t-norms.

Definition 15 *(Kuczma [8, p.130])*. Let $D \subset \mathbb{R}^n, n \in \mathbb{N}$ be a convex and open set. A function $f: D \longrightarrow \mathbb{R}$ is called convex if it satisfies the Jensen's functional inequality $f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2}$ for all $x, y \in D$.

Definition 16 *(Kuczma [8, p.130])*. Let $D \subset \mathbb{R}^n, n \in \mathbb{N}$ be a convex and open set. A function $f: D \longrightarrow \mathbb{R}$ is called concave if it satisfies the following functional inequality $f\left(\frac{x+y}{2}\right) \geq \frac{f(x)+f(y)}{2}$ for all $x, y \in D$.

Theorem 17 *(Kuczma [8, Theorem 7.1.1])*. For a function $f: D \longrightarrow \mathbb{R}$ the following statements are equivalent:

i. f is convex and continuous.
ii. For all $\lambda \in [0,1]$ and all $x, y \in D$ it holds
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{2}$$

The following characterization is true for continuous functions.

Theorem 18 *(Kuczma [8, Theorems 7.3.2 and 7.3.3])*. For a continuous function $f: [0,1] \longrightarrow \mathbb{R}$ the following statements are equivalent:

i. f is convex.
ii. f satisfies the inequality
$$f(y+\varepsilon) - f(y) \leq f(x+\varepsilon) - f(x), \tag{3}$$
for all $x, y \in [0,1]$ such that $y \leq x$ and all $\varepsilon > 0$ such that $x+\varepsilon, y+\varepsilon \in [0,1]$.

It is well-known that a function f is convex, if and only if, $-f$ is concave. Therefore, the analogous theorem can be formulated for concave functions.

Theorem 19. For a continuous function $f: [0,1] \longrightarrow \mathbb{R}$ the following statements are equivalent:

i. f is concave.
ii. f satisfies the inequality
$$f(y+\varepsilon) - f(y) \geq f(x+\varepsilon) - f(x), \tag{4}$$
for all $x, y \in [0,1]$ such that $y \leq x$ and all $\varepsilon > 0$ such that $x+\varepsilon, y+\varepsilon \in [0,1]$.

2.2.3 Fuzzy Implications

In this part we present main definitions connected with fuzzy implications.

Definition 20 *(Fodor and Roubens [4], Baczyński and Jayaram [1]).* A function $I: [0,1]^2 \longrightarrow [0,1]$ is called a fuzzy implication if it satisfies, for all $x,\ x_1,\ x_2, y, y_1, y_2 \in [0,1]$, the following conditions:

- if $x_1 \leq x_2$, then $I(x_1, y) \geq I(x_2, y)$
- if $y_1 \leq y_2$, then $I(x, y_1) \leq I(x, y_2)$
- $I(0,0) = 1$
- $I(1,1) = 1$
- $I(1,0) = 0$

Below, we cite one result that will be useful in the last part of our paper.

Theorem 21 *(Baczyński and Jayaram [1]).* Let $\phi: [0,1] \longrightarrow [0,1]$ be an increasing bijection. If I is a fuzzy implication, then the ϕ-conjugate of I given by formula $I_\phi(x,y) = \phi^{-1}(I(\phi(x), \phi(y)))$ for $x, y \in [0,1]$ is also a fuzzy implication.

Now, we present definitions of some families of fuzzy implications that will appear later.

Definition 22 *(Baczyński and Jayaram [1]).* A function $I: [0,1]^2 \longrightarrow [0,1]$ is called an R-implication if there exists a t-norm T such that

$$I(x,y) = \sup\{t \in [0,1]: T(x,t) \leq y\}, \quad \text{for } x, y \in [0,1] \tag{5}$$

If I is generated from a t-norm T, then it will be denoted by I_T.

Definition 23 *(Baczyński and Jayaram [1]).* A function $I: [0,1]^2 \longrightarrow [0,1]$ is called an (S,N)-implication if there exists a t-conorm S and a fuzzy negation N such that

$$I(x,y) = S(N(x), y), \text{ for } x, y \in [0,1]. \tag{6}$$

3 Special and Inversely Special Implications

As we mentioned before, the notion of directional monotonicity was introduced in 2015 (Bustince et al. [3]). However, earlier, in 1996, it appeared for fuzzy implications in the article by Hájek and Kohout [5], investigated in 2007 by Sainio et al. [11] and also in 2009 by Jayaram and Mesiar [6]. The authors suggested the following notion.

Definition 24 *(Sainio et al. [11]).* A fuzzy implication I is called special if

$$\forall_{\varepsilon>0}\ \forall_{x,y\in[0,1]}\ (x+\varepsilon, y+\varepsilon \in [0,1] \Rightarrow I(x,y) \leq I(x+\varepsilon, y+\varepsilon)). \tag{SP}$$

According to the Definition 1 we can say that special implications are $(1,1)$-increasing functions.

Below, we give some examples of special implications.

Example 25

1. The Łukasiewicz implication given by the formula
   $I_Ł(x,y) = \min\{1, 1-x+y\}$, for $x,y \in [0,1]$ (7)
   is a special implication (see [6]). Note that $I_Ł(x,y) = I_Ł(x+\varepsilon, y+\varepsilon)$ for $\varepsilon > 0$ and $x+\varepsilon, y+\varepsilon \in [0,1]$.
2. The Gödel implication given by the formula
   $I_G(x,y) = \begin{cases} 1, & x \le y \\ y, & x > y \end{cases}$, for $x,y \in [0,1]$,
   is special. Indeed, $I(x,y) = I(x+\varepsilon, y+\varepsilon)$ for $x \le y$ and suitable $\varepsilon > 0$. We also have $I(x,y) = y \le y + \varepsilon = I(x+\varepsilon, y+\varepsilon)$ for $x > y$ and proper $\varepsilon > 0$.

Analogously, we formulate the notion for fuzzy implications which are $(1,1)$-decreasing functions.

Definition 26 A fuzzy implication $I\colon [0,1]^2 \longrightarrow [0,1]$ is called inversely special if

$$\forall_{\varepsilon>0}\ \forall_{x,y\in[0,1]}\ (x+\varepsilon, y+\varepsilon \in [0,1] \Rightarrow I(x,y) \ge I(x+\varepsilon, y+\varepsilon)). \quad \text{(ISP)}$$

Below we show several examples of inversely special implications, which belong to different families of fuzzy implications.

Example 27

1. The Łukasiewicz implication $I_Ł$ is inveresly special (see Example 25).
2. Let S be a t-conorm, N the fuzzy negation given by
   $N(x) = \begin{cases} 0, & x = 1 \\ 1, & x < 1 \end{cases}$.
   Then the (S,N)-implication given by $I(x,y) = S(N(x),y) = \begin{cases} 1, x < 1 \\ y, x = 1 \end{cases}$ for $x,y \in [0,1]$ is inversely special. Indeed, for $x,y < 1$ and $\varepsilon > 0$ such that $x+\varepsilon < 1$ we have $1 = I(x,y) \ge I(x+\varepsilon, y+\varepsilon) = 1$. The condition (ISP) holds also for $x,y < 1$ such that $x+\varepsilon = 1$, since in this case $I(x,y) = 1 \ge y + \varepsilon = I(x+\varepsilon, y+\varepsilon)$. Note that this implication is also the R-implication generated from the drastic product t-norm T_D given by the formula $T_D(x,y) = \begin{cases} 0, & (x,y) \in [0,1)^2 \\ \min\{x,y\}, & \text{otherwise} \end{cases}$ for $x,y \in [0,1]$.
3. It is easy to check that the Rescher implication given by the formula
   $I_{RS}(x,y) = \begin{cases} 1, & x \le y \\ 0, & x > y \end{cases}$, for $x,y \in [0,1]$ is inversely special.
4. Note that the Gödel implication (see Example 25) is not inversely special. Let us take $x = 0.5$, $y = 0.3$ and $\varepsilon = 0.2$, then $I(x,y) = 0.3 < I(x+\varepsilon, y+\varepsilon) = 0.5$.

Lemma 28 Let I be a fuzzy implication. If I is special or inversely special, then it satisfies

- The identity principle i.e., $I(x,x) = 1$ for all $x \in [0,1]$ (IP)
- The left ordering property i.e., $\forall_{x,y\in[0,1]}\ (x \le y \Rightarrow I(x,y) = 1)$

Proof. We show it for inversely special implications (as for special ones it is similar). Let I be an inversely special implication and take $x \in [0,1)$ and $\varepsilon > 0$. Let us fix $\varepsilon = 1 - x > 0$, we have

$$1 \ge I(x,x) \ge I(x+\varepsilon, x+\varepsilon) = I(1,1) = 1$$

Of course $I(1,1) = 1$, hence $I(x,x) = 1$ for $x \in [0,1]$, so I satisfies (IP).

To show the second condition, let us take $x, y \in [0,1]$ such that $x \le y$, then

$$1 \ge I(x,y) \ge I(y,y) = 1$$

because of the monotonicity of I. Therefore, $I(x,y) = 1$

Note that the fuzzy implication I from Example 27 point 2 satisfies the left neutrality property, i.e.,

$I(1,y) = y$, for $y \in [0,1]$ (NP)

However, it does not satisfy the ordering property i.e., the following equality

$\forall_{x,y\in[0,1]}\ (x \le y \Leftrightarrow I(x,y) = 1)$ (OP)

Indeed, $I(x,y) = 1$ for $x = 0.9$ and $y = 0.5$. This makes a difference between fuzzy implications satisfying (ISP) and (SP). If a special implication satisfies (NP) then it satisfies (OP) as well (see [*6, Proposition 2.7*]). Here, as we have seen, it can be opposite.

For all fuzzy implications, the following result is true.

Theorem 29 *(cf. Jayaram and Mesiar [6, Theorem 9.6])*. For an increasing bijection $\phi\colon [0,1] \longrightarrow [0,1]$ the following statements are equivalent:

i. For each inversely special implication *I*, the implication I_ϕ is an inversely special fuzzy implication.
ii. ϕ is convex.

Proof. $(i. \Rightarrow ii.)$ We can take any fuzzy implication which satisfies (ISP), so let us consider the Łukasiewicz implication $I_Ł$. Assume that $(I_Ł)_\phi$ is inversely special for some increasing bijection ϕ. Let us fix arbitrarily $x, y \in [0,1]$ such that $x \ge y$ and take any $\varepsilon > 0$ such that $x + \varepsilon, y + \varepsilon \in [0,1]$. From (ISP) for $(I_Ł)_\phi$ we obtain

$$\begin{aligned}\phi^{-1}\big(1 - \phi(x) + \phi(y)\big) = (I_Ł)_\phi(x,y) &\ge (I_Ł)_\phi(x+\varepsilon, y+\varepsilon)\\ &= \phi^{-1}\big(1 - \phi(x+\varepsilon) + \phi(y+\varepsilon)\big)\end{aligned}$$

thus, by the monotonicity of ϕ^{-1}, we have

$$1 - \phi(x) + \phi(y) \geq 1 - \phi(x + \varepsilon) + \phi(y + \varepsilon)$$

hence

$$\phi(x + \varepsilon) - \phi(x) \geq \phi(y + \varepsilon) - \phi(y)$$

and ϕ is convex in virtue of Theorem 18.

$(ii. \Rightarrow i.)$ Since I is inversely special, then it satisfies the left ordering property. We show that I_ϕ satisfies it too. Let us take $x \in [0,1)$ and define $\varepsilon = 1 - \phi(x) > 0$. From (ISP) for I we obtain

$$1 \geq I\big(\phi(x), \phi(x)\big) \geq I(\phi(x) + \varepsilon, \phi(x) + \varepsilon) = I(1,1) = 1$$

Of course

$$I_\phi(x, x) = \phi^{-1}\Big(I\big(\phi(x), \phi(x)\big)\Big) = \phi^{-1}(1) = 1$$

Thus, $1 = I_\phi(x, x) \leq I_\phi(x, y) \leq 1$ for any $x, y \in [0,1]$ such that $x \leq y$, because of the monotonicity of the fuzzy implication I_ϕ, hence $I_\phi(x, y) = 1$ for $x \leq y$.

Therefore, it remains to show that I_ϕ is inversely special for $x, y \in [0,1]$ such that $x > y$. To do this let us fix arbitrarily $x, y \in [0,1)$ such that $x > y$ (the case when $x = 1$ is not applicable in the definition of (ISP)). We know that

$$I\big(\phi(x), \phi(y)\big) \geq I(\phi(x) + \delta, \phi(y) + \delta)$$

for any $\delta > 0$ such that $\phi(x) + \delta, \phi(y) + \delta \in [0,1]$. Let us take any $\varepsilon > 0$ such that $x + \varepsilon \leq 1$. Bijection ϕ is in particular continuous, so from our assumption on convexity and by Theorem 18 we have $\phi(y + \varepsilon) \geq \phi(y) + \phi(x + \varepsilon) - \phi(x)$. Now, for $\delta = \phi(x + \varepsilon) - \phi(x) > 0$ we have

$$\begin{aligned} I\big(\phi(x), \phi(y)\big) &\geq I(\phi(x + \varepsilon), \phi(y) + \phi(x + \varepsilon) - \phi(x)) \\ &\geq I\big(\phi(x + \varepsilon), \phi(y + \varepsilon)\big) \end{aligned}$$

Therefore $\phi^{-1}\Big(I\big(\phi(x), \phi(y)\big)\Big) \geq \phi^{-1}\Big(I\big(\phi(x + \varepsilon), \phi(y + \varepsilon)\big)\Big)$ and thus I_ϕ is inversely special.

4 Characterizations of Special R-implications

First, we cite characterizations of special implications that are R-implications generated from specific t-norms.

Theorem 30 *(Sainio et al. [11, Proposition 2])*. For a continuous Archimedean t-norm *T* the following statements are equivalent:

i. The R-implication I_T satisfies (SP).

ii. The continuous additive generator of T is a convex function.

Theorem 31 *(Sainio et al. [11, Theorem 2])*. For a continuous t-norm T the following statements are equivalent:

i. The R-implication I_T satisfies (SP).
ii. T is the ordinal sum of the summands $< a_\alpha, e_\alpha, T_\alpha >, \alpha \in A$, where each T_α is generated by a convex additive generator f_α.

In particular, when A is the empty set, then T is the minimum t-norm and I_T is the Gödel implication which is special. As a corollary, they received the following result.

Theorem 32 *(Sainio et al. [11, Corollary 2])*. For a left-continuous t-norm T the following statements are equivalent:

i. The R-implication I_T satisfies (SP).
ii. T is 1-Lipschitz.

As an easy corollary we receive the following fact.

Corollary 33 For a left-continuous t-norm T the following statements are equivalent:

i. The R-implication I_T satisfies (SP).
ii. T is a copula.

5 Characterizations of Inversely Special R-implications

In this section, we present new results for inversely special fuzzy implications which are in some sense equivalents of results from previous section. The following remark says about such R-implications generated from 1-Lipschitz t-norms.

Theorem 34 The Łukasiewicz implication given by (7) is the only one R-implication generated from a 1-Lipschitz t-norm that is inversely special.

Proof. Let us take a 1-Lipschitz t-norm T and consider the R-implication generated from it. For the simplificity let us denote it by I. First notice that for every R-implication we have

$$I(1, y) = \sup\{t \in [0,1] = T(1, t) \leq y\} = y$$

for y$\in$ [0,1]. From Theorem 32 we know that I is special. Let us take $x, y \in [0,1)$ such that $x > y$ and $\varepsilon = 1 - x > 0$. Since I satisfies (SP) we can write

$$I(x, y) \leq I(x + \varepsilon, y + \varepsilon) = I(1,1 - x + y) = 1 - x + y = I_Ł(x, y)$$

Also, I satisfies (ISP). Therefore

$$I(x,y) \geq I(x+\varepsilon, y+\varepsilon) = I(1,1-x+y) = 1-x+y = I_{Ł}(x,y)$$

for $x > y$. Therefore $I(x,y) = I_{Ł}(x,y)$ for all $x,y \in [0,1)$ such that $x > y$. From Lemma 28 we know that $I(x,y) = 1$ for $x \leq y$. Hence $I(x,y) = I_{Ł}(x,y)$ for all $x,y \in [0,1]$.

For some R-implications generated from continuous t-norms, we can formulate a characterization of inversely special implications in the analogous way to special ones (compare the following result with Theorem 30).

Theorem 35 For a continuous Archimedean t-norm T the following statements are equivalent:

i. The R-implication I_T satisfies (ISP).
ii. The continuous additive generator of T is a concave function.

Proof. $(i. \Rightarrow ii.)$ Let T be a continuous Archimedean t-norm and I_T be the R-implication generated from T. Also, let $f\colon [0,1] \longrightarrow [0,1]$ be the additive generator of T, i.e., $T(x,y) = f^{-1}(\min\{f(x)+f(y), f(0)\})$, for $x,y \in [0,1]$. Hence, by Theorem 2.5.21 in [1] we obtain

$$I_T(x,y) = f^{-1}(\max\{f(y)-f(x), 0\}), \text{ for all } x,y \in [0,1]$$

From Theorem 19 it is enough to show the condition (4). Let us fix arbitrarily $x,y \in [0,1]$ such that $x \geq y$. Then $f(x) \leq f(y)$, so $f(y)-f(x) \geq 0$ and hence

$$I_T(x,y) = f^{-1}\big(f(y)-f(x)\big)$$

for such x,y. Since I_T is inversely special, for any $\varepsilon > 0$ such that $x+\varepsilon, y+\varepsilon \in [0,1]$, we receive

$$I_T(x,y) \geq I_T(x+\varepsilon, y+\varepsilon)$$

so

$$f^{-1}\big(f(y)-f(x)\big) \geq f^{-1}\big(f(y+\varepsilon)-f(x+\varepsilon)\big)$$

f^{-1} is also a decreasing function, therefore

$$f(y)-f(x) \leq f(y+\varepsilon)-f(x+\varepsilon)$$

hence

$$f(y+\varepsilon)-f(y) \geq f(x+\varepsilon)-f(x)$$

thus, by Theorem 19, f is a concave function.

$(ii. \Rightarrow i.)$ Let us assume that f is a concave function and by Theorem 19 we have

$$f(y+\varepsilon)-f(y) \geq f(x+\varepsilon)-f(x)$$

for $x,y \in [0,1]$, $x \geq y$ and applicable $\varepsilon > 0$. Hence

$$f^{-1}\big(f(y)-f(x)\big) \geq f^{-1}\big(f(y+\varepsilon)-f(x+\varepsilon)\big)$$

thus

$$I_T(x,y) \geq I_T(x+\varepsilon, y+\varepsilon)$$

We know that $I_T(x,y)=1$ for $x<y$ and therefore I_T is inversely special.

Now we consider continuous t-norms (compare the following result with Theorem 31)

Theorem 36 For a continuous t-norm T the following statements are equivalent:

i. The R-implication I_T satisfies (ISP).
ii. T is continuous Archimedean with a concave generator.

Proof. $(i. \Rightarrow ii.)$ Let us take a continuous t-norm T and consider the R-implication I_T generated from this T. From Theorem 11 we know that T can be represented as an ordinal sum of continuous Archimedean t-norms. Then I_T is given by the following formula (see [1, Theorem 2.5.24]):

$$I_T(x,y)=\begin{cases} 1, & x \leq y \\ a_\alpha + (e_\alpha - a_\alpha)\cdot I_{T_\alpha}\left(\dfrac{x-a_\alpha}{e_\alpha-a_\alpha}, \dfrac{y-a_\alpha}{e_\alpha-a_\alpha}\right), & (x,y)\in[a_\alpha,e_\alpha]^2 \\ y, & \text{otherwise.} \end{cases}$$

Let us consider three cases with respect to the index set A.

1. If $A=\emptyset$, then $I_T=I_G$ (see Example 25 point 2). However, we have shown that I_G is special but not inversely special.
2. If $\bar{\bar{A}}=1$ and $a_\alpha=0, e_\alpha=1$, then $T=<0,1,T>$ and T is a continuous Archimedean t-norm. In this case, in virtue of Theorem 35, I_T is inversely special if and only if T has a concave generator.
3. In all other situations we consider two possibilities.
   a. There exists $\alpha 0 \in A$ such that $a_{\alpha 0}=0$ and $e_{\alpha 0}<1$. Let us take $x\in(e_{\alpha 0},1)$, $y\in(a_{\alpha 0},e_{\alpha 0})$, then there exists $\varepsilon>0$ such that $x+\varepsilon\in(e_{\alpha 0},1)$, $y+\varepsilon\in(a_{\alpha 0},e_{\alpha 0})$. Then $I_T(x,y)=y<y+\varepsilon=I_T(x+\varepsilon,y+\varepsilon)$, so I_T does not satisfy (ISP) in this case.
   b. $a_\alpha>0$, for all $\alpha\in A$. Let $a_{\alpha 0}=\min\{a_\alpha:\alpha\in A\}$ and $e_{\alpha 0}=\min\{e_\alpha:\alpha\in A\}$. Consider $x\in(a_{\alpha 0},e_{\alpha 0})$ such that $\frac{x}{2}<a_{\alpha 0}$. Then there exists $\varepsilon>0$ such that $x+\varepsilon\in(a_{\alpha 0},e_{\alpha 0})$ and $\frac{x+\varepsilon}{2}<a_{\alpha 0}$. Then $I_T\left(x,\frac{x}{2}\right)=\frac{x}{2}<\frac{x}{2}+\frac{\varepsilon}{2}=I_T(x+\varepsilon,\frac{x+\varepsilon}{2})$.

As we have shown, if I_T is represented by a proper ordinal sum (case 3) it is not inversely special. Therefore I_T is inversely special if and only if T is continuous Archimedean with a concave generator.

$(ii. \Rightarrow i.)$ This follows from Theorem 35.

Based on the above results, we can formulate the following fact.

Theorem 37 For a continuous t-norm T the following statements are equivalent:

i. The R-implication I_T satisfies (ISP).
ii. The R-implication I_T is ϕ-conjugate with the Łukasiewicz implication, where ϕ is convex.

Proof. $(i. \Rightarrow ii.)$ Let T be a continuous t-norm. From Theorem 36 we know that if the R-implication generated from a continuous t-norm T satisfies (ISP), then T is also Archimedean. Among all such t-norms there are only two classes – nilpotent and strict t-norms (see [7, Theorem 2.18]). Let f be the additive generator of T. If T is nilpotent, then $f(0) < \infty$ and if T is strict, then $f(0) = \infty$ (see [7, Proposition 3.29]). We also know from Theorem 36 that f is concave, so by Theorem 18 the condition (4), i.e., the inequality

$$f(y + \varepsilon) - f(y) \geq f(x + \varepsilon) - f(x)$$

is true for $x, y \in [0,1]$ such that $y \leq x$ and $\varepsilon > 0$ such that $x + \varepsilon, y + \varepsilon \in [0,1]$. Let $y = 0$ and take any $x \in (0,1)$ and $\varepsilon \in (0,1)$ such that $x + \varepsilon \in (0,1)$. From the above inequality we obtain that

$$f(0) \leq f(x) - f(x + \varepsilon) - f(\varepsilon)$$

Therefore, $f(0) < \infty$ (because $f(x), f(x + \varepsilon), f(\varepsilon) < \infty$) and T must be nilpotent. Hence I_T is ϕ-conjugate with the Łukasiewicz implication (see [1, Lemma 2.5.23]). Moreover, we can define ϕ in the following way $\phi(x) = 1 - \frac{f(x)}{f(0)}$ for $x \in [0,1]$. Also if f is concave, then of course ϕ is convex.

$(ii. \Rightarrow i.)$ For any increasing bijection ϕ, by Proposition 2.5.10 in [1], the function $(I_Ł)_\phi$ is a continuous R-implication generated from the continuous t-norm ϕ-conjugate with the Łukasiewicz t-norm, i.e., $(T_Ł)_\phi(x, y) = \phi^{-1}(\max\{\phi(x) + \phi(y) - 1,0\})$, $x, y \in [0,1]$. From Theorem 29 we know that $(I_Ł)_\phi$ satisfies (ISP), if ϕ is convex.

Now, we present a proposition which contains a little more general characterization of directional decreasing R-implications generated from continuous t-norms.

Proposition 38 Let $\varepsilon, \varepsilon_1, \varepsilon_2 > 0, \varepsilon_2 \leq \varepsilon \leq \varepsilon_1$, T be a continuous t-norm and I_T be the R-implication generated from T. Then the following statements are equivalent:

i. I_T is an inversely special implication.
ii. I_T is $(\varepsilon_1, \varepsilon)$-decreasing.
iii. I_T is $(\varepsilon, \varepsilon_2)$-decreasing.

Proof. $(i. \Rightarrow ii.)$ If an R-implication I_T is inversely special, then t-norm T is continuous Archimedean with an additive generator f. Therefore for $x, y \in [0,1]$, $x \geq y$ and applicable $\varepsilon > 0$ we can write the following inequality

$$f^{-1}\big(f(y)-f(x)\big) \geq f^{-1}\big(f(y+\varepsilon)-f(x+\varepsilon)\big)$$

thus for proper $\varepsilon_1 > 0$ we have

$$f(y)-f(x) \leq f(y+\varepsilon)-f(x+\varepsilon) \leq f(y+\varepsilon)-f(x+\varepsilon_1)$$

because f is strictly decreasing. Further, $f^{-1}\big(f(y)-f(x)\big) \geq f^{-1}\big(f(y+\varepsilon)-f(x+\varepsilon_1)\big)$, what means $I_T(x,y) \geq I_T(x+\varepsilon_1, y+\varepsilon)$ in this case. Moreover, we have

$$1 = I_T(x,y) \geq I_T(x+\varepsilon_1, y+\varepsilon)$$

for any $x < y$ and applicable $\varepsilon > 0$. Proofs $(ii. \Rightarrow i.), (iii. \Rightarrow i.)$ are obvious and $(i. \Rightarrow iii.)$ is analogous to the above one.

The similar result can be formulated for special implications and the proof is similar to the above one.

Proposition 39 Let $\varepsilon, \varepsilon_1, \varepsilon_2 > 0, \varepsilon_1 \leq \varepsilon \leq \varepsilon_2$, T be a continuous t-norm and I_T be the R-implication generated from T. Then the following statements are equivalent:

i. I_T is a special implication.
ii. I_T is $(\varepsilon_1, \varepsilon)$-increasing.
iii. I_T is $(\varepsilon, \varepsilon_2)$-increasing.

6 Other Classes of Inversely Special Implications

In this part we consider different families of fuzzy implications, i.e., (S,N)-implications, f-implications and g-implications. We will use the following theorems to characterize inversely special (S,N)-implications.

Theorem 40 *(Baczyński and Jayaram [1, Theorem 2.4.17])*. For a t-conorm S and a fuzzy negation N the following statements are equivalent:

i. $I_{S,N}$ is a continuous (S,N)-implication that satisfies (IP).
ii. S is a nilpotent t-conorm and $N \geq N_S$, where N_S is the natural negation of S (see Definition 2.3.1 in [1]).

Theorem 41 *(Baczyński and Jayaram [1, Theorem 2.4.20])*. For a function $I\colon [0,1]^2 \to [0,1]$ the following statements are equivalent.

i. I is an (S,N)-implication obtained from a nilpotent t-conorm S and its natural negation N_S.
ii. I is ϕ-conjugate with the Łukasiewicz implication.

Thanks to these results, we can formulate the following corollary.

Corollary 42 For a function $I\colon [0,1]^2 \to [0,1]$ the following statements are equivalent.

i. I is an inversely special (S,N)-implication obtained from a continuous t-conorm S and its natural negation N_S.
ii. I is ϕ-conjugate with the Łukasiewicz implication, where ϕ is convex.

Proof. $(i. \Rightarrow ii.)$ Let I be an inversely special (S,N)-implication obtained from a continuous t-conorm S and its natural negation N_S. If I satisfies (ISP), then by Lemma 28 it satisfies (IP). I is in particular continuous, so in virtue of Theorem 40 the t-conorm S is nilpotent. Now, from Theorem 41 and Theorem 29 we obtain the thesis.

$(ii. \Rightarrow i.)$ By Theorem 2.4.5 in [1] the function $(I_Ł)_\phi$ is the (S,N)-implication obtained from the ϕ-conjugate Łukasiewicz t-conorm $S_Ł$ (which is in particular continuous) and its natural negation. From Theorem 29 we know that $(I_Ł)_\phi$ satisfies (ISP), if ϕ is convex.

A fuzzy implication ϕ-conjugate with an R-implication generated from any t-norm is also an R-implication (see [1, Proposition 2.5.10]). Since all implications ϕ-conjugate with the Łukasiewicz implication are R-implications, (S,N)-implications satisfying (ISP) and generated from a continuous t-conorm and the natural negation of this t-conorm are a subclass of inversely special R-implications.

Now let us consider f-implications and g-implications. We will see there are no inversely special implications among them.

First, let us recall some definitions and their properties.

Definition 43 *(Baczyński and Jayaram [1, Definition 3.1.1])*. Let $f\colon [0,1] \to [0,1]$ be a strictly decreasing and continuous function with $f(1) = 0$. The function $I\colon [0,1]^2 \to [0,1]$ defined by

$$I(x,y) = f^{-1}\left(x \cdot f(y)\right)$$

for $x, y \in [0,1]$ with understanding $0 \cdot \infty = 0$, is called an f-generated implication. The function f itself is called an f-generator of the I. In such case, to emphasize the apparent relation, we will write I_f.

Theorem 44 *(Baczyński and Jayaram [1, Theorem 3.1.7])*. If f is an f-generator, then I_f does not satisfy (IP).

From the above theorem and Lemma 28 it is clear that all f-implications are not inversely special.

Corollary 45 There is no f-implication satisfying (ISP).

Definition 46 *(Baczyński and Jayaram [1, Definition 3.2.1]).* Let $g\colon [0,1] \longrightarrow [0,1]$ be a strictly increasing and continuous function with $g(0) = 0$. The function $I\colon [0,1]^2 \longrightarrow [0,1]$ defined by

$$I(x,y) = g^{-1}\left(\min\left\{\frac{1}{x}\cdot g(y), g(1)\right\}\right)$$

for $x, y \in [0,1]$, with the understanding $\frac{1}{0} = \infty$ and $\infty \cdot 0 = \infty$, is called a g-generated implication. The function g itself is called a g-generator of the I and we will write I_g instead of I.

Theorem 47 *(Baczyński and Jayaram [1, Theorem 3.2.9]).* If g is a g-generator, then the following statements are equivalent:

i. I_g satisfies (OP).
ii. I_g is a Goguen implication.

Now we can prove the following fact.

Proposition 48 There is no g-implication satisfying (ISP).

Proof. Let us suppose that there is a g-implication I_g which is inversely special. Then it satisfies the left ordering property. From Theorem 47 we know that if g-implication satisfies (OP), then it is the Goguen implication, which is not inversely special. That means that $I_g$does not satisfy the following condition:

$I_g(x,y) = 1 \Rightarrow x \leq y$ for $x, y \in [0,1]$.

Therefore, there exist $x, y \in [0,1]$ such that $I_g(x,y) = 1$ and $x > y$. Observe that $x < 1$ and $y > 0$, since $I_g(1,y) = y$ and $I_g(x,0)$=0 if $x > 0$. Thus there exists $\varepsilon > 0$ such that $x' = x - \varepsilon > 0,\ y' = y - \varepsilon \geq 0$ and $x' > y'$. We assumed that I_g satisfies (ISP) and hence $I_g(x',y') \geq I_g(x'+\varepsilon, y'+\varepsilon) = I_g(x,y) = 1$. Furthermore, we can take $\varepsilon = y$ and then $x' = x - y,\ y' = 0$ and we get

$$0 = I_g(x-y,0) = I_g(x',y') \geq I_g(x,y) = 1,$$

a contradiction. Therefore I_g cannot satisfy (ISP).

Conclusions

In this paper we have investigated special and inversely special implications, as directional monotonics, and we have provided some examples of them. We have characterized all inversely special R-implications generated from continuous t-norms. Also, we have considered other families of fuzzy implications. Our conclusion is that there are no inversely special implications other than R-implications in the set. Finally, we have shown some generalizations of inversely special implications as directional monotonic functions.

Acknowledgement

The author would like to thank Professor Michał Baczyński for his valuable suggestions and comments.

References

[1] M. Baczyński, B. Jayaram: Fuzzy implications, Studies in Fuzziness and Soft Computing, Vol. 231, Springer, Heidelberg, 2008

[2] G. Beliakov, A. Pradera, T. Calvo: Aggregation Functions: A Guide for Practitioners, Studies in Fuzziness and Soft Computing, Vol. 221, Springer, Heidelberg, 2007

[3] H. Bustince, J. Fernandez, A. Kolesárová, R. Mesiar: Directional monotonicity of fusion functions, European Journal of Operational Research 244, 2015, pp. 300-308

[4] J. Fodor, M. Roubens: Fuzzy Preference Modelling and Multicriteria Decision Support, Kluwer, Dordrecht, 1994

[5] P. Hájek, L. Kohout: Fuzzy Implications and Generalized Quantifiers, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems 4, 1996, pp. 225-233

[6] B. Jayaram, R. Mesiar: On Special Fuzzy Implications, Fuzzy Sets and Systems 160, 2009, pp. 2063-2085

[7] E. P. Klement, R. Mesiar, E. Pap: Triangular Norms, Kluwer, Dordrecht, 2000

[8] M. Kuczma: An Introduction to the Theory of Functional Equations and Inequalities, PWN-Polish Scientific Publishers & Silesian University, Warszawa, Kraków, Katowice, 1985

[9] C. H. Ling: Representation of Associative Functions, Publicationes Mathematicae Debrecen 12, 1965, pp. 189-212

[10] R. Moynihan: On τ_T Semigroups of Probability Distribution Functions II, Aequationes Mathematicae 17, 1978, pp. 19-40

[11] E. Sainio, E. Turunen, R. Mesiar: A Characterization of Fuzzy Implications Generated by Generalized Quantifiers, Fuzzy Sets and Systems 159, 2008, pp. 491-499

[12] T. Wilkin, G. Beliakov: Weakly Monotonic Averaging Functions. International Journal of Intelligent Systems 30, 2015, pp. 144-169

A Nutrition Adviser's Menu Planning for a Client Using a Linear Optimization Model

Lucie Schaynová

University of Ostrava, Faculty of Science, Department of Mathematics
30. dubna 22, 70103 Ostrava, Czech Republic
lucie.schaynova@osu.cz

Abstract: This paper presents a new linear optimization model which improves a nutritional adviser's work and prevents mistakes when preparing a diet plan for a client manually. The model takes the client's favourite or the adviser's recommended recipes into account, prevents unbalanced nutrition, respects the client's eating habits and habits of measuring when cooking, ensures recommendations for people from the Czech Republic and prevents wasting food items. The model also ensures that the client's daily recommended intake of nutrients is met, that certain nutrients are balanced in proportion when applicable, and that the energy intake is distributed during the whole day. The model involves linear constraints to ensure that two incompatible recipes are not used in the same meal and that a recipe is not used in an incompatible meal. A corresponding balanced feeding plan is produced for the client for several days. The solution will yield particular recipes for particular days with the exact amounts of the food items used. The final dietary plan for the client is optimal.

Keywords: linear programming; diet problem; nutrient requirement; menu planning; nutrition adviser

1 Introduction

The question of feeding people is a fundamental question for the entire planet: an estimated two-thirds of the world's population suffers from various degrees of nutrition deficiency (malnutrition[1]). This nutrition deficiency is caused by starvation, quantitative and qualitative insufficient nutrition as well as faulty and unbalanced nutrition. It is also linked to bad habits, such as overeating. People

[1] Malnutrition is a bad nutrition state of a client. It is caused by insufficient intake of basic nutrients (proteins, saccharides and fats) as well as vitamins, minerals and trace elements.

with gastrointestinal tract defects, absorption defects or digestive disorders, stress, alcoholism, smoking habits, etc. can also suffer from malnutrition. It is worth noting that sufficient food intake does not automatically mean sufficient intake of necessary nutritional factors, see [11] and [27].

According to Kleinwächterová and Brázdová [11], some causes of bad nutrition include: genetic and metabolic issues, exogenous factors linked to socio-economical status, nutrition (structure of nutrition, frequency of food intake, knowledge about nutrition, childhood food habits) and sports activities. According to recent research, biological leanness is related to hereditary factors, but fatness is not.

Kohout [12] and Urbánek [27] state that malnutrition can inhibit blood transportation, deteriorate muscles (the heart muscle reacts to malnutrition by weakening the active muscle matter of the myocardium), and causes shortness of breath, worsens gastrointestinal tract motility, decreases immunity, inhibits recovery, increases vulnerability to infectious complications, decreases the effectiveness of drugs, etc.

According to Rážová [20], the nourishment of the population in the Czech Republic has the following particular characteristics: unsuitable choice of food items (frequency, amounts, variety), high energetic intake, high intake of animal products (fats and proteins), bad ratio of nutrients in favour of saccharides (fibre), high intake of salt (smoked meat products), low intake of vegetable and fruit, bad water intake, etc.

This kind of bad nutrition causes various diseases, such as heart and vascular diseases, diabetes, intestinal cancers, etc., which often concern people living in economically developed countries and are characterized by overeating, sedentary lifestyle and stress. This is the reason why these diseases are called Civilizational. These diseases were uncommon in the past, see [11].

A regime of balanced nutrition combined with balanced energetic intake and necessary amounts of vitamins and minerals is generally accepted as having a protective effect. It constitutes the base for good health, quality of life as well as aiding in the prevention and treatment of many illnesses, see [12], [18], [20].

Nutrition from the perspective of linear programming is always about fulfilling all nutritional requirements of a larger group of people or the population from developing or industrial countries. The objective functions of the linear programming models are as follows: minimize climate impact through greenhouse gas emissions [4], [15]; minimize the difference between the optimal and current diet [4], [6], [16], [17]; minimize the cooking and preparation time of food [14]; and also minimize the cost [1], [3], [4], [7], [23]. The outcomes of the papers are recommendations or certain types of scenarios for people or government. To the author's best knowledge, there are no papers concerning the needs of an individual; not every person can follow the recommendations for the general

population. Only specialists certified in healthcare nutrition can treat individuals, but these specialists lack efficient tools to prepare individually-orientated diets.

The model in this article reflects the methods adopted by a nutrition adviser and improves their processes as well as optimizes their effectiveness. It paves the way for modern nutritional consultancy. The model prevents some mistakes that the advisor could make when creating a feeding plan. The model takes the national recommendations for people in the Czech Republic into account. The model uses complete recipes including techniques of their preparation (the advisor prefers boiling and stewing). In our previous article [22], we worked with food items without technological processing only. The model prevents food wastage, takes into account the system of measurement of the client (pinch, teaspoon, spoon, cup, etc.), and it creates a more balanced eating plan.

The nutritional adviser always uses software, but the feeding plan is composed manually. The adviser has to choose the food items, follow the client's preferences (which food items the client does not like or cannot eat), follow the recommendations, etc. The new feeding plan must also be reasonable and has to be acceptable for the client. Licenced programs usually work with about 14 nutrients. It takes more time if the adviser works with more nutrients. That is why the adviser does not work with the majority of nutrients and the creating of the feeding plan is based on the adviser's experience and practice. Further details can be found in [22].

The below constructed model will greatly help the adviser. The adviser will be sure that the client obtains the best feeding plan, all the needs of the client are satisfied, the plan does not harm the client and the recipes are meaningful.

2 The Problem

The nutrition adviser offers individual consultations to two types of the clients: clients whose physician recommended them to visit the adviser (have some malnutrition, high blood pressure or have some diseases that can be affected by proper nutrition), and clients who are simply interested in a healthy style (want to fix some nutritional details, lose weight, need support in doing sport). In both cases, it is important to work with the client's physician.

The task of the adviser is to analyse the client's consumed food items and beverages, measure the client's body (weight, fat, etc.), to determine the individual nutritional values, and to compute the feeding plan. Then the adviser presents the plan to the client and compares it with the client's current eating habits, see [19] and [20].

The adviser's methodology of examination consists of two parts.

1) Diagnostic part – anamnesis

The client informs the adviser about the client's personal data, job (sedentary job, working environment, possibilities and style of feeding), intolerance, allergy, eating habits, smoking, alcohol, past and present illnesses (hypertension, diabetes mellitus, liver disease, etc.). Some physical and biochemical examination (cachexy, swelling, power of muscles, total cholesterol, LDL and HDL cholesterol, glycemia, etc.) are available from the physician's data. The adviser is also interested in the client's family anamnesis – if there are any genetical risks, for example high blood pressure, familial hypercholesterolemia, diabetes mellitus, heart attack before the age of 60, tumours and everything that should be taken into account when creating the client's diet. See [12] and [29] for further details.

2) Analytical part

This part includes the measurement of height, weight, BMI, circumference of limbs, hipline, waistline, measurement of subcutaneous fat, visceral fat, the amount of the muscle mass, the amount of minerals in the client's bones, blood pressure and the resting heart rate. Then the adviser evaluates the client's body composition and takes into account the measurements and other factors, such as psychological or social, when calculating the nutrient requirements.

Next, the client has to prepare a list of all food items consumed during at least one week before the meeting, including the amounts of the items, technology of preparation, time of eating and physical activities. For further details, see [11] and [12].

Then the adviser determines the *ideal body weight* as follows [12]

$$\omega_m = 0.655\, h_m - 44.1\,, \qquad (1)$$
$$\omega_f = 0.593\, h_f - 38.8\,,$$

where ω_m or ω_f is the ideal weight of a man or a woman, respectively, in kilograms and h_m or h_f is the height of the man or the woman, respectively, in centimetres.

The adviser recognizes the *basal energy expenditure* and the *total daily energy*.

The basal energy expenditure is important to support all functions of the body. We can determine the energy by using the *indirect calorimetry*. This technique uses the measurement of oxygen consumption and carbon dioxide expenditure when the client is breathing over a period of time. The equipment to perform the measurements is uncommon, see [27].

That is why the basal energy expenditure is determined by using the *Harris-Benedikt equation*. The equation was established experimentally by indirect calorimetry measurement of many people. The corresponding equations are as follows

$$\beta_m = 4.184(66.473 + 13.751\omega_m + 5.003h_m - 6.755\alpha_m)\,, \qquad (2)$$
$$\beta_f = 4.184\big(655.095 + 9.563\omega_f + 1.849h_f - 4.675\alpha_f\big)\,,$$

where β_m or β_f is the basal energy of a man or a woman, respectively, in kilojoules (kJ) per day and α_m or α_f is the age of the man or the woman, respectively, in years. See [13] for different experimental calculations of basal energy.

When the adviser treats an obese client, the adviser has to use the *adjusted body weight* [27] instead of the ideal body weight in the Harris-Benedikt equation. This is due to the big difference between the current body weight and the ideal body weight, therefore the following is used

$$\omega' = 0.25\psi + \omega\,,$$

where ω' is the adjusted body weight, ψ is the real body weight and ω is the ideal body weight.

Apart from the basal energy expenditure, the *additional energy* corresponds to the demands made on the functioning of body activities including physical and psychological activity. According to [26], we can add it as follows.

We need to calculate the *factor of physical activity.* The calculation is generated from the list of the client's physical activities. It is calculated as the weighted average of relative times of activities performed by the client during a day; each activity has a specific weight (sleeping 0.95, resting 1.0, very easy work 1.5, hard work 7.0, see [12]). The relative time is the time (in hours) spent by the client to perform an activity divided by 24 hours. The weighted average is calculated for each day of the week and finally the average for the whole week is calculated. This one-week average is the factor of activity, denoted as ρ.

Then the total daily energy requirement can be calculated by using a device for monitoring the heart rate, or using the equation

$$\tau = \beta\rho + \delta, \qquad (3)$$

where τ is the total daily energy requirement in kJ, β is the basal energy expenditure, ρ is the factor of activity and δ is the postprandial thermogenesis. (The postprandial thermogenesis of a healthy client is $\delta = 919$ kJ, see [2]).

According to [11], energy is taken from macronutrients, such as proteins, fats and saccharides. Micronutrients include vitamins and minerals. There are two classes of vitamins: fat-soluble (A, D, E, K) and water-soluble (the others).

Provazník [19] states that each nutrient is of a particular importance. For example, sodium is responsible for osmotic pressure balance; cholesterol is a building nutrient of bile acids and steroid hormones; magnesium is important to construct the bones and to decrease the nervous muscle tension. Fat-soluble vitamins are not excreted by urine, so the client can be overdosed. Every nutrient is needed in a certain amount.

We adjust the total amount of energy according to the higher heating value. The physical higher heating value is the amount of energy which is lost by completely burning one gram of a nutrient in a calorimetric bomb. One gram of saccharides yields 17 kJ of energy, one gram of proteins yields 23 kJ and lipids around 38 kJ. The values are distinct from the physiological higher heating values, which are the amounts of energy the body can utilize. In the case of saccharides and lipids, the values are almost the same, but in the case of proteins the physiological value is 16.7 kJ. The nutrients should be composed so that the 15%, 30%, and 55% of the total daily energy intake comes from proteins, fats, and saccharides, respectively, see [26].

3 Mathematical Model

The aim is to design the diet plan for some period of time, so let us consider $D = 7$ days (Monday, Tuesday, Wednesday, etc.) denoted by $d = 1, \dots, D$. There will be some meals during each of the days. We will work with $K = 5$ meals (breakfast, first snack, lunch, second snack, dinner) per day denoted by $k = 1, \dots, K$. So we will have 35 meals during the week in total. Every meal will be cooked according to some recipes, so let us consider recipes $r = 1, \dots, R$. A recipe is a set of instructions and food items that describes how to prepare a meal. So let us consider food items $j = 1, \dots, n$ (such as chicken, potatoes, cheese, milk, etc.), including drink items (such as tea, mineral water, juice, etc.).

The recipe r uses food items $S_r \subset \{1, \dots, n\}$, where $S_r \neq \emptyset$ and $\bigcup_{r=1}^{R} S_r = \{1, \dots, n\}$. Some of the sets S_r can be singletons. The recipes can be composed individually. That depends on the client's habits and the client's or the adviser's preferences. For example, if the client is a vegan, we can use recipes just for vegans from a recipe book.

Each food item consists of some nutrients, so let us consider nutrients $i = 1, \dots, m$ (such as fats, saccharides, proteins, ect.). Consider a real non-negative matrix $\boldsymbol{A} = (a_{ij})$ where a_{ij} means the quantity of nutrient i in one unit of the food item j for all $i = 1, \dots, m$ and for all $j = 1, \dots, n$. The aim is to satisfy the recommended daily intakes of nutrients, which should be between some upper and lower bound. Denote the minimal and maximal recommended daily intakes of all nutrients by a non-negative vector $\boldsymbol{b}^{\min} = (b_i^{\min})$ and a non-negative vector $\boldsymbol{b}^{\max} = (b_i^{\max})$, respectively, with $i = 1, \dots, m$.

If the client suffers from some nutrition malfunction or is in danger of certain illnesses, the vectors $\boldsymbol{b}^{\min}$ and $\boldsymbol{b}^{\max}$ have to be modified, i.e. the values of the recommended daily intakes of certain nutrients have to be increased, decreased or have to be equal to zero.

Let us have a binary matrix $\boldsymbol{C}^{RR} = (c_{r_1 r_2}^{RR})$ for all $r_1, r_2 = 1, \dots, R$ which mean compatibility between recipes ($c_{r_1 r_2}^{RR} = 1$ if recipes r_1 and r_2 are compatible, i.e. can be used in the same meal, and $c_{r_1 r_2}^{RR} = 0$ otherwise) and binary matrix $\boldsymbol{C}^{KR} = (c_{kr}^{KR})$ for all $r = 1, \dots, R$ and for all $k = 1, \dots, K$, which means compatibility between meal k and recipe r ($c_{kr}^{KR} = 1$ if meal k and recipe r are compatible, i.e., meal prepared according to the recipe r can be served in the meal k, and $c_{kr}^{KR} = 0$ otherwise). Clearly, the matrix $\boldsymbol{C}^{RR}$ will be symmetric and with ones on its diagonal.

Let us have real non-negative matrices $\boldsymbol{M}^{\min} = (m_{rj}^{\min})$ and $\boldsymbol{M}^{\max} = (m_{rj}^{\max})$ for all $r = 1, \dots, R$ and for all $j = 1, \dots, n$, which means the minimal and maximal quantity of the food item j in the recipe r. The elements will be positive, $0 < m_{rj}^{\min} \leq m_{jr}^{\max}$, if $j \in S_r$, and zero, $m_{rj}^{\min} = m_{rj}^{\max} = 0$, if $j \notin S_r$.

Now we can proceed with the formulation of the mathematical model. Let z_{dkr} be a binary variable which means if the recipe r is used in the meal k of the day d ($z_{dkr} = 1$) or not ($z_{dkr} = 0$). Two incompatible recipes cannot be used in the same meal. We can express that by the following inequalities

$$z_{dkr_1} + z_{dkr_2} \leq 1\,, \tag{4}$$

for all $d = 1, \dots, D$, for all $k = 1, \dots, K$ and for all $r_1, r_2 = 1, \dots, R$ such that $c_{r_1 r_2}^{RR} = 0$.

We also do not want to use the recipe r if it is not compatible with the meal k, so we use the condition

$$z_{dkr} = 0\,, \tag{5}$$

for all $d = 1, \dots, D$, for all $k = 1, \dots, K$ and for all $r = 1, \dots, R$ such that $c_{kr}^{KR} = 0$.

We introduce the real non-negative variabes x_{dkrj}. The vaule x_{dkrj} means the amount of the food item j used in the meal k and the recipe r in day d. We need to satisfy the client's minimal and maximal daily recommended intake as follows

$$\sum_{k=1}^{K} \sum_{r=1}^{R} \sum_{j \in S_r} a_{ij} x_{dkrj} \geq b_i^{\min}, \tag{6}$$

$$\sum_{k=1}^{K} \sum_{r=1}^{R} \sum_{j \in S_r} a_{ij} x_{dkrj} \leq b_i^{\max}, \tag{7}$$

for all $i = 1, \dots, m$ and for all $d = 1, \dots, D$. Inequalities (6) and (7) are typical constraints of the classical Diet problem.

We want to use reasonable amounts of food items in the recipes so we add inequalities

$$x_{dkrj} \geq m_{rj}^{\min} z_{dkr} , \tag{8}$$

$$x_{dkrj} \leq m_{rj}^{\max} z_{dkr} , \tag{9}$$

for all $d = 1, \dots, D$, for all $k = 1, \dots, K$, for all $r = 1, \dots, R$ and for all $j \in S_r$.

There can be some nutrients which should be balanced in certain proportions. For example, according to [28], the ratio of the essential amino acids n-6:n-3 should be in the ratio 5:1. The proportion of the plant and animal proteins should be in the ratio 1:1, see [19]. Denote the set $I_1 = \{i_{11}, i_{12}, \dots, i_{1\mu_1}\}$ of nutrients which should be in the ratio $\zeta_{11}:\zeta_{12}:\ \dots\ :\zeta_{1\mu_1}$, set $I_2 = \{i_{21}, i_{22}, \dots, i_{2\mu_2}\}$ of nutrients which should be in the ratio $\zeta_{21}:\zeta_{22}:\ \dots\ :\zeta_{2\mu_2}$, etc., and set $I_\nu = \{i_{\nu 1}, i_{\nu 2}, \dots, i_{\nu\mu_\nu}\}$ of nutrients which should be in the ratio $\zeta_{\nu 1}:\zeta_{\nu 2}:\dots:\zeta_{\nu\mu_\nu}$. The nutrients can be in the ratios with some tolerances. Let $\varepsilon_{\iota\kappa}$ be the tolerance of the coefficient $\zeta_{\iota\kappa}$ for $\iota = 1, \dots, \nu$ and $\kappa = 1, \dots, \mu_\iota$. We assume that $0 < \varepsilon_{\iota\kappa} < \zeta_{\iota\kappa}$.

We can express that as follows

$$\begin{aligned}(\zeta_{\iota\kappa} - \varepsilon_{\iota\kappa}) \sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{j\in S_r} a_{i_{\iota\lambda}j} x_{dkrj} &\leq (\zeta_{\iota\lambda} + \varepsilon_{\iota\lambda}) \sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{j\in S_r} \\ (\zeta_{\iota\kappa} + \varepsilon_{\iota\kappa}) \sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{j\in S_r} a_{i_{\iota\lambda}j} x_{dkrj} &\geq (\zeta_{\iota\lambda} - \varepsilon_{\iota\lambda}) \sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{j\in S_r}\end{aligned} \tag{10}$$

(11)

for all $\iota = 1, \dots, \nu$, for all $\kappa = 1, \dots, \mu_\iota - 1$, for all $\lambda = \kappa + 1, \dots, \mu_\iota$ and for all $d = 1, \dots, D$.

In some situations, the variable x_{dkrj} should attain discrete values. For example, if $j = j_0$ is an egg of medium size, then the variables x_{dkrj_0} should be integer (the number of eggs). Or the client may use a system of measurement involving pinch (0.5 grams), cups (Figure 1 and Figure 2) with discrete cup system or spoons (see Figure 3) with discrete system of measurement. Then the variable x_{dkrj} should also be discrete. So denote $J \subseteq \{1, \dots, n\}$ the set of food items such that the corresponding variables x_{dkrj} should be integer. Then $x_{dkrj} \in \mathbb{Z}$ for all $j \in J$, for all $d = 1, \dots, D$, for all $k = 1, \dots, K$ and for all $r = 1, \dots, R$.

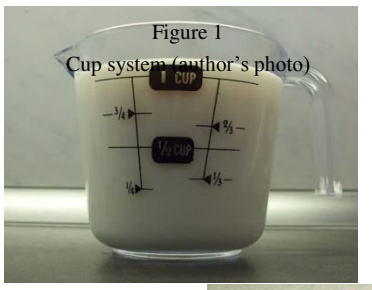

Figure 1
Cup system (author's photo)

Figure 3
Spoon system (author's photo)

We would like to avoid the situation of eating too much or too little in some meals. Denote the minimal and maximal energy intake as $b_1^{\min}$ and $b_1^{\max}$. Inspired by [21], we can naturally distribute the energy intake during the whole day among the meals, for example 20% of the total daily energy for breakfast, 12.5% for the first snack, 30% for lunch, 12.5% for the second snack and 25% for dinner. This depends on the feeding plan which the adviser is preparing. The desired energy intake distribution during the day is given by the non-negative vector $\boldsymbol{v}=(v_k)$ with $\sum_{k=1}^{K} v_k = 1$. Then we want to satisfy the inequalities

$$\sum_{r=1}^{R} \sum_{j \in S_r} a_{1j} x_{dkrj} \geq v_k \, b_1^{\min}, \tag{12}$$

$$\sum_{r=1}^{R} \sum_{j \in S_r} a_{1j} x_{dkrj} \leq v_k \, b_1^{\max}, \tag{13}$$

for all $d = 1, \ldots, D$ and for all $k = 1, \ldots, K$, where the nutrient no. 1 is energy.

The advisers do not usually care about wasting the food. We know that some food items are bought in packages of specific sizes. For example, we can buy yoghurt in packages of 100, 150, 200 or 500 grams, or eggs in packages of 6, 10, 15, 20 or 30 pieces.

Denote the set $\Xi = \{j_1, \dots, j_\vartheta\}$ of food items which are bought in packages of specific sizes. Let food item j_1 be bought in packages of sizes $\mathrm{P}_{j_1}^1, \mathrm{P}_{j_1}^2, \dots, \mathrm{P}_{j_1}^{\theta_{j_1}}$, where $0 < \mathrm{P}_{j_1}^1 < \mathrm{P}_{j_1}^2 < \dots < \mathrm{P}_{j_1}^{\theta_{j_1}}$, let food item j_2 be bought in packages of sizes $\mathrm{P}_{j_2}^1, \mathrm{P}_{j_2}^2, \dots, \mathrm{P}_{j_2}^{\theta_{j_2}}$, where $0 < \mathrm{P}_{j_2}^1 < \mathrm{P}_{j_2}^2 < \dots < \mathrm{P}_{j_2}^{\theta_{j_2}}$, etc., and let food item j_ϑ be bought in packages of sizes $\mathrm{P}_{j_\vartheta}^1, \mathrm{P}_{j_\vartheta}^2, \dots, \mathrm{P}_{j_\vartheta}^{\theta_{j_\vartheta}}$, where $0 < \mathrm{P}_{j_\vartheta}^1 < \mathrm{P}_{j_\vartheta}^2 < \dots < \mathrm{P}_{j_\vartheta}^{\theta_{j_\vartheta}}$. For $j \in \Xi$, we can use equations like

$$\sum_{d=1}^{D} \sum_{k=1}^{K} \sum_{\substack{r=1 \\ S_r \ni j}}^{R} x_{dkrj} = \sum_{\pi=1}^{\theta_j} (\mathrm{P}_j^\pi - \mathrm{P}_j^{\pi-1})\, \xi_j^\pi, \tag{14}$$

for all $j \in \Xi$, where ξ_j^π are new integer variables such that $0 \le \xi_j^1 \le \xi_j^2 \le \dots \le \xi_j^{\theta_j}$. For all $j \in \Xi$, we put $\mathrm{P}_{j0} = 0$.

Example: Food item j can be bought in packages of 100 grams, 150 grams and 180 grams, so let $\mathrm{P}_{j0} = 0$, $\mathrm{P}_{j1} = 100$, $\mathrm{P}_{j2} = 150$ and $\mathrm{P}_{j3} = 180$. Then we can use equations like

$$\sum_{d=1}^{D} \sum_{k=1}^{K} \sum_{\substack{r=1 \\ S_r \ni j}}^{R} x_{djkr} = 100\, \xi_j^{100} + 50\, \xi_j^{150} + 30\, \xi_j^{180},$$

where ξ_j^{100}, ξ_j^{150}, ξ_j^{180} are new integer variables such that $0 \le \xi_j^{100} \le \xi_j^{150} \le \xi_j^{180}$. The coefficient 30 by ξ_j^{180} means the difference between the size of the packages of 150 and 180 grams and 50 by ξ_j^{150} the difference between the size of the packages of 100 and 150 grams. We formulate analogous equations for each food item $j = 1, \dots, n$ that is supplied in packages.

To exclude the situation when some recipes are repeating during the week, we add inequalities

$$\sum_{d=d_0}^{d_0+6} \sum_{k=1}^{K} z_{dkr} \le 1\,, \tag{15}$$

for all $d_0 = 1, \dots, D-6$ and for all $r = 1, \dots, R$.

The adviser should follow the national nutrition recommendations for the population. According to Dostálová [5] and Hrnčířová [9], there are specific recommendations for people from the Czech Republic; what and how much they should eat or drink, including the recommendations about the intake of nutrients:

- some fermented food items every day,
- legumes at least two times a week,

- lean meat (300–400 grams) every week and a combination of poultry and veal,
- fish (400 grams) at least twice a week,
- animal viscera (liver, lungs, stomach, etc.) once every two weeks,
- handful of nuts a day (10 grams)
- 3 or 4 eggs a week
- vegetables at least 400 grams a day
- fruit 150–200 grams a day
- food of plant origin at least once a week [19]
- water intake at least 22 mililitres per 1 kilogram of personal weight where all minerals from mineral water should be between 150–500 miligrams per litre,
- alcohol: men wine/beer/spirits at most 250/500/60 mililitres, respectively, women at most 125/300/40 mililitres, respectively,
- sweets, smoked meat and other salted products eaten rarely,
- etc.

If the recipe r is used, then the corresponding food items $j \in S_r$ which the recipe consists of must be used. For that reason we introduce new binary variables y_{dkj} which mean whether the food item j is used in the meal k of the day d ($y_{dkj} = 1$) or not ($y_{dkj} = 0$). So we require that

$$y_{dkj} \geq z_{dkr} \,, \tag{16}$$

for all $d = 1, \dots, D$, for all $k = 1, \dots, K$, for all $r = 1, \dots, R$ and for all $j \in S_r$.

Conversely, if we use the food item j, then at least one recipe r including this food item must be used. Inequalities to express this condition are as follows

$$y_{dkj} \leq \sum_{\substack{r=1 \\ S_r \ni j}}^{R} z_{dkr} \,, \tag{17}$$

for all $d = 1, \dots, D$, for all $k = 1, \dots, K$ and for all $j = 1, \dots, n$.

We introduce sets $E_1, E_2, \dots, E_T$ of food items that are related as above, for example set E_1 of fermented food items, set E_2 of legumes, etc.

So let us consider the set of all fermented food items E_1. We know that the client should eat or drink some fermented food items every day. We express this condition by inequalities

$$\sum_{k=1}^{K} \sum_{j \in E_1} y_{dkj} \geq 1 \,, \tag{18}$$

for all $d = 1, \dots, D$. We should eat legumes from the set E_2 at least two times a week so we introduce the inequalities

$$\sum_{d=d_0}^{d_0+6} \sum_{k=1}^{K} \sum_{j \in E_2} y_{dkj} \geq 2 \,, \tag{19}$$

for all $d_0 = 1, \ldots, D-6$.

We should eat 300–400 grams of lean meat from the set E_3 every week and combine poultry and veal from E_{31} and E_{32} where the sets E_{31}, E_{32} are disjoint and $E_{31} \cup E_{32} \subseteq E_3$. This condition is expressed by inequalities

$$\sum_{d=d_0}^{d_0+6} \sum_{k=1}^{K} \sum_{r=1}^{R} \sum_{j \in E_3} x_{dkrj} \geq 300 \,, \tag{20}$$

$$\sum_{d=d_0}^{d_0+6} \sum_{k=1}^{K} \sum_{r=1}^{R} \sum_{j \in E_3} x_{dkrj} \leq 400 \,, \tag{21}$$

$$\sum_{d=d_0}^{d_0+6} \sum_{k=1}^{K} \sum_{j_1 \in E_{31}} y_{dkj_1} \geq 1 \,, \tag{22}$$

$$\sum_{d=d_0}^{d_0+6} \sum_{k=1}^{K} \sum_{j_2 \in E_{32}} y_{dkj_2} \geq 1 \,, \tag{23}$$

for all $d_0 = 1, \ldots, D-6$.

Furthermore, the client should eat food of plant origin, i.e. exclude meat during the whole day, at least once a week. So let the set E_4 include all meat items and let us add inequalities

$$\sum_{k=1}^{K} \sum_{j \in E_4} y_{dkj} \leq K \, |E_4| \, \eta_d \,, \tag{24}$$

$$\sum_{d=d_0}^{d_0+6} \eta_d \leq 6 \,, \tag{25}$$

for all $d = 1, \ldots, D$ and for all $d_0 = 1, \ldots, D-6$, where $|E_4|$ is the cardinality of the set E_4 and η_d are new binary variables.

Inequalities for the rest of the recommendations are analogous to (18)–(25).

Finally, we consider fluid intake. According to Zavadilová [28], adults should drink 22–38 mililitres of water per one kilogram of body weight in the weather with temperature between 22–37°C every day. The recommended fluid intake also depends on physical activity. Let E_T be the set of sparkling water, tea, juice, mineral water and other liquids. Then we add inequalities

$$\sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{j\in E_T} x_{dkrj} \geq b_m^{\min}, \tag{26}$$

$$\sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{j\in E_T} x_{dkrj} \leq b_m^{\max}, \tag{27}$$

for every $d = 1, \ldots, D$, where $b_m^{\min}$ or $b_m^{\max}$ is minimal or maximal amount of fluids per day, respectively.

The above conditions correspond to the recommendations of diet for the Czech Republic and the adviser can apply them to the clients who prefer a balanced diet plan. Of course, if the client cannot eat some food item, we do not include the food item into the sets $E_1, E_2, \ldots, E_T$.

The entire model is a mixed-integer linear programming model and consists of contraints (4)–(27). The model should use a large database of recipes. We can add an objective function (minimize the price of eaten food items or minimize the difference between the current and new bought food items) to the model, but this is not necessary for us now. We need to only find a feasible solution, so we minimize the zero objective function.

4 Results

Let us describe a particular client. We will not show the whole anamnestic and analytical part described in Section 2, but only the necessary fundamental data that we need to show our results of the model in Section 3.

The client is a woman, 26 years old and 173 cm tall. Using the equation (1) from Section 2 we have the ideal body weight $\omega_f = 64$ kilograms. Then we can calculate from equation (2) the basal energy expenditure $\beta_f = 6166$ kJ and from (3) the total energy $\tau = 10109$ kJ. Using the anamnestic and analytical part we can determine the amounts of the macro- and micronutrients. The macronutrients are as follows: 89 or 82 or 327 grams of proteins or fats or saccharides, respectively. The exact values can be in the tolerance of 5% so we can work with intervals. The amounts of microelements are simply inspired by [8], [24] and [25].

Table 1 presents the amounts of nutrients contained in food items we work with, the minimal and maximal recommended amounts of nutrients and a solution for one day in the Solution column. In the calculations we used 70 food items in 40 recipes and 31 nutrients.

In total, there are 43564 variables, out of which 22340 are integer, and 397642 constraints in the model.

Table 1
Contains the input data, calculated amounts of nutrients and final result per day

Nutrient	**Food items [100 g]**			**Recommended amounts**		
	Chicken	**Potato**	**…**	**Minimum**	**Solution**	**Maximum**
Energy [kJ]	694	322	…	9098	11006	11122
Proteins [g]	20	2	…	80	97	98
Fats [g]	10	0	…	73	80	90
Saccharides [g]	0	16	…	294	330	360
Fibre [g]	0	2	…	25	25	35
⋮	⋮	⋮	…	⋮	⋮	⋮
Vitamin [C]	2	15	…	75	220	230
⋮	⋮	⋮	⋮	⋮	⋮	⋮

The food plan found by our model for one day is presented in detail in Table 2.

Table 2
Optimal diet plan for one day

Meal	**Food items**
Breakfast	250 ml milk, 45 g oat flakes, 10 g almonds, 30 g orange, 7 g honey
Snack	100 g curd, 60 g apple, 30 g orange, 5 g linseed oil
Lunch	Soup: 75 g whole-wheat pasta, 30 g sweet corn, 25 g peas, 200 ml broth The main course: 130 g lentils, 150 g chicken, 5 g sunflower oil Salad: 25 g cucumber, 25 g potato, 35 g iceberg lettuce, 3 g olive oil, 1 pinch sesame seeds
Snack	60 g kaiser rolls (1 piece), 8 g margarine, 60 g cheese
Dinner	150 g slice a bread, egg spread (1 egg, 10 g margarine, 35 g curd, 3 g chives) Salad: 25 g tomato, 100 g bell pepper, 50 g cucumber
Drinking	250 ml fresh orange juice and 250 ml water, 2000 ml unsweetened tea

This model was solved by the optimization software FICO® Xpress Optimization Suite on a Windows XP SP3 computer with 1 GB RAM and Intel Atom 1.60 GHz CPU. The computation took about 45 seconds.

5 Discussion

The nutritional adviser is unable to create a diet for the whole week that respects the optimal amounts of all 31 nutrients every day.

If the adviser decides to use the presented mathematical model, then the adviser's work is reduced to assigning of food items to the recipes, assigning the recipes to the meals and setting minimal and maximal amounts of nutrients.

We would like to extend the model so that it includes not only the client but also the family members of the client. The reason is that, in practice, it is not easy to prepare different meals daily for everybody. And so in many cases it can happen that the client does not manage the diet properly or stops following the nutrition recommendations all together.

The model can also be extended to nutrition healthcare in hospitals and be useful for nutritional assistants, nutritional therapists and nutritionists (physicians specialized in artificial nutrition).

Conclusions

This article provides a new approach that will help improve the effectiveness for the nutritional adviser. We introduced a tool, which efficiently optimizes the adviser's work. It saves the adviser's time and effort, not only by supporting one client, but by supporting all the adviser's clients.

Acknowledgement

The author would like to thank Dr. David Bartl for useful discussions and comments which helped to improve the paper. This research was supported by the internal grant no. SGS12/PřF/2016–2017, Geometric mechanics, optimization and number theory. The use of the FICO® Xpress Optimization Suite under the FICO Academic Partner Program is gratefully acknowledged.

References

[1] Anderson, M., A., Earle, D., M.: Diet Planning in the Third World by Linear and Goal Programming. *The Journal of the Operational Research Society* **34.1** (1983) 9-16

[2] Bender, D., A.: *Introduction to Nutrition and Metabolism.* CRC Press, London, 2002

[3] Conforti, P., D'Amicis, A.: What is the cost of a healthy diet in terms of achieving RDAs? *Public Health Nutrition* **3.3** (2000) 367-373

[4] van Dooren, C., et al.: Combining Low Price, Low Climate Impact and High Nutritional Value in One Shopping Basket through Diet Optimization by Linear Programming. *Sustainability* **7.9** (2015) 12837-12855

[5] Dostálová, J., Dlouhý, P., Tláskal, P.: *Výživová doporučení pro obyvatelstvo České republiky* [on-line]. Společnost pro výživu, 2012, Available from: http://www.vyzivaspol.cz/vyzivova-doporuceni-pro-obyvatelstvo-ceske-republiky/ (cited 6 July 2016)

[6] Ferguson, E., L., et al.: Food-based dietary guidelines can be developed and tested using linear programming analysis. *The Journal of Nutrition* **134.4** (2004) 951-957

[7] Foytik, J.: Very low-cost nutritious diet plans designed by linear programming. *Journal of Nutrition Education* **13.2 (**1981) 63-66

[8] Heseker, H., Heseker, B.: *Die Nährwerttabelle: [über 40.000 Nährstoffangaben; einfache Handhabung; Tabellen zu Laktose, Fruktose, Gluten, Purin, Jod und trans-Fettsäuren]*. Umschau, Neustadt and der Weinstraße, 2016

[9] Hrnčířová, D., Rambousková, J., et al.: *Výživa a zdraví* [on-line]. Ministerstvo zemědělství, Odbor bezpečnosti potravin, Praha, 2012, Available from: http://www.bezpecnostpotravin.cz/UserFiles/publikace/Vyziva_a_zdravi.pdf (cited 16 August 2016)

[10] Kastnerová, M.: *Poradce pro výživu.* [*Nutrition Adviser*. In Czech.] Nová Forma, České Budějovice, 2011

[11] Kleinwächterová, H., Brázdová, Z.: *Výživový stav člověka a způsoby jeho zjišťování.* [*Nutritional Status of People and Methods of its Detection.* In Czech.] Institut pro další vzdělávání pracovníků ve zdravotnictví v Brně, Brno, 2001

[12] Kohout, P.: *Dokumentace a hodnocení nutričního stavu pacientů.* [*Documentation and Evaluation of Nutritional Status of Patients.* In Czech] Forsapi, Praha, 2011

[13] Legge, A.: How to Estimate Your Maintenace Calories. *Complete Human Performance* [on-line] Available from: http://www.completehumanperformance.com/calorie-needs/ (cited 28 August 2016)

[14] Leung, P., Wanitprapha, K., Quinn, L., A.: A recipe-based, diet-planning modelling system. *British Journal of Nutrition* **74.2** (1995) 151-162

[15] Macdiarmid, J., et al.: Sustainable diets for the future: can we contribute to reducing greenhouse gas emissions by eating a healthy diet? *The American Journal of Clinical Nutrition* **96.3** (2012) 632-639

[16] Maillot, M., Drewnowski, A.: Energy allowances for solid fats and added sugars in nutritionally adequate US diets estimated at 17-33% by a linear programming model. *The Journal of Nutrition* **141.2** (2011) 333-340

[17] Okubo, H., et al.: Designing optimal food intake patterns to achieve nutritional goals for Japanese adults through the use of linear programming optimization models. *Nutrition Journal* **14.57** (2015) 1-10

[18] Provazník, K., Komárek, L.: *Manuál prevence v lékařské praxi, VII. Doporučené preventivní postupy v primární péči.* [*Prevention Manual in Medicine Practice, VII. Recommended Preventive Procedures in Primal Healthcare*. In Czech.] Fortuna, Praha, 1999

[19] Provazník, K., Komárek, L., Janovská, J., Ošancová, K.: *Manuál prevence v lékařské praxi, II. Výživa.* [*Prevention Manual in Medicine Practice, II. Nutrition.* In Czech.] Fortuna, Praha, 1995

[20] Rážová, J.: *Metody a postupy v poradnách podpory zdraví.* [*Methods and Procedures in Health Care Advisory Work.* In Czech] Centrum zdraví Praha, Praha, 2001

[21] Reihserová, R.: *Jak poskládat stravu v průběhu dne?* [on-line]. Svět potravin. Available from: http://www.svet-potravin.cz/clanek.aspx?id=4177 (cited 7 July 2016)

[22] Schaynová, L.: A Client's Health from the Point of View of the Nutrition Adviser using Operational Research. In *Proceedings of the 34th International Conference on Mathematical Methods in Economics* 2016, Liberec: Technical University of Liberec (2016) 751-756

[23] Sklan, D., Dariel, I.: Diet planning for humans using mixed-integer linear programming. *British Journal of Nutrition* **70.1** (1993) 27-35

[24] Společnost pro výživu: *Referenční hodnoty pro příjem živin.* [*Dietary Reference Intakes for Nutrients.* In Czech] Výživa a servis s.r.o., Praha, 2011

[25] Svačina, Š., et al.: *Klinická dietologie.* [*Clinical Nutrition.* In Czech] Grada, Praha, 2008

[26] Trojan, S., Kittnar, O., Koudelová, J., et al.: *Lékařská fyziologie.* [*Medical Physiology.* In Czech] Avicenum, 1994

[27] Urbánek, L., Urbánková, P., et al.: *Klinická výživa v současné praxi.* [*Nutrition in Current Clinical Practice.* In Czech] Národní centrum ošetřovatelství a nelékařských zdravotnických oborů, Brno, 2008

[28] Zavadilová, V.: *Výživa a zdraví*. [*Nutrition and Health.* In Czech] Ostravská univerzita v Ostravě, Ostrava, 2014

[29] Zlatohlávek, L., et al.: *Klinická dietologie a výživa.* [*Clinical Nutrition and Dietetics.* In Czech] Current Media, Praha, 2016

Acoustic Simulations based on FVM Solution of the Helmholtz Equation

Izabela Riečanová, Angela Handlovičová

Department of Mathematics and Descriptive Geometry
Faculty of Civil Engineering, Slovak University of Technology in Bratislava
Radlinského 11, 810 05 Bratislava, Slovakia
riecanova@math.sk

Abstract: The main idea of the paper is an attempt to numerically simulate the data obtained by acoustic measurements. These measurements were performed in specialized acoustic laboratory. Their main idea was to study the reflection of different frequencies from boards with openings of various size and shape. The Finite volume method was used to make the simulations, where the Helmholtz equation is solved using the impedance boundary conditions. The results of the simulations are presented herein.

Keywords: measurement; Finite volume method; acoustic simulation; Fourier transform

1 Introduction - Acoustic Measurements

The main motivation of this work is to numerically simulate the values obtained by acoustic measurements. The measurements were performed in a specialized acoustic laboratory at the Faculty of Science at KU Leuven, in Belgium. The main idea is to study the reflection of different frequencies from boards with various openings. In Figure 1, there is the photo of one measuring experiment.

Figure 1
Photo of the measurement

As it is seen in the first figure, the measurements were performed in an anechoic room. Acoustically hard boards with openings of different size and shape were placed in the room and the impulse response measurements between the source and receiver were performed. The main focus was an analysis of sound reflection in frequency domain. Figure 2 shows simple scheme of the measurement.

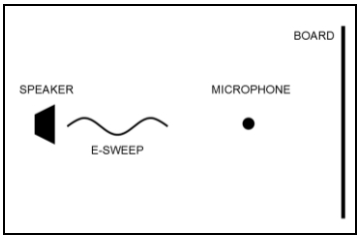

Figure 2
Scheme of the measurement

At known positions of the space, the speaker and microphone were placed. These positions were also varying. Exponential sweep, containing all audible frequencies was used as a test signal, sent from the loudspeaker and recorded by microphone, and impulse response (will be presented in next section) was calculated.

The main assumption was that if there is full board placed in the room (as in the Figure 2), all frequencies with wavelength smaller than the size of the board will be reflected. In case of a board with opening, a part of the sound energy with higher frequency content won't be reflected and will get through the panel, and the lower frequencies will fully reflect due to diffraction effects. If the opening is smaller, more of the frequencies are reflected. It is because the lower the frequency is, the bigger is its wavelength [9, 10]. When the frequency spectrum of the reflection was studied, our assumption was confirmed.

The main goal was the comparison of the data obtained from the measurements with numerical simulations implementing the Finite volume method.

2 Time Domain – Frequency Domain

This section describes the difference between the domains that the authors worked with and the conversion from one domain to the other.

The output from the acoustic measurements is the impulse response, which is shown in Figure 3.

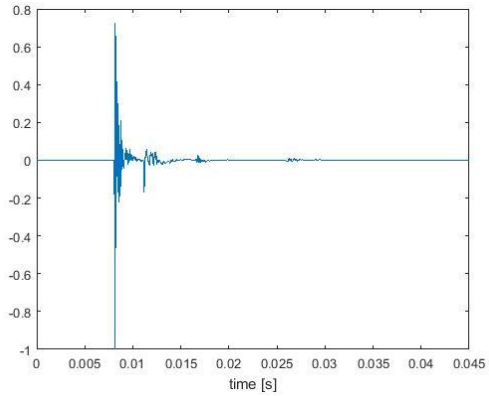

Figure 3
Impulse response of the space

The measurement data is in the time domain, which means that the particular signal is studied considering time. In Figure 3, the horizontal axis is the time measured in seconds and the vertical axis is the acoustic pressure measured in Pascal. This record contains the information about the behavior of all frequencies during whole time. It can be clearly seen, that the first and biggest peak in the graph is the direct sound arriving at the microphone, and the following smaller peaks are the sound reflections arriving with a time delay.

The numerical methods which are used for our simulations, work in the frequency domain. That means that the signal is studied with respect to frequency – during the computations a constant frequency is considered. To decompose the function of time into the frequencies, the Fourier transform was used.

The Fourier transform of a function of time is complex valued function of frequency

$$f : \mathbb{R} \to \mathbb{C}$$

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i t \xi} \, dt \quad . \tag{1}$$

In the equation 1, ξ is the frequency.

After the Fourier transform we obtain the data shown in the Figure 4.

On the horizontal axis the frequencies measured in Hertz. As can be seen there are also negative values of frequencies, which are the complex conjugate numbers of positive frequencies. These negative values were not important for us, but they are needed in case we would like to convert the data back to time domain.

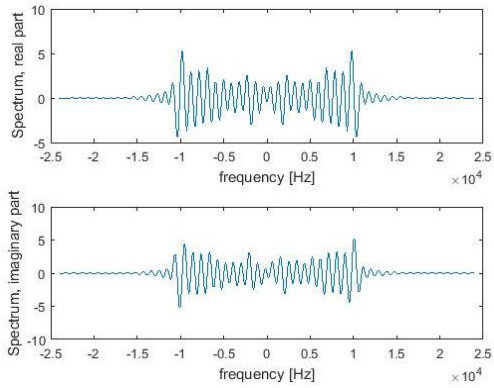

Figure 4
Frequency spectrum of impulse response after Fourier transform

The Magnitude of the complex number is the amplitude of acoustic pressure. Figure 5 shows this amplitude plotted on the logarithmic scale.

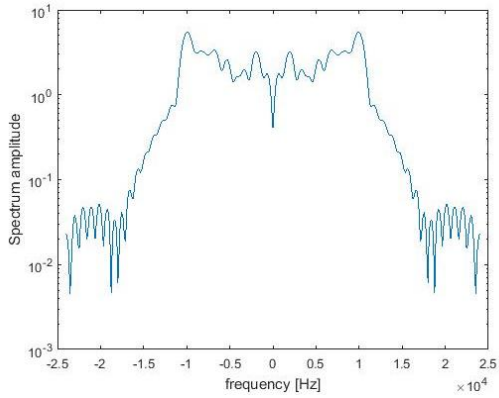

Figure 5
Amplitude of frequency domain signal plotted on logarithmic scale

We have taken the values from the frequency spectrum, which were used as the input data for the simulations. This was done by dividing the time-domain signal, so only a direct sound signal was obtained (Figure 6). We have applied the Fourier transform to this direct sound signal and computed data that was used in the program.

As the values in the frequency domain are variable (can be seen in Figure 4), it is not easy to choose the right frequency. If there is a Fourier transform done for either the direct sound or for the reflections, the frequency scaling is different so the interpolation had to be calculated. This approach may not be accurate, as the data oscillates, so this problem is left for further study.

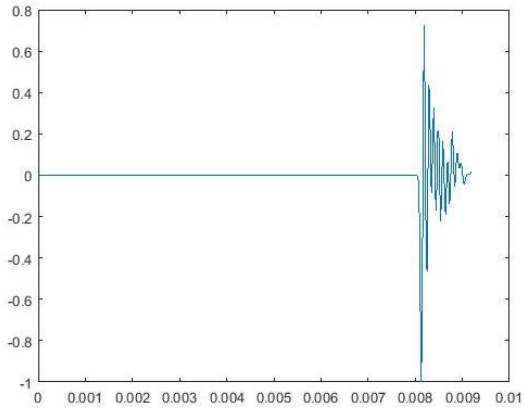

Figure 6
Time domain signal – direct sound only

3 Helmholtz Equation

Numerical methods in the field of acoustics solve the Helmholtz equation

$$\Delta A + k^2 A = 0. \tag{2}$$

Here Δ is the Laplace operator, A is the amplitude and k is the wavenumber (number of radians per unit distance). The wavenumber is given by

$$k = \frac{\omega}{c} = \frac{2\pi f}{c} \tag{3}$$

where ω is the angular frequency measured in radians per second, c is the phase velocity measured in meters per second, and f is the frequency.

The Helmholtz equation is related to the problems of steady-state oscillations. It is derived from the wave equation using the method of separating the variables, and it represents its time-independent form.

Because of its relation to the wave equation, the Helmholtz equation has use in various areas of physics, such as, electromagnetic radiation, elasticity or seismology. The main area of our interest is in acoustics. The algorithm based on the Finite volume method, presented later, is solving the Helmholtz equation.

4 Impedance Boundary Conditions

The boundary conditions which we mostly work with here are of the Robin type. They are called the impedance boundary conditions [1, 5]

$$ik\beta A + \frac{\partial A}{\partial n} = g. \tag{4}$$

Here A is already mentioned amplitude, i is imaginary unit and $\frac{\partial}{\partial n}$ is the normal derivative. The function g on the right hand side can be generally seen as the function of source. The parameter β is the relative surface admittance. When this parameter is set to $\beta = 0$, this represents the simulation of acoustically hard wall with maximum energy reflected. When we set $\beta = 1$, it simulates the wall with maximal sound absorption, i.e. the free space. It is important to note, that if we do the calculation considering the inward normal, the sign is opposite $\beta = -1$.

There must be mentioned that the Helmholtz equation with homogenous boundary conditions is an eigenvalue problem for the Laplacian [2]. When the situation $\lambda = k^2$ where λ is the eigenvalue, the solution to the equation is not unique. Moreover, its existence also depends on compatibility with the source function g.

This knowledge about the impedance boundary conditions, was implemented in our programs to simulate the aforementioned measurements.

5 Finite Volume Method

There are several numerical techniques for solving the Helmholtz equation and here we present the algorithm based on the Finite volume method [3, 4, 7, 8]. It is the method where the domain is discretized into cells called the finite volumes, which in our case were squares. The numerical solution u, i.e. the numerical value of amplitude, is a piecewise constant function with one constant value on each cell. The values of unknown function are calculated at discrete points of mesh usually called the representative points in the form of algebraic equations. Important feature of the method is the local conservativity of numerical fluxes, which means that the flux is conserved from one discretization cell to its neighbor.

Our case is three-dimensional, however we decided to simulate only the plane where the speaker and microphone were placed. Thus the problem reduces to two-dimensional which makes the computation easier and faster. The domain considered in the simulations was square, whereas its size was set to match the real situation.

After the volume discretization of the domain the grid of n^2 finite volumes is obtained (n is the number of discretizing points along one side of the domain). As the solution is complex valued function, it is in the form

$$u = u_r + iu_i. \tag{5}$$

For better calculations we have worked with the following form of (2)

$$-\Delta A - k^2 A = 0. \tag{6}$$

The steps of the method are the integration of the Helmholtz equation (6) over the finite volume p and applying the Green's theorem about the relationship between line and double integral. Thus the following equation for the numerical solution was obtained

$$-\int_{\partial p} \nabla u_p \boldsymbol{n_p} - \int_p k^2 u_p = 0 \tag{7}$$

where p is particular finite volume, u_p is its numerical solution on p, and $\boldsymbol{n_p}$ is the normal to the side of p. When approximating the integrals, we used the fact that numerical solution is constant function on each cell. To the normal derivative the difference approximation was applied and the following equation was obtained

$$\sum_{\sigma=1,\dots,4} |\sigma| \frac{u_p - u_q}{d_{pq}} - k^2 u_p m(p) = 0 \tag{8}$$

where σ is the edge of the cell ($|\sigma|$ is its length), $m(p)$ is the size of cell, u_q is the solution on the neighbor of particular finite volume, and d_{pq} denotes the distance between representative points of two adjacent finite volumes p and q (representative points were chosen the centers of finite volumes so the vector $X_p - X_q$, connecting the points, is perpendicular to the edge). If we suppose that our finite volumes are squares so that $\frac{|\sigma|}{d_{pq}} = 1$, we get the following

$$u_p\left(4 - k^2 m(p)\right) - \sum_{\sigma=1,\dots,4} u_q = 0. \tag{9}$$

This equation (9) is valid for interior finite volumes. Thus we compute the equations for each cell for both, the real and imaginary part. The system of equations with matrix of size $2n^2$ is obtained.

The equations for the finite volumes on the boundary or in the corner are slightly different, as these finite volumes have one or two exterior cells (Figure 7).

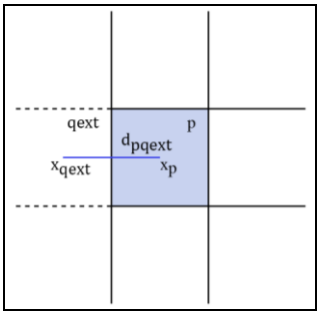

Figure 7
Representation of finite volume on the boundary

In these equations is the term u_{qext} denoting the numerical solution on exterior cell. As this value is not known, it was eliminated through the boundary conditions, which when written extra for real and imaginary part are

$$\frac{\partial u_r}{\partial n} - k\beta u_i = g_r$$

$$\frac{\partial u_i}{\partial n} - k\beta u_r = g_i \quad . \tag{10}$$

This was obtained by substituting (5) into (4). Using the boundary conditions the following formula was derived

$$u_{qext} = \frac{2hg + u_p(2 - ik\beta h)}{2 + ik\beta h} \tag{11}$$

which was used to solve this problem. *h* is the length of finite volume's side. The created system of equation was solved by method of LU decomposition and once it was done, the solution of the acoustic pressure was calculated at each point of the domain.

6 Numerical Simulations

This section presents the numerical simulation of acoustic measurement using the described knowledge. The simulated particular measurement was shown in Fig. 1, with the board with square opening placed in anechoic room. The way the boundary conditions were prescribed is shown in Figure 8.

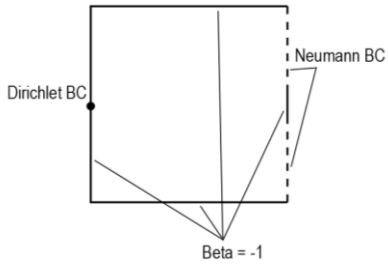

Figure 8
Representation of boundary conditions

For one finite volume on the boundary, the source function was prescribed with Dirichlet boundary conditions, where we put the values gained from the measurement. The relative surface admittance parameter was almost everywhere $\beta = -1$ to simulate the free space. Only on part of domain's right side it was $\beta = 0$ to simulate the board.

From the measurements we have known the data at the point where the microphone was placed. To make the simulation more precise we wanted to know the values at the position of loudspeaker. To do this we used the knowledge about the sound intensity and its decrease with distance considering the spherical sound wave [6]

$$I_2 = \frac{I_1}{4\pi r^2}. \tag{12}$$

Here I_1 is the sound intensity in known place measured in watt per square meter, and I_2 is decreased intensity in second place, where we want to know the value. Then we used the formula about the sound intensity and its relation to the sound pressure

$$I = \frac{p^2}{\rho_0 c} \tag{13}$$

where p is the sound pressure, $\rho_0 = 1.2\ kg/m^3$ is the density of air, and $c = 340.29\ m/s$ is the phase velocity. Using these formulas we have derived the new one, so we could increase the known measurement values and compute the data for the loudspeaker spot, which was used as Dirichlet boundary conditions. As each frequency decreases differently when propagating in air, this way of amplifying the data might not be correct. Solution would be an improvement of the measurement, where there would be more than just one microphone placed in the space. This way we would own the data from more points so we could calibrate the values used for our simulations.

The following figures show the results of the simulations for the frequency of 436 Hz (for the value of wavenumber $k = 8{,}1\ rad/m$). The value of real part of Dirichlet boundary conditions after recalculation using (12) and (13) was 4.772, and the value of imaginary part was 2.503. Figure 9 presents the real and imaginary part of the solution for $n = 30$ and Figure 10 shows the amplitude of acoustic pressure (magnitude of complex number). The size of domain was $1.5\ m$.

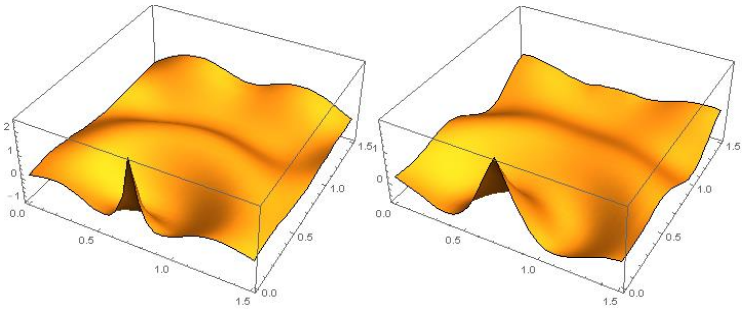

Figure 9

Numerical solution for the frequency of 436 *Hz*, left – real part, right – imaginary part

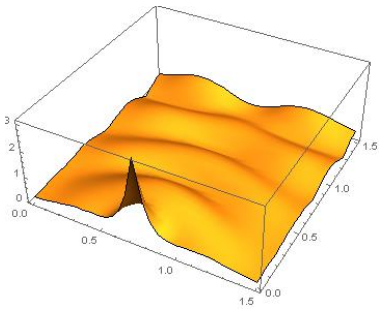

Figure 10

Numerical solution of acoustic pressure amplitude for the frequency of 436 *Hz*

Next are the results where the domain was diminished, so that the microphone lies exactly in the left side of the domain. This way no data had to be recalculated. The Dirichlet boundary conditions were put along whole left side. Figure 11 and 12 present the results for the same frequency again with $n = 30$.

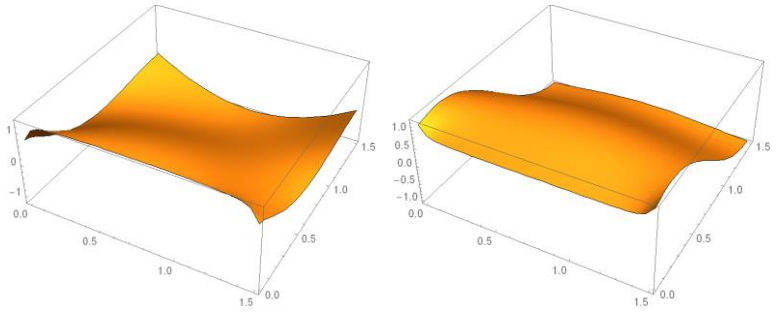

Figure 11

Numerical solution with diminished room for the frequency of 436 *Hz*, left – real part, right – imaginary part

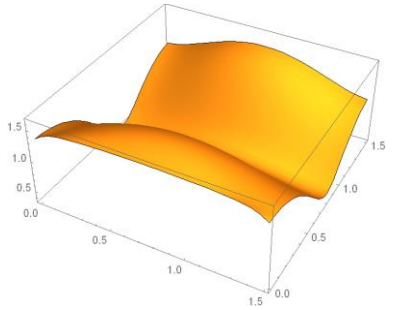

Figure 12

Numerical solution of acoustic pressure amplitude for the frequency of 436 *Hz*, diminished room

Conclusions

This work presents the results of acoustic simulations based on numerical methods, particularly the Finite volume method. The goal is use and process the data obtained by measurements performed in an acoustic laboratory and to possibly make comparisons with numerical simulations. Various problems appeared in the process. They could be solved by improving the measurements, which means placing more than one microphone in the room. In this way we could calibrate the data used in the simulations and secure their accuracy. It would also help us to compare the data from simulations with those from measurements. This is kept for future work.

Acknowledgement

This work was supported by VEGA 1/0728/15.

References

[1] Chandler-Wild S., Langdon S.: Boundary Element Methods for Acoustics. Lecture notes, University of Reading, Department of Mathematics, 2007

[2] Deuflhard P., Weiser M.: Adaptive Numerical Solution of PDEs. de Gruyter, Germany, 2012. ISBN 978-3-11-028310-5. e-ISBN 978-3-11-028311-2

[3] Eymard R., Gallouet T., Herbin R.: Finite volume methods. Ciarlet P. G. (ed.) et al., Handbook of numerical analysis. Vol. 7: Solution of equations in Rn (Part 3) Techniques of scientific computing (Part 3) Amsterdam: North-Holland/Elsevier, 713-1020, 2000

[4] Handlovičová A., Riečanová I.: Finite Volume Method Solution to the Complex 2D Helmholtz Equation with Impedance Boundary Conditions. Proceedings of the Congress of Information Technology, Computational and Experimental Physics, CITCEP 2015. Krakow: AGH, 2015, pp 81-87. ISBN 978-83-7464-838-7

[5] Handlovičová A, Riečanová I., Roozen N. B.: Rigid Piston Simulations of Acoustic Space Based on Finite Volume Method. Proceedings of Aplimat 2016, 15th Conference on Applied Mathematics. Bratislava: Faculty of Mechanical Engineering, Slovak University of Technology in Bratislava, 2016, pp 433-439. ISBN 978-80-227-4531-4

[6] http://www.sengpielaudio.com/calculator-SoundAndDistance.htm

[7] Riečanová I.: Complex-Valued Solution of Helmholtz Equation by Finite Volume Method. In ISCAMI 2015, book of abstracts. Bratislava: Faculty of Civil Engineering, Slovak University of Technology in Bratislava, 2015, pp 36. ISBN 978-80-227-4350-1

[8] Riečanová I.: Finite Volume Method Scheme for the Solution of Helmholtz Equation. Proceedings of Aplimat 2015, 14th Conference on Applied Mathematics. Bratislava: Faculty of Mechanical Engineering, Slovak University of Technology in Bratislava, 2015, pp 665-671. ISBN 978-80-227-4143-3

[9] Riečanová I.: Solution of the Helmholtz Equation by the Boundary Element Method and the Finite Volume Method. Proceedings of Advances in Architectural, Civil and Environmental Engineering: 24th Annual PhD Student Conference. Bratislava: Faculty of Civil Engineering, Slovak University of Technology in Bratislava, 2014, pp 29-36. ISBN 978-80-227-4301-3

[10] Rychtáriková M.: Room acoustical simulations in a multidisciplinary context. Slovak University of Technology in Bratislava, Faculty of Civil Engineering, 2010. ISBN 978-80-227-3422-6

Asymptotic stability of an evolutionary nonlinear Boltzmann-type equation

Roksana Brodnicka, Henryk Gacki

Institute of Mathematics, University of Silesia in Katowice, Bankowa 14, 40-007 Katowice, Poland,
rbrodnicka@o2.pl, Henryk.Gacki@us.edu.pl

In the paper a sufficient condition for the asymptotic stability with respect to total variation norm of semigroup generated by an abstract evolutionary non-linear Boltzmann-type equation in the space of signed measures with the right-hand side being a collision operator is presented. For this purpose a sufficient condition for the asymptotic stability of Markov semigroups acting on the space of signed measures for any distance ([4]), adapted to the total variation norm, joined with the maximum principle for this norm is used. The paper generalizes the result in [4] related to the same type of non-linear Boltzmann-type equation, where the asymptotic stability in the weaker norm, Kantorovich-Wasserstein, was investigated.

Keywords: Asymptotic stability, Markov operators, maximum principle for the total variation metric, nonlinear Boltzmann-type equation

1 Introduction

We are interested in the problem of the stability of solutions u of the following version of the Boltzmann equation

$$\frac{\partial u(t,x)}{\partial t} + u(t,x) = \int_x^\infty \frac{dy}{y} \int_0^y u(t,y-z)u(t,z)dz \qquad t \geq 0, \qquad x \geq 0, \tag{1}$$

with the additional conditions for $t \geq 0$

$$\int_0^\infty u(t,x)\,dx = \int_0^\infty xu(t,x)\,dx = 1, \tag{2}$$

which describes the law of conservation of mass and energy. Equation (1) was presented in the space $L^p(\mathbb{R}_+)$ with $p = 1,2$ and different weights (see [1], [3], [7]). Equation (1) was derived by J. A. Tjon and T. T. Wu from the Boltzmann equation using the Abel transformation (see [14]) and was later called by Barnsley and Cornille (see [1]) the *Tjon–Wu equation*.

Equation (1) governs the evolution of the density distribution function of the energy of particles imbedded in an ideal gas in the equilibrium stage (see [7], [8], [14]). The solution $u(t,\cdot)$ of the problem has an interpretation as a probability distribution function of the energy of particles in an ideal gas. In the time interval $(t,t+\Delta t)$ a particle changes its energy with the probability $\Delta t + o(\Delta t)$ and this change is described by the operator

$$(Pu)(x) = \int_x^\infty \frac{dy}{y} \int_0^y u(y-z)\,u(z)dz. \tag{3}$$

Hence, the change is equal to $[-u(t,x)+P(u(t,x))]\triangle t + o(\triangle t)$.

In order to understand the action of P consider three independent random variables ξ_1,ξ_2 and η, such that ξ_1,ξ_2 have the same density distribution function u and η is uniformly distributed on the interval $[0,1]$. Here we obtain that Pu is the density distribution function of the random variable

$$\eta(\xi_1+\xi_2). \tag{4}$$

This corresponds to the physical assumption that the energies of the particles before a collision are independent quantities and that a particle after collision takes η part of the sum of the energies of the colliding particles.

The assumption that η has the density distribution function of the form $\mathbf{1}_{[0,1]}$ is quite restrictive. In general, if η has the density distribution h, then the random variable (4) has the density distribution function

$$(Pv)(x) = \int_x^\infty h(\frac{x}{y})\frac{dy}{y} \int_0^y u(y-z)\,u(z)dz. \tag{5}$$

The problem of the asymptotic behaviour of solutions of the equation:

$$\frac{\partial u(t,x)}{\partial t} + u(t,x) = \int_x^\infty h(\frac{x}{y})\frac{dy}{y} \int_0^y u(y-z)\,u(z)dz \tag{6}$$

was investigated by A. Lasota and J. Traple in 1999 ([10], Theorem 1.1).

This version is more general than (1). In both versions there are no physical reasons which will allow us to assume that the distribution of energy of particles can be described only by density (so by the absolutely continuous measure).

Following this physical interpretation, Gacki in 2007 (see [4]) considered the evolutionary Boltzmann-type equation

$$\frac{d\psi}{dt} + \psi = P\psi \qquad \text{for} \qquad t \geq 0 \tag{7}$$

with the initial condition

$$\psi(0) = \psi_0, \tag{8}$$

where $\psi_0 \in \mathscr{M}_1(\mathbb{R}_+)$ and $\psi : \mathbb{R}_+ \to \mathscr{M}_{sig}(\mathbb{R}_+)$ is an unknown function. Moreover $P : \mathscr{M}_1(\mathbb{R}_+) \to \mathscr{M}_1(\mathbb{R}_+)$ is analogous to (5), but in this case P is an operator acting on the space of probability measures. The operator P will be described precisely in Section 3. By $\mathscr{M}_1(\mathbb{R}_+)$ and $\mathscr{M}_{sig}(\mathbb{R}_+)$ we denote the space of probability measures and the space of finite signed measures respectively. More precisely an operator P is acting on the subset $D \subset \mathscr{M}_1(\mathbb{R}_+)$ given by formula

$$D := \{\mu \in \mathscr{M}_1 : m_1(\mu) = 1\}, \quad \text{where} \quad m_1(\mu) = \int_0^\infty x\mu(dx). \tag{9}$$

Equation (7) was studied in the space $\mathscr{M}_1(\mathbb{R}_+)$. The operator P describes the collision of two particles in general situation.

In [4], the problem of the stability of solutions of a nonlinear Boltzmann-type equation (7) with the initial condition (8) was studied in Kantorovich-Wasserstein norm (see [4], [13]). The proof of the asymptotic stability is based on a property of the Kantorovich-Rubinstein norm in the space of probabilistic measures, which the author called *the maximum principle* (see [5]).

The purpose of our paper is to prove that the semigroup generated by the equation (7) with the initial condition (8) is asymptotically stable with respect to the total variation norm. The basic idea of our method is to apply technique related with the maximum principle for the total variation norm (see [2]).

The maximum principle method in studying the asymptotic stability of Markov semigroup with respect to various metrics was used in the papers [2], [4], [6], [9] and [10].

In order to make the paper self-contained all necessary definitions from the theory of Markov operators, dynamical systems and differential equations in Banach spaces are recalled at the beginning of Sections 2 and 3 respectively.

2 Preliminaries

Let (X,ρ) be a Polish space and let $\mathscr{B}_X$ be σ–algebra of its Borel. We denote by $\mathscr{M}$ the family of all finite (nonnegative) Borel measures on X. and by $\mathscr{M}_1$ we the subset of $\mathscr{M}$ such that $\mu(X) = 1$ for $\mu \in \mathscr{M}_1$. Now let

$$\mathscr{M}_{sig} = \{\mu_1 - \mu_2 : \ \mu_1, \mu_2 \in \mathscr{M}\},$$

be the space of *finite signed measures* endowed with the total variation norm $\|\cdot\|_T$ (under which it is a Banach space).

Fix an element c of X and for every real number $\alpha \geq 1$ we define sets $\mathscr{M}_{1,\alpha}$ and $\mathscr{M}_{sig,\alpha}$

$$\mathscr{M}_{1,\alpha} = \{\mu \in \mathscr{M}_1 : \ m_\alpha(\mu) < \infty\} \qquad \text{and} \qquad \mathscr{M}_{sig,\alpha} = \{\mu \in \mathscr{M}_{sig} : \ m_\alpha(\mu) < \infty\}$$

where

$$m_\alpha(\mu) = \int_X (\rho(x,c))^\alpha |\mu|(dx).$$

It is easy to verify that these spaces do not depend on the choice of c.

Denote by $B(x,r)$ a closed ball in X with center $x \in X$ and radius r. For $\mu \in \mathscr{M}_1$ define *the support of a measure* μ by

$$\operatorname{supp} \mu = \{x \in X : \ \mu(B(x,\varepsilon)) > 0 \ \text{ for every } \varepsilon > 0\}.$$

The support of a measure being a stationary solution will play an important role in the proof of the asymptotic stability of the equation (7). Every set $\mathscr{M}_{1,\alpha}$, for $\alpha \geq 1$ contains the subset of all measures $\mu \in \mathscr{M}_1$ with a compact support.

In the proof of the main result of this paper an important role is played by some property of the total variation norm, directly connected with the strong contractivity, which is called the maximum principle. The relation between contractivity and the maximum principle will be described below in Theorem 2.1.

The Maximum principle for total variation norm formulated as follows: Let $\mu_1, \mu_2 \in \mathscr{M}$. Then

$$\|\mu_1 - \mu_2\|_T = \|\mu_1\|_T + \|\mu_2\|_T \tag{10}$$

if and only if μ_1 and μ_2 are mutually singular (i.e. if there are two sets $A, B \in \mathscr{B}$ such that $A \cap B = \emptyset$, $A \cup B = X$ and $\mu_1(B) = \mu_2(A) = 0$). (For details see [2], p. 325).

We start with a definition of Markov operator

Definition 2.1. An operator $P : \mathscr{M} \to \mathscr{M}$ is called a *Markov operator* if it satisfies the following conditions:

(i) P is positively linear

$$P(\lambda_1\mu_1 + \lambda_2\mu_2) = \lambda_1 P\mu_1 + \lambda_2 P\mu_2$$

for $\lambda_1, \lambda_2 \geq 0$ and $\mu_1, \mu_2 \in \mathscr{M}$,

(ii) P preserves the measure of the space

$$P\mu(X) = \mu(X) \qquad \text{for} \qquad \mu \in \mathscr{M}. \tag{11}$$

Note that every Markov operator P can be uniquely extended as an operator to the space of signed measures.

In what follows we will understand by d the distance generated by the total variation norm on $\mathscr{M}_{sig}$. A Markov operator $P : \mathscr{M}_{sig} \to \mathscr{M}_{sig}$ is called *contracting* or *nonexpansive* with respect to d if

$$d(P\mu_1, P\mu_2) \leq d(\mu_1, \mu_2) \qquad \text{for} \qquad \mu_1, \mu_2 \in \mathscr{M}_{sig}. \tag{12}$$

A Markov operator $P : \mathscr{M}_{sig} \to \mathscr{M}_{sig}$ is called *strongly contracting* or *contractive in the class* $\widetilde{\mathscr{M}} \subset \mathscr{M}_{sig}$ with respect to d if

$$d(P\mu_1, P\mu_2) < d(\mu_1, \mu_2) \qquad \text{for} \qquad \mu_1, \mu_2 \in \widetilde{\mathscr{M}}. \tag{13}$$

Definition 2.2. We say that the measures $\mu, \nu \in \mathscr{M}$ *overlap supports* if there is no set $A \in \mathscr{B}$ such that

$$\mu(A) = 0 \text{ and } \nu(A^c) = 0$$

Contractivity of Markov operators in total variation plays an important role in investigation of asymptotics of solutions of equation (1). We have

Theorem 2.1. *Let P be a Markov operator. Assume that $P\mu_+, P\mu_-$ overlap supports for every nontrivial measure $\mu \in \mathscr{M}_{sig}$. Then Markov operator P is strongly contracting with respect to the distance d generated by the total variation norm.*

In the proof of this theorem, the crucial role is played by the inequality:

$$d(P\mu_+, P\mu_-) \leq ||P\mu_+||_T + ||P\mu_-||_T.$$

Applying the maximum principle to $P\mu_+$ and $P\mu_-$, we obtain the strong inequality. But we have

$$||P\mu_+|_T = ||\mu_+||_T \text{ and } ||P\mu_-||_T = ||\mu_-||_T,$$

so using the maximum principle once more (for μ_+ and μ_-), we directly obtain that P is strongly contracting. For details see [2], p. 326.

Now we recall few facts from the theory of dynamical systems.

Let T be a *nontrivial semigroup* of nonnegative real numbers i.e. $\{0\} \subsetneq T \subset \mathbb{R}_+$ and $t_1 + t_2 \in T$, $t_1 - t_2 \in T$ for $t_1, t_2 \in T$, $t_1 \geq t_2$.

A family of Markov operators $(P^t)_{t \in T}$ is called a *semigroup* if

$$P^{t+s} = P^t P^s \qquad \text{for} \qquad t, s \in T$$

and $P^0 = I$ where I is the identity operator.

A semigroup $(P^t)_{t \in T}$ is called a *semidynamical system* if the transformation $\mathscr{M}_{sig} \ni \mu \to P^t\mu \in \mathscr{M}_{sig}$ is continuous for every $t \in T$.

Remark 2.1. Every Markov operator $P : \mathscr{M}_{sig} \to \mathscr{M}_{sig}$ is continuous with respect to the total variation norm. Consequently, every semigroup $(P^t)_{t \in T}$ of Markov operators is a semidynamical system.

If a semidynamical system $(P^t)_{t \in T}$ is given, then for every fixed $\mu \in \mathscr{M}_{sig}$ the function $T \ni t \to P^t\mu \in \mathscr{M}_{sig}$ will be called a *trajectory* starting from μ and denoted

by $(P^t\mu)$. A point $\nu \in \mathscr{M}_{sig}$ is called a *limiting point* of a trajectory $(P^t\mu)$ if there exists a sequence (t_n), $t_n \in T$, such that $t_n \to \infty$ and

$$\lim_{n\to\infty} P^{t_n}\mu = \nu.$$

The set of all limiting points of the trajectory $(P^t\mu)$ will be denoted by $\Omega(\mu)$.

We say that a trajectory $(P^t\mu)$ is *sequentially compact* if for every sequence (t_n), $t_n \in T$, $t_n \to \infty$, there exists a subsequence (t_{k_n}) such that the sequence $(P^{t_{k_n}}\mu)$ is convergent to a point $\nu \in \mathscr{M}_{sig}$.

Remark 2.2. If the trajectory $(P^t\mu)$ is sequentially compact, then $\Omega(\mu)$ is a nonempty, sequentially compact set.

A point $\mu_* \in \mathscr{M}_{sig}$ is called *stationary* (or *invariant*) with respect to a semidynamical system $(P^t)_{t\in T}$ if

$$P^t\mu_* = \mu_* \qquad \text{for} \qquad t \in T. \tag{14}$$

A semidynamical system $(P^t)_{t\in T}$ is called *asymptotically stable* if there exists a stationary point $x_* \in X$ such that

$$\lim_{t\to\infty} P^t\mu = \mu_* \qquad \text{for} \qquad \mu \in \mathscr{M}_{sig}. \tag{15}$$

Remark 2.3. Since $(\mathscr{M}_{sig}, \|\cdot\|_T)$ is a Hausdorff space, an asymptotically stable dynamical system has exactly one stationary point.

We say that a Markov semigroup $(P^t)_{t\in T}$ is *contracting* or *nonexpansive semigroup with respect to the distance d generated by the total variation norm in the class* $\widetilde{\mathscr{M}} \subset \mathscr{M}_{sig}$ if the following condition holds

$$d(P^t\mu_1, P^t\mu_2) \le d(\mu_1, \mu_2) \qquad \mu_1, \mu_2 \in \widetilde{\mathscr{M}};\ t \in T. \tag{16}$$

A contracting semigroup $(P^t)_{t\in T}$ will be called *strongly contracting with respect to the distance d generated by the total variation norm in the class* $\widetilde{\mathscr{M}} \subset \mathscr{M}_{sig}$ if and only if for every $\mu_1, \mu_2 \in \widetilde{\mathscr{M}}$, $\mu_1 \neq \mu_2$ there is a number $t_0 \in T$ such that

$$d(P^{t_0}\mu_1, P^{t_0}\mu_2) < d(\mu_1, \mu_2).$$

Let $(P^t)_{t\in T}$ be a semidynamical system which has at least one sequentially compact trajectory and $\mathscr{Z}$ – the set of all $\mu \in \mathscr{M}_{sig}$ such that the trajectory $(P^t\mu)$ is sequentially compact. $\mathscr{Z}$ is a nonempty set, so

$$\Omega = \bigcup_{\mu\in\mathscr{Z}} \omega(\mu) \neq = \emptyset.$$

In the proof of the main result of this paper – Theorem 3.2 – we will use the following criterion for the asymptotic stability of trajectories

Theorem 2.2. *Let $x_* \in \Omega$ be fixed. Assume that for every $x \in \Omega$, $x \neq x_*$ there is $t(x) \in T$ such that*

$$d(S^{t(x)}x, S^{t(x)}x_*) < d(x, x_*). \tag{17}$$

Further assume that the semidynamical system $(S^t)_{t\in T}$ is nonexpansive with respect to distance d, i.e.,

$$d(S^t x, S^t y) \leq d(x,y) \qquad for \qquad x,y \in \mathscr{M}_{sig} \qquad and \qquad t \in T. \tag{18}$$

Then x_* is a stationary point of $(S^t)_{t\in T}$ and

$$\lim_{t\to\infty} d(S^t z, x_*) = 0 \qquad for \qquad z \in Z. \tag{19}$$

where Z is the set of all $z \in \mathscr{M}_{sig}$ such that the trajectory $(S^t z)$ is compact.

This criterion is a special case, adapted to the distance generated by the total variation norm, of the more general result (for any distance), which may be found in [4], p. 28–30.

3 Main result - asymptotic stability

In this section we show that the equation (7) may by considered in a convex closed subset of a vector space of signed measures. This approach seems to be quite natural and it is related to the classical results concerning the semigroups and differential equations on convex subsets of Banach spaces (see [3], [11]).

Let $(E, \|\cdot\|)$ be a Banach space and let $\tilde{D}$ be a closed, convex, nonempty subset of E. In the space E we consider *an evolutionary differential equation*

$$\frac{du}{dt} = -u + \tilde{P}u \qquad \text{for} \qquad t \in \mathbb{R}_+ \tag{20}$$

with the initial condition

$$u(0) = u_0, \qquad u_0 \in \tilde{D}, \tag{21}$$

where $\tilde{P} : \tilde{D} \to \tilde{D}$ is a given operator.

A function $u : \mathbb{R}_+ \to E$ is called a solution of problem (20), (21) if it is strongly differentiable on $\mathbb{R}_+$, $u(t) \in \tilde{D}$ for all $t \in \mathbb{R}_+$ and u satisfies relations (20), (21).

We start from the following theorem which is usually stated in the case $E = \tilde{D}$.

Theorem 3.1. *Assume that the operator $\tilde{P} : \tilde{D} \to \tilde{D}$ satisfies the Lipschitz condition*

$$\|\tilde{P}v - \tilde{P}w\| \leq l\,\|v- = w\| \qquad for \qquad u,w \in \tilde{D}, \tag{22}$$

where l is a nonnegative constant. Then for every $u_0 \in \tilde{D}$ there exists a unique solution u of problem (20), (21).

The standard proof of the Theorem 3.1 is based on the fact, that a function $u : \mathbb{R}_+ \to \tilde{D}$ is the solution of (20), (21) if and only if it is continuous and satisfies the integral equation

$$u(t) = e^{-t} u_0 + \int_0^t e^{-(t-s)} \tilde{P} u(s)\, ds \qquad \text{for} \qquad t \in \mathbb{R}_+. \tag{23}$$

Due to completeness of $\tilde{D}$ the integral on the right hand side is well defined and equation (23) may be solved by the method of successive approximations.

Observe that, thanks to the properties of $\tilde{D}$, for every $u_0 \in \tilde{D}$ and every continuous function $u : \mathbb{R}_+ \to \tilde{D}$ the right hand side of (23) is also a function with values in $\tilde{D}$.

The solutions of (23) generate a semigroup of operators $(\tilde{P}^{\,t})_{t\geq 0}$ on $\tilde{D}$ given by the formula

$$\tilde{P}^{\,t} u_0 = u(t) \qquad \text{for} \qquad t \in \mathbb{R}_+, \qquad u_0 \in \tilde{D}. \tag{24}$$

We can now come to the main result of the paper – a sufficient condition for the asymptotic stability of solutions of the equation (7) with respect to the total variation metric.

At the beginning we return to equation (7) and give the precise definition of P.

We start from recalling that the *convolution of measures* $\mu, \nu \in \mathscr{M}_{sig}$ is a unique measure $\mu * \nu$ satisfying

$$(\mu * \nu)(A) := \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} 1_A(x+y) \mu(dx) \nu(dy) \qquad \text{for} \qquad A \in \mathscr{B}_X. \tag{25}$$

(see [9]).

A linear operator $P_{*2} : \mathscr{M}_{sig} \mapsto \mathscr{M}_{sig}$ is defined by

$$P_{*2}\mu := \mu * \mu \qquad \text{for } \mu \in \mathscr{M}_{sig}. \tag{26}$$

It is easy to verify that $P_{*2}(\mathscr{M}_1) \subset \mathscr{M}_1$. Moreover the maps $P_{*2}\big|_{\mathscr{M}_1}$ have a simple probabilistic interpretation. Namely, if ξ_1, ξ_2 are independent random variables with the same distribution μ, then $P_{*2}\mu$ is the distribution of the sum $\xi_1 + \xi_2$.

The second class of operators we are going to study is related to the multiplication of random variables (see [9]). The formal definition is as follows. Given two measures $\mu, \nu \in \mathscr{M}_{sig}$, we define the *elementary product* $\mu \circ \nu$ by the formula

$$(\mu \circ \nu)(A) : \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} 1_A(xy) \mu(dx) \nu(dy) \qquad \text{for} \qquad A \in \mathscr{B}_{\mathbb{R}_+}. \tag{27}$$

For fixed $\varphi \in \mathscr{M}_1$ we define the linear operator $P_\varphi : \mathscr{M}_{sig} \to \mathscr{M}_{sig}$ by the formula

$$P_\varphi \mu := \varphi \circ \mu \qquad \text{for} \qquad \mu \in \mathscr{M}_{sig}. \tag{28}$$

Again, as in the case of convolution, from this definition it follows that $P_\varphi(\mathscr{M}_1) \subset \mathscr{M}_1$. For $\mu \in \mathscr{M}_1$ the measure $P_\varphi \mu$ has an immediate probabilistic interpretation. If φ and μ are the distributions of random variables ξ and η respectively, then $P_\varphi \mu$ is the distribution of the product $\xi \eta$.

Now we may return to the equation (7) and give the precise definition of P. Namely we define

$$P := P_\varphi P_{*2}, \tag{29}$$

where $\varphi \in \mathscr{M}_1$ and $m_1(\varphi) = \frac{1}{2}$. From equality (29) it follows that $P(\mathscr{M}_1) \subset \mathscr{M}_1$. Further using (26) and (28) it is easy to verify that for $\mu \in D$

$$m_1(P_{*2}\mu) = 2 \qquad \text{and} \qquad m_1(P_\varphi \mu) = \frac{1}{2}, \tag{30}$$

where D is defined by the formula (9).

From the definition of the set D and operator P, we obtain the following properties:

1. The set D is a convex subset of $\mathscr{M}_{\mathrm{sig},1}$.
2. The set D with distance d is a complete metric space.
3. If $\varphi \in \mathscr{M}_1$ and $m_1(\varphi) = 1/2, m_1(\nu_0) = 1$, then $P(D) \subset D$.

Equation (7) together with the initial condition (8) may be considered in a convex subset D of the vector space of signed measures. From the properties (1), (2), (3) and the results of [3] it follows immediately that for every $\psi_0 \in D$ the initial value problem (7), (8) has exactly one solution ψ satisfying the integral equation

$$\psi(t) = e^{-t}\,\psi_0 + \int_0^t e^{-(t-s)}\,P\,\psi(s)\,ds \qquad \text{for } t \in \mathbb{R}_+. \tag{31}$$

Corollary 3.1. *If $\varphi \in \mathscr{M}_1$ and $m_1(\varphi) = 1/2$ then for every $\psi_0 \in D$ there exists an unique solution ψ of problem* (7), (8).

The solutions of (31) generate a semigroup of Markov operators $(P^t)_{t\geq 0}$ on D given by

$$\psi(t) = P^t\,\psi_0 \qquad \text{for } t \in \mathbb{R}_+,\ \psi_0 \in D. \tag{32}$$

Now using criterion for the asymptotic stability of trajectories Theorem 2.2 jointly with the maximum principle for total variation metric from the Theorem 2.1, we can easily derive the following main result of this paper:

Theorem 3.2. *Let P be the operator given by (29). Moreover, let φ be a probability measure with $m_1(\varphi) = 1/2$ and let 0 be accumulation point of* $\operatorname{supp}\varphi$. *If P has a fixed point $\psi_* \in D$ such that*

$$\operatorname{supp}\psi_* = \mathbb{R}_+, \tag{33}$$

then

$$\lim_{t\to\infty} \|\psi(t) - \psi_*\|_T = 0 \tag{34}$$

for every compact solution ψ *of* (7), (8).

Proof. It is sufficient to verify condition (17) of Theorem 2.2.

From (31) it follows immediately that

$$\begin{aligned} \|P^t \psi_0 - \psi_*\|_T &\leq e^{-t} \|\psi_0 - \psi_*\|_T \\ &+ \int_0^t e^{-(t-s)} \|P^s \psi_0 - \psi_*\|_T \, ds \qquad \text{for } \psi_0 \in D \text{ and } t > 0. \end{aligned}$$

This may be rewritten in the form

$$\begin{aligned} \|P^t \psi_0 - \psi_*\|_T &\leq e^{-t} \|\psi_0 - \psi_*\|_T + (1 - e^{-t}) \|\psi_0 - \psi_*\|_T \\ &= \|\psi_0 - \psi_*\|_T \qquad \text{for } \psi_0 \in D \text{ and } t > 0. \end{aligned} \tag{35}$$

Condition (33) is equivalent to the fact that the measures $P^t \psi_0, \psi_* \in D$ overlap supports for $t > 0$ and $\psi_0 \in D$. Applying Theorem 2.1 for P^t, we will get that Markov operator P^t is strongly contracting. Consequently, in (35) we have a strict inequality, because $P^t(\psi_*) = \psi_*$. An application of Theorem 2.2 completes the proof.

□

Remark 1. We showed that if equation (7) has a stationary solution μ_* such that $\operatorname{supp} \mu_* = \mathbb{R}_+$, then this measure is asymptotically stable. The positivity of u_* plays an important role in the proof of the stability. Namely, it allows us to apply the maximum principle in order to show that the total variation distance between u_* and an arbitrary solution u is decreasing in time.

Moreover, in [4] p. 34. the following result was shown:

Let φ *be a probability measure and let* $m_1(\varphi) = \frac{1}{2}$. *Assume that:*

(i) There is $\sigma_0 > 0$ *such that*

$$(0, \sigma_0) \subset \operatorname{supp} \varphi. \tag{36}$$

(ii) The operator P has a fixed point $v \in \mathscr{M}$ *such that* $v \neq \delta_0$.
Then

$$\operatorname{supp} v = \mathbb{R}_+. \tag{37}$$

From above it follows that the assumption (33) can be replaced by the more effective condition (36).

Observe that in the case of the classical linear Tjon–Wu type equation (1) the measure φ is absolutely continuous with density $\mathbf{1}_{[0,1]}$. Moreover, $u_*(t,x) := \exp(-x)$ is the density function of the stationary solution of (1). This is a simple illustration of the situation described by Theorem 3.2.

Moreover, the condition (33) is not particularly restrictive because in Lasota's and Traple's paper (see [12]) it has been proved that the stationary solution ϕ_* has the following property: Either ψ_* is supported at one point or $supp\,\psi_* = \mathbb{R}_+$. The first case holds if and only if $\varphi = \delta_{\frac{1}{2}}$. But this case is forgettable as a physical model of particle collisions because it is more restrictive than the model described by the classical Tjon-Wu equation.

Remark 2. It is worth noting that:

1. Every solution of the equation $P\mu = \mu$ is a stationary solution of equation (7).
2. We have many possibilities to apply the criterion written in Theorem 3.2. For example, if we consider the equation (7) with the following assumption:

$$2m_r(\varphi) < 1, \text{ where } r > 1,$$

then for every $\psi_0 \in D$ the solution of (7) and (8) is compact (see [4]).

Summary

The Boltzmann equation in the general form gives us information about time, position and velocity of particles in the dilute gas. This equation is a base for many mathematical models of colliding particles.

In particular, in [2] authors described the homogeneous model where a small number of particles is introduced into a gas which contains many more particles, at equilibrium. The solution of the considered in [2] equation in the time t informs us about an energy state of the introduced particles in t.

In present paper authors consider the homogeneous model in the dilute gas with a possibility of collisions of two particles. The solution of the equation describing this model, (7), in time t, gives an information about an energy change between colliding particles in t.

In the future, it is planned to describe the mathematical model of colliding particles with a possibility of collisions of arbitrary many particles. Moreover, the external forces may exist.

Acknowledgements. The Authors are indebted to Joanna Zwierzyńska for her valuable remarks and editorial help.

References

[1] M. F. Barnsley and H. Cornille: General solution of a Boltzmann equation and the formation of Maxwellian tails, Proc. Roy. London A 374, (1981), 371–400.

[2] R. Brodnicka and H. Gacki: Asymptotic stability of a linear Boltzmann-type equation, Appl. Math., 41 (2014), 323–334.

[3] M. G. Crandall: Differential equations on convex sets, J. Math. Soc. Japan 22 (1970), 443–455.

[4] H. Gacki: Applications of the Kantorovich-Rubinstein maximum principle in the theory of Markov semigroups, Dissertationes Math. 448 (2007), 1–59.

[5] H. Gacki and A. Lasota: A nonlinear version of the Kantorovich-Rubinstein maximum principle, Nonlinear Anal. 52 (2003), 117–125.

[6] H. Gacki: On the Kantorovich-Rubinstein maximum principle for the Fortet-Mourier norm, Ann. Pol. Math. 86.2 (2005), 107–121.

[7] A. Lasota and M. C. Mackey: Chaos, Fractals, and Noise, Springer-Verlag, Berlin 1994.

[8] A. Lasota: Invariant principle for discrete time dynamical systems, Univ. Jagellonicae Acta Math. (1994), 111–127.

[9] A. Lasota: Asymptotic stability of some nonlinear Boltzmann-type equations, J. Math. Anal. Appl. 268 (2002), 291–309.

[10] A. Lasota and J. Traple: An application of the Kantorovich-Rubinstein maximum principle in the theory of the Tjon-Wu equation, J. Differential Equations 159 (1999), 578–596.

[11] A. Lasota and J. Traple: Asymptotic stability of differential equations on convex sets, J. Dynamics and Differential Equations 15 (2003), 335–355.

[12] A. Lasota and J. Traple: Properties of stationary solutions of a generalized Tjon-Wu equation, J. Math. Anal. Appl. 335 No. 1, (2007), 669–682.

[13] S. T. Rachev: Probability Metrics and the Stability of Stochastic Models, John Wiley and Sons, New York 1991.

[14] J. A. Tjon and T. T. Wu: Numerical aspects of the approach to a Maxwellian equation, Phys. Rev. A. 19 (1979), 883–888.

On Modification of Population-Based Approach Used in Adaptive Differential Evolution Algorithm

Petr Bujok

Department of Computer Science, University of Ostrava
30. dubna 22, 70103 Ostrava, Czech Republic
petr.bujok@osu.cz

A new approach for the mutation operation in the differential evolution (DE) algorithm is introduced. The aim of this technique is to enhance the mutation strategy to avoid the local minimum area. The proposed method is implemented to five state-of-the-art DE variants and the standard DE variant DE/rand/1/bin. Twelve DE variants are compared on CEC 2015 problems at four dimension levels. The results show that the proposed method is able to increase the performance of the original DE variants in the significant part of the test problems.

Keywords: Global optimization problem; differential evolution; auxiliary population; experimental comparison; CEC 2015 test suite

1 Introduction

A single-objective global optimization problem with bound constraints is defined as follows. The cost function to be minimized is $f(\vec{x})$, $\vec{x} = (x_1, x_2, \ldots, x_D) \in \mathbb{R}^D$. The domain of feasible solutions Ω is constrained by bounds, a lower limit (a_j) and an upper limit (b_j), $\Omega = \prod_{j=1}^{D}[a_j, b_j]$, $a_j < b_j$, $j = 1, 2, \ldots, D$. The global minimum point $\vec{x}^*$ satisfying condition $f(\vec{x}^*) \leq f(\vec{x})$, $\forall \vec{x} \in \Omega$ is the solution of the problem. The fitness is inversely proportional to the cost function, in the case of the minimization problem.

There is no deterministic algorithm which solves the global optimization problem in a polynomial time in general. Evolutionary algorithms are mostly able to provide an acceptable solution in a reasonable computational time. However, even very sophisticated adaptive evolutionary algorithms cannot guarantee the finding of an acceptable solution in the finite computational time. There are optimization problems, where the state-of-the-art evolutionary algorithms fail. That is why new concepts applied to evolutionary algorithms are intensively studied. Differential evolution (DE) is one of the leading paradigms among the evolutionary algorithms.

In this paper, a new approach for the mutation operation in the DE algorithm is introduced. This approach is focused on the problem when the optimization algorithms

get stuck in the local minimum. The newly proposed approach in the operation of mutation is applied in several well known DE or adaptive DE variants. Selected DE variants and the corresponding counterparts are used to solve the problems of CEC 2015 test suite.

The rest of the paper is organized in the following manner. The basic scheme of the DE algorithm, a brief description of the selected DE variants and some related works are shown in Section 2, 2.1 and 2.2. The new approach for DE mutation is proposed in Section 3. Settings of experiments and their results are given in Section 4 and 5. Finally, conclusions are made in the last Section.

2 Differential Evolution Algorithm

Differential evolution introduced by Storn and Price in [11] is a population-based evolutionary algorithm for problems with a real-valued cost function. The population P of the size N is developed step-by-step from generation to generation.

DE uses evolutionary operators, i.e. mutation, crossover, and selection that are applied in the development of a new generation of P. The DE algorithm is shown in a pseudo-code in Algorithm 1.

Algorithm 1 Differential evolution algorithm

```
initialize population P = {x_1, x_2, ..., x_N}
evaluate f(x_i),    i = 1, 2, ..., N
while stopping condition not reached do
  for i = 1, 2, ..., N do
    create a new trial vector y (mutation and crossover)
    evaluate f(y)
    if f(y) ≤ f(x_i) then
      insert y into Q
    else
      insert x_i into Q
    end if
  end for
  P ← Q
end while
```

The new trial point $\vec{y}$ is created from a mutant point $\vec{u}$ generated by using a kind of the mutation and from the current point of the population $\vec{x}_i$ by the application of the crossover. A combination of the mutation and the crossover variant is usually called a DE strategy. The mutation can cause that a mutant point $\vec{u}$ moves out of the domain Ω. In such a case, the values of $u_j \notin [a_j, b_j]$ are turned over into Ω by using transformation $u_j \leftarrow 2 \times a_j - u_j$ or $u_j \leftarrow 2 \times b_j - u_j$ for the violated component. A better point from the pair of $\vec{x}_i$, $\vec{y}$, based on the value of the cost function, is selected for the new generation (Q).

The most frequently used mutation strategy in DE is rand/1 and it is generated as follows:

$$\vec{u} = \vec{x}_{r1} + F(\vec{x}_{r2} - \vec{x}_{r2}) \tag{1}$$

where $\vec{x}_{r1}, \vec{x}_{r2}, \vec{x}_{r3}$ are three mutually distinct points taken randomly from population P, not coinciding with the current $\vec{x}_i$, and $F > 0$ is a control parameter. The crossover operator constructs the trial vector $\vec{y}$ from current individual $\vec{x}_i$ and the mutant vector $\vec{u}$. There are two types of a crossover and a binomial crossover replaces the elements of vector $\vec{x}_i$ using the following rule:

$$\vec{y}_{i,j} = \begin{cases} \vec{u}_{i,j} & \textit{if } rand_j(0,1) \leq CR \textit{ or } j = rand(1,D) \\ \vec{x}_{i,j} & \textit{otherwise}. \end{cases} \tag{2}$$

where $rand(1,D)$ is random vector uniformly distributed in $[0,1)$ and $CR \in [0,1]$ is a control parameter influencing the number of elements to be exchanged by the crossover. Eq. (2) ensures that at least one element of $\vec{x}_i$ is changed, even if CR= 0. The variant of DE using mutation (1) and binomial crossover, in abbreviation DE/rand/1/bin, is the most frequently used DE strategy in applications.

The DE algorithm has several control parameters whose settings significantly influence the ability to solve different optimization problems. These control parameters and their settings have been intensively studied in recent years. A comprehensive summary of an advanced results in DE research is available in [2, 3, 7], where several kinds of the mutation and the crossover were listed and some adaptive or self-adaptive DE variants were described. Swagantan et al.proposed a wide survey of state of the art of differential evolution algorithm. [4]. No strategy, i.e. a combination of a mutation and crossover variant, is able to outperform all remaining strategies in the case of all optimization problems. This fact corresponds with the so-called No-free-lunch theorem [15]. On the other hand, adaptive variants of DE enable to change DE control parameters during the run of the algorithm to the current problem without trial-and-error tuning of the control parameters [9, 16].

2.1 Adaptive DE Variants

The self-adaptive DE variants (jDE [1], JADE [18], SaDE [10], and EPSDE [6]) are currently considered as the state-of-the-art DE variants and the performance of novel DE variants is compared with these state-of-the-art DE variants in currently appearing studies. These variants are also included in this experiment along with a variant of composite trial vector generation strategies and the control parameters that have been recently published (CoDE) [12].

A simple and efficient adaptive DE variant (mostly called jDE in literature) was proposed by Brest et al. [1]. It uses the DE/rand/1/bin with an evolutionary self-adaptation of F and CR. The pair of these control parameters is encoded with each individual of the population and survives if an individual is successful, i.e. if it

generates such a trial vector which is inserted into the next generation. The values of F and CR are initialized randomly for each point $\vec{x}_i$ in population P and survive with the individuals in the population, but they can be randomly mutated in each generation with given probabilities τ_1 and τ_2.

The differential evolution algorithm with strategy adaptation (SaDE) was introduced by Qin and Suganthan. A more sophisticated and a more efficient variant was proposed later in [10] and it is used in our experimental comparison. Four mutation strategies (rand/1/bin, rand/2/bin, rand-to-best/2/bin, and current-to-rand/1) for creating new trial vectors are stored in a strategy pool. Each strategy has a probability to be selected and applied and these probabilities are updated after each LP generations.

JADE is an algorithm of the adaptive differential evolution, introduced by Zhang and Sanderson in [18]. The original DE concept is extended with three different improvements - current-to-pbest mutation strategy, adaptive control of parameters F and CR, and archive A. The archive A is initialized as an empty set. In every generation, parent individuals are replaced by a better offspring, $(f(\vec{y}) \leq f(\vec{x}_i))$, individuals are put into the archive. After every generation the archive size is reduced to N individuals by randomly dropping surplus individuals.

In the EPSDE adaptive variant [6], an ensemble of mutation strategies and parameter values is applied. The mutation strategies and the values of control parameters are chosen from pools. The combination of the strategies and the parameters in the pools should have diverse characteristics, so that they can exhibit distinct performance during different stages of evolution when dealing with a particular problem. The triplet of (strategy, F, CR) is encoded along with each individual of the population. If the parent vector produces a successful offspring vector, this triplet survives with the trial vector for the next generation and it is also stored. Otherwise, the triplet is randomly reinitialized.

The last adaptive DE variant used in the experiments is DE with composite trial vector generation strategies and control parameters, CoDE, presented by Wang et al. [12]. The results showed that CoDE is at least competitive with the algorithms in the comparison. The CoDE combines three well-studied trial vector strategies with three control parameter settings in a random way to generate trial vectors. The strategies are rand/1/bin, rand/2/bin, and current-to-rand/1 and all the three strategies are applied when generating a new vector (select the offspring vector with the least function from the triplet).

2.2 Related Works

Zamuda and Brest [17] introduced a detail analysis of controlling of the F, CR parameters in a new adaptive DE variant (SPSRDEMMS) derived from jDE [1]. Multi-mutation strategy mechanism is applied to the population size reduction. Applied population-size reduction mechanism decreases a population size to half of a certain size in three defined stages. Moreover, the population of individuals is divided into two sub-populations, superior and inferior ones, with respect to the

function values. Further, a migration of the best individual from the better part to the worse part is occasionally performed. The analysis shows the influence of very small changes of τ_1 and τ_2 on the efficiency of SPSRDEMMS.

Wang et al. proposed in 2016 [13] a new DE variant with enhanced covariance matrix crossover called CPI-DE. In this algorithm, a covariance matrix is computed using the mean vector of the search distribution and it is further updated after each generation. This covariance matrix enables to generate new individuals in the Eigen coordinate system (principal components). Two new individuals are generated for each parent-vector, one in the standard coordinate system and the second one in Eigen coordinate system. The best one from triplet of parent, children1 and children2 is used in next generation. CPI-DE is able to increase performance in two classic DE variant and three state-of-the-art algorithms as it shown on CEC 2013 and CEC2014 test suites at dimension levels $D = 30$ and $D = 50$. CPI-JADE variant is further compared with three various JADE versions applying another covariance-matrix approaches.

Wang et al. study a restrained condition of selecting individuals for mutation in DE [14]. This condition ensures that a randomly selected individuals for mutation are mutually different and that there is no degenerated (zero-valued) differential vector. Authors performed experimental comparison of this restrained condition in six classic DE variants with various mutation and seven state-of-the-art DE algorithm on two benchmark sets (CEC 2005 and CEC 2013). Results show that three out of six classic DE variants perform significantly better when restrained condition is violated. In the case of adaptive DE, only CoDE variant without this restrained condition perform significantly better than the original CoDE algorithm.

Piotrowski in [8] summarized a population size control. Author widely compared several fixed and flexible population size setting in ten various DE algorithms on two various test benchmark sets. Author recommended fixed population size $N = 100$ for artificial problems with smaller dimensions, for middle dimension (approximately $D = 30$) a fixed settings $N = 3D$ or $N = 5D$ are the best choice. For dimension $D = 50$, the best performing population size was found $N = 5D$ or $N = 100$. For the real-world problems with various dimensionality ($1 \leq D \leq 240$) the population size should be set $N = 100$ or $N = 50$. In the case of flexible population size, JADE with reduction of N based on (un-)successes in previous generations was the best performing algorithm. If there is no improvement in 4 generations, 1 % of poorer individuals are removed, and vice versa, if the population is improved, 1 % of newly generated individuals are added.

3 Enhanced Approach for DE Mutation Strategy

A new approach to increase the efficiency of DE algorithm is proposed. The main idea is to enhance the mutation strategy to avoid the local minimum area. A significant problem is that when the population contains some potentially good solutions the population is slowly moved towards these good individuals. Sometimes this fact causes that the solution is detected quickly, but often only a poor local solution is

found. The solution to avoid the population from getting stuck in the local solution area could be in our new DE mutation strategy approach.

Algorithm 2 Differential evolution with enhanced mutation strategy

initialize population $P = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$
evaluate $f(\vec{x}_i), \quad i = 1, 2, \ldots, N$
initialize auxiliary population $R = \{\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_{Nr}\}$, $Nr = N * pr$, $pr \in [0,\ 1]$
while stopping condition not reached **do**
  for $i = 1, 2, \ldots, N$ **do**
    create a mutation vector $\vec{u}$ from P and R (using 3 rules)
    produce a new trial point $\vec{y}$
    evaluate $f(\vec{y})$
    if $f(\vec{y}) \leq f(\vec{x}_i)$ **then**
      insert $\vec{y}$ into Q
    else
      insert $\vec{x}_i$ into Q
      if point from R was used **then**
        reinitialize currently used point from R
      end if
    end if
  end for
  $P \leftarrow Q$
end while

Besides the population P of the N potential solutions another population R is initialized and kept in the initialized form during the search process. Sometimes, one of the vectors in the mutation strategy could be selected from the auxiliary population R, to leave the local solution area.

The situation in Figure 1 illustrates when the points of the population P (circles filled white) get stuck in the local solution area. In spite of the stuck P, the points of the auxiliary population R (circles filled black) are located out from the local solution area and they could move the points of P.

This mutation approach has several parameters whose settings significantly influence the efficiency of this method. The first parameter is the size of the auxiliary population R, denoted Nr. The value of this parameter should be taken from the interval $[1, N]$. A bigger size of R means more places to move in the area Ω. We suppose to compute the value of Nr as a proportion of the population P, $Nr = N \cdot pr$, where pr is a real number from $[0, 1]$.

The next parameter specifies which point(s) of the actually used mutation strategy will be selected from Nr. We suppose a very simple idea based on three rules:

1. only one point in the mutation strategy could be selected from R,
2. if the best point is used in the mutation strategy, it is never taken from R,
3. the first point in the equation of the mutation strategy is never taken from R.

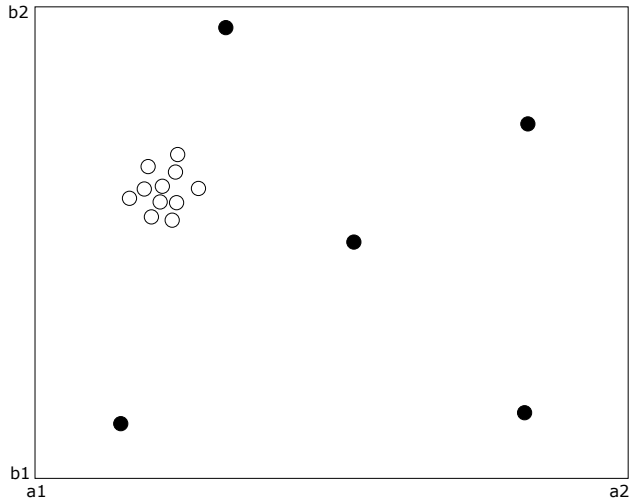

Figure 1
Population located in the local solution area and point of the auxiliary population R

For example, in rand/1 mutation strategy (1) only one of the points in bracket ($\vec{x}_{r2}$ or $\vec{x}_{r3}$) could be taken from R.

The last setting of the proposed method specifies how often the auxiliary population is used. There are many possible ways how to set this parameter, we suppose the simplest one. The point in the mutation strategy which is given to be from R is selected from the union of both populations $P \cup R$ in each step of each generation. It means that only the size of the auxiliary population, pr, could control the frequency of using points from R. The bigger the pr value is, the more frequently points from R are used and vice versa.

When the point $\vec{z}_j$ from the auxiliary population R was used and generate a successful trial individual $(f(\vec{y}) \leq f(\vec{x}_i))$, then $\vec{z}_j$ survives in R. On the other hand, when a new trial individual is generated using some point of R and it is unsuccessful $(f(\vec{y}) > f(\vec{x}_i))$, the used point from R is randomly reinitialized in Ω to move to a better place.

4 Experiments and Settings

The aim of the experiments in this paper is to verify the performance of the proposed enhanced approach in mutation strategy. We apply the proposed method in six DE variants, one standard DE with fixed settings and five adaptive DE variants mentioned in Section 2.1. All algorithms in the experiment are implemented in the Matlab environment.

The test suite of 15 problems was proposed for a special session on Real-Parameter Numerical Optimization, a part of the Congress on Evolutionary Computation (CEC)

2015. This session was intended as a competition of optimization algorithms where new variants of algorithms are introduced. The functions are described in [3] including the experimental settings required for the competition. The source code of the functions is also available on the web site given in this report.

It is expected that CEC test suite will currently become one of the most relevant benchmarks required for publishing new single-objective optimization algorithms. The test functions CEC 2015 are divided into four categories, based on its difficulty (from easy to hard): unimodal functions (F1, F2), multimodal functions (F3 - F5), hybrid functions (F6 - F8) and composition functions (F9 - F15).

Our tests were carried out at four levels of dimension, $D = 10, 30, 50, 100$, with 51 independent runs per each test function. Each point in the population is evaluated by the cost function. The function-error value is computed as the difference between the function value of the current point and the known function value in the global minimum point. The run of the algorithm stops if the prescribed amount of function evaluation $MaxFES = D \times 10^4$ is reached or if the minimum function error in the population is less than 1×10^{-8}. Such an error value is considered sufficient for an acceptable approximation of the correct solution. The search range (domain) for all the test functions is $[-100, 100]^D$.

The parameters of the state-of-the-art DE variants are set to the recommended values. The values of the jDE variant parameters, $\tau_1 = 0.1$, $\tau_2 = 0.1$, the learning period of the SaDE variant, $LP = 50$, the learning period of the EPSDE variant, $LP = N = 100$. The population size of P is set to $N = 100$ for state-of-the-art algorithms and $N = 30$ for standard DE/rand/1/bin variant. The control parameters of F and CR are in standard DE set $F = 0.8$, $CR = 0.8$. The size of the auxiliary population R is set $Nr = N \times pr$, where $pr = 0.05$. Then the value is $Nr = 5$ for the state-of-the-art variants and $Nr = 2$ for the standard DE/rand/1/bin variant.

5 Results

The median values of twelve compared algorithms (six original DE variants and six corresponding counterparts based on the proposed approach) are presented in Tables 3-14. The original variants are denoted based on the abbreviation mentioned above and the names of the new DE variants are enhanced by the text -mut. The median values were computed from 51 independent runs for each algorithm, function and dimension level. The median value for a better variant from the pair original-new is printed in bold.

The performance of the proposed method is compared by Wilcoxon two-sample test. The results of these tests are shown in the column Sign.. The symbol + denotes that new approach outperforms the original DE variant significantly, the symbol - marks better performance of the original DE variant and the symbol $\approx$ is used when there is no significant difference between the DE variants. When the DE variant of the pair is significantly better, the median value is printed in bold and underlined.

For a better comparison of the performance of the proposed approach, the number

of the wins and losses are counted in Table 1. These numbers are counted independently for each pair of algorithm and each dimension level. Based on the percent values of wins we can observe that the enhanced mutation method has the least efficiency for the JADE (30 %) and EPSDE (33 %) variants. Contrary, the most efficiency of the new approach is for the standard DE (62 %), for jDE (48 %) and CoDE (47 %) variants. The overall performance of all 12 algorithms was compared

Table 1
The number of wins and losses of adaptive DE variants and corresponding counterparts

Alg.	$D = 10$	$D = 30$	$D = 50$	$D = 100$	$\sum$	%
CoDE	5	5	7	6	23	38.3
CoDE-mut	6	8	6	8	28	**46.7**
DE	5	6	3	4	18	30
DE-mut	7	8	11	11	37	**61.7**
EPSDE	7	7	7	8	29	**48.3**
EPSDE-mut	4	5	6	5	20	33.3
JADE	4	8	7	12	31	**51.7**
JADE-mut	5	5	6	2	18	30
jDE	4	5	5	5	19	31.7
jDE-mut	7	8	7	7	29	**48.3**
SaDE	4	8	5	7	24	40
SaDE-mut	6	5	8	8	27	**45**

Table 2
Mean ranks from Friedman-rank test results for all the algorithms in comparison

Alg.	$D = 10$	$D = 30$	$D = 50$	$D = 100$	Mean
JADE	4.1	4.7	4.7	4.5	4.5
JADE-mut	3.8	4.9	4.7	5.3	4.7
jDE-mut	5.5	5.5	5.4	4.9	5.3
jDE	5.6	5.7	5.6	4.9	5.5
EPSDE	7.1	5.7	5.5	6	6.1
EPSDE-mut	7.6	5.7	5.6	6.2	6.3
SaDE-mut	5.3	7.6	6.7	6.7	6.6
SaDE	5.5	7.3	7.1	6.7	6.6
CoDE-mut	8.9	6.7	7.4	7.5	7.6
CoDE	9	7.4	7.5	7.5	7.8
DE-mut	7.7	8.4	8.1	8.6	8.2
DE	7.9	8.4	9.7	9.2	8.8

using Friedman test for medians of function-error values. The null hypothesis on the equal performance of the algorithms was rejected, the achieved p value for rejection was $p < 5 \times 10^{-7}$. Mean ranks of the algorithms are presented in Table 2. Note that the algorithm winning uniquely in all the problems has the mean rank 1 and another algorithm being a unique looser in all the problems has the mean rank 12. In the last column of Table 2, the average mean rank is computed for all dimensions. It is

obvious that the least mean rank is for the original JADE variant. Interesting is the fact that all pairs of the original DE and the corresponding counterpart are together. Such a newly proposed DE variant outperforms the original algorithm in four out of six cases.

In some problems with many local-solution areas, DE algorithm often gets stuck. In these situations, the individuals of P are located in very small (local-solution) area and standard DE algorithm is not able to jumped-out. When the individual from the proposed auxiliary R population is in a mutation occasionally applied, it promises to use individuals out of the local solution area. On the other side, provided results show in some problems, that using the auxiliary memory decreases the convergence of DE. It could be caused by the fact that pr is set to fixed value. Some adaptive mechanism for pr value could be helpful.

The value of the fundamental control parameter pr influences the efficiency of algorithm. For smaller pr values, very small auxiliary R population of initialized individuals is used. This means that probability to select some individual of R in the mutation is very small and moving the population from local solution area is rarely. On the other side, when pr is set to big value (i.e. close to 1), the initialized points from R are applied more frequently (almost in each mutation) and the convergence of the algorithm is decreased.

Conclusions

The experimental comparison showed that the newly proposed enhanced mutation variant increased the performance of five state-of-the-art DE variants and also of the standard DE variant in some of the test problems. The number of wins of the new variants was smaller in the case of very efficiency JADE (30 %) and EPSDE (33 %) variants. The best performance was detected in the standard DE/rand/1/bin (62 %), jDE (48 %) or CoDE (47 %) variants.

Based on the Friedman rank test for median values, we can see that the best mean rank was acquired by the original JADE variant. No one DE variant has the best or the worst results for all problems and dimension levels. This fact only validates the No-Free-Lunch theorem [15].

This first study of the proposed mutation method showed that only several initialized individuals kept during the search process could significantly increase the performance of DE. There are many possible parameters to study and improve proposed technique. The study of the control parameters of the auxiliary population remains a challenge for further research.

Acknowledgement

This work was supported by the University of Ostrava from the project SGS08/UVAFM/2016.

Table 3
Medians of function values from 51 runs for DE/rand1/bin and DE/rand1/bin-mut variants and the results of Wilcoxon signed rank test for $D = 10, 30$

	$D = 10$			$D = 30$		
F	DE	DE-mut	Sign.	DE	DE-mut	Sign.
1	6.27E-02	**7.95E-07**	+	2.52E+06	**61214.9**	+
2	0	0	=	171.082	**7.86E-06**	+
3	20.3169	**20.2226**	+	20.9627	**20.5971**	+
4	20.3463	**9.90912**	+	204.622	**45.0517**	+
5	982.36	**411.36**	+	6857.85	**3978.27**	+
6	**0.41629**	1.61939	-	**3789.85**	6690.13	-
7	**0.14615**	0.19463	≈	5.49817	**3.02143**	+
8	**0.32203**	0.468	-	1103.15	**946.748**	≈
9	100.019	**100.018**	≈	**106.192**	106.686	-
10	143.109	**143.108**	≈	**693.532**	735.922	≈
11	**4.31265**	4.35499	≈	**410.239**	459.271	-
12	112.403	**112.239**	≈	112.553	**109.451**	+
13	**0.09273**	0.09715	-	**0.01032**	0.01068	-
14	6677.01	6677.01	≈	**42626.4**	43511.7	-
15	100	100	≈	100	100	≈

Table 4
Medians of function values from 51 runs for DE/rand1/bin and DE/rand1/bin-mut variants and the results of Wilcoxon signed rank test for $D = 50, 100$

	$D = 50$			$D = 100$		
F	DE	DE-mut	Sign.	DE	DE-mut	Sign.
1	4.79E+07	**348268**	+	6.21E+07	**1.34E+06**	+
2	121.174	**3.58E-03**	+	369.562	**36.8171**	+
3	21.1391	**20.8247**	+	21.3243	**21.2112**	+
4	403.504	**87.5563**	+	910.326	**226.249**	+
5	13041.1	**8637.04**	+	30157.3	**25113.5**	+
6	158580	**78616**	+	1.93E+07	**305807**	+
7	42.5321	**41.8693**	≈	137.958	**127.193**	≈
8	25810.7	**20548**	+	3.21E+06	**127927**	+
9	103.113	**102.228**	+	110.593	**108.105**	+
10	3792.33	**1266.27**	+	10286.9	**4194.43**	+
11	**442.227**	709.328	-	**760.799**	1779.12	-
12	201.536	**117.477**	+	200.406	**116.659**	+
13	**0.02498**	0.02607	-	**0.06267**	0.06543	-
14	**52669.4**	52682	-	**108853**	108887	-
15	100	100	≈	**100**	104.402	-

Table 5
Medians of function values from 51 runs for CoDE and CoDE-mut variants and the results of Wilcoxon signed rank test for $D = 10, 30$

	$D = 10$			$D = 30$		
F	CoDE	CoDE-mut	Sign.	CoDE	CoDE-mut	Sign.
1	0	0	≈	**4103.29**	8799.17	-
2	0	0	≈	0	0	≈
3	20.1112	**20.0963**	+	20.6215	**20.5205**	+
4	10.8867	**10.3145**	+	111.928	**101.869**	+
5	443.423	**431.602**	≈	4562.86	**4120.39**	+
6	27.4738	**25.4821**	≈	**758.52**	778.078	≈
7	1.11645	**0.94282**	+	8.44592	**8.41286**	≈
8	**2.32**	2.39408	≈	265.22	**251.065**	≈
9	**100.474**	100.596	-	**107.496**	107.712	-
10	149.322	**147.102**	+	**535.597**	537.875	≈
11	**3.98753**	4.10085	≈	410.262	**301.232**	+
12	**112.796**	112.996	-	110.897	**110.783**	≈
13	**0.09345**	0.09356	≈	**0.0111**	0.01115	≈
14	6662.87	6662.87	≈	42854.9	**42838.2**	≈
15	100	100	≈	100	100	≈

Table 6
Medians of function values from 51 runs for CoDE and CoDE-mut variants and the results of Wilcoxon signed rank test for $D = 50, 100$

	$D = 50$			$D = 100$		
F	CoDE	CoDE-mut	Sign.	CoDE	CoDE-mut	Sign.
1	**277480**	362414	-	1.36E+06	**1.27E+06**	≈
2	**1.96E-08**	1.02E-06	-	**1.67E-06**	2.51E-05	-
3	20.8712	**20.8052**	+	21.1895	**21.1579**	+
4	268.033	**250.965**	+	721.089	**686.159**	+
5	9764.8	**9185.22**	+	25291.2	**25060.5**	+
6	**4122.22**	4687.87	≈	316670	**296073**	≈
7	41.5762	**41.54**	≈	**135.926**	142.072	-
8	**1283.72**	1289.14	≈	**91204.4**	119871	-
9	**102.967**	103.066	-	110.522	**110.429**	≈
10	1877.67	**1732.36**	+	**2946.49**	3029.96	≈
11	**403.616**	417.833	-	**917.103**	1049.05	-
12	117.295	**117.154**	≈	**115.211**	115.435	≈
13	**0.0284**	0.02858	-	0.07484	**0.0704**	+
14	52680.6	52680.6	≈	108891	**108885**	≈
15	100	100	≈	100	100	≈

Table 7
Medians of function values from 51 runs for EPSDE and EPSDE-mut variants and the results of Wilcoxon signed rank test for $D = 10, 30$

	$D = 10$			$D = 30$		
F	EPSDE	EPSDE-mut	Sign.	EPSDE	EPSDE-mut	Sign.
1	0	0	≈	**1125.5**	2224.95	≈
2	0	0	≈	0	0	≈
3	20.1312	**20.1168**	≈	**20.6018**	20.6295	≈
4	11.4291	**11.2616**	≈	124.249	**123.511**	≈
5	**467.604**	495.116	≈	**4885.59**	4989.62	≈
6	**15.3399**	19.4987	≈	**941.091**	990.867	≈
7	0.4456	**0.44256**	≈	**7.08087**	7.22593	≈
8	**0.64177**	0.77123	≈	**333.416**	391.655	≈
9	**100.002**	100.003	-	105.659	**105.444**	≈
10	**141.556**	141.67	≈	**544.848**	551.536	≈
11	3.33224	**3.31824**	≈	404.124	404.124	≈
12	**112.07**	112.185	≈	110.261	**110.226**	≈
13	0.09273	0.09273	≈	0.01036	**0.01032**	≈
14	**6670.66**	6677.01	≈	42793	**42784.3**	≈
15	100	100	≈	100	100	≈

Table 8
Medians of function values from 51 runs for EPSDE and EPSDE-mut variants and the results of Wilcoxon signed rank test for $D = 50, 100$

	$D = 50$			$D = 100$		
F	EPSDE	EPSDE-mut	Sign.	EPSDE	EPSDE-mut	Sign.
1	**175045**	177178	≈	**587919**	609679	≈
2	0	0	≈	0	0	≈
3	20.8877	**20.8681**	≈	**21.2029**	21.2111	≈
4	**286.909**	288.666	≈	**779.656**	779.749	≈
5	10383.8	**10305.2**	≈	27304.9	**27242.8**	≈
6	**2172.48**	2702.65	≈	203293	**175716**	≈
7	**41.0566**	41.1552	≈	136.885	**135.937**	≈
8	**869.352**	1086.5	≈	**35186**	39968.2	≈
9	**102.456**	102.502	≈	**108.614**	108.674	≈
10	1003.44	**938.6**	≈	**3020.22**	3281.25	≈
11	439.034	**437.629**	≈	**854.065**	862.279	≈
12	201.536	**117.442**	+	200.406	**117.141**	≈
13	0.02516	**0.02502**	≈	0.06269	**0.06239**	≈
14	**52662.7**	52678.2	≈	**108865**	108870	≈
15	100	100	≈	100	100	≈

Table 9
Medians of function values from 51 runs for JADE and JADE-mut variants and the results of Wilcoxon signed rank test for $D = 10, 30$

	$D = 10$			$D = 30$		
F	JADE	JADE-mut	Sign.	JADE	JADE-mut	Sign.
1	0	0	≈	**1.31028**	4.04967	-
2	0	0	≈	0	0	≈
3	20.0579	**20.052**	≈	**20.2779**	20.2812	≈
4	3.55754	**3.48658**	≈	26.5575	**25.3809**	≈
5	**52.7562**	57.1781	≈	**1699.69**	1734.4	≈
6	0.41629	**0.41629**	≈	**941.019**	1019.68	≈
7	**0.30991**	0.31690	≈	7.87472	**7.66545**	+
8	**0.53595**	0.57577	≈	**219.249**	251.085	≈
9	100	100	≈	106.547	**106.395**	≈
10	143.108	143.108	≈	**715.256**	742.593	≈
11	**2.97343**	3.02492	≈	**408.796**	408.851	≈
12	111.798	**111.767**	≈	**108.972**	109.142	≈
13	0.09273	**0.09273**	≈	0.01049	**0.01041**	≈
14	6670.66	6670.66	≈	43620	**43410.6**	≈
15	100	100	≈	100	100	≈

Table 10
Medians of function values from 51 runs for JADE and JADE-mut variants and the results of Wilcoxon signed rank test for $D = 50, 100$

	$D = 50$			$D = 100$		
F	JADE	JADE-mut	Sign.	JADE	JADE-mut	Sign.
1	7272.18	**5756.84**	≈	**76376.1**	86061.3	≈
2	0	0	≈	0	0	≈
3	**20.3602**	20.3676	≈	**20.4627**	20.4655	≈
4	**51.7394**	57.042	-	**154.44**	157.685	≈
5	3532.94	**3486.93**	≈	**10313.6**	10331	≈
6	2661.89	**2537.13**	≈	**11887.2**	13762	≈
7	42.5838	**42.2447**	≈	**116.745**	127.462	≈
8	**1197.8**	1215.5	≈	**3835.97**	3945.31	≈
9	**102.66**	102.765	≈	**110.555**	111.139	-
10	1760.11	**1572.17**	+	**4361.36**	4597.21	≈
11	522.613	**513.155**	≈	**1259.63**	1298.47	≈
12	**116.342**	116.451	≈	116.311	**115.426**	≈
13	**0.0255**	0.02554	≈	0.0633	**0.06324**	≈
14	**52678.2**	52680.6	≈	**108881**	108887	≈
15	100	100	≈	**101.302**	101.463	≈

Table 11
Medians of function values from 51 runs for jDE and jDE-mut variants and the results of Wilcoxon signed rank test for $D = 10, 30$

	$D = 10$			$D = 30$		
F	jDE	jDE-mut	Sign.	jDE	jDE-mut	Sign.
1	**0**	5.88E-08	-	34246.4	**31656.3**	≈
2	0	0	≈	0	0	≈
3	20.0689	**20.0629**	≈	**20.3096**	20.3108	≈
4	5.45472	**5.39326**	≈	**39.9106**	40.8473	≈
5	231.063	**195.111**	≈	2428.06	**2422.21**	≈
6	9.86734	**9.59008**	≈	**1018.45**	1550.9	-
7	0.36721	0.36984	≈	8.56411	**8.40703**	≈
8	0.47653	0.44016	≈	219.256	**199.384**	≈
9	100.007	100.01	≈	106.102	**106.072**	≈
10	141.618	141.548	≈	637.044	**635.476**	≈
11	3.19005	3.18598	≈	**404.124**	408.696	≈
12	112.458	112.554	≈	109.716	**109.524**	+
13	0.09273	0.09273	≈	**0.01033**	0.01035	≈
14	6662.87	6662.87	≈	43419.2	**43410.6**	≈
15	100	100	≈	100	100	≈

Table 12
Medians of function values from 51 runs for jDE and jDE-mut variants and the results of Wilcoxon signed rank test for $D = 50, 100$

	$D = 50$			$D = 100$		
F	jDE	jDE-mut	Sign.	jDE	jDE-mut	Sign.
1	396624	**374644**	≈	1.43E+06	**1.42E+06**	≈
2	0	0	≈	0	0	≈
3	20.4252	**20.4199**	≈	20.6601	**20.6526**	≈
4	83.9694	**83.924**	≈	225.627	**199.583**	+
5	**4638.29**	4641.82	≈	12710.9	**12530.1**	≈
6	**30558.7**	38689.4	≈	**188422**	226244	-
7	**43.0974**	44.4463	≈	**134.681**	135.603	-
8	**8003.51**	9033.64	≈	79939.2	**76601.8**	≈
9	**102.194**	102.199	≈	**107.077**	107.225	≈
10	1428.24	**1325.64**	≈	**3791.2**	3863.28	≈
11	488.976	**467.289**	≈	1079.89	**946.754**	≈
12	116.835	**116.831**	≈	114.994	**114.9**	≈
13	0.02497	**0.02491**	≈	**0.06174**	0.0619	≈
14	52661	52661	≈	108887	108887	≈
15	100	100	≈	100	100	≈

Table 13
Medians of function values from 51 runs for SaDE and SaDE-mut variants and the results of Wilcoxon signed rank test for $D = 10, 30$

	$D = 10$			$D = 30$		
F	SaDE	SaDE-mut	Sign.	SaDE	SaDE-mut	Sign.
1	0	0	≈	**7216.46**	7438.64	≈
2	0	0	≈	0	0	≈
3	**20.0823**	20.0842	≈	**20.385**	20.3967	≈
4	4.83555	**4.44725**	≈	34.7638	**33.6348**	≈
5	180.634	**161.037**	≈	**2674.22**	2725.79	≈
6	**4.68252**	6.05311	≈	**1255.06**	1457.19	≈
7	0.38712	**0.33234**	≈	8.19227	**8.09298**	≈
8	1.26452	**1.11676**	+	**302.761**	410.456	≈
9	100	100	≈	106.853	**106.519**	≈
10	**141.655**	143.109	≈	814.879	**798.48**	≈
11	**3.16052**	3.18455	≈	**417.978**	426.424	≈
12	111.984	**111.979**	≈	**109.505**	109.614	≈
13	0.09273	0.09273	≈	0.01044	**0.01041**	≈
14	6677.01	**6670.66**	≈	**43610.6**	43806.1	≈
15	100	100	≈	100	100	≈

Table 14
Medians of function values from 51 runs for SaDE and SaDE-mut variants and the results of Wilcoxon signed rank test for $D = 50, 100$

	$D = 50$			$D = 100$		
F	SaDE	SaDE-mut	Sign.	SaDE	SaDE-mut	Sign.
1	**86925.6**	89696.9	≈	**419702**	513455	-
2	0	0	≈	1.88E-06	**2.22E-07**	+
3	**20.5758**	20.577	≈	**20.8954**	20.9055	-
4	92.14	**80.3453**	+	262.669	**190.037**	+
5	6012.55	**5785.05**	≈	17364.8	**17195.9**	+
6	**15052**	16486.5	≈	**84926.9**	87087.2	-
7	49.6038	**49.5846**	≈	**113.103**	142.63	≈
8	**3334.2**	3442.65	≈	**26663**	37648.8	-
9	102.431	**102.334**	+	108.714	**108.285**	+
10	**1734.43**	1879.22	≈	**3801.22**	4080.34	-
11	681.752	**655.88**	+	1860.02	**1765.77**	+
12	116.769	**116.624**	≈	115.608	**115.312**	+
13	0.02553	**0.02545**	≈	0.06383	**0.06324**	+
14	52698	**52693.6**	≈	108912	**108895**	+
15	100	100	≈	**101.902**	101.976	-

References

[1] J. Brest, S. Greiner, B. Boškovič, M. Mernik and V. Žumer: Self-adapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems, *IEEE Transactions on Evolutionary Computation*, vol. 10, pp. 646–657, 2006.

[2] P. Bujok and J, Tvrdík: A Comparison of Various Strategies in Differential Evolution, In: *R. Matoušek (Ed.) MENDEL 2011 - 17th International Conference On Soft Computing, Brno, Czech Republic*, pp. 48–55, 2011.

[3] S. Das and P. N. Suganthan: Differential evolution: A survey of the state-of-the-art, *IEEE Transactions on Evolutionary Computation*, vol. 15, pp. 27–54, 2011.

[4] S. Das, S. S. Mullick, P. N. Suganthan: Recent advances in differential evolution An updated survey, *Swarm and Evolutionary Computation*, Vol. 27, pp. 1–30, 2016.

[5] J. J. Liang, P. N. Suganthan, and Q. Chen: Problem definitions and evaluation criteria for the CEC 2015 competition on learning-based real-parameter single objective optimization, Computational Intelligence Laboratory, Zhengzhou University, Zhengzhou China and Nanyang Technological University, Tech. Rep., 2014.

[6] R. Mallipeddi, P. N. Suganthan, Q. K. Pan and M. F. Tasgetiren: Differential evolution algorithm with ensemble of parameters and mutation strategies, *Applied Soft Computing*, vol. 11, pp. 1679–1696, 2011.

[7] F. Neri and V. Tirronen: Recent advances in differential evolution: a survey and experimental analysis, *Artificial Intelligence Review*, vol. 33, pp. 61–106, 2010.

[8] A. P. Piotrowski: Review of Differential Evolution population size, *Swarm and Evolutionary Computation*, Vol. 32, pp. 1–24, 2017.

[9] W. Qian and A. Li: Adaptive differential evolution algorithm for multiobjective optimization problems, *Applied Mathematics and Computation*, vol. 201, no. 12, pp. 431–440, 2008.

[10] A. K. Qin, V. L. Huang, and P. N. Suganthan: Differential evolution algorithm with strategy adaptation for global numerical optimization, *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 2, pp. 398-417, 2009.

[11] R. Storn and K. V. Price: Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces, *Journal of Global Optimization*, vol. 11, pp. 341–359, 1997.

[12] Y. Wang, Z. Cai, Q. Zhang: Differential evolution with composite trial vector generation strategies and control parameters. *IEEE Transactions on Evolutionary Computation*, vol. 15, pp. 55-66, 2011.

[13] Y. Wang, Z.-Z. Liu, J. Li, H.-X. Li, G. G. Yen: Utilizing cumulative population distribution information in differential evolution, *Applied Soft Computing*, Vol. 48, pp. 329–346, 2016.

[14] Y. Wang, Z.-Z. Liu, J. Li, H. X. Li and J. Wang: On the selection of solutions for mutation in differential evolution. *Frontiers of Computer Science*. In press, DOI: 10.1007/s11704-016-5353-5.

[15] D. H. Wolpert and W. G. Macready: No free lunch theorems for optimization, *IEEE Transactions on Evolutionary Computation*, vol. 1, pp. 67–82, 1997.

[16] Z. Yang, K. Tang, and X. Yao: Self-adaptive differential evolution with neighborhood search, *IEEE Congress on Evolutionary Computation, CEC 2008*, pp. 1110–1116, 2008.

[17] A. Zamuda and J. Brest: Self-adaptive control parameters? randomization frequency and propagations in differential evolution, *Swarm and Evolutionary Computation*, vol. 25, pp. 72–99, 2015.

[18] J. Zhang and A. C. Sanderson: JADE: Adaptive differential evolution with optional external archive, *IEEE Transactions on Evolutionary Computation*, vol. 13, pp. 945–958, 2009.

Computer-aided Diagnostics of Schizophrenia: Comparison of Different Feature Extraction Methods

Radomír Kůs, Daniel Schwarz

Institute of Biostatistics and Analyses, Masaryk University
Kamenice 3, 62500 Brno, Czech Republic

E-mail: radomir.kus@mail.muni.cz, schwarz@iba.muni.cz

Abstract: Receiving an early diagnosis of schizophrenia is a crucial step towards its treatment. However, in current thinking, the diagnosis is based on time-consuming criteria, burdened with subjectivity. Hence, objective and more reliable therapeutic tests are desirable for the clinical practice of Psychiatry. Since schizophrenia is characterized by progressive brain volume changes during the course of the disease, many studies have recently turned attention to machine learning and brain morphometric techniques serving as tools for computer-aided diagnosis of schizophrenia based on neuroimaging data. In our study, the methodology is applied to distinguish between 52 first-episode schizophrenia patients and 52 healthy volunteers on the basis of T1-weighted magnetic resonance images of their brains preprocessed by the means of voxel-based and deformation-based morphometry. The proposed classification schemes vary in the feature extraction and selection steps. Namely, Mann-Whitney testing is implemented as a simple univariate approach playing the role of a comparator to multivariate methods such as inter-subject PCA, the K-SVD algorithm, and pattern-based morphometry. The highest classification accuracy, 70%, is reached with the pattern-based morphometry technique. The study points out the difference between univariate and multivariate approaches towards neuroimaging data. Additionally, the contrast between feature extraction capabilities of voxel-based and deformation-based morphometry is demonstrated.

Keywords: feature extraction; computer-aided diagnosis; schizophrenia; brain morphometry; voxel-based morphometry; deformation-based morphometry; magnetic resonance imaging; classification; machine learning

1 Introduction

As schizophrenia worsens with the progression of the disease [1], its early and accurate diagnosis can be beneficial for patient prognosis and overall treatment strategies [2]. Unfortunately, since psychiatry deals with mental states of patients,

its measurement techniques (such as Schneiderian First-Rank Symptoms) evaluate general symptoms common to a variety of mental disorders rather than the specific ones.

In the case of schizophrenia, the final verdict is partly based on observing patient's actions and noting the constellation of patient's symptoms, partly on psychiatric rating systems, and diagnostic classification and rating scales. Thus, most of the diagnoses are dependent on a subjective perspective and judgment of the psychiatrist assessing a patient [3], leading to the situation when more sophisticated methods, taking into account more aspects than a naked eye does, are desired. For instance, the changes in the brain morphology are only subtle during the first episode, and hence often indistinguishable even by an experienced psychiatrist.

Consequently, many studies have recently turned attention to Magnetic Resonance Imaging (MRI) as it can be utilized to explore the structure of the brain and to understand better the neurobiology of brain disorders. Moreover, neuroimaging data can be to a lesser or greater extent [4] successfully exploited as an input data to machine learning techniques which attempt to reduce or completely eliminate the need for human intuition in the analysis of the neuroimaging data.

2 Schizophrenia through the Optics of MRI

As the evidence of pathological changes in the brain morphology of schizophrenia patients exists [1], the researchers have started exploiting MRI data as a base to the disorder diagnosis. Should the schizophrenia-related manifestation in brain structure be profoundly understood, predictions pertaining to patients diagnoses could be made on the basis of an individual brain MR scan.

However, such an approach faces two confronting requirements where, on the one hand, all the information crucial for the classification must be retained in the data. On the other hand, a successful diagnosis – in terms of an accurate classification – is feasible only when the dimensionality of the problem is properly reduced as the brain image classification algorithms fail to operate on data exhibiting an adverse ratio between its dimensionality and the number of acquired samples [5].

Many studies have attempted to find the connection between schizophrenia neuropathology and brain structure. Although various brain regions have been identified, the results are not entirely consistent [6-9]. Nevertheless, even though the general consensus upon what brain structures are affected in schizophrenia is yet to be achieved, it has been revealed that structural changes happen in both gray [9] and white matter [10] and that these changes are not bound to a specific region but rather they are distributed throughout the entire brain following spatially complex and unknown patterns.

Thus, reduction of brain image data dimensionality using regions of interest (ROIs) methods may be misleading as they are prone to human error due to manual brain segmentation [11]. In contrast, automated and whole-brain morphometry methods, such as voxel-based morphometry (VBM) and deformation-based morphometry (DBM), are two main concepts used by the neuroimaging community for assessing MRI brain scans without the need to limit the analysis to arbitrarily predefined anatomical hypotheses or ROIs.

Although both of the methods are designed to assess the brain structure, they differ in workflow and interpretation. Whereas VBM segments the images in order to generate gray matter (GM) maps and uses low-dimensional registration, DBM utilizes full brain scans and the employed registration algorithm is high-dimensional [12-13]. The images resulting from those techniques are interpreted as local GM volume and local brain volume changes respectively. Their application is therefore advised to be chosen accordingly, with the knowledge of the disease process [14].

A strong critique of those techniques stems from the fact that, as they deal with brain images on a voxel-to-voxel basis, they neglect multivariate group differences [15]. Advantageously, multivariate approaches known from machine learning can be employed to conjointly account for voxels interactions [16] hence extracting complex patterns suitable for schizophrenia classification, leaving the brain morphometry methods a place among data preprocessing tools rather than feature extraction techniques.

3 Research Problem and Proposed Methodology

Typically, studies on computer-aided diagnostics of schizophrenia employ several machine learning algorithms in order to achieve the highest classification accuracy. However, as classification performance considerably depends on a preceding feature extraction step, an equal effort should be made in finding what algorithms are the most suitable for each application domain.

Thus, the purpose of this study is to elicit conditions under which one feature extraction method outperforms the other and vice versa. Namely, multivariate machine learning methods are put in contrast with univariate statistics. The comparison takes place on a dataset of first-episode schizophrenia patients preprocessed by the means of VBM and DBM separately, allowing us also to comment on whether and how the brain morphometry techniques influence the ensuing feature extraction step.

The whole study is organized accordingly to a well-known scheme in schizophrenia classification, starting by presenting a dataset and its preprocessing (Section 3.1), describing employed feature extraction algorithms (Section 3.2) and

a classification pipeline utilized to evaluate the performance of all the algorithms (Section 3.3).

As we aim at analyzing and comparing feature extraction algorithms, we include a short experiment about the anticipated behavior of the algorithms performed on a synthetic dataset (Section 4.1). Next, an elaboration on parameters of the algorithms and the process of their tuning is stated (Section 4.2). Last, the classification results are introduced (Section 5), followed by a commentary on capabilities of the feature extraction algorithms (Section 6).

3.1 Datasets

3.1.1 Subjects

The datasets consisted of 104 individual T1-weighted MRI whole-head scans, where exactly one-half of the scans belonged to 52 first-episode schizophrenia patients (FES) who were recruited at the Department of Psychiatry, Masaryk University in Brno. The patients were all male with the mean age of 24 years (±5.1). The diagnosis was based on diagnostic interviews regarding patient's history, substance abuse, etc., and evaluated using the Positive and Negative Syndrome Scale (PANSS [17]). A senior psychiatrist reviewed the tests and, in compliance with International Statistical Classification of Disease and Related Health Problems (ICD-10), established the diagnosis. Additionally, the patients were physically examined and, given specific criteria such as suffering from another neurological disease, substance dependence, etc. were met, excluded from the study.

The other 52 scans were acquired from volunteering healthy controls (HC) whose mean age (24 ± 3.1 years) and handedness matched with the patients.

3.1.2 Acquisition & Preprocessing

The images were obtained using a 1.5 T MR scanner with a resolution of 160×512×512 voxels per scan and, subsequently, using the VBM8 toolbox available in the SPM8 Matlab software package, they were corrected for bias-field inhomogeneity and spatially normalized by affine co-registration to the standard SPM T1 template.

Acquired and co-registered images were preprocessed correspondingly to VBM and DBM approaches resulting in two datasets in here referred to as GM Densities and Volume Changes.

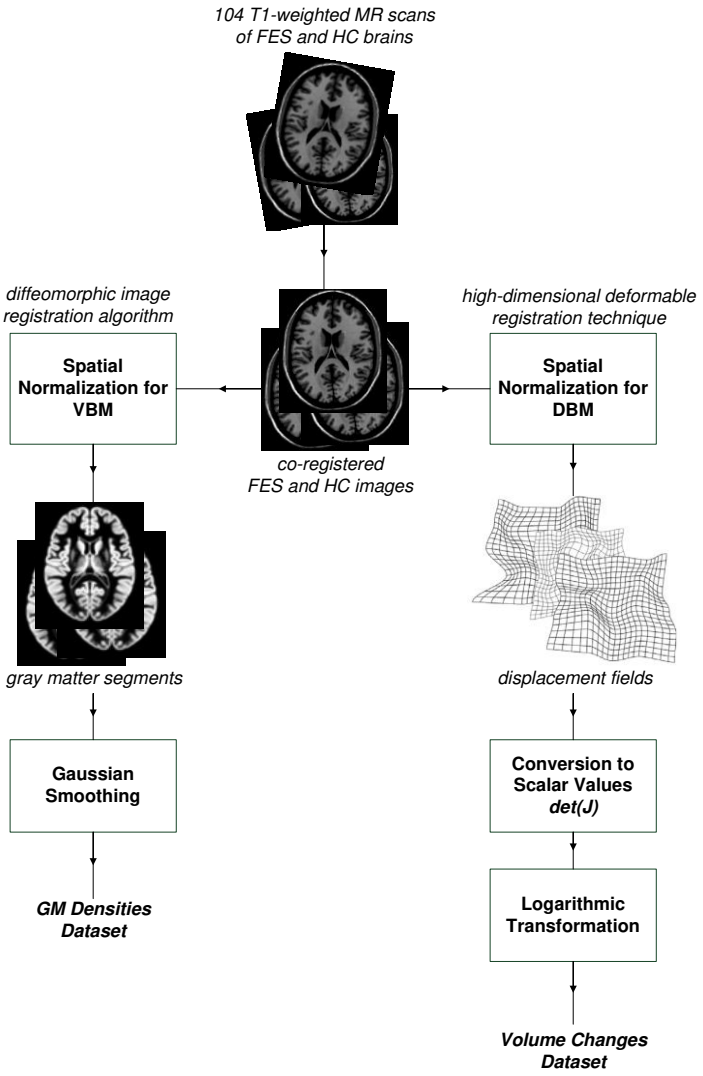

Figure 1
Scheme of the Datasets Preprocessing

In order to create the GM densities dataset, additional steps needed to be performed following the VBM pipeline. After the affine registration of the T1-weighted images, the images were non-linearly registered using fast diffeomorphic image registration algorithm (DARTEL [18]). Resulting GM tissue segments were modulated with the determinant of Jacobian matrices of the deformations to account for registration related changes in local volumes.

Subsequently, the modulated GM segment images were smoothed with the 8 mm FWHM Gaussian kernel to enable inter-subject comparisons.

As the Volume Changes dataset resulted from the DBM method, it was based on an additional spatial normalization which output displacement fields referring to volume adjustments needed for each image to match the template. Thus, after the images were normalized to the same stereotactic space, a high-dimensional deformable registration [19] was performed. The obtained 3-D displacement fields were converted to scalars by computing the Jacobian determinants at each voxel. Additionally, the scalar values were logarithmically transformed in order to distribute the values symmetrically around zero instead of an asymmetric distribution of solely positive values which the determinant of Jacobian matrix normally yields.

For better illustration, the datasets preprocessing is schematically depicted in Figure 1. The same datasets have been successfully used in a previous study [11].

3.2 Feature Extraction Methods

3.2.1 Univariate statistics (Mann-Whitney testing)

In order to reveal structural differences between schizophrenic and healthy brains, both VBM and DBM utilize a voxel-wise comparison between the groups, i.e. they employ univariate statistical analysis [12-13]. Therefore, Mann-Whitney testing (MW) was implemented as a univariate approach playing the role of a comparator to multivariate feature extraction methods.

MW indicates whether the tested variables come from the same distribution. Applying MW on each voxel, we selected those voxels which statistically belonged to different populations, i.e. they were important for distinguishing between FES and HC.

In general, when testing multiple hypotheses, one should correct for the number of false discoveries either with the familywise error rate (FWER [20]) or the false discovery rate (FDR [21]) corrections. However, those techniques are often too stringent [13]. Moreover, statistical significance does not necessarily imply discriminative power. Therefore, we regarded the resulting p-values as a selection criterion rather than a level of significance. In other words, we manually set the threshold for p-values dividing the voxels to those which were to be incorporated into classification and which were to be disregarded.

3.2.2 Intersubject PCA (isPCA)

Principal component analysis (PCA [22]) is a classic multivariate procedure seeking a transformation converting data to a set of orthogonal principal

components ordered according to the amount of variance they explain in the original data. However, PCA requires a covariance matrix of descriptors to be computed which, in the case of our data, was not feasible since the number of voxels in each image was over a half of a million.

Fortunately, it has been proven [23-24] that the eigenvectors v_j, corresponding to new components, can be computed from the eigenvectors w_j of a covariance matrix of subjects as:

$$v_j = \frac{X^T w_j}{\sqrt{q_j (N-1)}}$$

greatly reducing the demands on computation. The matrix X^T represents a transposed data matrix containing N subjects and q_j are the eigenvalues of the intersubject covariance matrix. Such a method, later named intersubject PCA (isPCA [25]), allowed us to preserve all the dataset variability using solely $N–1$ eigenvectors.

The feature space dimensionality can be progressively reduced by disregarding some of the new components. Since the eigenvectors are sorted in an ascending order of explained data variance, at first sight it may be tempting to get rid of the last ones. However, the amount of explained variance does not necessarily imply schizophrenia-related differences between FES and HC and therefore just as the first component can be, for instance, related to differences in liquor, the last component might be crucial for recognizing the proper affiliation of the subject.

Thus, before removing components from the ensuing analysis, we sorted the components according to their discriminative power measured by the level of their significance once tested with the subjects projected into the new feature space spanned by the components.

3.2.3 K-SVD

The aim of K-SVD [26] is to find the best sparse representation of the images x_i captured in X by solving

$$\min_{\Phi,C} \| X - \Phi C \|_2^F \ \textit{subject to} \ \| c_i \|_0 \leq s$$

where $i \in \{1,..,N\}$ and $\| . \|^F$ is the Frobenius norm.

Firstly, the dictionary Φ is initialized with l_2-normalized columns. The subsequent optimization process iteratively alternates between the sparse coding phase, when the optimization of each sparse coefficient vector c_i takes place, and the dictionary update phase. Here, for every atom in the dictionary, an error matrix representing the error of discarding the atom from the dictionary is computed, restricted to the

columns that correspond to non-zero sparse coefficients and finally it is decomposed using singular value decomposition (SVD). The update of both the dictionary and the loading matrix C is dependable on the matrices resulting from the SVD factorization.

When applied to brain imaging data, resulting atoms in the dictionary represent complex morphological patterns revealed by the algorithm in the brain scans. As the sparsity constraint s controls for the maximum number of atoms utilized to compose each image, its adjustment allows for extraction of small regions as well as global patterns [27].

Again, in order to gain the best set of atoms for schizophrenia diagnostic inference, the atoms can be sorted and the least discriminative ones can be discarded. However, due to the optimization process, we decided not to discard any atoms once they were learned.

3.2.4 Pattern-based Morphometry (PBM)

Although the K-SVD algorithm has emerged relatively recently, it has already been incorporated into a new methodology, pattern-based morphometry (PBM [27]). Despite its name, it is not a morphometry preprocessing technique as VBM and DBM described above. Instead, it provides a new perspective to multivariate pattern extraction using K-SVD, which is why we categorize it as a feature extraction method.

Unlike in the above-mentioned case where the dictionary is built upon data matrix of images, PBM introduces the idea of generating atoms from the so-called difference images. The generation of a difference images matrix is diagrammatically depicted in Figure 2.

For each image, using the Euclidean distance, a set of its k-nearest neighbors with a different affiliation is found. In other words, for an image a belonging to the group A (e.g. FES), its k most similar images belonging to the group B (e.g. HC) are searched for and vice-versa. Subsequently, the images are subtracted from their neighbors N. In the end, the resulting difference images matrices D_A and D_B are put together into a single matrix X. Assuming the images are in columns, the new matrix will have k-times more columns than the original matrix. At this point, the extracted atoms straightforwardly represent structural changes between FES and HC.

We created the new dictionary accordingly and used it in the same manner as with the K-SVD algorithm.

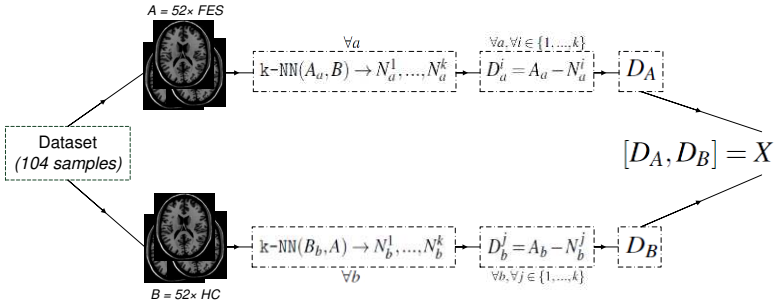

Figure 2
Generation of the Difference Images Matrix

3.3 Classification Pipeline

In order to evaluate the algorithms in real situations, they were incorporated into a classification pipeline (Figure 3).

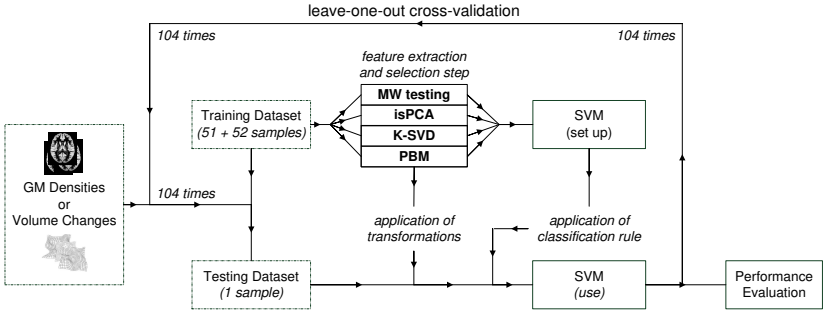

Figure 3
Classification Pipeline

Leave-one-out cross-validation scheme was used to assess the performance of the linear SVM classifier based on features extracted from the GM densities or the Volume changes datasets. Note that the feature extraction and selection steps were performed for each iteration of the cross-validation.

The rationale of the design was that alternating only the datasets and feature extraction methods in otherwise rigidly fixed pipeline settings facilitated their later comparison.

In terms of the performance evaluation, classification accuracy, sensitivity and specificity were used as its metrics.

4 Preliminary Experiments

4.1 Anticipated Behavior – a Toy Example

Before proceeding to classification on real datasets, we created a synthetic dataset consisting of 2-D images of 10 hand-drawn circles and 10 hand-drawn triangles in order to illustrate the difference between the behavior of univariate and multivariate feature extraction approaches.

Each image in the synthetic dataset consisted of 50,184 pixels with values ranging from 0 to 255. Figure 4 shows the pixels selected by MW and the most discriminative patterns revealed by isPCA, K-SVD, and PBM respectively from left to right. The gray scale patterns are displayed in colors, where yellow represents the most and dark blue the least significant pixels.

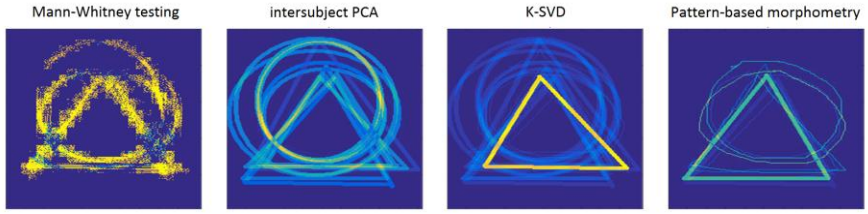

Figure 4
Features Extracted from the Synthetic Dataset

The toy example underlines what is known from the theory. Whereas univariate statistics dismantled geometric shapes into pixels, multivariate methods were capable of recognizing complex patterns[1] while dealing with the same data. Moreover, in the case of K-SVD and PBM, the most discriminative feature resembled a representative from the group of triangles.

4.2 Parameters Tuning

The last step preceding final classification was tuning the parameters of the employed algorithms as their proper settings enhance the classification performance. Table 1, summarizes the parameters included for tuning.

As their influence on the classification was unknown, we evaluated the classification cross-validated accuracies for various parameters settings equidistantly distributed over the parameter space in a way to capture a behavior of each of the parameters separately for the GM Densities and the Volume

[1] whole circles and triangles

Changes datasets, hence mapping the parameter spaces. In order to reduce computational costs, random projection (RP), with a random matrix suggested in [28], was utilized to reduce the dimensionality of the problem.

Table 1
List of Parameters of the Feature Extraction Algorithms

Algorithm	Parameter	Token
MW	p-values threshold	t
isPCA	number of retained components	c
K-SVD	number of atoms	$a_{k\text{-}svd}$
	sparsity constraint	$s_{k\text{-}svd}$
PBM	number of atoms	a_{pbm}
	sparsity constraint	s_{pbm}
	number of nearest neighbors	k

4.3 Final Parameters Settings

The parameters settings reaching the highest classification accuracies are displayed in Table 2. As sparsity constraint and the number of nearest neighbors did not exhibit any trend, we set them, in accordance with [27], to 5 and 3 correspondingly.

Table 2
List of Final Parameters Settings for Both of the Datasets

Algorithm	Token	Value	
		GM Densities	*Volume Changes*
MW	t	0.01	0.05
isPCA	c	11	84
K-SVD	$a_{k\text{-}svd}$	1	103
	$s_{k\text{-}svd}$	5	5
PBM	a_{pbm}	1	309
	s_{pbm}	5	5
	k	3	3

5 Classification Results

All tested feature extraction algorithms with their final parameters settings being put through the classification pipeline. Cross-validated classification accuracies along with sensitivities and specificities for each of the methods and both the datasets are shown in Table 3.

On average, the classification methods with the use of Volume Changes features resulting from DBM outperformed the classification methods using the GM Densities features resulting from VBM. Comparing the classification algorithms, the highest accuracy, slightly over 70%, was attained by PBM.

Whereas the across-datasets performance increased for all multivariate methods when switched from GM Densities to Volume Changes, it diminished for the univariate statistics.

Table 3
Classification Performance on Both Datasets

	GM Densities			***Volume Changes***		
Algorithm	Accuracy [%]	Sensitivity [%]	Specificity [%]	Accuracy [%]	Sensitivity [%]	Specificity [%]
MW	67.31	63.46	71.15	66.35	65.38	67.30
isPCA	68.27	63.46	73.08	69.23	71.15	67.31
K-SVD	65.38	63.46	67.31	69.23	69.23	69.23
PBM	64.42	63.46	65.38	**70.19**	**69.23**	**71.15**

6 Discussion

The main distinction we would like to stress here is the difference for the results in the different types of features: GM Densities and Volume Changes. Considering the datasets are two modalities of the same data, we were able to evaluate the differences between the VBM and DBM approaches to the preprocessing of the MRI data.

The most noteworthy piece of information stems from the parameter settings, indicating the number of features (components, atoms) that are optimal for the classification. In the case of the GM Densities dataset, the best classification results were achieved with the minimum of features retained. On the contrary, the Volume Changes dataset yielded the best results when the number of features was set at its highest values.

Also, the most discriminative isPCA component of GM Densities captured 12.5 times more variance of the original data than the one calculated from the covariance matrix corresponding to deformations. When comparing components with the most variance explaining the ratio was approximately 2.5. Such findings are in correspondence with [25], where isPCA components are evaluated in more detail.

Furthermore, for the Volume Changes dataset, multivariate approaches slightly outperformed univariate MW, serving as a mere feature selection. However, the differences are not statistically significant.

All the aforementioned behavior indicates that Volume Changes concealed more sophisticated patterns, than can be discovered, disregarding voxel-to-voxel interactions. Consequently, our results confirm that whereas VBM serves mainly for extracting information about changes on a local scale, DBM preserves information from a wider region.

At this point, it should be stressed that accuracies around 70% are insufficient for clinical practice. Nevertheless, our findings can serve as guidelines to those dealing with unknown parameter spaces. With VBM, the best parameter settings in terms of the number of retained features will most likely lay among low values. In the case of DBM, the opposite statement is the most probable.

We also suggest that studies utilizing DBM as a preprocessing tool should reach for multivariate feature extraction approaches as they appear to be superior on such data. Interestingly, a novel PBM technique provided superb results in comparison to others and thus it should be considered as a valid candidate when deciding on a method of extracting brain differences patterns. Moreover, PBM improves the ratio of the number of subjects over the number of features, as it generates a dataset consisting of more images.

Conclusions

This work presented an analysis of two brain morphometry techniques and various feature extraction methods often utilized in the computer-aided diagnostics of schizophrenia. The methodology was incorporated into a classification pipeline and applied to distinguish between first-episode patients and healthy controls on the basis of magnetic resonance images of their brains. First, each method was thoroughly examined in order to explore its parameters and their influence on the classification. Then, the methods were evaluated in terms of classification performance. Our findings confirmed the distinction between VBM and DBM and resulted in recommendations on the numbers of retained features. We also showed that by applying multivariate machine learning techniques, such as, PBM on data preprocessed with the DBM approach have beneficial effects on classification results.

References

[1] N. E. van Haren, W. Cahn, H. E. H. Pol, and R. S. Kahn: The Course of Brain Abnormalities in Schizophrenia: Can We Slow the Progression?, Journal of Psychopharmacology, Vol. 26, No. 5 suppl, 2012, pp. 8-14

[2] D. O. Perkins, H. Gu, K. Boteva, and J. A. Lieberman: Relationship Between Duration of Untreated Psychosis and Outcome in First-Episode Schizophrenia: A Critical Review and Meta-Analysis, American Journal of Psychiatry, Vol. 162, No. 10, 2005, pp. 1785-1804

[3] S. M. Lawrie, B. Olabi, J. Hall, and A. M. McIntosh: Do We Have Any Solid Evidence of Clinical Utility about the Pathophysiology of Schizophrenia?, World Psychiatry, Vol. 10, No. 1, 2011, pp. 19-31

[4] G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli: Using Support Vector Machine to Identify Imaging Biomarkers of Neurological and Psychiatric Disease: A Critical Review, Neuroscience & Biobehavioral Reviews, Vol. 36, No. 4, 2012, pp. 1140-52

[5] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller: Introduction to Machine Learning for Brain Imaging, NeuroImage, Vol. 56, No. 2, 2011, pp. 387-99

[6] I. C. Wright, S. Rabe-Hesketh, P. W. Woodruff, A. S. David, R. M. Murray, and E. T. Bullmore: Meta-Analysis of Regional Brain Volumes in Schizophrenia, The American Journal of Psychiatry, Vol. 157, No. 1, 2000, pp. 16-25

[7] E. Antonova, T. Sharma, R. Morris, and V. Kumari: The Relationship between Brain Structure and Neurocognition in Schizophrenia: A Selective Review, Schizophrenia Research, Vol. 70, No. 2-3, 2004, pp. 117-45

[8] R. Honea, T. J. Crow, D. Passingham, and C. . Mackay: Regional Deficits in Brain Volume in Schizophrenia: A Meta-Analysis of Voxel-Based Morphometry Studies, American Journal of Psychiatry, Vol. 162, No. 12, 2005, pp. 2233-45

[9] A. M. Shepherd, K. R. Laurens, S. L. Matheson, V. J. Carr, and M. J. Green: Systematic Meta-Review and Quality Assessment of the Structural Brain Alterations in Schizophrenia, Neuroscience & Biobehavioral Reviews, Vol. 36, No. 4, 2012, pp. 1342-56

[10] D. Antonius, V. Prudent, Y. Rebani, D. D'Angelo, B. A. Ardekani, D. Malaspina, and M. J. Hoptman: White Matter Integrity and Lack of Insight in Schizophrenia and Schizoaffective Disorder, Schizophrenia Research, Vol. 128, No. 1-3, 2011, pp. 76-82

[11] D. Schwarz, and T. Kašpárek: Brain Morphometry of MR Images for Automated Classification of First-Episode Schizophrenia, Information Fusion, Vol. 19, 2014, pp. 97-102

[12] C. Gaser, I. Nenadic, B. R. Buchsbaum, E. A. Hazlett, and M. S. Buchsbaum: Deformation-based Morphometry and Its Relation to Conventional Volumetry of Brain Lateral Ventricles in MRI, NeuroImage, Vol. 13, No. 6, 2001, pp. 1140-45

[13] A. Mechelli, C. J. Price, K. J. Friston, and J. Ashburner: Voxel-based Morphometry of the Human Brain: Methods and Applications, Current Medical Imaging Reviews, Vol. 1, No. 2, 2005, pp. 105-13

[14] C. Scanlon, S. G. Mueller, D. Tosun, I. Cheong, P. Garcia, J. Barakos, M. W. Weiner, and K. D. Laxer: Impact of Methodologic Choice for Automatic Detection of Different Aspects of Brain Atrophy by Using Temporal Lobe Epilepsy as a Model, American Journal of Neuroradiology, Vol. 32, No. 9, 2011, pp. 1669-76

[15] C. Davatzikos: Why Voxel-Based Morphometric Analysis Should Be Used with Great Caution When Characterizing Group Differences, NeuroImage, Vol. 23, No. 1, 2004, pp. 17-20

[16] E. Zarogianni, T. W. J. Moorhead, and S. M. Lawrie: Towards the Identification of Imaging Biomarkers in Schizophrenia, Using Multivariate Pattern Classification at a Single-Subject Level, NeuroImage: Clinical, Vol. 3, 2013, pp. 279-89

[17] S. R. Kay, A. Fiszbein, and L. A. Opler: The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia, Schizophrenia Bulletin, Vol. 13, No. 2, 1987, pp. 261-76

[18] J. Ashburner: A Fast Diffeomorphic Image Registration Algorithm, NeuroImage, Vol. 38, No. 1, 2007, pp. 95-113

[19] D. Schwarz, T. Kašpárek, I. Provazník, and J. Jarkovský: A Deformable Registration Method for Automated Morphometry of MRI Brain Images in Neuropsychiatric Research, IEEE Transactions on Medical Imaging, Vol. 26, No. 4, 2007, pp. 452-61

[20] H.-Y. Kim, Statistical notes for clinical researchers: post-hoc multiple comparisons, Restorative Dentistry & Endodontics, Vol. 40, No. 2, 2015, pp. 172-76

[21] Y. Benjamini, and Y. Hochberg: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society, Vol. 57, No. 1, 1995, pp. 289-300

[22] I. T. Jolliffe: Principal Component Analysis, Springer, New York, 2012

[23] C. E. Thomaz, J. P. Boardman, S. Counsell, D. L. G. Hill, J. V. Hajnal, A. D. Edwards, M. A. Rutherford, D. F. Gillies, and D. Rueckert: A Multivariate Statistical Analysis of the Developing Human Brain in Preterm Infants, Image and Vision Computing, Vol. 25, No. 6, 2007, pp. 981-94

[24] O. Demirci, V. P. Clark, V. A. Magnotta, N. C. Andreasen, J. Lauriello, K. A. Kiehl, G. D. Pearlson, and V. D. Calhoun: A Review of Challenges in the Use of fMRI for Disease Classification / Characterization and A Projection Pursuit Application from A Multi-Site fMRI Schizophrenia Study, Brain Imaging and Behavior, Vol. 2, No. 3, 2008, pp. 207-26

[25] E. Janoušová, D. Schwarz, and T. Kašpárek: Combining Various Types of Classifiers and Features Extracted from Magnetic Resonance Imaging Data

in Schizophrenia Recognition, Psychiatry Research: Neuroimaging, Vol. 232, No. 3, 2015, pp. 237-49

[26] M. Aharon, M. Elad, and A. Bruckstein: K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE Transactions on Signal Processing, Vol. 54, No. 11, 2006, pp. 4311-22

[27] B. Gaonkar, K. Pohl, and C. Davatzikos: Pattern Based Morphometry, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011, Lecture Notes in Computer Science, Vol. 6892, No. Pt 2, 2011, pp. 459-66

[28] D. Achlioptas: Database-Friendly Random Projections: Johnson-Lindenstrauss with Binary Coins, Journal of Computer and System Sciences, Vol. 66, No. 4, 2003, pp. 671-87