General Triangle Parallel Robot (GTPR) Basic Features of a New Robot Type - Kinematics and related Application Issues

János Somló

Óbuda University, Budapest, Hungary
somlo@uni-obuda.hu

Abstract: A new robot type is proposed herein, which is named, General Triangle Parallel Robot (GTPR). This robot differs from the widely used Delta robots (Clavel Delta robot) in the respect that it's basic triangle (and the similar working triangle) may be any general triangle. The method of generation of GTPR is given in the present paper. The general method of determination of inverse transformation problem for GTPR is outlined. So, any working process may easily be realized, by any GTPR. Because GTPR covers a much wider class of devices, as the special case the Delta robot, GTPR may yield a number of advantages. This robot type may be advantageous from the point of view of simplicity and effectiveness of constructions realizing different applications (for example stepping). It is a nice feature that the solution of kinematics problems is extremely simple. For these robots, the drive allocations may be solved very effectively. It is possible to construct such triangles, which are advantageous, from the point of view of static forces, Etc. Sometimes these robots are referred as GTP(S)R, where (S) indicates the author (Somlo).

Keywords: Parallel robots; Delta robots; Clavel; General triangle parallel (Somlo) robot; Inverse transformation; Direct transformation; Stepping robots

1 Introduction

Parallel robots are more and more promising, in the solution of recent application problems of robotics technology. One of the fields, where these robot types may promise a breakthrough, are the stepping robots.

In the present paper, we propose a new type of parallel robots which is, in fact, the generalization of the well-known **Clavel robot** (see later).

We name the new robot type, **General Triangle Parallel (Somlo) Robot.**

Parallel robots own several outstanding features which make them suitable for the solution of a number of applications, over the use of serial robots. These features are:

- They are cheap compared with serial robots of the same class
- The motors moving the parallel arms can be allocated together, on the same platforms, close to each other in preferable places, leaving space for gear trains, etc.
- The kinematics problems may easily be solved and applied to the solution of motion planning problems
- The forces on the arms are distributed. That is three components are present instead of one in the case of serial robots. Accordingly, rather favourable proportions of arm masses to other masses may be realized.
- Sometimes, parallel robots promise better solution for special tasks than the serial ones. These kind of tasks are for example, the solution of stepping motions.

1.1 History Basics

The history of parallel robots was considered including several authors, Bonev [3] is one of the most spanning. The recent development of the parallel robots (Delta robots) is based on the pioneering work of professor R. Clavel [1, 2]. The Delta robot idea and construction, belonging to him, is the most popular direction.

Monography of J.-P. Merlet [4] is a summary of most important results.

A picture of the wide choice of parallel robots can be obtained from the material provided by ParalleMIC (the Parallel Mechanism Information Center). Recently, more than 26 companies are producing parallel robots (source: www.parallemic.org / WhosWho/ Comp Robo.html).

The construction variants of parallel robots are very wide. But the dominant variant is the Delta robot of the Clavel-type. This robot uses DC or AC servos rotating the thing-rods and shin parallelograms (or Kardan mechanisms) moving a point of the work triangles (see below).

The development of Delta robots began its story at the end of 80s. The solution of the inverse and direct kinematics problems was necessary for the work and was solved. A number of publications are available on this topic. New, rather sophisticated results were obtained by Zsombor Murray [5, 6]. Based on these results, software for direct and inverse problem solutions were developed which are widely available [7].

Parallel robots are more and more promising in the solution of recent application problems of robotics technology.

A huge number of homemade devices have been developed. LEGO parallel robots can be made, too. At the cheapest end, there is the Novint Falcon which is a 3D joystick. But, in fact, this is a perfect Delta robot for as little as, 250 Euros.

The construction variants of parallel robots are diverse. But the dominant variant is the Delta robot. This robot uses DC or AC servos driven motors rotating the thing-rods and shin parallelograms (or Kardan mechanisms) moving a point of the work triangles.

The software developments were restricted to Delta type robots. Delta type robots have equilateral triangles as basis and work triangles. These are special cases of general triangles parallel robots discussed in the recent article. The software developments, until now, where developed for equilateral cases because of no need for other cases.

The practical applications, until now, are mostly solved by Delta type robots. The research work dealing with other types are sparse. The robots Tripteron and Quadrupteron were proposed by the Laboratoire de robotique Laval (Quebec, Canada). These are different than the general triangle parallel robots discussed below.

GTPR promises great advantages, in many fields of robot applications. It is especially true, for example, for stepping devices, based on parallel robots. The advantages are from better opportunities of manipulation of work spaces.

Zsombor Murrey in works [5, 6] solved the inverse and direct kinematics problem of Delta robots. These results are not restricted to Delta robots. The authors of the above papers emphasized that his results are rather general, but did not indicate the way of extended use. The present paper solves this task and propose and some practical methodologies for the inverse and direct kinematics determination, for general triangle parallel robots.

2 Delta Robot (Clavel-Type)

On Figure 1 a Delta robot construction is given. This is very similar to the construction given in R. Clavel’s US Patent description [1]

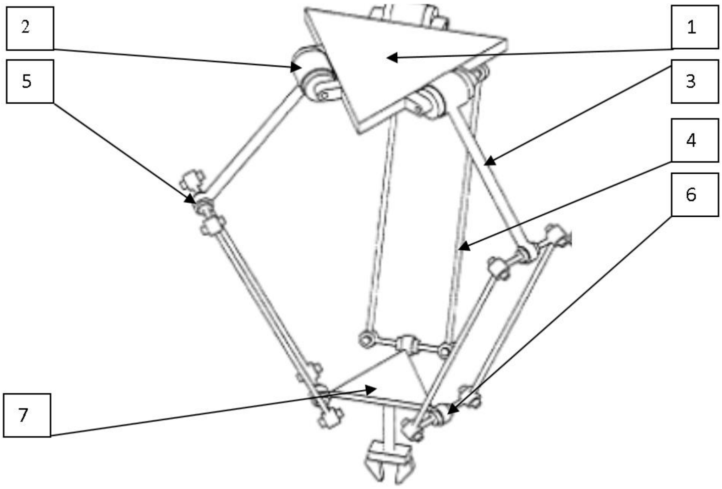


Figure 1
A Delta robot structure

Here:

1- upper platform (basic triangle)
2- driving arm axis
3- driving arm
4- parallelogram
5- driving arm, parallelogram joint
6- parallelogram, lower platform (working triangle) connecting joint
7- lower (working) platform

The simplest imagination of the working of these devices is:

There are 3 arms rotating in vertical plane. The centre points of their rotation are in the edges of an equilateral triangle in x, y plane.

The end points of the arms are coupled with the coupling elements of the parallelogram mechanism. The construction provides that the working point (for example, the centre of the working triangle) moves in a plane parallel with the basic triangle. The motion is activated by the rotation of the arms. Because the arms are coupled with working parallelogram units which are coupled with the working triangle in a special way at the motion of the arms the lower end of the parallelograms may only move in a parallel with the basic triangle plane.

When the arms are moving the working triangle edges move but the working triangle stay similar (and the sides parallel) with the basic triangle (see: Figure 2).

2.1 Inverse Transformation

Let us consider a Delta robot. In Figure 2 the upper view is given, as it can be seen, from the direction of z axis.

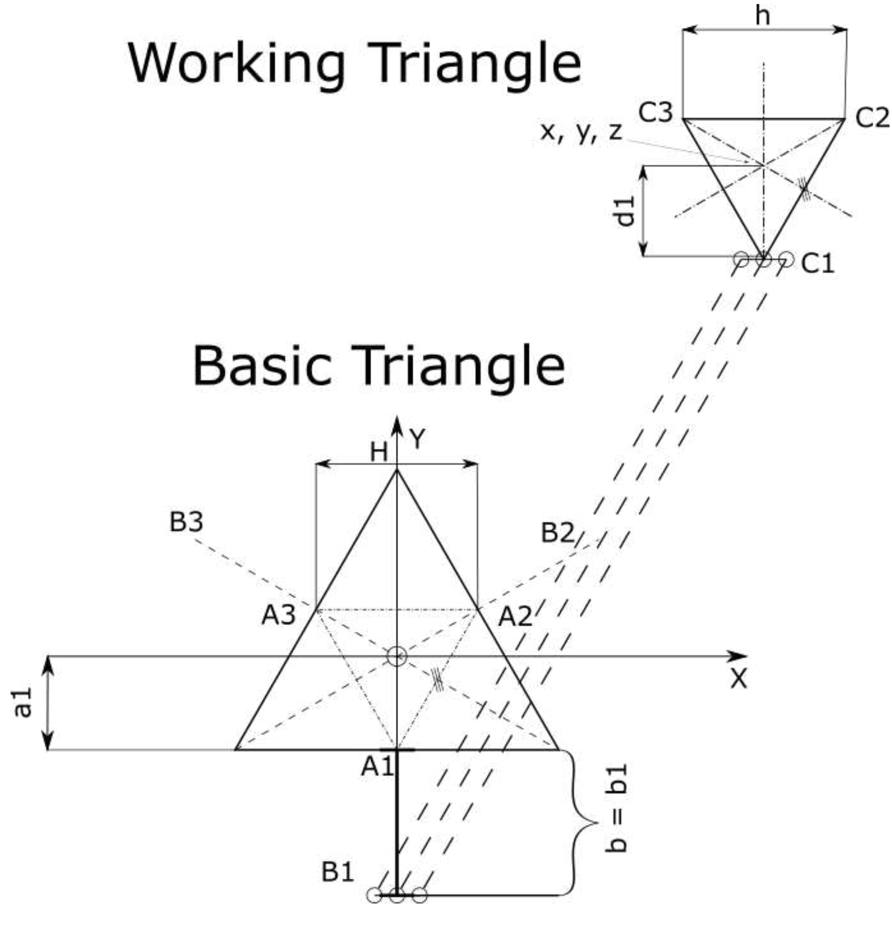


Figure 2

The basic and working triangles

The parameter values are as indicated in Figure 2.

The driving arms move the working point through the motions of points B1, B2, B3, which are the results of rotation in vertical plane as it is demonstrated in Fig. 3. On this figure the motion of the 1st arm is shown.

Let first consider a point when the first driving arm is in horizontal position. Let at this point q1= φ1=0

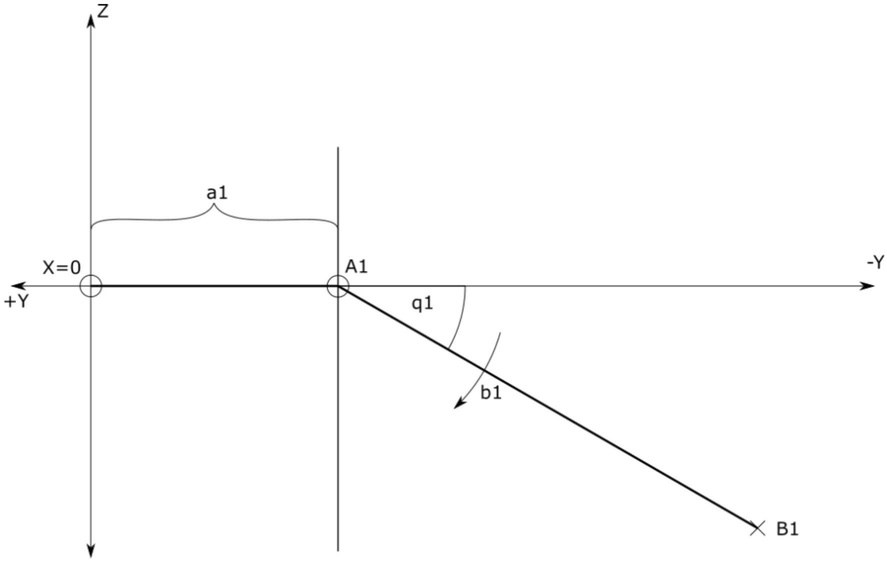


Figure 3
Rotation of the 1. arm

Changing φ1 we have for the B1 point:

$$B1(x) = 0$$

$$B1(y) = -a1 - b1\cos\varphi 1$$

$$B1(z) = -b1\sin\varphi 1$$

For point C1:

$$C1(x) = x$$

$$C1(y) = y - d1$$

$$C1(z) = z$$

The distance of B1C1 is c1. That is:

$$c1^2 = x^2 + [a1 - b1\cos\varphi 1 - y + d1]^2 + [-b1\sin\varphi 1 - z]^2$$

From this:

$$2(-a1+d1-y).b1.\cos\varphi 1 + 2.b1.z.\sin\varphi 1 + x^2 - c1^2 + b1^2 + z^2 + (-a1+d+y)^2 = 0 \quad (1)$$

Equation (1) contains φ1 which is the **solution of inverse transformation** for the first arm.

Williams [9] discusses the solution of equation (1) in the form

$$E_i\cos\varphi i + F_i\sin\varphi i + G_i = 0 \quad (2)$$

In this case:

$E_i = 2(-a1+d1-y).b1$

$F_i = 2.b1.z$

$G_i = x^2 - c1^2 + b1^2 + z^2 + (a1+d-y)^2$

In [9] the tangent half-angle substitution is used and closed formulas are obtained for determination of the φ1 value. For the determination of φ2, φ3 values transformation of the x, y, z quantities are necessary.

Solving the equations for any x, y, z value we get the proper

φ1, φ2, φ3 values, that is the results of the inverse transformations.

The proper motion of the parallel robot may be realized if for given x, y, z values the φ1, then similarly the φ2, φ3 values are determined and input as the proper command for the driving motors.

For the inverse transformation, in the literature, together with the above, a number of other methods is also outlined. Computer programs for the solution are available, too [7].

One of the possible solutions is based on the following fact.

Point B1 moves on a circle in the yz plane. A sphere with radius "c1" and with centre in point x, y-d1 and z intersects the yz plane in a circle with radius $\sqrt{c1^{2-}x^2}$. The common point (points) of the two circles give the solution of the inverse problem.

This solution may be named the "Two Circle Intersection" method [12].

3 General Triangle Parallel Robot (GTPR)

In the present paper we propose a new type of parallel robots.

The difference of this from Delta robot (Clavel type) is that the basic triangle and accordingly the similar working triangle is any general triangle.

So, the Delta robot is only a special case of the GTPR when the triangle is an equilateral one.

At the general triangle parallel robots, we restrict our attention to a construction variant of parallel robots where (as in most of the applied in practice robots) **rotating driving arms and special parallelogram mechanism moved by them are applied and the rotation of the arms result the working triangle motion in parallel with the basic triangle plane. The working triangles are similar to the basic triangles. The working points may be the centre of the working triangle or any other point of the working triangle.**

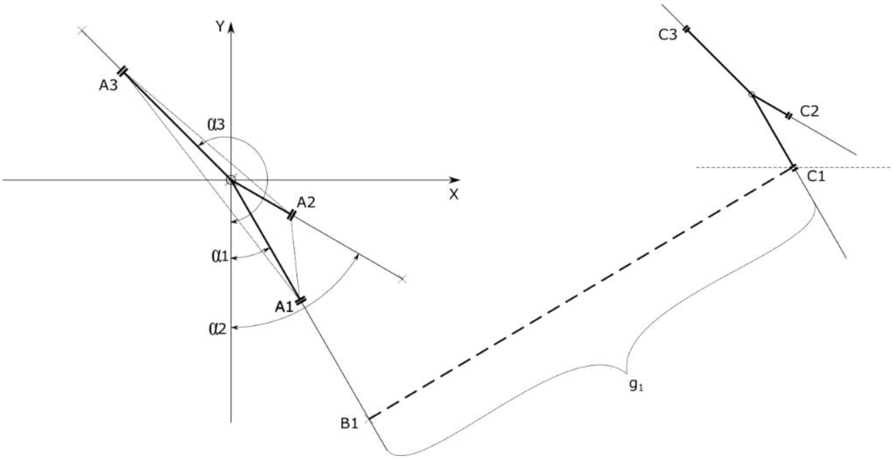


Figure 4
Generation of GTPR

On Figure 4 we show how a GTPR is generated.

Point "O" is the centre of the coordinates system x, y, z.

The 1st arm of the robot is rotating in a vertical plane. (Vertical plane is perpendicular to x, y plane.) The plane of rotation goes through "O". The angle of the rotation plane is α1. The arm rotation centre distance from "O" is s1. The rotation arm length is r1.

The 2nd and 3rd arms have similar structure as it is shown in Figure 4. The corresponding angles are α2 and α3.

The triangles A1A2A3 and C1C2C3 are similar. Any corresponding geometrical elements of these has proportional length.

We use the D(N,M) operator to indicate lengths. For example:

D(O,A1)= s1 the distance to the first rotation axis from the centre point O.

Other examples are:

D(W,C1)= p.D(O,A1); D(C1,C2)=p.D(A1,A2); etc.

As it was mentioned, the basic and working triangles are similar.

The value "p" is the coefficient of proportion.

Any GTPR may be determined by giving the following data:

α , α2, α3, s1, s2, s3, r1, r2, r3, g1, g2, g3, p

Where g1, g2, g3 are the lengths of the parallelograms. (lengths of the central rods).

For example, for the Delta robot analysed above:

α1=0, α2=120^o , α3=240^o

s1=s2=s3=($\sqrt{3}$/3).H (H is the side of the equilateral triangle)

r1=r2=r3=r

c1=g1=g2=g3=g

h=p.H

d1=p.a1

p is the given value of proportion coefficient.

3.1 Determination and Symbol of the GTPR

According to the above, a GTPR may be determined as follow:

Let us allocate in xy plane a general (basic) triangle with edges in points A1, A2, A3. Let in the vertical plane including the O,A1; O,A2; O,A3 sections, around points A1, A2, A3 arms with the proper lengths rotate.

The arm is joined with parallelogram mechanism in proper way. The other "end" of the parallelogram has it joint axis on the working triangle. The working triangle is similar to the basic triangle. The parallelogram mechanism provides that the corresponding sides of basic and working triangles stay parallel during the motion of the working point.

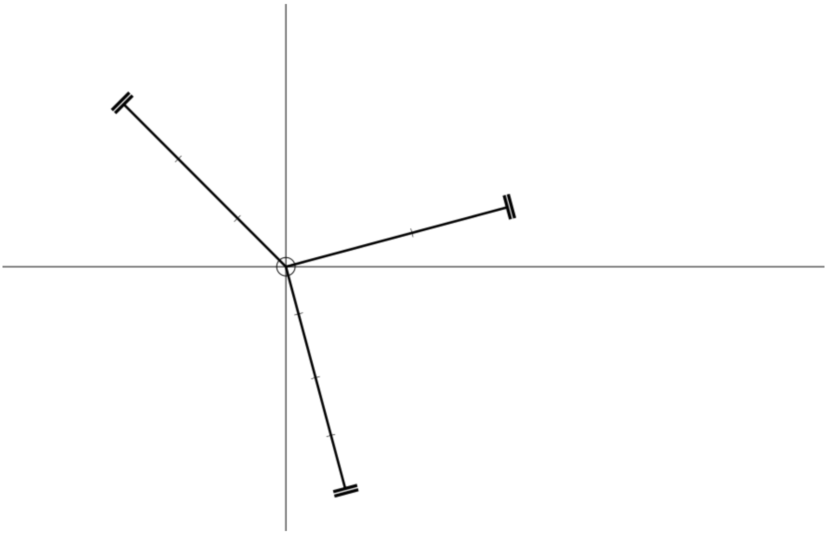


Figure 5

The symbol of GTPR

To symbolize the GTPR we use a pattern given in Figure 5.

3.2 Inverse Transformation for the GTPR

For the inverse transformation of GTPR the same approach can be used as for Delta robot above.

Let us analyse the basic features of GTPRs.

The goal of the actions is to change the working point positions as required.

This is given by the coordinates of vector W.

The components of these vectors are:

Wx=x; Wy=y; Wz=z

The coordinate system origin is O.

That is: Ox=Oy=Oz=0

At the given arrangement we consider motions where the z coordinates of motions are negative. That is, we consider motions in the lover half space.

Aj, Bj (j=1, 2, 3) are points on the driving arms. The j indicates the proper arm.

C1, C2, C3 are the edge points of the working triangle.

Aj, Bj, C1, C2, C3 are vectors is x, y, z space

We use for the coordinates of these vectors:

Ajx, Ajy, Ajz, Bjx, Bjy, Bjz, (j=1, 2, 3).

C1x, C1y, C1z; C2x, C2y, C2z; C3x, C3y, C3z

As it was mentioned:

r1, r2, r3, s1, s2, s3 for given GTPR are given, as well as,

the parallel mechanism bar lengths g1, g2, g3.

The determination of the robots also includes the

α1, α2, α3 values.

Using the distance operator, we have:

D(B1,C1)=g1

D(B2,C2)=g2

D(B3,C3)=g3

Figure 6 shows how an arm moves. It is shown how a rotating centre point A1 is allocated. We consider it as a point of the basic triangle.

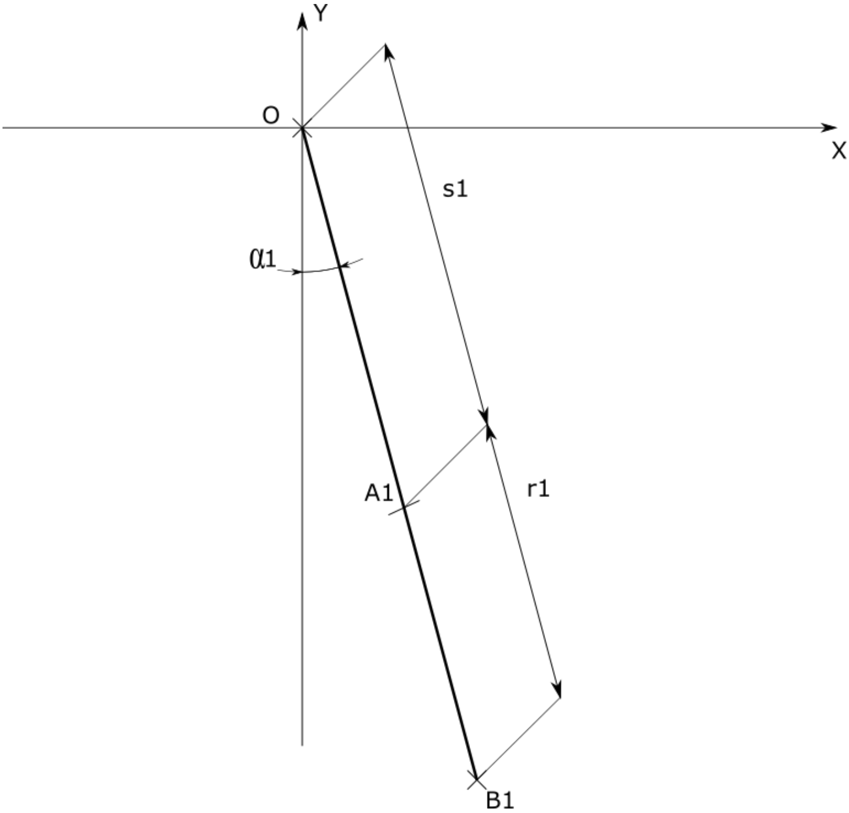


Figure 6
Rotation of an arm

The arms rotate in vertical plane. So we have:

$Bjx= (sj+rj.\cos\varphi j).\sin\alpha j$

$Bjy=-(sj+rj.\cos\varphi j).\cos\alpha j.$

$Bjz=rj.\sin\varphi j$

For points on the working triangle we have:

$C1x= x+ps1\sin\alpha 1$

$C1y= y-ps1\cos\alpha 1$

$C1z=z$

Similarly:

$C2x=x+ps2\sin\alpha 2$

$C2y=y-ps2\cos\alpha 2$

$C2z=z$

and

$$C3x=x+ps3\sin\alpha 3$$

$$C3y=y-ps3\cos\alpha 3$$

$$C3z=z$$

According to the above we have:

$$D(Bj,Cj)=gj \qquad (j=1, 2, 3)$$

That is

$$gj^2=[(sj+rj\cos\varphi j).\sin\alpha j-(x+psi\sin\alpha j)]^2+[-(sj+rj\cos\varphi j)\sin\alpha j- -(y-p.s1\cos\alpha j)]^2+(rj\sin\varphi j-z)^2 \qquad (3)$$

Rearranging Equation (3) we get:

$$[Uj+rj\sin\alpha j\cos\varphi j]^2+[Vj+rj\cos\alpha j\cos\varphi j]^2+(rj\sin\varphi j-z)^2=gj^2 \qquad (4)$$

Where:

$$Uj=sj\sin\alpha j-x-psj\sin\alpha j$$

$$Vj=-sj\cos\alpha j-y-psj\cos\alpha j$$

Performing the operation, we get:

$$Uj^2+Vj^2+z^2+(2Uj\sin\alpha j+2Vj\cos\alpha j)\cos\varphi j-2rjz\sin\varphi j+rj^2\cos^2\alpha j\cos^2\varphi j+ +rj^2\sin^2\varphi j=gj^2 \qquad (5)$$

$$j=1, 2, 3$$

Equation (5) may be solved with suitable nonlinear problem solving methods and software.

The obtained results give the inverse transformation and make possible to realize the required motions.

In the given case a simpler way is possible, too.

The allocation of robots in world coordinate systems is free. So, it is always possible to choose α1=0.

This causes that the second order terms of trigonometrical functions in Equation (5) are eliminated, instead of second order trigonometrical expressions in the equation, rj^2 term appear.

In this case

$$Uj=-x$$

$$Vj=-sj-y-psj=-y+(1+p)sj$$

Equation (5) becomes:

$$2Vj\cos\varphi j-rjz\sin\varphi j+Vj^2+z^2+rj^2-gj^2=0 \qquad (6)$$

Equation (6) can be solved as Williams [9] proposed (see: Equation (2))

In Equation (2)

$E_i = 2Vj$

$F_i = -rjz$

$G_i = Vj^2 + z^2 + rj^2 - gj^2$

Transforming input values to get φ2 and φ3

Solving Equation (6) we get the inverse transformation values for arm 1.

This is φ1.

Now, let us outline how the solutions for the other 2 arm may be obtained.

Let us consider Figure 7. It is very easy to recognize that for an arm allocated in a plane having α2, α3 allocation angle the only difference in determining φ2, φ3 is that the relative position of x, y, z is different than for the first arm.

Let the task is to realize:

$x = x_{0,}\ y = y_{0,}\ z = z_0$

Let us consider Figure 7. Perform the following transformations for the second arm:

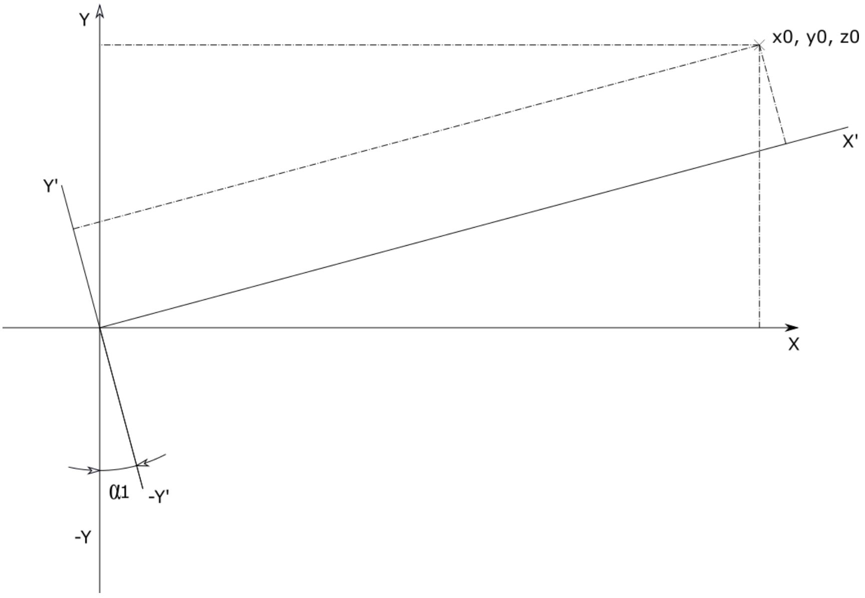


Figure 7

Input values transformations

$x = x_1 = x_0\cos\alpha 2 + y_0 \sin\alpha 2$

$y = y_1 = y_0\cos\alpha 2 - x_0 \sin\alpha 2$ (7)

$z = z_1 = z_0$

Substituting (7), solving (6) we can determine φ2 exactly in the way as it was in the case of α1.

The same is valid when α3 figures instead of α2.

In this way we have got a model for the determination of φ2. φ3.

The inverse transformation problem is solved.

3.2 Direct Transformation for the GTPR

The solution of inverse transformation is the basic task for working actions realization for any robot. For GTPR it may be solved as outlined above.

Sometimes, for example, in analyse of force relations direct transformation, is needed, too.

Solving the direct transformation task the following idea may be used.

Let us consider Figure 8. This Figure is for Delta robot but exactly the same approach is possible for GTPR. We determine so called virtual spheres centre points in the following way. From vectors B1, B2, B3. we extract vectors parallel with C1,OW; C2,OW; C3,OW of the same length. We got the central points of virtual spheres (B1v, B2v, B3v). Spheres with radiuses g1,g2, g3 with centres in B1v,B2v, B3v meet in one point. This is x, y, z.

For any φ1,φ 2, φ3 point from the equations of the three spheres the unique x, y, z value can be determined. This is the solution of the direct kinematics problem.

More about the solution of tree sphere intersection problem may be found in [9].

We determine the so-called virtual spheres centre points in the following way.

Formalizing the above, we introduce in point Bj (j=1, 2, 3) vectors parallel with OWC1, OWC2, OWC3 respectively.

$Bjvx=(sj+rj\cos\varphi j-psj)\sin\alpha j$

$Bjvy=((sj+rj\cos\varphi j-psj)\cos\alpha j$ (8)

$Bjvz=rj\sin\varphi j$

Having this centre points three spheres equations can be described:

$(x-Bjvx)^2+(y-Bjvy)^2+(z-Bjvz)^2=gj^2$ (9)

j=1, 2, 3

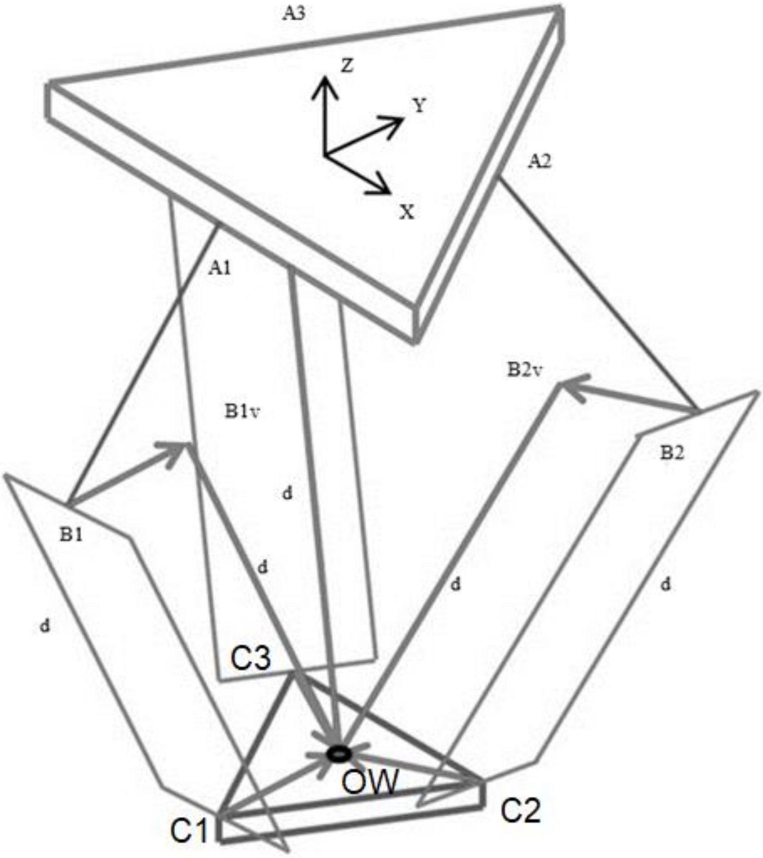


Figure 8
Virtual points for direct transformation

At given φ1, φ2, φ3 values the terms in Equations (9) are given quantities. The three spheres meet only in one x, y, z point.

The determination of this point gives the solution of the inverse transformation problem.

Sometimes this problem is named “The Three Spheres Intersection” problem.

The solution for this is proposed, for example, in Williams [9].

4 Application Examples

4.1 Increase of the Working Space in Different Directions and Grasping Oriented Robots

In Figure 9 (a) we show a triangle which results in the increase of the working space in the y direction. These robots have increased workspace in the given directions compared with Delta robots. In Figure 9 (b) we show basic triangle which can be very favourable when grasping tasks are to be solved.

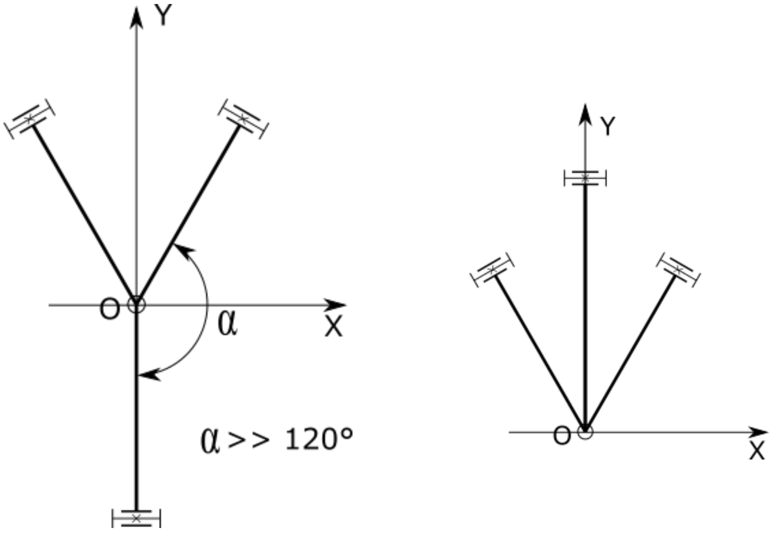


Figure 9a Figure 9b

Extended in y direction work space and grasping tasks solving robots

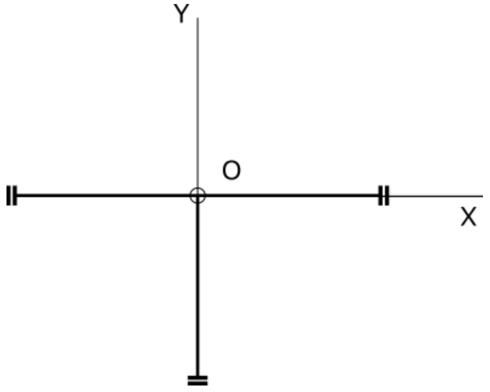


Figure 10

Stepping robot application

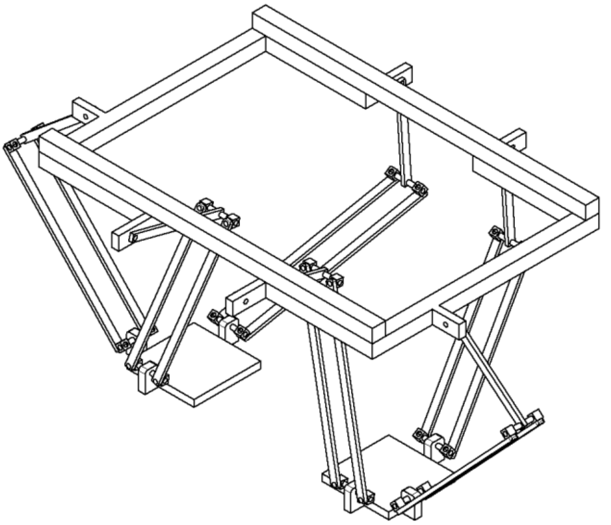


Figure 11
Stepping robot

4.2 Stepping Purpose GTPR

In Figure 10 a GTPR schema is given, which can be very useful in stepping robot applications. In Figure 11 a stepping robot is shown realized, using the GTPR shown in Figure 10.

It would be possible to go on with the application examples, but the benefit of introducing GTPR is that instead of only equilateral triangles it opens a very wide horizon, for different applications, with different requirements. The possible variants of these schemas is very high and diverse.

References

[1] Clavel, R. (1990) U.S. Patent No. 4,976,582. Washington, DC: U.S. Patent and Trademark Office

[2] Clavel, R. (1991) Conceptiond'un robot Parallelerapide 4 degre's de Liberte. PhD thesis EPFL Laussane, Switzerland

[3] Bonev, I. (2014) Delta Parallel Robot, The Story of Success, The Parallel Mechanisms Information Center, (http://www.parallemic.org)

[4] Merlet, J. P. (2012) Parallel robots (Vol. 74) Springer Science & Business Media

[5] Zsombor-Murray, P. J. (2004) Descriptive geometric, kinematicanalysis of Clavel's "Delta" Robot, Centre of Intelligent Machines, McGill University

[6] Zsombor-Murray, P. J. (2009) An improved approach to the kinematics of Clavel's DELTA robot. IntelligentMachines, McGill University

[7] MarginallyClever Software (2012) Delta robot, Forward/Inverse Kinematics, www.marginallyclever.com/other/samples/fkik-test.html

[8] M. Zenkl, Design of Phantom Delta Robot Construction (2017) Budapest, Diploma Work, Óbuda University

[9] Robert L. Williams II, The Delta Parallel Robot: Kinematics Solutions. Mechanical Engineering, Ohio University, October 2016

[10] D. Varga, Fantom Delta Robot Precision Measurement (2017) Budapest, Diploma Work, Óbuda University

[11] Somlo J., Paniti I., Rudas I. (2017) Léptető szerkezet humanoid robothoz (Stepping device for humanoid robots) Hungarian Intellectual Property Office, P1600241, Budapest, 2017.10.30.

[12] Somlo J,. Varga D., Zenkl M., Miko B., (2018) The "Phantom" Delta Robot. A New Device for Parallel Robot Investigation. In Acta Polytechnica Hungarica, Vol. 15, No. 4, pp. 143-160; DOI: 10.12700/APH.15.4.2018.4.8

[13] J. Somlo, B. Lantos, P. T. Cat (1997) Advanced Robot Control. Academic Press, Budapest, Hungary, 1997, P.: 425

[14] Shin, McKey (1991) Minimum Cost Trajectory Planning for Industrial Robots. Control Dynamic Systems. Academic Press. pp. 345-403

[15] B. Valcsák (2017) Determination of inverse and direct transformation for parallel robots using geometrical approach. 2017, Budapest, Diploma Work, Óbuda University

Evaluation of Short-Term Relationships between Selected Investment Funds and the Capital Market in Poland

Aleksandra Matuszewska-Janica[1], Dorota Żebrowska-Suchodolska[1], Grzegorz Mentel[2]

[1]Department of Econometrics and Statistics
Faculty of Applied Informatics and Mathematics
Warsaw University of Life Science – SGGW
Nowoursynowska 166 St., 02-787 Warsaw
e-mails: aleksandra_matuszewska@sggw.pl;
dorota_zebrowska_suchodolska@sggw.pl

[2]Department of Quantitative Methods
Faculty of Management
Rzeszow University of Technology
Powstancow Warszawy 12 St., 35-959 Rzeszow
e-mail: gmentel@prz.edu.pl

Abstract: Investment funds (FIO) as collective investment institutions, place their funds in the stock exchange, thus, participate in financing enterprises. It is true that this share is at the level of several percentage points, but asset transfers at this level can significantly affect the valuation of assets. On the other hand, through the valuation of assets, the stock exchange may also affect the value of participation units of the funds investing in shares. Therefore, the relationship between investment funds and the stock exchange becomes bi-directional. The main aim of the analysis herein is to examine the interaction between the valuation of participation units of selected FIO and the capital market. The analysis includes the share funds existing since 2003. The reference point is the Warsaw Stock Exchange (four main indices: WIG, WIG20, mWIG40 and sWIG80), where these funds invest their cash. The presented analysis is carried out in two stages. The first, an assessment of the interaction between changes in the valuation of fund units and changes in the quotations of the four main WSE indices in Warsaw: WIG, WIG20, mWIG40 and sWIG80. The Granger causality test is used for this purpose. In the second stage, the funds were classified, considering the results of the causality test and portfolio structure. For the clustering, we applied k-means methods. The obtained results indicate two main findings. First, the vast majority of funds, in relation to WIG (whole market index) and WIG20 (blue-chip index) are characterized by causality, which can be described as bilateral (Index↔FIO). In turn, the FIO relationship with the mWIG40 (medium companies index) and sWIG80 (small companies index) can be described as one-sided (Index→FIO). Such a situation is undoubtedly the outcome of the fact that the blue-chip equities have a *significant share in the funds' portfolios. Second, the results of cluster analysis point, were that obtained clusters of funds are more diversified because of the structure of the portfolios than the interactions with the main stock exchange indices.*

Keywords: open-end mutual funds; stock exchange; short-term relationships; Granger causality

1 Introduction

Investment funds, as collective investment institutions, placing their funds on the stock exchange, thus participate in financing enterprises. It is true that this share is at the level of several percent, but asset transfers at this level significantly affect the valuation of assets. On the other hand, through the valuation of assets, the stock exchange may also affect the value of participation units of funds investing in shares. Therefore, the relationship between investment funds and the stock exchange becomes reciprocal.

The purpose of the conducted analysis was, therefore, to check the interaction between the valuation of participation units of selected FIO and the capital market. The analysis includes the Share Funds existing since 2003. The reference point is the Warsaw Stock Exchange, where these Funds invest.

The influence of the behavior of the stock exchange on the valuation of joint-stock units of investment funds is undisputed. As mentioned earlier, the funds invest their funds mainly on the Polish stock exchange, therefore the valuation of the fund units is strictly dependent on the quotations of the assets included in the portfolio. The analyzed funds invest primarily in assets that are taken into account in the construction of major indices (WIG20, mWIG40 and sWIG80). This leads to changes in the indices expressed in changes in indices that will affect the price of participation units.

The relationships between the valuation of FIOs and index quotes are complex. As mentioned earlier, it is not one-way. The influence of the valuation of units on the quotation of indices can be explained in several ways. In a synthetic way, it was taken by [4]. She points out that the influence of funds on the stock exchange can be explained on the basis of such phenomena as: immediate impact, institutional herding or long-term trends. Investment funds have an immediate impact on the price of shares due to the ability to sell significant blocks of shares. Such dependencies were examined, among others [34, 41] or [19]. Institutional herding takes place when one fund buys some value, it is very likely that they will imitate it. This phenomenon was examined, among others [25, 43] or [14]. Investment funds affect stock prices in the long term due to the use of a passive strategy or a gradual increase in their assets. He even mentions a report by the Deutsche Bundesbank [8].

2 Review of the Literature

Relations between financial time series (to which undoubtedly belong both indexes and series with the valuation of investment fund units) are considered in the field of both short-term and long-term changes. A review of analyzes related to relations for capital markets was presented, inter alia, in the works of: [3] [24].

Long-term relations are largely analyzed in relation to the concept of cointegration [17, 11, 12]. This type of analysis has been widely used since the early nineties of the last century, and its importance has been emphasized, among others, at work [22]. A review of literature in this field is presented, for example, in the works of: [13] and [39]. The concept of cointegration in the analysis of the relationship between "funds" and the capital market was used, among others, in works: [1, 5, 18, 35] or [2].

As a complement to the cointegration analysis, the short-term compounds are tested using the Granger causality test. One of the possibilities is to run this test based on the Vector Error Correction Model (VECM) [21]. The existence of such relations between the "funds" and the local stock market was demonstrated, inter alia, in the works [35] or [1]. Short-term dependencies are also analyzed using fixed-line financial series, such rates of return or in terms of volatility [26]. Testing causality between the returns on exchange indices and aggregate mutual funds flow is presented, among others, in the works [44] for monthly data and exchanges in Hong Kong and Singapore, [17] quarterly data and the Portuguese exchange and [20] for quarterly data and the S&P500 index. In turn, analysis for data daily was presented in the paper [33], where the authors showed a relation between the funds flow and the Indian exchange. The study of relationships in the field of variability is presented in the paper [6].

The relationship between the inflow of capital to investment funds and the rates of return from the market has been explained in the literature, among others using feedback trading hypothesis. Investors invest their financial surpluses in funds when stock prices rise, which delays the rise in asset prices. Research for the US market was conducted, among others [42] (using monthly data from 1984-1993) or [37]. Using a similar methodology, they did not observe return relationships between the flows of investment funds and share prices. Completely different results were obtained by [36] [9], who used the monthly rate of return of funds to test the Granger causality test. Dependencies for other markets were indicated by [1] (for the Greek market) or [34] (for the Korean market).

In Polish literature, there are few papers describing the relationship between the value of fund assets and changes in stock market indices. The analysis of the dynamics of the value of assets can be found in Satoły's work [38]. It focuses on the period from June 2006 to March 2010 analyzing capital flows both at the end of the month and in the quarter. Satoła concludes that the development of the investment funds market is one of the factors of the development of the financial market. The upturn in the stock market attracts capital to investment funds, and this has a positive impact on the development of the entire market. The rise in the value of assets in 2007 according to Satoła was influenced by both the bull market and psychological factors. The decisions made at that time were a reaction to changes in stock market indices.

3 Investment Fund Markets in Poland

The first open-end investment fund was established in Poland in 1992 (the First Polish American Trust Fund Pioneer), and another one only three years later. From 20 funds in 1997, the number of funds in 2004 increased more than seven times. The development of the investment funds market was influenced by further legal regulations and an increase in investment awareness of Poles.

The Act on Trust Funds of August 28, 1997 changed the nomenclature of funds (the trust fund was changed into an open-end investment fund) and gave the fund legal personality. In 1999, the first closed fund was created. The amendment of the Act in 2000 enabled the sale of fund units outside brokerage houses and through individuals. In this year, funds were created that invest outside of Poland and global funds (investing in Poland and abroad). In 2001, the bond and money market markets developed.

The entry into force of the Act on investment funds in 2004 had a major impact on the current shape of the investment fund market. The adaptation of Polish law to the regulations in force in the European Union unified the principles of fund management, information obligations of open funds and the rules for selling shares by foreign funds based in the EU. In 2004, the Management Board of Funds and Assets was also set up to carry out activities related to the operation of the funds and the standards of their operation.

At the end of 2004, the net assets of investment funds amounted to 37.43 (PLN bn), and at the end of December, the net asset value of the share-based funds alone was 29.63 (PLN bn).

Comparing the Polish and European Fund Markets, the share of fund assets registered in Poland in European Fund assets is small (around 0.5%). Assets of funds are constantly growing, their share in relation to GDP increases, and the structure of the market from the point of view of their division into shares, bonds, money market, etc. reflects the structure of the European market [45].

One of the classification of funds is the division behind the Chamber of Fund and Asset Management, which lists the following types [30, 31, 32]: absolute return funds, equity funds, private equity funds, debt funds, cash and cash funds, mixed funds, real estate, capital protection, raw material market and securitization. In this structure, equity funds are in fourth place, having approximately 11-12% share in the investment fund market. This percentage gives net asset value at the level of

PLN 33.17 billion as at the end of December 2017, even with a negative balance of sales (Table 1). A similar share of equity funds in the entire market does not mean that the net asset value remains unchanged. On the contrary, this value is gradually growing, except during the financial crisis, and the development of other types of funds gives a stable position of equity funds against the market.

Table 1

Net asset value of particular types of funds (as at the end of December 20017 and 2016)

Net asset value of individual types of funds (PLN million)	December 2016	December 2017
absolute return funds	14999	15060
equity funds	28234	33166
non-public asset funds	103628	100597
debt funds	43487	47566
cash and cash funds	32185	41186
mixed funds	24651	32544
real estate funds	2265	2418
capital protection funds	2442	no data
raw materials market funds	1082	1060
securitization funds	5951	5382
together	258922	278979

Source: own study

4 Analyzed Fund Characteristics

The subject of the research was 15 equity funds operating in Poland since 2003. 13 of them belong to Universal Funds, while 2 are funds investing in shares of small and medium-sized companies (Investor Top 25 Malych Spółek, Rockbridge Akcji Dynamiczne Spółek).

Table 2 lists the names of the funds covered by the study and the Company that manages them.

Table 2

Funds accepted for the study

LP	Name of the fund	TFI	Sign
1	Arka BZ WBK Akcji Polskich	BZ WBK Towarzystwo Funduszy Inwestycyjnych S.A.	Arka
2	Aviva Investors Polskich Akcji	Aviva Investors Poland Towarzystwo Funduszy Inwestycyjnych S.A.	Aviva
3	Esaliens Akcji	Esaliens Towarzystwo Funduszy Inwestycyjnych S.A. (dawniej Legg Mason TFI S.A.)	Esaliens

4	Investor Akcji	Investors Towarzystwo Funduszy Inwestycyjnych S.A.	Investor1
5	Investor Akcji Spółek Dywidendowych	Investors Towarzystwo Funduszy Inwestycyjnych S.A.	Investor2
6	Investor Top 25 Małych Spółek	Investors Towarzystwo Funduszy Inwestycyjnych S.A.	InvestorT
7	Millennium Akcji	Millennium Towarzystwo Funduszy Inwestycyjnych S.A.	Millennium
8	NN Akcji	NN Investment Partners Towarzystwo Funduszy Inwestycyjnych S.A.	NN
9	Novo Akcji	OPERA Towarzystwo Funduszy Inwestycyjnych S.A.	Novo
10	Pioneer Akcji Polskich	Pioneer Pekao Towarzystwo Funduszy Inwestycyjnych S.A.	Pioneer
11	PZU Akcji Krakowiak	Towarzystwo Funduszy Inwestycyjnych PZU S.A.	PZU
12	Rockbridge Akcji	Rockbridge Towarzystwo Funduszy Inwestycyjnych S.A.	Rockbridge1
13	Rockbridge Akcji Dynamicznych Spółek	Rockbridge Towarzystwo Funduszy Inwestycyjnych S.A.	Rockbridge2
14	Skarbiec Akcja	Skarbiec Towarzystwo Funduszy Inwestycyjnych S.A.	Skarbiec
15	UniKorona Akcje	Union Investment Towarzystwo Funduszy Inwestycyjnych S.A.	UniKorona

Source: Authors' own

When analyzing funds in terms of net assets (Figure 1), one can notice a large difference in value. The funds with the largest market shares can be: NN Akcji, Arka BZW BK Akcji Polskich, Esaliens Akcji, Aviva Investors Polskich Akcji or PZU Akcji Krakowiak. The lowest net asset value is represented by Novo Akcji, Rockbridge Akcji Dynamicznych Spółek, Rockbridge Akcji and Investor Akcji Spółek Dywidendowych, which have a nearly nine times smaller share of net assets than the largest funds.

The type of fund, in accordance with the Act on investment funds of May 27, 2004, already imposes certain restrictions on the manner of investing. In addition, the funds in their prospectuses specify this in more detail, giving the upper or lower percentage of investing funds in shares. As of June 2017, it is shown in Figure 2.

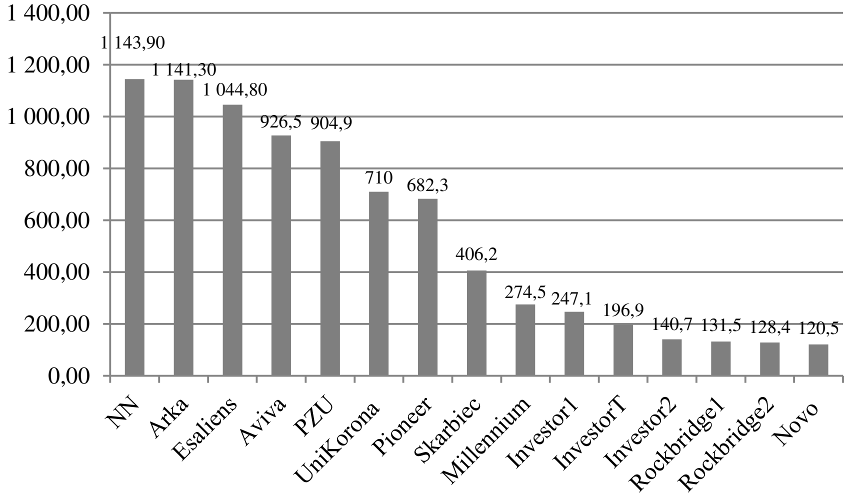


Figure 1

Net asset value in PLN million (as of January 2018)

Source: own study

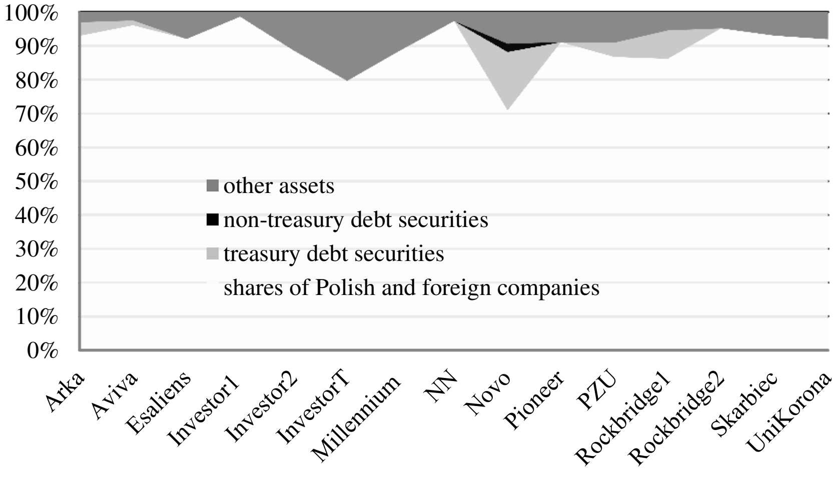


Figure 2

Asset Class (as at June 2017)

Source: own study

Funds from 70.94% (in the case of the Novo Akcji fund) to 98.64% (in the case of the Investor Akcji fund) invest their assets in shares. Only five funds (Novo Akcji, Rockbridge Akcji, Pzu Akcji Krakowiak, Arka BZ WBK Akcji Polskich, Aviva Investors Polskich Akcji) had debt securities in their portfolios. They constituted from 1.41% in the case of Aviva Investors Polskich Akcji up to 17.2% for Novo Akcji. The Novo Akcji fund was the only one that held debt securities in its portfolio. The last component of the portfolio was other assets, i.e. deposits, derivatives, currencies and cash. Their share ranged from 1.36% to 11.35%.

Focusing on the shares that accounted for the largest share in the portfolio, it can be seen that large funds invested primarily a significant part of their equity assets in large companies included in the WIG20 index (Figure 3). The share for the Arka BZW BK Akcji Polskich fund was 52.11%. A small share in the portfolio of companies included in the WIG20 occurred, which is understandable, for the Rockbridge Akcji Dynamicznych Spólek (1.82%) and Investor Top 25 Małych Spółek (1.15%) funds. However, the situation was similar in the case of the Investor Akcji fund (companies from WIG20 constituted only 3.16% of the share assets). In addition to investing in large companies, the fund portfolio also includes medium-sized companies (from 9.41% to 39.99%) and small (from 4.34% to 24.91%). In addition, the portfolios of funds consist of shares that are not part of the listed indices. The share of such assets in the portfolio is from 12.47% to 58.57% in the case of the Investor Akcji fund.

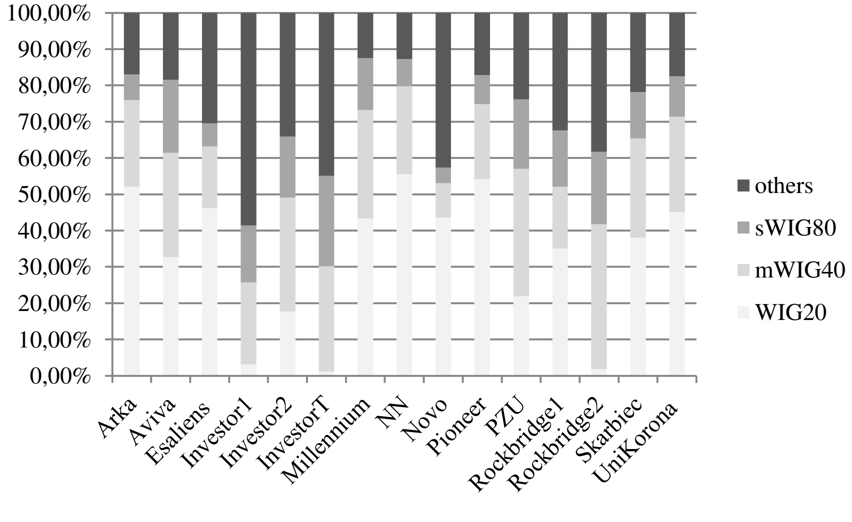


Figure 3

Structure of assets (as at June 2017)

Source: own study

5 Description and Methodology of the Test

The presented analysis consisted of two stages. The first was an assessment of the interaction between changes in the valuation of fund units and changes in the quotations of the four main WSE indices in Warsaw: WIG, WIG20, mWIG40 and sWIG80. The causality test in the Granger sense was used for this purpose. In the second stage, the funds were classified taking into account the results of the causality test and portfolio structure.

Granger causality is defined in following way [16] [7]:

If $\text{MSE}(\check{Y}_t \mid U_{t-1}) < \text{MSE}(\check{Y}_t \mid U_{t-1} \setminus X_{t-1})$ then to $X \rightarrow Y$

where: U_{t-1} – set of previous information that is available at the moment t, X_t – set of previous information that is available at the moment t, X_t is a subset of U_t: $X_t \subset U_t$, Y_t – present value of the variable Y $(Y_t \subset U_t)$, $\check{Y}_t$ – unbiased forecast of variable Y, MSE – mean square error of *ex post* forecast.

In other words, we can have interpreted it that changes of variable X „cause" the changes of variable Y when we can better predict Y using X. Such situation we mark it further as $X \rightarrow Y$, where arrow points to the direction of causality. The reverse relation is defined in an analogous way $Y \rightarrow X$.

When both relations occur simultaneously, i.e. $X \rightarrow Y$ and $Y \rightarrow X$, feedback or mutual causality is referred to and denotes as: $X \leftrightarrow Y$.

The parameters of the two models are estimated in the first step. First is the unrestricted model in the form:

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{q} \beta_j x_{t-j} + \varepsilon_t \tag{1}$$

and the second, restricted model (with assumption that parameters β_j are equal zero) is as follows:

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t \tag{2}$$

where: x_t- value of variable X in period t; y_t - value of variable Y in period t; α_0, α_i, β_j - parameters of the regressions.

The presented version of the test concerns stationary time series, therefore the rate of return of the analyzed series is used in the study:

$$y_t = \ln\frac{Y_t}{Y_{t-1}} \text{ and } x_t = \ln\frac{X_t}{X_{t-1}} \tag{3}$$

In the causal relationship WIG→FIO for the variable Y_t, we accept the valuation of one investment fund and the variable X_t for the quotation of one of the indices.

On the other hand, for the FIO → WIG relationship, the variable Y_t accepts the quotations of one of the indices and for the variable X_t, the valuation of one investment fund.

The set of hypotheses is formulated in this test as follows:

H_0: $\beta_j = 0$ for $j = 1, 2, ..., q$, changes in the *X* process do not cause changes in the *Y* process (short indication: $\neg X \rightarrow Y$)

H_1: $\beta_j \neq 0$ for $j = 1, 2, ..., q$, changes in the *X* process cause changes in the *Y* process ($X \rightarrow Y$)

When we have a large number of observation the test statistics is as follows (see [28], p. 177-178):

$$W_G = \frac{RRSS - URSS}{URSS} \cdot T \tag{4}$$

where: *URSS* - residual sum of squares from the unrestricted equation (1); *RRSS* - residual sum of squares from the restricted equation (2).

The causality test is carried out for a different number of *p* lags from 1 to 10, with the assumption that the maximum delay of the variables x_t and y_t,: *p* and *q* are equal ($p=q$). The rejection of the null hypothesis for a major numbers of lags will be interpreted as a situation where the changes in the quotation of one asset (*X*) contribute more to changes in the quotation of the second asset (*Y*), which in short we will consider as a stronger "causality effect", stronger reaction of the *Y* processes for changes in the *X* processes.

For the clustering was applied *k*-means method [29] [15]) and STATISTICA software. In turn, the procedure for conducting cluster analysis were taken from the work of [40]. Data was standardized and as a distance measure it was applied Euclidean distance. The *k*-means method is one of the most widely applied methods for data clustering. It consists of dividing the analyzed sample of objects into predefined number of cluster. This method consists in dividing the analyzed group of objects into predefined number of classes. In the first phase of analysis, objects (states) were divided into different number of clusters: groups: from 2 to 12 (k = 2, 3, ..., 7). Then, based on silhouette index (SI, see [23]), the best divisions were selected. Walesiak reports that values over 0.5 designated that reasonable structure has been found. then the number of clusters is acceptable.

Diagnostic variables:

Z_{1i} - the number of rejections H_0 in the Granger causality test, when the FIO→WIG relationship is tested (in the case of the i^{th} fund);

Z_{2i} - the number of rejections H_0 in the Granger causality test, when the FIO→WIG20 relationship is tested (in the case of the i^{th} fund);

Z_{3i} - the number of rejections H_0 in the Granger causality test, when the FIO→mWIG40 relationship is tested (in the case of the i^{th} fund);

Z_{4i} - the number of rejections H_0 in the Granger causality test, when the FIO→sWIG80 relationship is tested (in the case of the i th fund);

Z_{5i} - share in the portfolio of i th fund of securities other than companies listed on the WSE;

Z_{6i} - share in the portfolio of the i th fund of companies listed in the WIG20 index;

Z_{7i} - share in the portfolio of the i th fund of companies listed in the mWIG40 index;

Z_{8i} - share in the portfolio of the i th equity fund companies in the sWIG80 index;

Z_{9i} - share in the portfolio of the i th fund of companies listed on the GPG but not included in the WIG20 or mWIG40 or sWIG80 indexes.

Z_{1i}-Z_{4i} variables can have values from 0 to 10, and Z_{5i}-Z_{9i} variables from 0% to 100%. To simplify the analysis, we adopted the composition of the portfolio for June 2017.

6 Results

The results of the Granger causality test (see Table 3) indicate that the changes in the values of the stock exchange indices cause changes in the valuation of investment fund units. What is the expected fact? Because, as mentioned earlier, the valuation of the fund depends, among other things, on the value of the assets included in the investment portfolio of the fund. On the other hand, the analyzed portfolios include companies that are listed within stock indices to a great extent. In addition, attention should be paid to the fact that the valuation of units is also affected by the interest in their purchase or sale, and this largely depends on the situation on the stock exchange. In other words, the bull market on the stock exchange contributes to an increase in interest in the investing in the FIOs and the bear market will cause a decrease in such interest.

Table 3

Results of the Granger causality test: lags (p) for which H_0 is rejected at the significance level of 0.05

	FIO→ WIG	WIG →FIO	FIO→ WIG20	WIG20 →FIO	FIO→ mWIG40	mWIG40 →FIO	FIO→ sWIG80	sWIG80 →FIO
Arka	9;10	1-10	5-10	1-10		1-10	2;5	1-10
Inwestor1	3-5;7-9	1-10	1-9	1-10		1-10	1-4;8-10	1-10
NN	3;4	1-10	3-10	1-10		1-10	1-3;5-7	1-10
Novo		1-10	9	1-10	1-4	1-10	1-10	1-10
PZU	1;3;4;6	1-10	1-10	1-10	4;6-8	1-10	6;10;	1-10
Skarbiec	4	1-10	2;4;6;8-10	1-10	5;7;8	1-10	1;2;6;8	1-10

UniKorona		1-10	3;5;6;8-10	1-10	1-4;10	1-10	1-10	1-10
Aviva	2;5-10	1-10	2-10	1-10		1-10		1-10
InvestorT	2-10	1-10	1-10	1-10	2;6;7	1-10	1-3;5;10	1-10
Esaliens	3;5-10	1-10	2-10	1-10	3-8	1-10	4;9;10	1-10
Inwestor2		1-10	1-10	1-10		1-10	1;2;5-8	1-10
Millenium	2;5-8	1-10	1-10	1-10	1;3;4	1-10	4;6	1-10
Rockbridge1	3;8;10	1-10	1-3;6;8;10	1-10	3	1-10	2;3;6;7	1-10
Rockbridge2	1-9	1-7	1-4;8	1;5-8	2;5	1-10		1-10
Pioneer	2-10	1-10	1-10	1-10	1;4-6	1-10	5;6	1-10

Source: own study

The Index→FIO designation indicates the results of the causality test where the changes in the index prices were the reason (in the Granger sense) of changes in the valuation of FIOs.
The FIO→Index designation indicates the results of the causality test where changes in the valuation of FIOs were the cause (in the Granger sense) of changes in the index prices.

The presented results indicate that changes in the values of the analyzed funds units may also be the cause (in the Granger sense) for changes in the index prices. The maximum number of rejections H_0 (10 times) is obtained for three funds: Pioneer, InvestorT and Rockbridge2, and for two others (Millenium and Aviva) 9 rejections in the case of the FIO impact on the WIG index., The results indicate a significant stronger impact of changes in the value of FIO units on WIG20 index quotations in comparison to the other indices. There are at least 9 H_0 rejections for 13 funds out of 15 (9 rejections for Investor1, Arka, NN, Skarbiec, Millenium, Inwestor2 and Rockbridge2 and 10 rejections for Esaliens, Pioneer, PZU, Aviva, InvestorT and Rockbridge2). UniKorona and Novo are funds for which we diagnose with a lack of causality with respect to the mWIG40 and sWIG80 indexes which at least 9 H_0 rejections. Interestingly, in the case of these funds, we have at most two H_0 rejections in the FIO→WIG20 relationship and no H_0 rejections in the FIO→WIG relationship. UniKorona and Novo are the only funds which do not affect changes in the WIG index.

It is worth noting that for the major FIOs (8 out of 15) the rejection of the null hypothesis $\neg$FIO→WIG20 takes place already for the first lag (p=1) in the case of the WIG20 index. They are mainly funds with small capitalization, less than 300 million PLN (Millenium, InvestorT Inwestor1, Inwestor2, Rockbridge1, Rockbridge2). The other two (PZU and Pioneer) can be defined as medium-sized funds (with the portfolio value above 600 million PLN, but not exceeding 1000 million PLN). However, it should be noted that the shares of managing authorities of these portfolios (PZU and PEKAO SA) constitute a significant share in the composition of the WIG20 index. In turn, for the three largest funds (with the portfolio value above 1000 million PLN) the impact of the FIO valuation on the index quotation is visible only after at least two periods. This phenomenon is explained by the fact that money management in the smaller funds is more flexible (i.e. it is easier to withdraw them from the market) than in the case of large funds.

On the other hand, in the case of PZU and Pioneer funds, we expect that their valuation is strongly related to the valuation of shares in the managing institutions of these funds. In turn, they have a large share in the WIG20 index, therefore the companies price of these institutions is closely related to the WIG20 index. Such an interpretation may be a research hypothesis for the next study.

The last stage of the analysis is funds clustering applying the *k*-means method. The results are presented in the Table 4. The best division is found in 5 groups. The values of the silhouette index in this case is 0.768, which indicates a very good division (with the so-called strong class structure). We obtain three one-object clusters (Unikorona in group 1, Investor1 in group 2 and Novo in group 4). Two clusters include six funds each of them. There are such funds as Arka, NN, Skarbiec, Esaliens, Millenium and Pioneer in group 3. This group is characterized by a significant number of H_0 rejections in the FIO→WIG20 relation (about 9 rejections on average) and in the FIO→WIG relation (about 7 rejections on average). On the other hand, in the FIO→mWIG40 relation we notice a low number of rejections (the average is 0.3). H_0 is rejected 3 times on average in the FIO→sWIG80 relation. This group was characterized by a fairly high share of WIG20 companies (48.3% on average) and a relatively low share of companies not listed in the main WSE indices (18.6% on average) in comparison to other clusters. Group 5 includes following funds: PZU, Aviva, InvestorT, Inwestor2, Rockbridge1, Rockbridge2. In this group, we observe a significant number of H_0 rejections in the FIO→WIG and FIO→WIG20 relations (7.2 and 9.7 on average respectively).

Table 4
Group average

FIO in a group	Group number	FIO→WIG	FIO→WIG20	FIO→mWIG40	FIO→sWIG80	Share of assets other than shares	Participation in WIG20	Participation in mWIG40	Participation in sWIG80	Share of other shares
UniKorona	1	0	1	9	10	8.0%	45.2%	26.2%	11.2%	17.5%
Inwestor1	2	2	9	0	7	1.4%	3.2%	22.5%	15.7%	58.6%
Arka, NN Skarbiec, Esaliens, Millenium, Pioneer	3	7	9.3	0.3	2.7	7.6%	48.3%	23.9%	9.3%	18.6%
Novo	4	0	2	10	10	29.1%	43.6%	9.4%	4.3%	42.6%
PZU, Aviva, InvestorT, Inwestor2 Rockbridge1, Rockbridge2	5	7.2	9.7	1.3	1.3	12.9%	18.4%	30.3%	19.4%	32.0%

Source: Own Study

Conclusions

The presented results indicate that Fund Groups significantly differentiate more because of the structure of the Fund's portfolios, than the relationships and the interactions with the main stock exchange indices. The vast majority of Funds in relation to WIG and WIG20 indexes are characterized by causality, which can be described as bilateral (Index↔FIO). The influence of the changes in the values of the stock exchange indices on the valuation of investment fund (Index→FIO) units is evident. Primarily, because of the fact that the valuation of the Fund depends on the value of the assets included in the investment portfolio of the Fund, among other things. In turn, we also diagnosed the reverse relation (FIO→Index). Such situations can be explained by several factors, i.e.: immediate impact, institutional herding or long-term trends. It is almost worthless, that the analyzed FIO's portfolios include companies that are listed within stock indices, to a great extent and cannot remain inconsequential. On the other hand, the FIO relationship with the mWIG40 and sWIG80 indexes can be described as, one-sided (Index→FIO). While changes in index prices affect changes in the valuation of Funds, the reverse relation is observed to a limited extent. Such situations can be explained by the policies pursued by the Funds. They adjust to the whole market (stock exchange) trends as described by major indices, in our analysis, they are WIG and WIG20.

References

[1] Alexakis C., Niarchos N., Patra T., Poshakwale S.: The dynamics between stock returns and mutual fund flows: Empirical evidence from the Greek market. International Review of Financial Analysis, Vol. 14(5) 2005, pp. 559-569, https://doi.org/10.1016/j.irfa.2004.10.019

[2] Aydoğan B., Vardar G., Tunç G.: The Interaction of Mutual Fund Flows and Stock Returns: Evidence From The Turkish Capital Market. Ege Academic Review, 14(2) 2014, pp. 163-173

[3] Bailey W., Choi J. J.: International market linkages, Journal of Economics and Business, Vol. 55, 2005, pp. 399-404

[4] Boyte-White C.: How Mutual Funds Affect Stock Prices, Inwestopedia 2015, https://www.investopedia.com/articles/investing/, d.d.22.05.2018

[5] Burucu H., Contuk F. Y.: The dynamics between mutual funds flows and stock returns: empirical evidence from the turkey markets, International Journal of Economics and Finance Studies, Vol. 3(1) 2011, pp. 95-109

[6] Cao C., Chang E. C., Wang Y.: An empirical analysis of the dynamic relationship between mutual fund flow and market return volatility, Journal of Banking & Finance, Vol. 32(10) 2008, pp. 2111-2123, https://doi.org/10.1016/j.jbankfin.2007.12.035

[7] Charemza W. W., Deadman D. F.: New Directions In Econometric Practice, Second Edition, Books, Edward Elgar Publishing, number 1139, April 1997

[8] Deutsche Bundesbank. Monthly Report January 2013, Current developments in the mutual funds market: demand, structural changes and investor behavior

[9] Edwards F., Zhang X.: Mutual Funds and Stock and Bond Market Stability, Jolcrrznl of Finnilcia1 Services Resenrch, 13, 1998, pp. 257-282

[10] Enders W.: Applied econometric time series, John Wiley & Sons, 2008

[11] Engle R., Granger C.: Long Run Economic Readings in Cointegration, Oxford University Press, New York 1991

[12] Engle R. F., Granger C. W. J.: Cointegration and error correction mechanism: Representation, Estimation and Testing, Econometrica, 55, 1987, pp. 251-276

[13] Francis B. B., Leachman L. L.: Superexogeneity and the dynamic linkages among international equity markets, Journal of International Money and Finance, Vol. 17, 1998, pp. 475-492

[14] Frey S., Herbst P., Walter A.: Measuring Mutual Fund Herding: A Structural Approach, mimeo 2007

[15] Gatnar E., Walesiak M. (ed.): Metody statystycznej analizy wielowymiarowej w badaniach marketingowych, Wydawnictwo AE we Wrocławiu, Wrocław 2004

[16] Granger C. W. J.: Investigating Causal Relations by Econometric Models and Cross-spectral Methods, Econometrica, 37 (3) 1969, pp. 424-438, DOI:10.2307/1912791

[17] Granger C. W. J.: Some Properties of Time Series Data and Their Use in Econometric Model Specification, Journal of Econometrics 16, 1981, pp. 121-130

[18] Hossain M. S., Rahman A. M., Rajib M. S. U.: Dynamics of Mutual Funds in Relation to Stock Market: A Vector Autoregressive Causality Analysis, International Journal of Economics and Financial Issues, Vol. 3(1), 2012, pp. 191-201

[19] Hsiehy M. F., Yangz T. Y., Yangx Y. T., Lee J. S.: Evidence of Herding and Positive Feedback Trading for Mutual Funds in Emerging Asian Countries, Quantitative Finance, 11(3) 2011, pp. 423-435, https://doi.org/10.1080/14697688.2010.506882

[20] Jank S.: Mutual fund flows, expected returns, and the real economy, Journal of Banking & Finance, Vol. 36(11) 2012, pp. 3060-3070, https://doi.org/10.1016/j.jbankfin.2012.07.004

[21] Johansen S.: Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models, Econometrica: Journal of the

Econometric Society, Vol. 59(6) 1991, pp. 1551-1580, DOI: 10.2307/2938278

[22] Kasa K.: Common stochastic trends in international stock markets, Journal of Monetary Economics, Vol. 29, 1992, pp. 95-124

[23] Kaufman L., Rousseeuw P. J.: Finding Groups in Data: an Introduction to Cluster Analysis, Wiley, New York 1990

[24] Kearney C., Lucey B. M.: International equity market integration: Theory, evidence and implications, International Review of Financial Analysis, Vol. 13, 2004, pp. 571-583

[25] Lakonishok J., Shleifer A., Vishny R.: The impact of institutional trading on stock pricing, Journal of Financial Economics 32, 1992, pp. 23-43

[26] Li H., Majerowska E.: Testing stock market linkages for Poland and Hungary: A multivariate GARCH approach, Research in International Business and Finance, Vol. 22, 2008, pp. 247-266

[27] Lobão J., Levi A.: The relation between mutual fund flows, stock returns and macroeconomic variables: evidence from Portugal, Portuguese Journal of Finance, Management and Accounting, Vol. 2(4) 2016, pp. 54-75

[28] Maddala G. S.: Introduction to econometrics (Vol. 2), New York: Macmillan, 1992

[29] McQueen J. B.: Some methods for classification and analysis of multivariate observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley 1967

[30] Mentel G., Brożyna J., Szetela B.: Evaluation of the effectiveness of investment fund deposits in Poland in a time of crisis, Journal of International Studies, 10(2), 2017, pp. 46-60, DOI:10.14254/2071-8330.2017/10-2/3

[31] Mentel G., Horváthová Z.: Factors of Efficiency of Open Investment Funds in 1997-2015, Economics and Sociology, Vol. 9, No. 1, 2016, pp. 101-113, DOI: 10.14254/2071-789X.2016/9-1/7

[32] Mentel G., Szetela B., Tvaronavičienė M.: Qualifications of Managers vs. Effectiveness of Investment Funds in Poland, Economics and Sociology, Vol. 9, No. 2, 2016, pp. 126-136, DOI: 10.14254/2071-789X.2016/9-2/8

[33] Naik P. K., Padhi P.: An Empirical Evidence of Dynamic Interaction between Institutional Fund Flows and Stock Market Returns in India, Indian Journal of Finance, Vol. 9(4) 2015, pp. 21-32, DOI: 10.17010/ijf/2015/v9i4/71455

[34] Oha N. Y., Parwada J. T.: Relations between Mutual Funds Flows and Stock Market Returns in Korea, International Financial Markets, Institutions and Money, 17, 2007, pp. 140-151

[35] Pojanavatee S.: Cointegration and causality analysis of dynamic linkage between stock market and equity mutual funds in Australia, Cogent Economics & Finance, Vol. 2(1) 2014), 918855 (on-line) https://doi.org/10.1080/23322039.2014.918855

[36] Potter M.: The dynamic relationship between security returns and mutual fund flows, University of Massachusetts–Amherst, PhD dissertation 1996

[37] Ramelona E., Kleiman P., Gruenstein D.: Market returns and mutual fund flows, FRBNY Economic Policy Review, 33-52, Federal Reserve Bank of New York 1997

[38] Satoła Ł.: Dynamika wartości aktywów na polskim rynku funduszy inwestycyjnych, Zeszyty Naukowe Polityki Europejskie, Finanse i Marketing, Nr 4 (53) 2010

[39] Syriopoulos T.: Dynamic linkages between emerging European and developed stock markets: Has the EMU any impact?, International Review of Financial Analysis, Vol. 16, 2007, pp. 41-60

[40] Walesiak M.: Rekomendacje w zakresie strategii postępowania w procesie klasyfikacji zbioru obiektów, in: A. Zeliaś (ed.), Przestrzenno-czasowe modelowanie zjawisk gospodarczych. Materiały z XXVII Ogólnopolskiego Seminarium Naukowego, Wydawnictwo AE w Krakowie, Kraków 2006, pp. 185-203

[41] Walter A., Weber F. M.: Herding in the German Mutual Funds Industry, European Financial Management, 12(3) 2006, pp. 375-406, https://doi.org/10.1111/j.1354- 7798.2006.00325.x

[42] Warther V. A.: Aggregate mutual fund flows and security returns, Journal of Financial Economics, 39, 1995, pp. 209-235

[43] Wermers R.: Mutual fund herding and the impact on stock prices, Journal of Finance 54, 1999, pp. 581-622

[44] Yangbo B., Wickramanayake J., Watson J., Tsigos S.: The relationship between mutual fund flows and stock market returns: a comparative empirical analysis, Corporate Ownership & Control, Volume 8(1) 2010, pp. 785-799, http://dx.doi.org/10.22495/cocv8i1c8p4

[45] Żebrowska-Suchodolska D., Karpio A.: Polski rynek otwartych funduszy inwestycyjnych na tle rynku europejskiego, Zeszyty Naukowe SGGW w Warszawie Scientific Journals Nr 4(53) 2010, Polityki Europejskie, Finanse i Marketing Nr 4(53) 2010, pp. 322-331

Lean Principles Application in the Automotive Industry

György Czifra[1], Peter Szabó[2], Miroslava Mĺkva[2], Jaromíra Vaňová[2]

[1] Óbuda University, Bánki Donát Faculty of Mechanical and Safety Engineering, Népszínház u. 8, H-1081 Budapest, Hungary,
E-mail: czifra.gyorgy@bgk.uni-obuda.hu

[2] Faculty of Materials Science and Technology in Trnava, Slovak University of Technology in Bratislava, ul. J. Bottu č. 2781/ 25, 917 24 Trnava, Slovak Republic, E-mail: peter.szabo@stuba.sk, miroslava.mlkva@stuba.sk, jaromira.vanova@stuba.sk

Abstract: In today's competitive environment, the strategic goal of most organizations, is not only to survive, but also to move forward. It is therefore important to examine business processes and minimize waste. Anything that does not add value and contributes to unnecessarily spent funds can be considered to be a waste, so the goal of each community should be to get rid of activities that negatively affect its effective functioning. The aim of the paper is to highlight the possibilities of using Lean principles in the automotive industry. The purpose of using lean manufacturing methods is to eliminate identified shortcomings and wastage, to ensure smoother production and to meet customer requirements, and to reduce business costs and increase its competitiveness on the market. The article describes selected industrial engineering methods that deal with waste elimination (time, transport, waiting, movement, inventory), followed by an example of a specific application of Value Stream Mapping (Yamazummi chart) in automotive companies in Slovakia. This example gives space for discussion on the next direction of Lean principles usage in practice and potential benefits for the future.

Keywords: Lean Management; Lean Principles; Values Stream Mapping; Yamazumi Chart

1 Introduction

Today's global and turbulent environment is characterized by rapidly changing conditions. Businesses are thinking deeply about how to innovate their production processes, looking for more efficiency and quality, but they often lack specific projects and strategies to put these requirements into practice. Balog and Straka [1] notes that businesses focus its efforts on the future development of the company, have business strategy prepared, which should be focused on the unique production or the services with high added value.

One of the possibilities is the lean manufacturing strategy. Lean manufacturing aims to reduce costs by removing value-added activities. Based on the Toyota manufacturing system, many lean manufacturing tools (e.g. just in time, value stream mapping, etc.) are widely used in industrial manufacturing, including the automotive industry [2].

2 Theoretical Background – Lean Management and Lean Factory

Basics of Lean we first observed in the 50s and 60s of the 20th Century in the management of the Japanese automaker, Toyota. The founder and owner of a comprehensive methodology is considered to be James P. Womack with his publication (from 1990) *The Machine That Changed the World: The story of Lean Production*. It is precisely J. Womack, who is considered to be the author who introduced the concept of Lean, as Toyota's production system [3].

The notion of lean has also expanded to business management as the term “Lean Management”. In the literary sources we can meet with multiple definitions for lean management.

Bosenberg and Metzen [4] defines Lean Management as *"complex system that includes the entire enterprise. Human is staged to the middle of the business. Its elements consist of funded spiritual principles, working principles with new insights into organization, integrating strategies for solving the core business problems, scientific-engineering methods, as well as a number of pragmatic working tools."*

According to Chauhan et al., [5] lean manufacturing is a systematic method for the elimination of waste ("Muda"), wastes created through overburden ("Muri") and wastes created through unevenness in workloads ("Mura") within a manufacturing process.

According to Svozilová [6] Lean is a combination of principles and methods aimed at identifying and eliminating activities which do not bring any value in the process of product or service creation. These activities ultimately are waste products or waste.

The concept of lean is based on the production of a flexible response to customer and demand. Every employee has a great responsibility for the quality and production process. Decision-making competencies are decentralized in the lean manufacturing system so every worker has the right to interrupt production in the production process. Lean production management is heavily focused on maximizing customer satisfaction, which is in direct contradiction with traditional "Taylor" principles of mass production [7].

Košturiak and Frolík [8], say: *"The business's sophistication means doing just the things that are needed, doing it right now, doing it faster than others, and spending less money at the same time. Lean is about increasing the company's performance by producing more than our competitors on a given site, adding a higher added value to a given number of people and devices than others, that we are making more orders at a given time, that we need less time for each business process. The lean principle of the company is to do exactly what our customer wants with a minimum number of activities that do not increase the price of the product or service. Being lean means making more money, earning it faster and making less effort."*

The authors Sundar et al. [9] in their article write: Lean Manufacturing is considered to be a waste reduction technique as suggested by many authors, but in practice lean manufacturing maximize the value of the product through minimization of waste. Lean principles define the value of the product/service as perceived by the customer and then making the flow in-line with the customer pull and striving for perfection through continuous improvement to eliminate waste by sorting out Value Added activity (VA) and Non-Value Added activity (NVA). The sources for the NVA activity wastes are Transportation, Inventory, Motion Waiting, Overproduction, Over processing and Defects.

The above definitions show that the foundation of Lean Enterprise is the elimination of waste, thus losses that do not add value for the customer.

If an enterprise wants to live a lean philosophy, it cannot understand it in a limited way, it has to connect it to other sectors of its activities, starting with product development, logistics, production, as well as the slim administration itself and the overall slim understanding of the business itself. Basic elements of Lean Enterprise are shown on Figure 1.

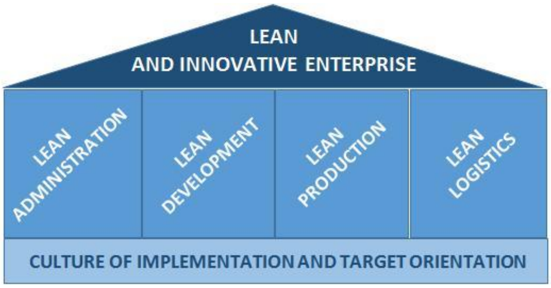

LEAN
AND INNOVATIVE ENTERPRISE
LEAN ADMINISTRATION
LEAN DEVELOPMENT
LEAN PRODUCTION
LEAN LOGISTICS
CULTURE OF IMPLEMENTATION AND TARGET ORIENTATION

Figure 1

Basic elements of Lean Enterprise [10]

The basic principles of lean, include [11]:

1. *Understanding the concept of value from a customer perspective* – it is necessary to know what is really the value for the final customer and that is important for the organization.
2. *Value flow analysis* – as we suddenly understand the value as understood by the customer, it is important to define correctly the value flows in the production process (again from the customer's point of view). It is necessary to determine the steps that add value and which do not need to be eliminated.
3. *Fluent flow* – wherever possible, a smooth flow of material is required without unnecessary buffers and w.i.p. production.
4. *Pull system application* – the organization does not produce to warehouse; production is conditional to customer requirements.
5. *Perfection* – it is necessary to reduce or completely eliminate wastage, the creation of value for the final customer should be the same or higher than its expectations.

1.1 Lean Production Methods

Lean manufacturing methods are the cornerstone of success for lean manufacturing in enterprises. There are several methods of lean manufacturing, the company can utilize their full range - in general, the given methods can also be understood as management or process development methods. However, it is necessary to realize that the actual implementation of individual instruments does not necessarily have to bring the desired effect to the company, as long as they do not absorb the lean manufacturing philosophy itself. It consists in the active participation of top management, on-site production management, active involvement of all employees in the enterprise in continuous improvement processes, not only at their workplace, but wherever appropriate. Lean manufacturing can be introduced with a large number of instruments, but it is important to remember that these tools are only helping tools of a sophisticated philosophy. The power of used tools in this way is a combination of their use.

1.1.1 Value Stream Mapping

Value Stream Mapping (VSM) or *Value Stream Analysis* is an analytical technique that is one of the basic methods of lean manufacturing philosophy. Based on Abdulmalek and Rajgopal [2] this map is used to identify sources of waste and to identify lean tools for reducing the waste.

Value flow is a complex package of activities that is ultimately added to the end value for the customer in the process of transforming materials to products. However, it includes activities that add and not add value to the final product [12].

VSM serves to describe processes that add and add value to both manufacturing, service, and administrative structures. This methodology is based on the preparation of a physical map so that all team members are directly involved in the process. Creation can be compared to a controlled brainstorming where the structure of work is defined by product flow through production and by identifying all the significant factors affecting its final form and condition [13].

There are 2 basic flows within the transformation process: 1) information flow – orders against the value flow (from customer to receipt of input material) and so on, 2) material flow – products flowing in the value flow direction (from input material to product delivery to the customer) [13].

To compile the value flow map, it is necessary to measure times and steps individually and do not rely on standard times. The best solution is when the recorder observes all the steps and writes all the data, as well as the delays that may occur.

The information obtained then serves to build the map itself (Figure 2), which is divided into three parts:

- The top of the map is used for information flow
- The middle of the map is used for material flow
- The bottom of the map is used for process time recording

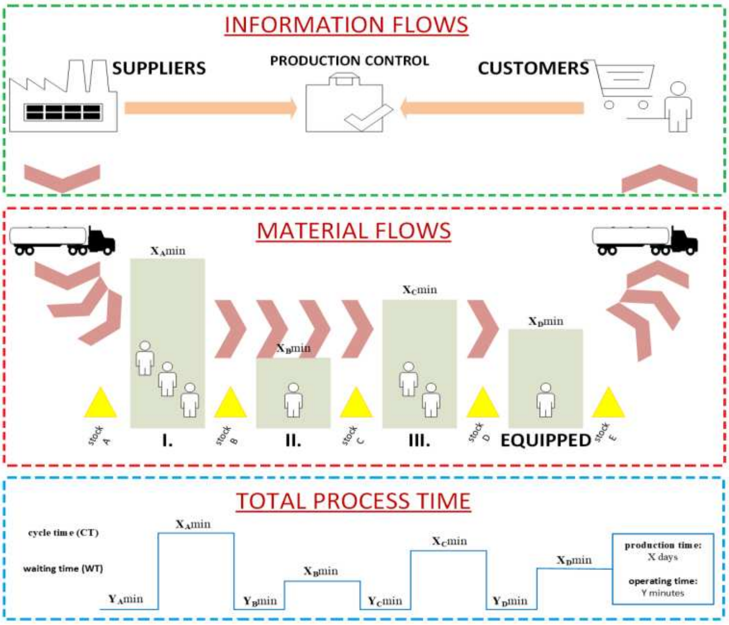


Figure 2

Value Stream Mapping (source: own processing based on [10] [14])

The information we receive is then written to the bottom of the map in the time zone that tells us how much time the material has been on the road and how long it has been processed.

Value stream mapping helps to identify possible losses bottlenecks, weaknesses and reasons for inefficient flows anywhere in the enterprise.

1.1.2 Takt Time and Cycle Time

To achieve a customer driven value stream it is important to design the production or manufacturing system to be consistent with the pace at which the customer is demanding a part or product. This pace is often referred to as the “Takt Time” [15].

Takt time or clock time is the basic indicator of lean manufacturing. Indicates the speed (tempo) that the parts are to be manufactured to meet the order. The purpose of the clock time is to precisely match production with demand. Takt time (TT) is calculated as the share of available production time and the number of pieces the customer requests [16].

Cycle time is a theoretical value. The real time value that each operator (factory worker) or machine needs to complete is a cycle time. The actual process is slowed down by ineffective actions and abnormalities that prolong our time or cycle time. The basic idea of a lean production is to approach the cycle time to the tact time so that we can "captivate to produce." The cycle time must be less than the cycle time [16].

Yamazumi chart is a Japanese method designed to visualize the time data of activities identified in the analyzed process. The data is displayed in the form of a bar graph with a color resolution of activities based on their inclusion in the identified categories.

According to Semjon & Evin [17], the Yamazumi chart is: *“a folded column graph showing the balance of the load cycle time between several operators on the assembly line. It may be made for one or more assembly line products.”* Sabadka et al., [18] characterize Yamazumi chart as a bar chart that shows the total cycle time for each operator when performing their process in the production flow.

In order to plan and solve the problems and then make Yamazumi, the company needs to know the assembly process in detail. It is not enough to know only the important mounting details, but they need to know the storage length of the component or how long does it take to take it out of the shelf. Also, be aware of the long way for the component. It is necessary to record how many seconds are needed to achieve the tool, such as the working times of presses, screwdrivers, welders, laminators, and the like. You should also be aware that the employee is following the work process or not. It is very important to understand the details of the work before it can improve the process while creating a Yamazumi chart. It allows you to create seamless, safe and efficient processes. Visualizes the current situation and highlights the critical points on the assembly operations. Helps solve problems and improves current conditions. [17]

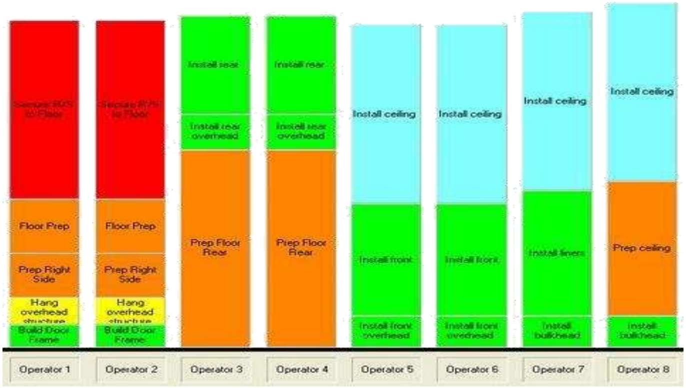


Figure 3
Yamazumi Chart Preview [17]

In practice, the Yamazumi chart can be applied on magnetic boards that are accessible and located near assembly lines. It is very important to have customer audits where it is possible to see the whole assembly process and follow it in successive sequences. Yamazumi presents a certain overview of the assembly sequence, where each part of the step is color-rendered. The graph is shown in Figure 3.

The Yamazumi graph defines and color-differentiates the type of work [17]:

Green designation – Value added work, changing form, properties and value of the product.

Orange labeling – work required, work without added value, but necessary to change shape, properties and value of the product.

Red labeling – work without added value, does not change the form, properties and value of the product.

Yellow labeling – optional work is not performed on each product and depends on the specifications – product type distinction.

Blue labeling – Various work is performed on each product, but its range and duration depend on the specifications.

1.1.3 Lean Layout

Layout, as the layout of production, non-production and storage capacities in the company is undoubtedly one of the most important tasks because it has a direct impact not only on the company economy but also affects safety at work and social environment of the company [19].

Lean Layout is applying the concepts and principles of Lean to something larger than one process. Lean layout is a method of building a space-saving workplace with smooth material flows and productive production.

We consider the layout as the way the process factors are organized in the production process, and how they are redistributed into the individual process activities. Inappropriate layout can lead to too long, confusing and unpredictable flow, customer waiting, long process time, inflexible activity and high cost [20].

With a wrong arrangement of the layout we can observe the different kinds of waste, such as unnecessary handling and shipping, mistakes in production planning, stockpiling, and high rate of w.i.p., long lead time, uneven flow of semi-finished products, low standardization [21].

The economy of work movements requires the observance of principles such as the use of the shortest distances; to logically place the material, tools, controls and the like in the functional area of the worker; the most used and heaviest objects are placed at the height of the work surface; the material should be stored to allow quick and easy grip. [8]

According to [20], the most important features of the lean layout to be taken into account when forming it are:

- Own security – All processes which are potentially dangerous to the customer or employee must be secure.
- The length of the flow of material, information or customers should be proportionate action. This usually means minimizing the distance that must pass. However, this is not always the case, for example in a supermarket.
- Flow transparency – Every material and customer flow should be well marked, clear and obvious to every employee and customer.
- Conditions of employees – Employees should be placed away from the noisy and unpleasant parts of the plant.
- Management coordination – Supervision and communication can help staff to be deployed and use appropriate communication facilities.
- Accessibility of maintenance – All machinery and equipment should be accessible for cleaning and maintenance.

- Usage of space – Space should be used appropriately. This usually means minimizing space. However, for example, luxury hotels need to create a sense of importance and luxury.
- Long-term flexibility - the layout needs to be changed periodically. A good layout is designed to flexibly change with future needs.

1.1.4 Lean Controlling

Management activities in corporate governance are derived from strategy definition (in the form of a strategic plan), followed by operational plans (marketing, sales, finance, human resources, innovation, manufacturing etc.) control future deviations from the stated objectives is the responsibility of controlling bodies (sometimes as part of financial department), which are usually responsible for the management of the risks arising from the deviations found in the small and medium-sized firm as well. Since each plan (strategic and operational) must have defined goals, controlling is also focused on the detection and subsequent management of specific strategic and operational objectives. [22]

Many companies in the manufacturing and process industry have fully embraced the lean management philosophy; their finance and controlling departments, however, are still stuck in a 1920s approach to standard cost calculation. Outdated structures and systems need to adapt to changing requirements and reflect new challenges in production control.

Lean Controlling increases transparency while reducing complexity at the same time. It is based on the following premises [23]:

- Cost optimization with a focus on customer benefit
- More efficient and robust processes with a higher degree of automation
- Leveraging of information that is truly relevant to the steering and controlling process
- A holistic approach to control the value chain based on flow metrics

Lean controlling means transmitting ideas and fundamental elements of lean manufacturing to controlling. These include, in particular, the realization of kaizen thinking, the reduction of the depth of production (Outsourcing), the closer relations with the suppliers as well as the strong process orientation of the company.

New controlling objects are processes at different hierarchical levels (e.g., major or partial processes) and should also become subject(s) of planning and management activities. Implementation of Kaizen also affects the content of controlling. Instead of deviations, the prevention of waste or the assessment of continuous improvement in the production process will be the focus of controlling activities [24].

In the context of lean controlling, controlling efficiency is basically the relationship between controlling outputs and the objectively spent time and cost involved. Lean controlling creates transparency through Lean Produktion & Lean Office. The main task of lean controlling is to highlight the "Black Hole" costs of waste and non-value creation activities, as well as to improve through lean methods [25].

The most important change must be in thinking: Englishmen call it "mindset change". The intent of Lean controlling is to ask: is the statistics, records or report required by us, or we add value to it? If so, we will do our best, if not, to prevent this requirement from arising in the future, so that we can concentrate on a real role: to uncover sources of losses and eliminate them through controlling tools [26].

Benefits resulting from the application of lean controlling methods for individual business areas [23]:

- For accounting – optimize cash flows
- For operational controlling – full transparency in the calculation of materials and production costs
- For strategic controlling – embed strategy within budgeting and forecasting processes
- For lean transformation – drive organizational change towards a culture of continuous improvement

2 Data and Research Methodology

In terms of lean manufacturing, we have mentioned some of the lean manufacturing methods used in automotive companies in Slovakia in previous chapters. The results presented are based on the research tasks carried out in industrial enterprises belonging to the automotive segment and their supply chains. The survey was conducted on a sample of 17 enterprises selected by random selection. The goal of the questionnaire was to determine the level of knowledge and use of lean methods in order to make production processes more efficient. With respect to uniform standards and standards used in automotive, we consider the differences in the application of these methods to be negligible.

The results are shown in Figure 4. We approached this insight based on the long-term analysis and experience from industrial organizations within cooperation and based on a survey conducted in the framework of the research projects.

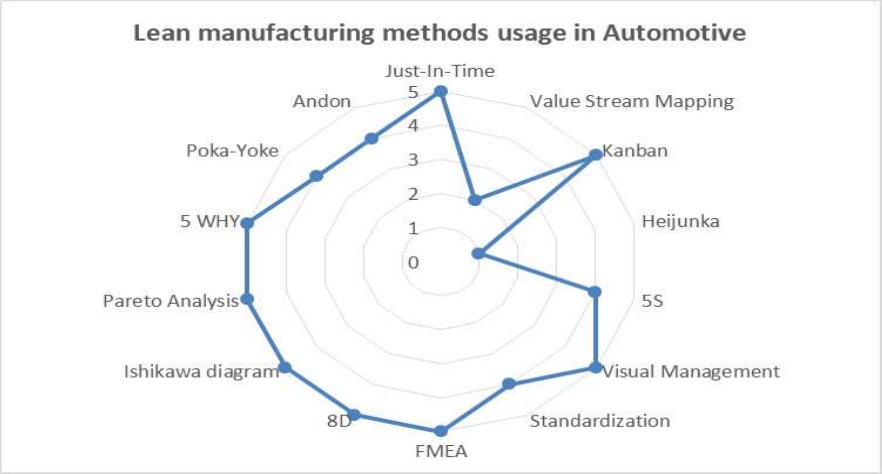


Figure 4
Lean manufacturing methods usage in Automotive (source: own processing based on several surveys)

For each method, we stated value of 1-5, the values mean:

5 It is used throughout the company on a regular basis

4 It is used regularly in the selected department

3 Used in part, insufficiently, without planning or evaluation

2 Used in the past, but no longer

1 Not used at all and never used

From the graph (Figure 4) we have shown that methods such as Ishikawa Diagram, Pareto Analysis, 5W, 8D, FMEA Visual Management, Just in Time and Kanban are used in the automotive industry sufficiently, which may either be their simplicity of application, or relation to the requirements of standards that are obligatory on the automotive industry. Method Heijunka the surveyed enterprises due to its nature, is not used. Method 5S and its related Standardization is used in enterprises, but in some cases only in selected workplaces. The VSM method is poorly utilized in the enterprises surveyed.

In the next part of the case study, we point out the importance and need for applying this method in enterprises, as this method serves to map the value flow and makes it possible to identify the causes of unnecessary waste of resources (time, human labor, material, information or financial). VSM helps to identify possible losses, bottlenecks, and inefficient flows anywhere in the business.

3 Application of Lean Manufacturing Methods – Company Case Study

No production is 100% fluent and does not work without downtime. It means waiting, that is, waste in production, which can lead to failure to meet the required supply of manufactured pieces and stop production at the customer. This part of the paper will deal with the analysis of the use of lean manufacturing methods for component manufacturing - characterized as assembly. Analyzed production can be characterized as manual - without the use of machinery. The analysis was processed using the VSM method, with a map of the current material and information flow of production shown in Figure 5.

The map (see on Figure 5) shows the production process at individual assembly stations on two parallel lines, where two components are made, forming one unit after their subsequent assembly. For each station, the following basic parameters are selected: TT (Target Cycle Time), CT (Cycle Time), Utilization, Shifts and WIP (Work in progress - processed production, pieces). TT and TCT are given by calculation, Cycle time represents the real time of the operation calculated using a predetermined time method (MOST + tracking time on lines). These data provide an overview of the difference between Target Operating Time (TCT) and Real Time on Line (CT).

These data inform about what are the time losses and differences between the different assembly station. Percentage utilization of operators on the line shows how real workloads are performed by operators at the station concerned.

Results from VSM have revealed areas of production that cause losses and waste and which need to be further targeted. These are the following:

- The high proportion of non-value adding (NVA) time for the assembly process - the linearity difference (TCT) and the real time of operation (CT) is high.
- This also means low / too high percentage of operators on the line.
- Stocks of finished components are high - the second line has no inventory.

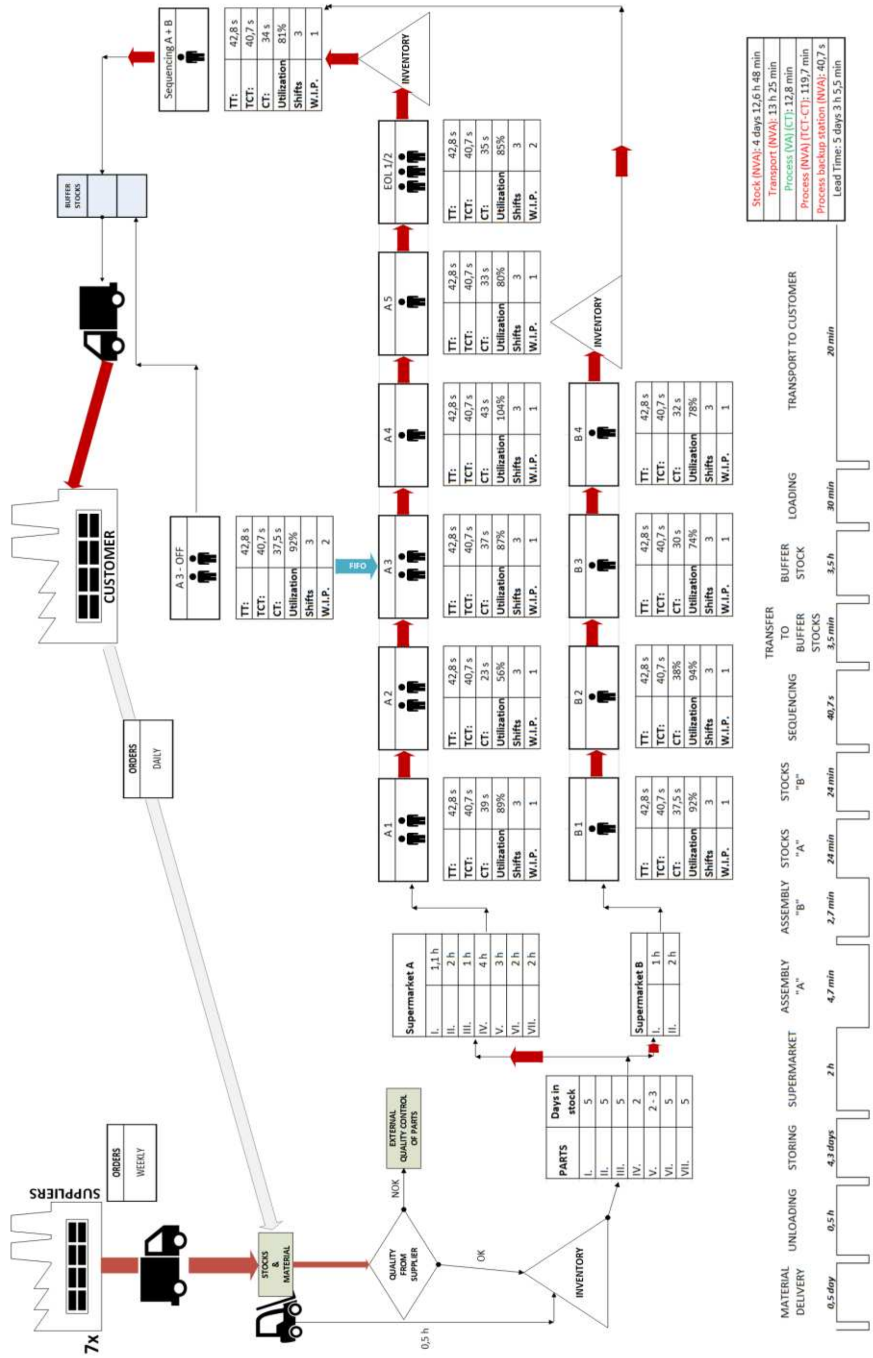


Figure 5
Value Stream Mapping – original state (source: own processing based on company internal materials)

In particular, we focused directly on the lines, i.e. to identify the cause of the high difference CT versus TCT. Due to the difference in times, it is necessary to analyze the possibility of balancing operations between stations, consequently the need to re-evaluate the number of operators on both lines, their operational utilization, and also the type of activities performed by the operators in the operation.

A detailed timing analysis of activities at individual stations was transferred to the graphical form - the Yamazumi chart (Figure 6). This chart shows all the times of the operations performed within the operation, redistributed between components "60" and "40" (when two operators are working on the station). By cumulating the individual activity times, we receive the cycling time of the entire operation (CT). The individual activities are color-coded (NVA activities - red, VA activities - green and necessary work, but no added value - yellow).

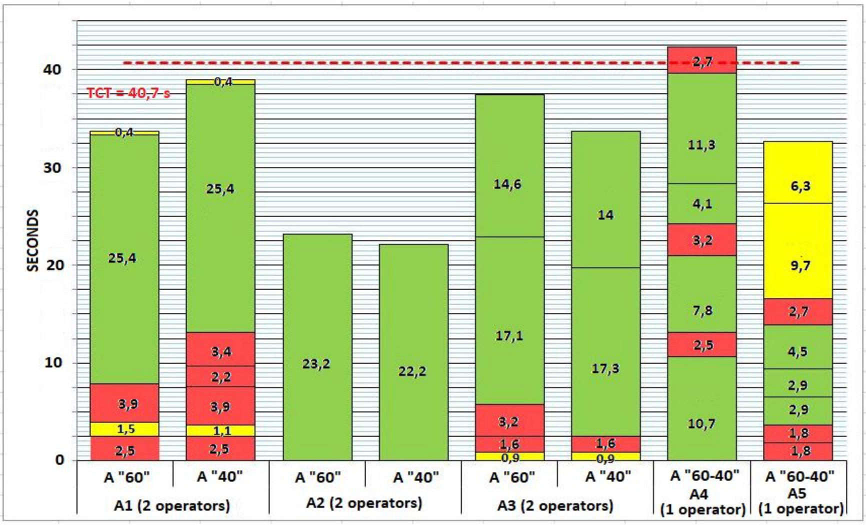


Figure 6
Yamazumi chart of production line A – original state (source: own processing based on company internal materials)

The chart provides an overview of adding value activities and activities that do not add value to the assembled component. The chart and individual actions show that the station operator A1 takes a lot of time walking the material and scanning material - NVA activities. At station A2, operators do not perform any NVA activities. It is necessary to review this operation and see if it is possible to prosecute it to perform only one operator on both components. The A4 station is the Bottleneck station (the longest installation station, heavy duty). Several sub-activities are carried out on it, as on station A5, which has a certain time-off against the production cycle.

On-site observation confirmed that at station A1, operators must walk longer distances during assembly time, especially for parts. The station A2 are compared to the operators of the completed operation in advance, and are waiting to complete the cycle. A detailed analysis of operations and tracking on the line allowed us to identify the activities that need to be focused on streamlining the production process (e.g. walking on material), as well as potential options for better unbalance of individual line A operations (in particular A2 and A4 operations). By analyzing the individual activities, we identified the potential improvement points that are shown in the new, optimized VSM diagram (see Figure 7). The changes listed below lead to the elimination of ineffective and non-creasing time values and to better utilization of the operators' workforce on the line.

Optimization of time on line A will not be done without changing the original layout of the production line, due to the shortening of operator walks, the demanding translation of parts and the saving of human work. The change of layout is based on the deficiencies and wastage found on the A1 station, which was needed outside the line - as an offline station. For this reason, assembly from station A2 has moved to station A1. As a result, station A2 became empty station and was omitted from the new layout. A FIFO stack for individual parts was placed between the A1 OFF and A1 stations.

Following the optimized VSM, a modified Yamazumi chart (Figure 8) has been developed to illustrate the time structure of all operations in the structure of times of effective operations (green color), times of required operations without added value (yellow color) and times of ineffective actions (red color).

It is clear from the graph that after the proposed reorganization of the work of the individual workplaces, the use of the operators' work is increased, in addition, the inefficient times for the assembly of individual parts have been significantly reduced, leading to an increase in the overall efficiency of production on line A.

The proposed new status brings benefits not only in reducing NVA times and achieve standard use of the work of operators, but also in reducing the number of operators required to achieve the production volume. At A1 Station OFF, unnecessary walking was reduced and the time required for installation was saved. Due to the time saved at this station, a time reserve for the line tact time was created.

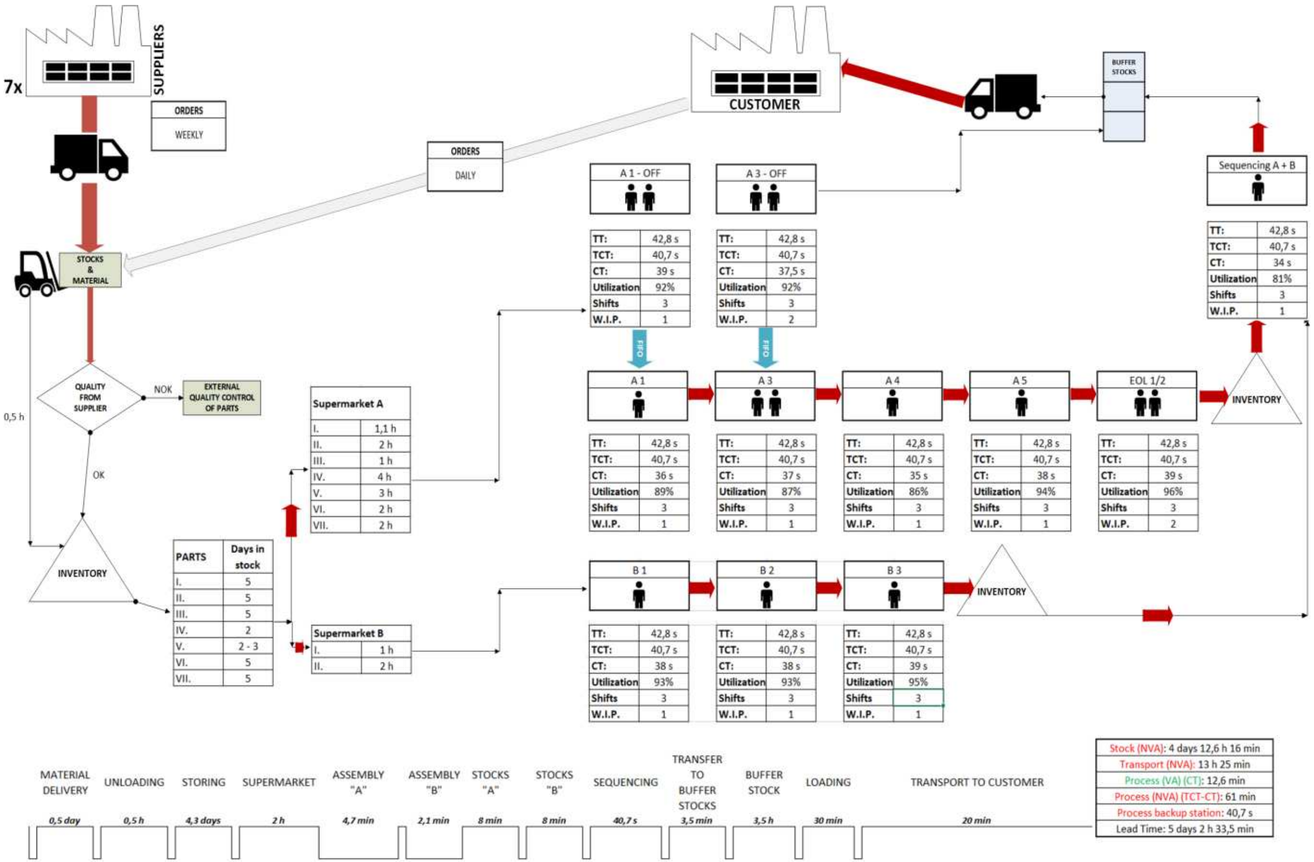


Figure 7

Value Stream Mapping – proposed state (source: own processing based on company internal materials)

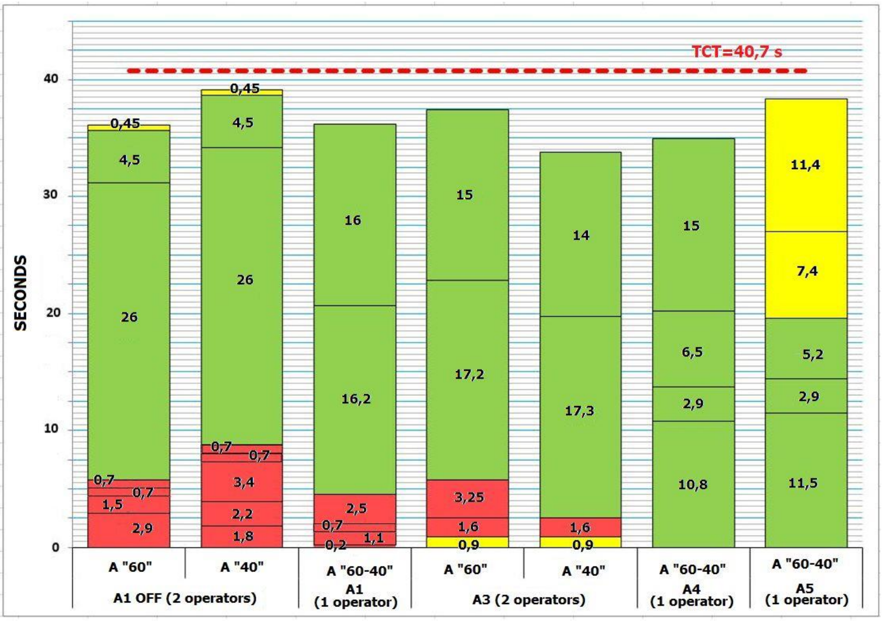


Figure 8

Yamazumi chart of production line A – proposed state (source: own processing based on company internal materials)

4 Results and Discussion

The main objective of the overall design of streamlining is to eliminate identified shortcomings and waste, to ensure smoother production and meet customer requirements, and to reduce the business costs of all 7 basic types of waste, the proposed solutions reduce unnecessary movements, handling, unnecessary processes, and excess inventory. In addition, the solution proposal has been able to unbalance the use of operators' working hours at individual stations of assembly lines. The result is the most balanced utilization of individual operators to the required values (85-95%).

In general, the benefits of lean manufacturing are reduced inventory, less process waste, less rework, reduced lead time, financial saving and increase process understanding.

The main benefits of the proposal for making the layout more efficient are the follows:

- Reduction of unnecessary walking and manipulation at the current assembly stations of both lines

- Reduction of surplus stocks of finished products
- Improving ergonomic conditions when performing operations on stations A1 and A5.

The proposed layout streamlining is a change that leads to a reduction in the time that does not add value to the resulting product, to ensuring that the standard values of the operator's percentage usage are achieved, and to reducing the number of operators and reducing the production time.

Conclusion

The implementation of Lean Manufacturing, in business, is a strategic decision of the company’s Top Management. Management makes this decision based on certain expectations of this type of change. Lean production is not just a set of methods and tools, it offers much more. Lean manufacturing has to be understood as a philosophy and discipline that must be rooted in every employee, from senior management to the manufacturing workers on the line. This idea is based on continuous improvement (Kaizen) - the company should constantly strive to make production more efficient, improve processes, improve working conditions and all activities that lead to the enhancement and achievement of better lean production.

Acknowledgement

This paper was supported by the Grant Agency of the Ministry of Education, Science, and Research of Slovak Republic Projects VEGA 1/0348/17 “*The impact of the coexistence of different generations of employees on the sustainable performance of organizations”* and also VEGA 1/0218/16 "*The model of the implementation of controlling as a management tool within small and medium enterprises in the engineering and electronics industries*“.

References

[1] BALOG, Michal; STRAKA, Martin. Application of the logistics principles for the company Omega, sro in crisis time. *Acta logistica*, 2014, 1.1: 17-21

[2] ABDULMALEK, F.A.; RAJGOPAL, J., (2007) Analyzing the benefits of lean manufacturing and value stream mapping via simulation: A process sector case study. *International Journal of production economics*, 2007, 107.1: 223-236, https://doi.org/10.1016/j.ijpe.2006.09.009

[3] SAYER, N. and WILLIAMS, B. (2011) *Lean For Dummies.* Hoboken: John Wiley & Sons, Inc., 2011, ISBN: 978-1-118-05118-4

[4] BOSENBERG D., METZEN H., (1997) *Lean Manažment: Náskok pomocou štíhlych konceptov.* Bratislava, Vydavateľstvo SLOVO. 272s. ISBN 85711-16-8

[5] CHAUHAN, P., et al., (2015) Application of Lean Manufacturing Principles for Process Time Reduction–A case of Conveyor Pulley Manufacturing, In: *Proceedings of 5th National Conference on Recent Advances in Manufacturing (RAM-2015)* 2015, pp. 423-428

[6] SVOZILOVÁ, A., (2011) *Zlepšování podnikových procesů*. Praha: Grada, ISBN 978-80-247-3938-0

[7] KEŘKOVSKÝ, M., (2009) *Moderní přístupy k řízení výroby.* Vydavateľstvo H.C.Beck, Praha 2009, 2. vydanie, ISBN 978-80-7400-119-2

[8] KOŠTURIAK, J., FROLÍK. Z., (2006) *Štíhlý a inovativní podnik*. Praha : Alfa Publishing, s.r.o., 2006, ISBN 80-86851-38-9

[9] SUNDAR, R.; BALAJI, A. N.; KUMAR, RM Satheesh. A review on lean manufacturing implementation techniques. *Procedia Engineering*, 2014, 97: 1875-1885, https://doi.org/10.1016/j.proeng.2014.12.341

[10] DLABAČ, J. (2014) Štíhly materiálový tok. In: *Štíhla výroba a logistika* [online]. Praha: Strojárenský mesačník MM. Kód článku: 140430. s.14 [cit. 2018-04-03] Dostupné na internete: http://www.mmspektrum.com/clanek/stihly-materialovy-a-hodnotovy-tok.html

[11] WOMACK, J. P. – JONES, D. T.: *Lean Thinking*. 2nd Edition, Free Press New York, 2003, ISBN 0-7432-4927-5

[12] ONOFREJOVÁ, D. (2015) Mapovanie hodnotového toku a významnosť vo výrobnom procese [online]. Transfer inovácií. [cit. 2018-04-15] Retrieved from: https://www.sjf.tuke.sk/transferinovacii/ pages/archiv/ transfer/32-2015/pdf/195-198.pdf

[13] SALAJ, M. (2010) *Mapovanie hodnotového toku – value stream mapping* [online] [cit. 2018-05-30] Retrieved from: http://www.leanportal.sk/Files/Modely/Mapy%20hodnotovych%20tokov.pdf

[14] *Mapovanie hodnotového toku. Value Stream Mappping* [online] [cit. 2018-06-29] Retrieved from https://www.ipaslovakia.sk/files/6314-najlepsiepraktikymetodymapovanietokuhodnot

[15] FEKETE, M., HULVEJ, J. (2013) *„Humanizing“ Takt Time and Productivity in the Labor – Intesive Manufactoring Systems* [online] Zadar, Croatia – International Conference 2013 [cit. 2018-05-16] ISBN 978-961-6914-02-4. Dostupné na internete: http://www.toknowpress.net/ISBN/978-961-6914-02-4/papers/ML13-245.pdf

[16] ČERVINKA, M. (2013) *Takt time* [online] Portál Štíhla výroba [cit. 2018-05-16] Dostupné na internete: http://www.stihlavyroba.sk/2013/04/takt-time.html

[17] SEMJON V. and EVIN. E., (2009) Increasing the productivity of the assembly line by balancing of assembly stations using the Yamazumi method. *Transfer Inovácií* 13, 2009, 73-77

[18] SABADKA, Dušan, et al. Optimization of Production Processes Using the Yamazumi Method. *Advances in Science and Technology. Research Journal*, 2017, 11.4: 175-182, DOI: https://doi.org/ 10.12913 /22998624 /80921

[19] ŠIMKO, D. Methods of distribution, layout and hungarian method. METHOD. *Acta logistica*, 2016, 3.1: 9-13

[20] *Lean Layout* [online] IPA Slovakia [cit. 2018-05-17] Retrieved from: https://www.ipaslovakia.sk/sk/ipa-slovnik/lean-layout

[21] MAŠÍN, I., VYTLAČIL, M. (1996) *Cesty k vyšší produktivitě*. Liberec: Institut průmyslového inženýrství. s. 55-56, ISBN 80-902235-0-8

[22] HAVLÍČEK, K. (2011) *Manažment & controlling malé a střední firmy*. 1. vydání. Praha: Vysoká škola finanční a správní, o.p.s., 2011, 212 s. Edice EUPRESS. ISBN 978-80-7408-056-2

[23] *The efficient way to control Lean Management* [online] [cit. 2018-05-17] Retrieved from https://www.camelot-mc.com/en/client-services/finance-performance-management/lean-controlling/

[24] *Lean Controlling* [online] [cit. 2018-05-17] Retrieved from: http://www.betriebswirtschaft-lernen.net/erklaerung/lean-controlling/

[25] *Lean Controlling. Integraler Bestandteil einer Lean Tansformation* [online] [cit. 2018-05-17] Retrieved from: https://www.iqxperts-consulting.com/lean-controlling.html

[26] *Mitől lesz egy megoldás lean? És mi köze ennek a controllinghoz? What makes a solution lean? And what has this to do the controlling?* [online] [cit. 2018-06-29] Retrieved from https://www.controllingportal.hu/ mitol_lesz_egy_megoldas/

Properties of the cross-product of Bessel and modified Bessel functions of the first kind

Anikó Szakál[1]

[1] Óbuda University, University Research and Innovation Center, Budapest, Hungary
e-mail: szakal@uni-obuda.hu

Abstract: *In this note our aim is to present two new integral representations for the cross-product of Bessel and modified Bessel functions of the first kind, and to point out that this cross-product is in fact the solution of a fourth-order linear homogeneous Bessel-type differential equation. Moreover, we point out that an inequality by Ashbaugh and Benguria as well as of Ashbaugh and Laugesen, involving the cross-product of Bessel functions, can be shown by using the method of Lagrange multipliers.*

Keywords: *Bessel functions; modified Bessel functions; Wronski determinant; contour integral; Hankel integral; fourth order differential equation; asymptotics; Lagrange multipliers.*

MSC (2010): 33C10.

1 Introduction

Let J_ν and I_ν denote the Bessel and modified Bessel functions of the first kind. Motivated by their appearance as eigenvalues in the clamped plate problem for the ball, Ashbaugh and Benguria have conjectured that the positive zeros of the function

$$z \mapsto \Phi_\nu(z) = J_\nu(z)I'_\nu(z) - J'_\nu(z)I_\nu(z)$$

increase with ν on $\left[-\frac{1}{2},\infty\right)$. Lorch [5] verified this conjecture and presented some other properties of the zeros of the above cross-product of Bessel and modified Bessel functions. His result has been used in [2] by Ashbaugh and Benguria related to Rayleigh's conjecture for the clamped plate and its generalization to three dimensions. In [1] the authors extended the above result of Lorch and proved that in fact the positive zeros of the above cross-product or Wronskian increase with ν on $(0,\infty)$. Motivated by the above results, in this note we make a further contribution to the subject and our aim is to present two new integral representations for the cross-product of Bessel and modified Bessel functions of the first kind. Moreover, we point out that this cross-product is the solution of a Bessel-type fourth order differential equation and its asymptotic expansion for large arguments can be obtained from known results on hypergeometric functions. Finally, we present an alternative proof of an inequality by Ashbaugh and Benguria [2] as well as of Ashbaugh and Laugesen [3], involving the cross-product of Bessel functions, by using the classical method of Lagrange multipliers.

2 Integral representations of the cross-product of Bessel functions

By using the known recurrence relations

$$zJ'_\nu(z) - \nu J_\nu(z) = -zJ_{\nu+1}(z)$$

and

$$zI'_\nu(z) - \nu I_\nu(z) = zI_{\nu+1}(z),$$

the cross-product $J_\nu(z)I'_\nu(z) - J'_\nu(z)I_\nu(z)$ actually can be written as

$$\Phi_\nu(z) = J_{\nu+1}(z)I_\nu(z) + J_\nu(z)I_{\nu+1}(z).$$

It has been shown that the cross–product $\Phi_\nu(z)$ possesses the series form [1, p. 821, Lemma 2]

$$\Phi_\nu(z) = 2\sum_{n\geq 0} \frac{(-1)^n \left(\frac{z}{2}\right)^{2\nu+4n+1}}{n!\Gamma(\nu+n+1)\Gamma(\nu+2n+2)}, \qquad \nu > -1, z \in \mathbb{C}. \tag{2.1}$$

However, by the Legendre duplication formula

$$\Gamma(2w) = \frac{2^{2w-1}}{\sqrt{\pi}}\Gamma(w)\Gamma\left(w+\tfrac{1}{2}\right), \qquad \Re(w) > 0,$$

transforming the denominator in (2.1) we get

$$\begin{aligned}\Phi_\nu(z) &= 2\sum_{n\geq 0} \frac{(-1)^n \left(\frac{z}{2}\right)^{2\nu+4n+1}}{n!\Gamma(\nu+n+1)\Gamma(\nu+2n+2)} \\ &= \frac{\sqrt{\pi}z^{2\nu+1}}{2^{3\nu+1}\Gamma(\nu+1)\Gamma\left(\frac{\nu}{2}+1\right)\Gamma\left(\frac{\nu}{2}+\frac{3}{2}\right)} \sum_{n\geq 0} \frac{\left(-\frac{z^4}{64}\right)^n}{n!\,(\nu+1)_n\left(\frac{\nu}{2}+1\right)_n\left(\frac{\nu}{2}+\frac{3}{2}\right)_n} \\ &= \frac{z^{2\nu+1}}{2^{2\nu}\Gamma(\nu+1)\Gamma(\nu+2)}\,{}_0F_3\left(\frac{\nu}{2}+1, \frac{\nu}{2}+\frac{3}{2}, \nu+1; -\frac{z^4}{64}\right), \end{aligned} \tag{2.2}$$

where the multiplicative constant in front of the generalized hypergeometric term we infer by another use of Legendre's formula.

Next, consider the line integral form of the generalized hypergeometric function [6, 16.5.1], adopted to our situation:

$$\begin{aligned}&{}_0F_3\left(\frac{\nu}{2}+1, \frac{\nu}{2}+\frac{3}{2}, \nu+1; -\frac{z^4}{64}\right) \\ &\quad = \frac{\Gamma(\nu+1)\Gamma(\nu+2)}{2^{\nu+1}\mathrm{i}\sqrt{\pi}} \int_{\mathscr{L}} \frac{\Gamma(-s)\left(\frac{z^4}{64}\right)^s \,\mathrm{d}s}{\Gamma(\nu+1+s)\Gamma\left(\frac{\nu}{2}+1+s\right)\Gamma\left(\frac{\nu}{2}+\frac{3}{2}+s\right)},\end{aligned}$$

where $\mathscr{L}$ is a contour that starts at infinity on a line parallel to the positive real axis, encircles the nonnegative integers in the negative sense, and ends at infinity on another line parallel to the positive real axis. After some routine transformations we arrive at

Theorem 1. *For all $\nu > -1, z \neq 0$ there holds the integral representation*

$$\Phi_\nu(z) = \frac{1}{\mathrm{i}\sqrt{\pi}} \frac{z^{2\nu+1}}{2^{3\nu+1}} \int_{\mathscr{L}} \frac{\Gamma(-s)\left(\dfrac{z^4}{64}\right)^s \mathrm{d}s}{\Gamma(\nu+1+s)\Gamma\left(\frac{\nu}{2}+1+s\right)\Gamma\left(\frac{\nu}{2}+\frac{3}{2}+s\right)}. \tag{2.3}$$

In turn, having in mind the Hankel loop-integral formula for the reciprocal Gamma function [6, 5.9.2]

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi\mathrm{i}} \int_{-\infty}^{(0+)} \mathrm{e}^t t^{-z} \mathrm{d}t, \qquad z \in \mathbb{C},$$

where the integration path starts at infinity on the real axis, encircling 0 in a positive sense, and returning to infinity along the real axis, respecting the cut along the positive real axis. In turn, this formula is equivalent with the Bromwich–Wagner type complex line integral

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi\mathrm{i}} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \mathrm{e}^s s^{-z} \mathrm{d}s, \qquad c > 0. \tag{2.4}$$

Indeed, consider the Fourier–integral

$$\frac{\mathrm{e}^c}{2\pi} \int_{\mathbb{R}} (c+\mathrm{i}t)^{-z} \mathrm{e}^{\mathrm{i}t} \mathrm{d}t, \qquad c > 0.$$

The integrand has one branch point $t = \mathrm{i}c$ in the upper half–plane. Taking the branch cut $B = [\mathrm{i}c, \mathrm{i}\infty)$ we deform the contour of integration so that it runs counterclockwise from $\mathrm{i}\infty$ to $\mathrm{i}\infty$ around B. Combined with the definition of the Gamma function, this will give an expression proportional to $\Gamma(1-z)\sin(\pi z)$. The Euler's reflection formula and the change of variable $s \mapsto c + \mathrm{i}t$ finishes the derivation of (2.4).

Theorem 2. *For all $\nu > -1$, $c > 0$ and $z \in \mathbb{C}$, we have*

$$\Phi_\nu(z) = \frac{z}{2\pi\mathrm{i}} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \mathrm{e}^t t^{-2} J_\nu\left(\frac{z^2}{2t}\right) \mathrm{d}t\,. \tag{2.5}$$

To prove this, inserting $1/\Gamma(\nu+2n+2)$ expressed *via* (2.4) into $\Phi_\nu(z)$, we get

$$\begin{aligned}
\Phi_\nu(z) &= \frac{1}{\pi\mathrm{i}} \sum_{n\geq 0} \frac{(-1)^n \left(\frac{z}{2}\right)^{2\nu+4n+1}}{n!\,\Gamma(\nu+n+1)} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \mathrm{e}^t t^{-\nu-2n-2} \mathrm{d}t \\
&= \frac{1}{\pi\mathrm{i}} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \mathrm{e}^t t^{-\nu-2} \sum_{n\geq 0} \frac{(-1)^n \left(\frac{z}{2}\right)^{2\nu+4n+1}}{n!\,\Gamma(\nu+n+1) t^{2n}} \mathrm{d}t \\
&= \frac{1}{\pi\mathrm{i}} \left(\frac{z}{2}\right)^{2\nu+1} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \mathrm{e}^t t^{-\nu-2} \sum_{n\geq 0} \frac{\left(-\dfrac{z^4}{16t^2}\right)^n}{n!\,\Gamma(\nu+n+1)} \mathrm{d}t,
\end{aligned}$$

which is equivalent to the assertion, since the rest is obvious.

3 A fourth-order Bessel-type differential equation

The Bessel function of the first kind J_ν is a particular solution of the second-order linear homogeneous Bessel differential equation, while the modified Bessel function of the first kind I_ν is a particular solution of the second-order linear homogeneous modified Bessel differential equation. In this section we would like to point out that their Wronskian, that is, the cross-product $J_\nu(z)I'_\nu(z) - J'_\nu(z)I_\nu(z)$ is a particular solution of the following fourth-order linear homogeneous Bessel-type differential equation

$$z^4 w''''(z) + 4z^3 w'''(z) + (1-4\nu^2)(z^2 w''(z) + zw'(z)) + (4\nu^2 - 1 + 4z^4)w(z) = 0. \tag{3.1}$$

This can be verified by using the fact that J_ν and I_ν are solutions of Bessel and modified Bessel differential equations or we can use the method of Frobenius and seek the solution of (3.1) in form of a power series and arrive to (2.2). If we write the equation (3.1) in the form

$$w''''(z) + \frac{4}{z}w'''(z) + (1-4\nu^2)\left(\frac{w''(z)}{z^2} + \frac{w'(z)}{z^3}\right) + \left(\frac{4\nu^2-1}{z^4} + 4\right)w(z) = 0, \tag{3.2}$$

then this equation has a regular singularity at the origin and an irregular singularity at the point at infinity, all other points of the complex plane are regular or ordinary points for the differential equation. Note that the classical Bessel and modified Bessel differential equations have the same classification. A calculation shows that the Frobenius indicial roots for the regular singularity of the differential equation (3.2) at the origin 0 are $\{-1, 1, 1-2\nu, 1+2\nu\}$. The application of the Frobenius power series method yields four linearly independent series solutions of (3.2), each with infinite radius of convergence in the complex plane. If we use the transformation $q(z) = \sqrt{z}w(z)$, then (3.1) will become

$$z^4 q''''(z) + 2z^3 q'''(z) - \left(4\nu^2 + \frac{1}{2}\right)z^2 q''(z) + \frac{3}{2}zq'(z) + \left(4z^4 + \frac{21}{16}\right)q(z) = 0,$$

which according to the Wolfram Alpha software has the general solution

$$\begin{aligned} q_\nu(z) =& c_1 \cdot z^{-\frac{1}{2}}{}_0F_3\left(\frac{1}{2}, \frac{1}{2} - \frac{\nu}{2}, \frac{\nu}{2} + \frac{1}{2}; -\frac{z^4}{64}\right) \\ &+ c_2 \cdot z^{\frac{3}{2}}{}_0F_3\left(\frac{3}{2}, 1 - \frac{\nu}{2}, \frac{\nu}{2} + 1; -\frac{z^4}{64}\right) \\ &+ c_3 \cdot z^{\frac{3}{2}-2\nu}{}_0F_3\left(1-\nu, 1-\frac{\nu}{2}, \frac{3}{2} - \frac{\nu}{2}; -\frac{z^4}{64}\right) \\ &+ c_4 \cdot z^{\frac{3}{2}+2\nu}{}_0F_3\left(\frac{\nu}{2}+1, \frac{\nu}{2} + \frac{3}{2}, \nu+1; -\frac{z^4}{64}\right). \end{aligned}$$

We can see that this is in agreement with our knowledge on equation (3.1). More precisely, the powers of z in the above general solution, that is,

$$\left\{-\frac{1}{2},\frac{3}{2},\frac{3}{2}-2\nu,\frac{3}{2}+2\nu\right\}$$

correspond exactly to Frobenius indices, that is, they are

$$\frac{1}{2}+\{-1,1,1-2\nu,1+2\nu\}.$$

In view of (2.2), this shows that indeed the cross-product $\Phi_\nu(z)$ is a particular solution of the fourth-order linear homogeneous Bessel-type differential equation (3.1).

Asymptotic series expansion for large arguments for the cross-product $\Phi_\nu(z)$ can be obtained by using the well-known asymptotic series of $J_\nu(z)$, $J'_\nu(z)$, $I_\nu(z)$ and $I'_\nu(z)$ for large arguments. However, because of the ${}_0F_3$ representation of the cross-product $\Phi_\nu(z)$, it is more convenient to use the asymptotic expansion of hypergeometric functions. Since for $|z|\to\infty$

$${}_0F_3(a,b,c;z)=\frac{\Gamma(a)\Gamma(b)\Gamma(c)}{4\sqrt{2}\pi\sqrt{\pi}}e^{4\sqrt[4]{z}}z^{\frac{1}{4}\left(\frac{3}{2}-a-b-c\right)}\left(1+\mathscr{O}\left(\frac{1}{\sqrt[4]{z}}\right)\right),$$

in view of (2.2) we get for $|z|\to\infty$

$$\Phi_\nu(z)=\frac{e^{z\sqrt{2\mathrm{i}}}}{2^{\nu+\frac{3}{2}}\pi^2}\left(\frac{z\sqrt{\mathrm{i}}}{2\sqrt{2}}\right)^{2-2\nu}\left(1+\mathscr{O}\left(\frac{1}{z\sqrt{2\mathrm{i}}}\right)\right).$$

4 An inequality by Ashbaugh and Benguria for the cross-product of Bessel functions

Let

$$f_\nu(x)=x^{2\nu+1}\left(\frac{J_{\nu+1}(x)}{J_\nu(x)}+\frac{I_{\nu+1}(x)}{I_\nu(x)}\right)$$

and consider the expression $F_\nu(a)=f_\nu(k_{\nu,1}a)+f_\nu(k_{\nu,1}b)$, where $a^n+b^n=1$, $\nu=n/2-1$ and $k_{\nu,1}$ denotes the first positive zero of f_ν, that is, of Φ_ν. Ashbaugh and Benguria [2] proved that for $n\in\{2,3\}$, $a^n+b^n=1$ and $j_{\nu,1}/k_{\nu,1}<b<1$, where $j_{\nu,1}$ is the first positive zero of J_ν, the inequality

$$F_\nu(a)=f_\nu(k_{\nu,1}a)+f_\nu(k_{\nu,1}b)<0 \tag{4.1}$$

is valid. In this section our aim is to show the following result.

Theorem 3. *The inequality* (4.1) *holds true for* $n\geq 4$, $a^n+b^n=1$ *and* $a,b\in(0,1)$.

For this, we consider the function

$$L_\nu(a,b,\lambda)=f_\nu(k_{\nu,1}a)+f_\nu(k_{\nu,1}b)+\lambda(1-a^n-b^n)$$

and employ the classical method of Lagrange multipliers to find the critical value of $F_\nu(a)$. The system

$$\begin{cases} \dfrac{\partial L_\nu(a,b,\lambda)}{\partial a} = k_{\nu,1} f_\nu'(k_{\nu,1}a) - n\lambda a^{n-1} = 0 \\ \dfrac{\partial L_\nu(a,b,\lambda)}{\partial b} = k_{\nu,1} f_\nu'(k_{\nu,1}b) - n\lambda b^{n-1} = 0 \\ \dfrac{\partial L_\nu(a,b,\lambda)}{\partial \lambda} = 1 - a^n - b^n = 0 \end{cases}$$

gives the stationary points of the Lagrange function $L_\nu(a,b,\lambda)$. Combining the first two equations we get

$$\frac{f_\nu'(a)}{a^{n-1}} = \frac{f_\nu'(b)}{b^{n-1}}.$$

On the other hand, by using the Mittag-Leffler expansions for Bessel and modified Bessel functions of the first kind, we have that the function

$$x \mapsto \frac{f_\nu'(x)}{x^{2\nu+1}} = 2 + \frac{J_{\nu+1}^2(x)}{J_\nu^2(x)} - \frac{I_{\nu+1}^2(x)}{I_\nu^2(x)} = 2 + \left(\sum_{n\geq 1} \frac{2x}{j_{\nu,n}^2 - x^2}\right)^2 - \left(\sum_{n\geq 1} \frac{2x}{j_{\nu,n}^2 + x^2}\right)^2$$

is increasing on $(0, j_{\nu,1})$ since

$$\left(\sum_{n\geq 1} \frac{2x}{j_{\nu,n}^2 - x^2}\right)^2 - \left(\sum_{n\geq 1} \frac{2x}{j_{\nu,n}^2 + x^2}\right)^2 = \sum_{n\geq 1} \frac{4 j_{\nu,n}^2 x}{j_{\nu,n}^4 - x^4} \sum_{n\geq 1} \frac{4x^3}{j_{\nu,n}^4 - x^4}$$

increases with x on $(0, j_{\nu,1})$ as a product of two increasing and positive functions of x. Here $j_{\nu,n}$ denotes the nth positive zero of J_ν. Therefore, whenever $a, b \in (0,1) \subset (0, j_{\nu,1})$, they should be equal and then $a = b = 2^{-1/n}$.

Now, in view of the infinite product representation (see [1, 4]) of $\Phi_\nu(x)$ as well as of $\Pi_\nu(x) = J_\nu(x) I_\nu(x)$ we get

$$f_\nu(x) = \frac{x^{2\nu+2}}{\nu+1} \prod_{n\geq 1} \frac{\gamma_{\nu,n}^2 - x^4}{j_{\nu,n}^4 - x^4} \frac{j_{\nu,n}^4}{\gamma_{\nu,n}^2},$$

where $\gamma_{\nu,n}$ denotes the nth positive zero of $\Phi_\nu(\sqrt{x})$. According to [4, Theorem 1] all the zeros of $\Phi_\nu(\sqrt{x})$ are real and thus if we consider the value $f_\nu(k_{\nu,1}2^{-1/n})$, then its sign depends only on the difference $\Delta = j_{\nu,1}^4 - k_{\nu,1}^4 \cdot 2^{-4/n}$, since the other members of the infinite product are all positive. But, Δ is negative, since according to [3] we have $2^{1/n} j_{\nu,1} < k_{\nu,1}$ for $n \geq 4$. This implies that

$$f_\nu(k_{\nu,1}2^{-1/n}) < 0$$

for $n \geq 4$.

On the other hand, F_ν can be estimated from above by the maximum of its critical values and its two marginal values. In our particular case, see the Lagrange multipliers, it follows that for all $a \in [0,1]$ we get

$$F_\nu(a) \leq \max\left\{F_\nu(0), F_\nu(1), F_\nu\left(2^{-1/n}\right)\right\}.$$

Note that $F_\nu(0) = F_\nu(1) = 0$ and due to the fact that $f_\nu(k_{\nu,1}2^{-1/n}) < 0$ for $n \geq 4$, it follows that $F_\nu(a) \leq 0$ for all $a \in [0,1]$. If there is an $a_0 \in (0, 2^{-1/n}]$ such that $F_\nu(a_0) = 0$, by the last relation (and again by Lagrange multipliers) we have necessarily that F_ν is identically zero on $[0, a_0]$, which is not possible.

Thus, indeed $F_\nu(a) < 0$ for $n \geq 4$, $a^n + b^n = 1$ and $a, b \in (0,1)$. Moreover, since f_ν is increasing on $(0, j_{\nu,1})$ for each $\nu > 0$, it follows that for $n \geq 4$, $a^n + b^n = 1$ and $a, b \in (0,1)$ we have that $f_\nu(2^{1/n} j_{\nu,1}a) < f_\nu(k_{\nu,1}a)$ and $f_\nu(2^{1/n} j_{\nu,1}b) < f_\nu(k_{\nu,1}b)$ and in view of (4.1) this in turn implies the following result.

Theorem 4. *The inequality*

$$f_\nu(2^{1/n} j_{\nu,1}a) + f_\nu(2^{1/n} j_{\nu,1}b) < 0 \tag{4.2}$$

holds true for each $n \geq 4$, $a^n + b^n = 1$ *and* $a, b \in (0,1)$.

Note that inequality (4.2) was proved by Ashbaugh and Laugesen [3, eq. (5.3)] in the case when $n \geq 4$, $a^n + b^n = 1$ and $0 < a < 2^{-1/n}$.

References

[1] H.A. Al-Kharsani, Á. Baricz, T.K. Pogány, Starlikeness of a cross-product of Bessel functions. *J. Math. Inequal.* 10(3) (2016) 819–827.

[2] M.S. Ashbaugh, R.D. Benguria, On Rayleigh's conjecture for the clamped plate and its generalization to three dimensions, *Duke Math. J.* 78(1) (1995) 1–17.

[3] M.S. Ashbaugh, R.S. Laugesen, Fundamental tones and buckling loads of clamped plates, *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* 23(2) (1996) 383–402.

[4] Á. Baricz, A. Szakál, R. Szász, N. Yağmur, Radii of starlikeness and convexity of a product and cross-product of Bessel functions, *Results Math.* 73(2) (2018) Art. 62, 34 pp.

[5] L. Lorch, Monotonicity of the zeros of a cross-product of Bessel functions, *Methods Appl. Anal.* 1(1) (1994) 75–80.

[6] F.W.J. Olver, D.W. Lozier, R.F. Boisvert, C.W. Clark (Eds.), *NIST Handbook of Mathematical Functions*, Cambridge Univ. Press, Cambridge, 2010.

State-space Analysis of the Interval Merging Binary Tree

István Finta

Nokia, Bell Labs
H-1083, Budapest, Bókay street 36-42
istvan.finta@nokia-bell-labs.com

Sándor Szénási

Óbuda University
H-1034, Budapest, Bécsi street 96/b
szenasi.sandor@nik.uni-obuda.hu

Abstract: In the course of transmission through networks a particular packet, like a Storm tuple or a performance/fault management (PM/FM) report in XML format of an Operation Support System (OSS) application, data loss/out of order arrival/duplication phenomena may cause the packet not to arrive at the destination, arrive exactly once or to arrive in several copies. These anomalies have to be handled both on the lower and higher level network or application layers to an extent depending on the later usage. The efficiency of handling depends on the applied data structures.

To detect packet loss and duplication, a special, tree-like data structure was proposed earlier, the Interval Merging Binary Tree (IMBT). We analyzed IMBT from several perspectives and we compared its performance with other well-known tree variants, under various circumstances. However, in contrast to a completely balanced binary search tree, it is impossible to associate to the newly developed data structure a one dimensional function, dependent on the number of input keys, to determine for instance the average cost of an operation. Nevertheless, for further development, it is essential in case of any data structure, to determine the actual boundaries of its applicability.

In this contribution we explore the state space of IMBT in order to be able to classify the data structure regarding the input pattern during the later performance analysis. We used in the modeling Fibonacci sequences, bipartite multi-graphs and combination tables.

Keywords: data structure; balanced binary tree; bipartite graph; fibonacci sequence; state space; combination table;

Introduction

Performance management is an OSS application in which performance measurement records, generated periodically by network elements, are processed in order to assess the performance of the network. Each record consists of performance-related counters (key + value) each describing a specific aspect of the performance within a period. The periodicity of records makes it possible to associate incremental keys to the individual counters from the records, where the value is the content of the counter itself. The percentage of lost records is typically very low, therefore relatively few counters are lost in the transmission. Counters are converted into Key Performance Indicators (KPI-s) via Extract Transform Load (ETL) functionality for which we used Storm [2], a stream processing engine. Because of the at-least-once processing pattern of Storm, duplicated and out-of-order keys might occur. Packet loss and duplication of raw measurement data will lead to errors when aggregating counters into KPI-s, conveying a wrong perspective about the performance of the network, therefore loss and duplication cannot be tolerated in this particular use case.

In order to decide in real time whether a counter is identified by a key has already arrived or not and to insert it if not, we need a space-efficient data structure which is fast searchable and allows fast insertion of keys. After careful considerations, we ruled out a number of alternatives. The examined alternatives were external databases, Bloom filter [3], Balanced BSTs[4] [5], hash tables [4]. External databases turned out to be too slow. We ruled out Bloom filters because it allows false positives: for a key reported to be present we know only with certain probability its true presence in the data structure. This uncertainty is not allowed in our case. Balanced BSTs were ruled out due to their linearly increasing space need, which is proportional with the number of handled keys by them so far. The proportionally increasing space need was a drawback regarding hash tables as well. Additionally the need for periodical 're-hashing' in an upper-unbound environment would also significantly decrease the computation performance. Finally we arrived to proposing an efficient data structure and associated algorithms that we called Interval Merging Binary Tree (IMBT)[1].

In the unpublished [6] we have examined several tree layout instances and extreme scenarios for the arrival pattern of keys. Additionally we have deducted the formulas regarding the cost of SEARCH operation, as the basis of other operations, like INSERT or REMOVE. We have examined both theoretically and experimentally the performance of IMBT for an exponential distribution of the key arrival pattern [7]. Until now if only N, the number of IMBT handled keys, was given, we could not estimate nor even model accurately the state space of IMBT. State space modelling can facilitate the mapping of the statistical distribution-based input patterns into the IMBT state classes, if these exists at all. In the more general interpretation of state space we mean an N-dependent numerical value that characterizes the BST, and with normalization by N a statement can be made regarding the cost of operations. In case of traditional BSTs if the tree arrangement is given, then we can easily determine that N-dependent value which is the base of metrics like average time complexity of SEARCH operation etc. However, in contrast to traditional

BSTs, the IMBT state space is a multivalued function of N.
Therefore the analysis is divided into the following sections, through which we will unveil the aspects affecting the N multivalued dependency.
In section *Basics of the Interval Merging Binary Tree* we briefly introduce the IMBT data structure. From the description it will be clear that the analysis of the tree can be split into two independent aspects. In section *Interval State Space* we will show the relation between the possible number of arrangements of intervals across the tree and the Integer Partitions [8]. This is the first aspect.
The second aspect will be introduced in section *Traversal Strategy Based Weight Classes*. In this section we will describe the relationship between IMBT and a privileged tree arrangement, the completely balanced binary search tree. Here we will highlight the relationship between the Fibonacci sequences [9] and the number of comparisons required to reach a set of intervals within IMBT. In case of not limiting the examination to the completely balanced trees, according to Caylay's theorem [10] n^{n-2} different tree arrangements should be considered, where n is the number of nodes in a tree, which is impractical and turns out not to be needed.
In section *Bipartite Graphs and Combination Tables* we combine the two approaches into one model. During the combination we would like to determine the possible number of different values, which represents in fact the state space. In case when we just simply multiply the number of integer partitions of N with the different number of "step classes", then we get many duplicate values. That is, the state space would be highly overestimated. To mitigate this, we will introduce $G(I,W)$ bipartite graphs as a representation. In the course of matrix representation of the graphs we will apply a simplification and we can show that the result is nothing else than a combination table. The degrees of freedom of a combination table is a huge number. Regarding the enumeration of non-conform combination tables, or $G(I,W)$ graphs in our case, there are available results like [11], or [12], but as will be shown in our case both sides of the table increase deterministically, according to integer partitions and Fibonacci sequences. In our work we will also apply an additional equal transformation, like in the previous two references, to be able to formulate the criterion to get such sum of two members multiplications where the duplicates are minimized or zero. Therefore our result can be considered as an upper bound of the state space of IMBT in case when N is given.

Basics of the Interval Merging Binary Tree

IMBT is a data structure of disjoint sets, organized into a tree. The speciality of the sets is that each must contain all the keys between the greatest and the lowest value of a particular set. Sometimes these type of sets are called integer interval, hence we named the data structure interval merging binary tree, where merging refers to the operation of immediate merging 2 disjoint sets that become joint as a result of an incoming key.

As stated in the *Introduction*, we assume an input stream of keys where the key is a sequence number. Keys are arriving mostly ordered respective to the sequence number. The task is to filter out those entries that arrived already once, meaning that the sequence number has had already this value in an earlier key instance. Additional boundary conditions regarding the arrival pattern apply:

1. upper unbounded range: there is no upper bound of the sequence numbers apart from the limit of the binary representation of this field,
2. lower unbounded range: at any point in time a new key can arrive to the system with a sequence number lower than any sequence number encountered so far,
3. there are long, contiguous intervals of keys with relatively few 'gaps' (missing keys) in between,
4. after a while almost all keys arrive,
5. key duplication (i.e. same key arrived at least twice) on the arrival side is possible due to some reason.

Let's suppose that keys arrive to IMBT in the following order:

$...k_0, k_{-1}, k_2, k_3, k_7, k_5, k_4, k_6, k_{-2}, ...$

According to a naive approach all elements should be stored in a hash or in a binary search tree which is easily searchable, but still the binary search tree or the hash remains an upside-downside open system with infinite storage requirements when keys can arrive with infinite delay.

The first tweak to the naive approach is to represent the arrived keys as pairs. So, elements will be stored like the following:

$(k_0, k_0), (k_{-1}, k_{-1}), (k_2, k_2), (k_3, k_3), (k_7, k_7), (k_5, k_5), (k_4, k_4), (k_6, k_6), (k_{-2}, k_{-2}).$

At first sight it looks like that we did not win anything, but only doubled the memory footprint. The second tweak is not to automatically put newly arrived elements at the end, but rather to organize the elements in an ordered fashion, filtering at the same time duplicates found during the ordering process. This can be conceptually a sequence of 3 operations: insert at the end, order by key and a filter to skips the entry if it is already found:

$(k_{-2}, k_{-2}), (k_{-1}, k_{-1}), (k_0, k_0), (k_2, k_2), (k_3, k_3), (k_4, k_4), (k_5, k_5), (k_6, k_6), (k_7, k_7).$

The third tweak is to add an operation that we call interval merging: every pair of neighbour values is checked and if the values are consecutive, the two pairs are converted into one, where the first value of the resulting pair is the first value of the first pair and the second value of the resulting pair is the second value of the second pair. The skeleton code is available in [1].

In the following we describe the operation of the algorithm for our small data set:

- k_0 arrives, our data structure will store the following element:

  (k_0, k_0)

- k_{-1} arrives, our data structure will store the following element:

  (k_{-1},k_0)

- k_2 arrives, our data structure will store the following elements:

  $(k_{-1},k_0),(k_2,k_2)$

- k_3 arrives, our data structure will store the following elements:

  $(k_{-1},k_0),(k_2,k_3)$

- k_7 arrives, our data structure will store the following elements:

  $(k_{-1},k_0),(k_2,k_3),(k_7,k_7)$

- k_5 arrives, our data structure will store the following elements:

  $(k_{-1},k_0),(k_2,k_3),(k_5,k_5),(k_7,k_7)$

- k_4 arrives, our data structure will store the following elements:

  $(k_{-1},k_0),(k_2,k_4),(k_5,k_5),(k_7,k_7)$

  Then

  $(k_{-1},k_0),(k_2,k_5),(k_7,k_7)$

- k_6 arrives, our data structure will store the following elements:

  $(k_{-1},k_0),(k_2,k_6),(k_7,k_7)$

  Then

  $(k_{-1},k_0),(k_2,k_7)$

- k_{-2} arrives, our data structure will store the following element:

  $(k_{-2},k_0),(k_2,k_7)$

So, at the end storing only two intervals are required to represent 9 arrived keys. In case of we would organize these intervals into a binary tree then, as mentioned in the Introduction, the IMBT search operation state space would be influenced from two different aspects:

- the length of the intervals,
- the steps/comparison required to find that interval, that is the position of the interval within the tree.

In the following section we will examine the role of the intervals in the state space analysis of IMBT.

Interval State Space of IMBT

Fig.1, Fig.2 and Fig.3 indicate various types of evolutions of the tree as a function of the incoming keys, where in all cases we have 4 input packets.

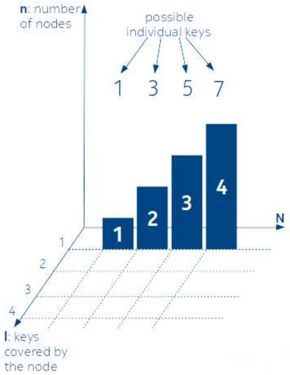


Figure 1
IMBT interval evolving when no direct neighbour exists.
On the figure *N* represents the *T* time as well. By looking to the figure from the right side, the remaining axes display a histogram of the intervals in different moments.

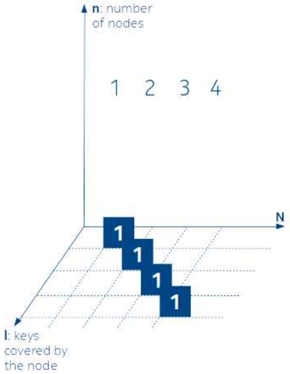


Figure 2
IMBT interval evolving when the keys are subsequent

As it is visible in case of four keys ($N = 4$), based on the possible number of neighbours, the following scenarios can be distinguished:

- None of the keys are neighbour of each other, like Fig.1,
- Two of them are neighbours and the other two are not,
- Two of them are neighbours and the remaining ones as well,

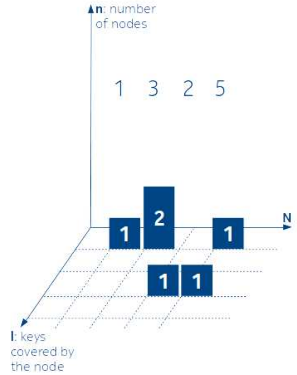


Figure 3
IMBT interval evolving when there are both neighbour and stand alone keys

– Three of them are neighbour and one is not, like Fig.3,

– All the keys are neighbour of each other, like Fig.2.

Therefore we can say that according to Hardy and Ramanujan [8]:

Theorem 1. the number of possible interval states in case of IMBT, at $T = N$ time, is equal with the number of ways N can be written as a sum of positive integers:

$$\lim_{N\to\infty} p(N) \approx \frac{1}{4N\sqrt{3}} e^{\pi\sqrt{2N/3}} \tag{1}$$

We can identify the addends of the sum as the individual interval lengths of the nodes in the IMBT. In this case for the average interval lengths a, considering the list items above, we get the following values, respectively: $4/1 = 4$, $4/2 = 2$, $4/2 = 2$, $4/3$, $4/4 = 1$. As it is visible there are two equivalent values: 2. Therefore it is generally true that the number of integer partitions is a rough upper estimation regarding the possible number different averages for a given input size N.
Additionally $p(N)$ does not say anything about the weight of the intervals based on their position in the tree. Since the same decomposition may led to very differently weighted arrangements it matters if for instance the intervals of 8 is written in e.g..: $1+1+4+1+1$ or $4+1+1+1+1$ or $1+1+1+1+4$ form. Supposing that the intervals are organized into a balanced binary search tree, the cost of the search operation in the first case is the most favourable, and in the last case is the least favourable. To account for these differences, in *Traversal Strategy Based Weight Classes* we will factor the transversal strategies in our analysis.

Traversal Strategy Based Weight Classes

In Fig.4 the arrows with number represent the j^{th} comparison during the SEARCH operations. Here we would like to mention that, for the sake of simplicity, during the comparisons the less or equal will be considered as one atomic step. The dark

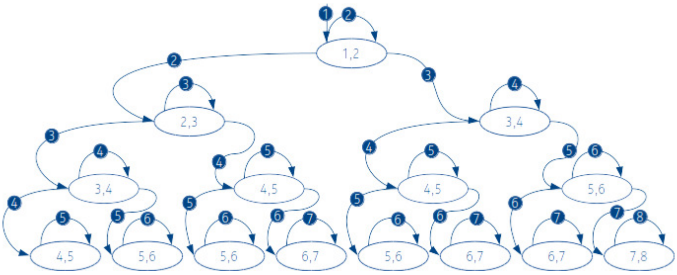


Figure 4
IMBT weight classes caused by the traversal strategy

background of the number expresses that the result of the comparison can be positive (that is, the node covers more than one key). In this figure, instead of boundaries of the intervals, the same information is displayed in the nodes as on the arrows.

As we can see if the key to be searched for is equal with the left hand value of the root node then exactly one comparison will be performed. If the key to be searched for is in between the left and right hand values of the root node or equal with the right hand value then two comparisons will be performed.

If the key is greater than the right hand value of the root node and falls into the interval of the right hand child's left and right hand values then three or four comparisons are required, depending on the exact value.

If the key is less than the left hand value of the root node and falls into the interval of the right hand child's left and right hand values then two or three comparisons are required, depending on the exact value.

By continuing the examination of the distribution of the different classes of intervals, based on the required number comparisons, we can recognize the following rules, when the intervals are organized into a completely balanced tree.

Considering the root (first) level there is one such interval where $(1,2)$ comparison can occur. In the second level there is one interval where $(2,3)$ and one interval where $(3,4)$ comparison(s) can occur. Finally, in the third level the cumulated number of intervals where $(1,2)$ and $(2,3)$ comparison(s) can occur is unchanged. However, the number of $(3,4)$ comparison intervals is increasing from one to two. Additionally two $(4,5)$ and one $(5,6)$ comparison intervals appear.

By the cumulative number of types, as more and more layers are taken into account, we will get the pattern described in Table-1. Examining carefully the lists we can realize that

Theorem 2. the central element of each row composed from cumulative number of weight types is the Fibonacci sequence itself. The numbers in the lists (lines in this case), preceding the central elements, are also the evolving Fibonacci sequences themselves. The rest of the numbers must satisfy the requirement that the sum of the numbers is equal with $2^n - 1$ in every n^{th} line.

However, another rule also can be recognized there:

Theorem 3. the numbers in a line from Table-2 are equal to the sum of the two preceding numbers of the previous line.

Table 1
Distribution of weight classes in case of the IMBT is completely balanced.
The Fig.4 snapshot is marked with bold.

Total number of nodes in IMBT	Distance from the root [number of comparisons regarding the left hand value]										
	1	2	3	4	5	6	7	8	9	10	11
1	1										
3	1	1	1								
7	1	1	2	2	1						
15	**1**	**1**	**2**	**3**	**4**	**3**	**1**				
31	1	1	2	3	5	7	7	4	1		
63	1	1	2	3	5	8	12	14	11	5	1

Table 2
Fibonacci sequences in the cumulated weight classes

1										
1	**1**	1								
1	1	**2**	2	1						
1	1	2	**3**	4	3	1				
1	1	2	3	**5**	7	7	4	1		
1	1	2	3	5	**8**	12	14	11	5	1

Until now we have shown that there are two distinct aspects influencing the state space of IMBT. One is if how many ways the number of keys can be decomposed into integer partitions.
The second aspect is represented by the weight classes. It is based on the number of nodes and depends on the associated traversal strategy.
Now, to be able to determine the combined number of input pattern classes somehow we have to put these components together. In *Bipartite Graphs and Combination Tables on the modeling of IMBT State Space* we will present this combination procedure and the resulting mathematical models.

Bipartite Graphs and Combination Tables on the modeling of IMBT State Space

To be able to start the combined analysis we will perform the following mappings. Let's denote the length of the interval belonging to an n_i node from the IMBT by $l_i \in L$, where L is a multi-set. Then we map the set of same length of intervals onto $i_1, i_2, ..., i_k \in I$ elements. This means that by having the $L = \{l_1, l_2, ..., l_n\}$ lengths, where the values of $l_h = l_i = ... = l_j$ is equal, then this fact results in one new element, i_p, in the I set. That is the following $l_h \rightarrow i_p, l_i \rightarrow i_p, l_j \rightarrow i_p$ surjection is performed in case of $l_h = l_i = l_j$. Therefore $k \leq n$.

Let's denote the number of comparisons required to achieve the left hand value of an arbitrary n_i node by $s_i \in S$, where S is a multi-set. Then let's map the traversal strategy based identical comparison weight types onto $w_1, w_2, ..., w_j \in W$ elements. This means that by having the $S = \{s_1, s_2, ..., s_n\}$ lengths, where the values of $s_h = s_i = ... = s_j$ are equal, then this fact results in one new element, w_p, in the W set. That is, the following $s_h \rightarrow w_p, s_i \rightarrow w_p, s_j \rightarrow w_p$ surjection is performed in case of $s_h = s_i = s_j$. Therefore $k \leq n$.

Since the newly defined I and W are two disjoint sets we can consider them as the vertices of a $G(I,W)$ bipartite (multi-)graph. We will assign degrees to each vertex in the following manner:
The degree of each i_i vertex is equivalent with the number of those particular interval lengths. According to this in case of $l_h = l_i = l_j$ the degree of the associated i_i vertex is $d(i_i) = 3$.
The degree of each w_i vertex is equivalent with the number of those particular weight types in the search tree.
Therefore we can write that

Theorem 4. $\sum_{i=1}^{j} d(w_i) = \sum_{i=1}^{k} d(i_i) = n = |E|$, where $E = \{e_1, ..., e_n\}$ is the set of the e_i edges of $G(I,W)$.

The fact that the above two sets, I and W, are the independently different classifications of the same nodes of the IMBT implies that the sum of the degrees of the vertices in both sets is equal to n. □

Let's consider an IMBT arrangement/configuration where $n = 4$, and both I and W sets contain one-one vertex with degree two, and two additional vertices with degree one-one. So, $d(i_1) = d(w_1) = 2$ and $d(i_2) = d(i_3) = d(w_2) = d(w_3) = 1$. At this moment regarding N we can only say that $N \geq n$.
It is obvious that to get the above I set two of the lengths must be equal, eg. $l_1 = l_2$, and the other must differ from both $l_1 = l_2 \neq l_3$, $l_1 = l_2 \neq l_4$ and $l_3 \neq l_4$.

Definition: Those L interval length multi-sets are called *interval lengths ratio base class*es, denoted by L^b, in which

> at least one l_i exists which is co-prime to all the other l_j, such that $i \neq j$ supposing that $l_i \neq l_j$, or
>
> if $l_i = l_j$ for all $i \neq j$, than $l_i = l_j = ... = l_k = prime\ number$.

That is, L is an L^b if

$$\exists l_i \in L \mid (\forall i \neq j \wedge l_i \neq l_j \Rightarrow gcd(l_i, l_j) = 1) \vee (\forall i \neq j \Rightarrow l_i = l_j = prime\_number). \tag{2}$$

If $L = \{l_1, l_2, l_3, l_4\}$ is an interval lengths ratio base class, that is $L = L^b$, then L^b determines all the $N_1, N_2, ...$, which differ from each other by only an integer factor for a given $(L^b, n = |L^b|)$ pair. This representation/decomposition is unique, except for the order of the factors:

$$\begin{aligned} N_x &= (d(i_1)\times l_1 \times x) \;+\; (d(i_2)\times l_3 \times x) \;+\; (d(i_3)\times l_4 \times x) \\ &= x \times (d(i_1)\times l_1 \;+\; d(i_2)\times l_3 \;+\; d(i_3)\times l_4). \end{aligned} \tag{3}$$

where $x \in \{1,2,3,...\}$. If n is given that is the maximum information we can get regarding N.

In Fig.5 all the different possible configurations are shown for the above $G(I,W)$, where $|I| = |W| = 3$ and $|E| = n = 4$. That is, there are three-three vertices on both sides of the G graph.

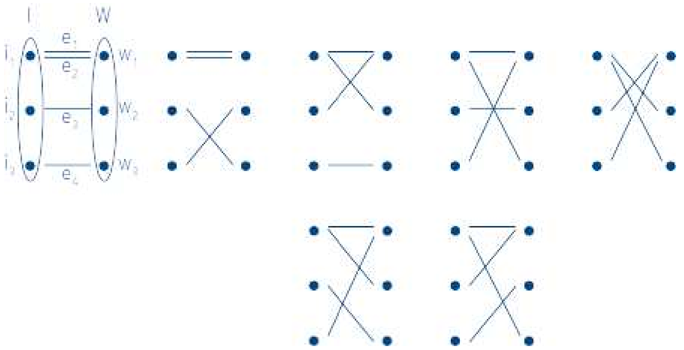


Figure 5
G(I,W), where $|I| = |W| = 3$ and $n = 4$

At this stage we can claim that $N \geq 4$. If we are aware of the $l_1, l_2, l_3, l_4 \in L$ values, e.g.: $l_1 = l_2 = 1, l_3 = 2$ and $l_4 = 3$ and therefore $L_1 = L^b$ then we can say that $N_1 = 7$. However, $N_2 = 14, N_3 = 21, ...$ and $L_2, L_3 \neq L^b$.

As it is visible from the Fig.5 there are seven different possible configuration. Regarding the number of possible configurations, in case of a given (L,n) pair, till now we have a mathematical model as $G(I,W)$ is a bipartite graph. We can formulate the following

Theorem 5. The simplified adjacency matrix representation of a $G(I,W)$, which is derived from an IMBT according to the above process, corresponds to a contingency table.

Let's assume an $G(I,W)$ bipartite graph derived from an IMBT. Let's prepare the adjacency matrix of $G(I,W)$, where parallel edges are allowed, in the following manner. Since $G(I,W)$ is a bipartite graph there are no edges between the vertices belonging to the same vertex set. Then we will apply the following simplification: instead of enumerating all the points from both sets on the right side and the top of the adjacency matrix merely the points from I will be displayed with the associated $d(i_i)$ values on the right side. On the top of the matrix only the points from W will be displayed with the associated $d(w_j)$ and values.
The edges appear as numerical entries in the cells of the matrix. The value of a

particular cell represents the number of edges between the i_i and w_j points. However the $d(i_i)$ and $d(w_j)$ values are constraints regarding the sum of a given i row and j column.
From *Theorem 4* we know that the sum of cells in a row is equivalent with the degree of that particular vertex. The same is true for all columns. Therefore the sum of sums of every row is equivalent with the sum of sums of every column. This feature of the simplified adjacency matrix is corresponding to a contingency or combination table, which may contain discrete samples of the same multitude from two different points of view. □

In Fig.6 the simplified adjacency matrix representation of the $G(I,W)$ graphs from the Fig.5 is shown. Since the edges do not appear directly, the simplified adjacency

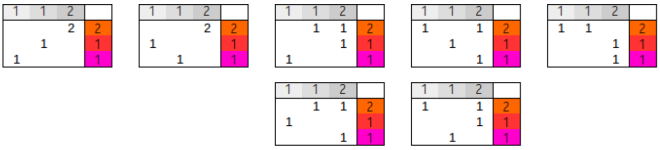


Figure 6
Simplified adjacency matrix of G(I,W)

matrix remains unchanged in case when two different, e_k and e_l edges that are not sharing on any vertices on any of their ends, are mutually replaced with each other. This holds also for the case when neighbours edges, sharing on a multi-degree vertex, replace their non-sharing ends with each other.
Therefore from this simplified adjacency matrix like it is still hard to establish the formal condition of states being different, that is the total weight of the IMBT. The number of states for a given (I,W) is the different number of total weights of the IMBT.

Nevertheless, we can apply the following transformation without violating the validity of the transformed model. During the transformation we are composing so called domains in the matrix in a way that every row(or column) with value $d(i_i)$ (or $d(w_j)$) will be substituted with $d(i_i)$ (or $d(w_j)$) rows(or columns), where the constraint value of each row is '1'. Therefore the 1×1 cells, which are in the cross of the $d(i_i)$ row and the $d(w_j)$ column, will be replaced by such a domain that consists of $d(i_i) \times d(w_j)$ cells.
In Fig.7 the domain composition of the above $G(I,W)$ is visible, where the domains are marked/surrounded by dotted lines.

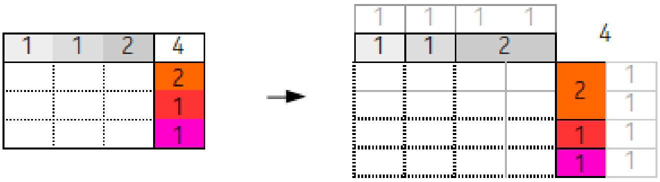


Figure 7
G(I,W) simplified adjacency matrix transformation to domain representation

In Fig.8 the domain transformed matrix representation of the Fig.5 examples are shown. The numbers with blue background mark the related $G(I,W)$ from the examples. Let's denote the set of all the $G(I,W)$ graphs belonging to the same partition of N by P_{N,L_i}. From the *Interval State Space Section* we know that $i \in \{1...p(N)\}$. A particular $G_k(I,W) \in P_{N,L_i}$ expresses the n members sum of two members products, where the members of the products are from the L_i and W sets respectively. Therefore the $G_k(I,W) \in P_{N,L_i}$ determined sum of products can be mapped onto the IMBT state space. Now we will define the subset of P_{N,L_i}, denoted by P^s_{N,L_i}, according to the following.

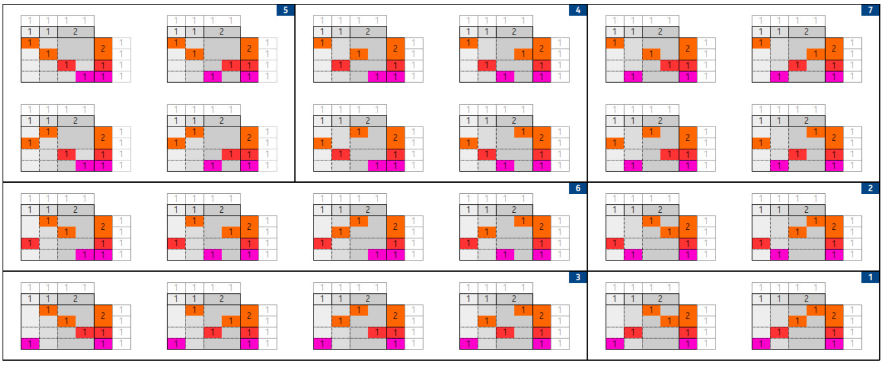


Figure 8
G(I,W) examples with domain representation.
The numbers with blue background marks the related G(I,W) from the above examples.

P^s_{N,L_i} is the subset of the P_{N,L_i} set that contains the maximum number of $G_i(I,W)$ graphs from P_{N,L_i}, so that in the $G(I,W)$ associated transformed matrices the sum of cells are different for all the (G_i,G_j) $i \neq j$ pairs in at least 4 domains.

Theorem 6. $|P^s_{N,L_i}|$ is an upper bound regarding the possible number of IMBT states belonging to an $N \rightarrow L_i$ partition.

Let's consider in the following lengths $l_1, l_2, ...l_n$ and steps $s_1, s_2, ...s_n$. Let's additionally assume that there are i elements from both the l's and s's where the associated lengths and steps are equivalent with each other. Additionally there are two additional j and k elements from both L and S where the associated values are the same and $i+j+k \leq n$. Let the associated value of the i elements be $v_i = 2$, $v_j = 3$ and $v_k = 4$. Then there will be such a $G_1(I,W)$ and $G_2(I,W)$ bipartite graphs that are identical in every other pairings regarding the member of the products except the $G_1 \rightarrow r_1 = ... + l_{i,i} \times s_{i,i} + l_{j,1} \times s_{j,1} + ... + l_{j,j} \times s_{k,1} + l_{k,1} \times s_{j,j}$ and $G_2 \rightarrow r_2 = ... + l_{j,1} \times s_{i,i} + l_{i,i} \times s_{j,1} + ... + l_{k,1} \times s_{j,j} + l_{j,j} \times s_{k,1}$. In this case (G_1, G_1) pair satisfies the above condition regarding the sum of domains, however the associated r_1 and r_1 results are identical, therefore this represents the same state of IMBT. □

From the above we can formulate the following

Theorem 7. The upper bound of the IMBT state-space in case of knowing only N, and the same n number of lengths are always sorted into the same tree structure no matter whatever it is:

$$IMBT_{States}(N) = |P^s_{N,L_1} \cup P^s_{N,L_2} \cup \ldots \cup P^s_{N,L_{p(N)}}| \leq \sum_{i=1}^{p(N)} |P^s_{N,L_i}|. \tag{4}$$

In case of a completely balanced IMBT the degrees belonging to a particular w_i are equivalent with the corresponding number from the corresponding line of Table-2. For instance in case of $n = 7$ we can identify the third line of Table-2. Therefore we know that the number of different weights is 5. And the seven nodes are sorted into five classes according to the followings $d(w_1) = 1, d(w_2) = 1, d(w_3) = 2, d(w_4) = 2, d(w_5) = 1$.

Conclusions

In this contribution we have introduced a special tree structure, the IMBT. Then we have pointed out the aspects contributing to the state space of this data structure and we provided an upper bound for the cardinality of this state space.
Now we have a mathematical model through we can perform measurements and an assessment of a concrete $G(I,W)$ representation to which a series of keys tends. This might be a possible classifier regarding the statistical distribution of the key arrival process. In the following we plan to determine some correlation/combination tables for different distributions.

References

[1] Finta, I., Farkas, L., Sergyán, Sz., Szénási, S.: Interval Merging Binary Tree, ICA3PP 2017, Helsinki, Finland, August 21-23, 2017
DOI:10.1007/978-3-319-65482-9

[2] STORM - A distributed real-time computation system, `http://storm.apache.org/documentation/Home.html`, last visited 2019-01-02

[3] Bloom, B. H.: Space/time trade-offs in hash coding with allowable errors, Communications of the ACM, Volume 13 Issue 7, pp 422-426, New York, NY, USA, July 1970.

[4] Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein, C.: Introduction to Algorithms (3rd ed.). MIT Press and McGraw-Hill, 2009.
ISBN:0-262-03384-4

[5] Bayer, R.: Symmetric binary B-Trees: Data structure and maintenance algorithms, Acta Informatica, Volume 1, Issue 4, pp. 290-306, 1972.
DOI:10.1007/BF00289509

[6] Finta, I., Farkas, L., Szénási, S.: Parametric Analysis of Interval Merging Binary Tree, Digital Communications and Networks, Initial submission: October

25th, 2017
ISSN: 23528648

[7] Finta, I., Élias, G., Illés, J.: Packet Loss and Duplication Handling in Stream Processing Environment, CINTI 2018, Budapest, Hungary, November 21-22, 2018
DOI:10.1007/978-3-319-65482-9

[8] Hardy, G.H., Ramanujan, S.: Asymptotic Formulae in Combinatory Analysis, Proceedings of the London Mathematical Society, 1918

[9] Bóna, M.: A Walk Through Combinatorics: An Introduction to Enumeration and Graph Theory. pp. 145-164, World Scientific Publishing, 2002
ISBN 981-02-4900-4.

[10] Cayley, A.: A Theorem on Trees. Quarterly Journal of Pure and Applied Mathematics 23, pp. 376-378, 1889

[11] Barvionk, A.: Enumerating Contingency Tables via Random Permanents, Combinatorics, Probability and Computing, Volume 17, pp. 1-19, 2008
DOI:10.1017/S0963548307008668

[12] Barvinok, A., Luria, A., Samorodnitsky, A., Yong, A.: An approximation algorithm for counting contingency tables, Random Structures Algorithms 37 (2010), no. 1, pp. 25-66, 2010
DOI:10.1002/rsa.20301
arXiv:0803.3948

Security Information System, Based on Fingerprint Biometrics

Komlen Lalović

ITS–Information Technology School, 34 Cara Dušana Street, 11080 Belgrade, Serbia, komlen.lalovic@its.edu.rs

Ivan Tot

University of Defense, 33 Pavla Jurišića Sturma Street, 11000 Belgrade, Serbia, ivan.tot@va.mod.gov.rs

Aleksandra Arsić

Mathematical Institute, 36 Kneza Mihaila Street, 11001 Belgrade, Serbia, aleksandra@mi.sanu.ac.rs

Milan Škarić

Enon Solutions, 86 Omladinskih Brigada Street, 11070 Belgrade, Serbia, milan@enonsolutions.com

Abstract: This work presents a novel security information system based on fingerprint biometrics. It combines cancelable biometrics, security algorithms such as RSA and AES, both synchronous and asynchronous for data encryption. By implementing novel devices developed on the Raspberry Pi Platform and wireless communication, 100% accuracy in real terms, will be enabled and FAR and FRR will be decreased to a minimum - as it were, zero. At the core of the information system are devices and software algorithms implemented for biometric identification of maternity, as a dual fingerprint scanner that provides data of mother and new born baby fingerprints and in the further process, guarantees maternity of a new born baby with 100% accuracy. All this is developed as a novel method of identity determination, based on baby fingerprint minutiae, instead of current systems that are prone to many errors. This system will prevent any possible replacement or identity theft in every maternity ward.

Keywords: Biometric; Fingerprint; Security; Software; Information system

1 Introduction

This paper presents work in the field of advanced security systems, widely used in modern society and in protection systems. Biometry is the scientific discipline and techniques for measuring and analyzing biological characteristics of people.

We have solved the issue and societal problem of the timely determination and logging of a newborn babies minutiae, thereby protecting them with a crypto algorithm enrolled with cancelable biometrics. Our information system prevents any possible theft or misplacement of the baby's identity, providing 100% accuracy in its determination and showcases a device which scans, stores the encrypted personal data, with one goal - to provide validated parenthood for each newborn baby, in a crowded hospital.

The information system (IS) completely removes the fear that almost every mother has during birthing, and it also removes the question "Have I brought home my baby… for sure?". This method presents a new implementation of information technology security in the public health system and raises it to a new level.

It also links the patented device for biometric identification of newborn babies with two existing software algorithms and the necessary writing procedures, so that this new approach can be carried out. Not only does it solve a huge problem - possible theft or misplacement a newborn baby's identity, but it also removes a fear that women have, when giving birth.

This is not just problem in Serbia, it is global problem, baby switching has happened in almost every country in the world. According to Brandon Gille, USA study, of 4 million new born babies, 28,000 from them had been misplaced! [20]

There are no similar solutions at this time. There is some attempts to solve this kind of problem, with foot prints, in USA, but it is an ink print not digitally processed.

2 Problem Solving

The information system possesses two main software algorithms that will be presented herein and which provide all of the device functionalities. These functionalities will be illustrated in figures, which point out how our model was built, designed and developed. The paper will present possible advantages and benefits that are a qualitative leap in the public health care system, precisely in maternity wards all over the world.

By implementing this new system, it is possible to establish a Wi-Fi communication and storage types, for fingerprints scanned from both the mother and the baby together at the moment of birth, and to generate a unique identity reference which is encrypted and will guarantee parenthood over every newborn baby with 100% accuracy. Wi-Fi is required during scan process or a cable connection within 30 minutes of scanning, since device has a data store timeout for those 30 seconds because of security issues.

Considering experiment, results have been presented and listed in several publications, where most relevant is under SJEE. It contains experiment with newly born baby and the scan process of all hand fingers with different type of scanners. [5]

This invention belongs to the field of Applied Information Technology. Biometrics systems and device with its functionality is similar and close to a dual fingerprint scanner with two fields - first the scanning process takes place (both the mother's and the baby's fingers), then the device makes a unique reference which will be the ID for each mother/baby relationship for every newborn baby in the hospital.

3 Technical Overview

The main technical problems which are solved by this IS are the following:

The design and development of a new device based on our patented device [11] incorporates a dual fingerprint scanner, for scanning fingers of both the mother and the baby at the moment of birth. The device cross view will be similar to the current classic fingerprint scanners; therefore, the tablets would have two fields for scanning fingers of two different people (the mother and the baby). After this process, the device will encrypt and store the data. The device cross view is given in the following pictures.

The real technology improvement and contribution is that the device is highly practical and easy to use and control. The maintenance of the device is easy, classic and similar to other fingerprint scanners. Beside its common purpose and scanning two fingers of different people at the same time, it will provide a unique ID reference (like Primary Key PK) which will be the basis for every pair of the scanned mother and baby.

The realized information system (IS) presents the optimal solution for this type of work, which defines the strict procedures that need to be followed. The IS will also implement the IT technology in the public health system. Current biometric devices can scan one or more fingers from **one** person, then repeat with another, but there are no fingerprint scanners which may scan fingers of two different people at the same time on a single device. Especially, there are no devices that may make a unique reference while scanning, which will further be connected to the record of scanned fingerprints previously stored data. [4] [7]

This is provided by the **Device for biometric identifying of Maternity,** in the functionality development of which the encrypted data that will carry information about two people (the mother and the baby) and the linked unique ID reference will be implemented.

Our device has two fields for scanning fingers of two different people at the same time. This is crucial improvement in the technology for solving problems.

However, the question that may emerge concerns the time-saving feature needed for the process scanning of both people and all of the back-end Information Technology (*IT*) in background, of the future system that now is now scanning for one person.

Our device provides improvements in economy and the time spent during the process of scanning, enabling the possibility for reduction of costs for each new device, with the advantage of less time needed for processing data obtained from the image of the finger scanner. The device provides the optimal solution for resource usage in the case of processing data acquired during the process of fingerprint scanning, primarily considering the memory usage and activity of the Central Processing Unit (*CPU*). Figure 1 presents a completely new information system with devices and communication. All devices are synchronized with the server and database which stores the data, using cancelable biometrics and using a cable or wireless connection. Devices do not communicate one to another.

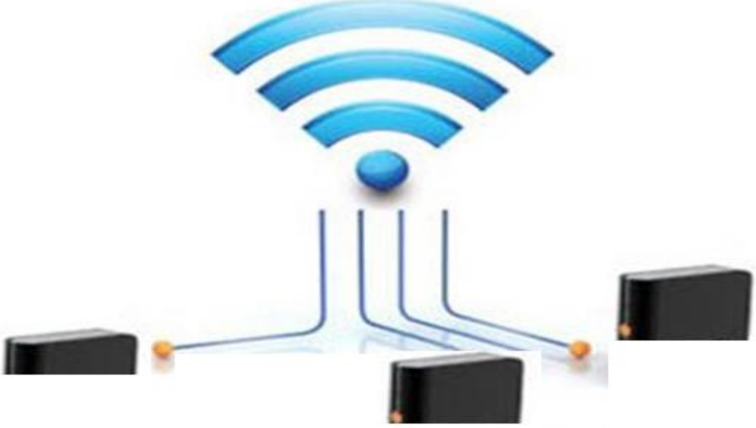


Figure 1

Information system and communication between components

4 Information System

The Information System created for this new baby identification method, based on the fingerprints, consists of obtaining, encrypting, storing and verifying data. The next step follows the procedures that need to be performed. Figure 2 shows the Class diagram for this Information System.

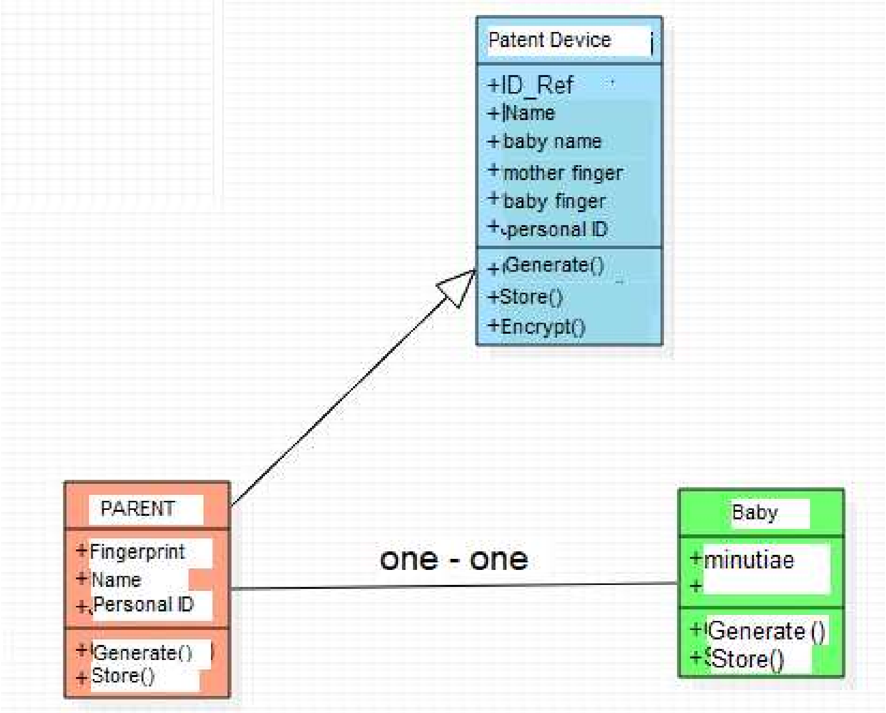


Figure 2
Class diagram of Information system

4.1 Can Baby Fingerprints Be Obtained?

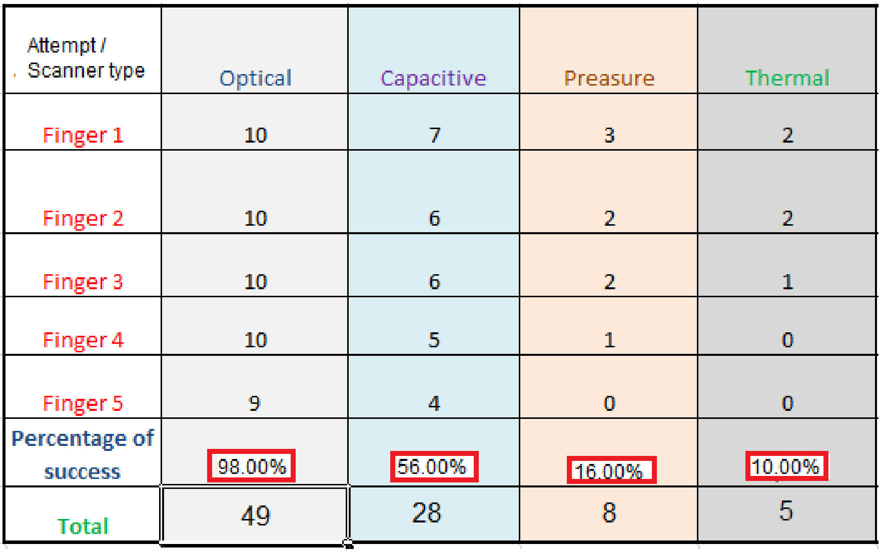
The device that we have invented has features, such as, a real time dual biometric fingerprint scanner that has two fields for scanning fingerprints of the mother and the baby or two and more, at the very moment of birth. The first field is larger, with classic scan resolution of 500 dpi; the second field is physically smaller but has larger scan resolution – a minimum of 1000 dpi, so it can produce scans of a the baby’s fingerprint, that is relatively small. [8] [12] [13]

Our research is based on the scientific fact that human fingerprints are formed during the prenatal period for every fetus/baby and that they remain constant in shape of minutiae during the entire life of the individual. [1] [3] [9] [10]

The most important fact to mention, is that babies who are born prematurely, during the 8th and especially by the end of the 7th month of pregnancy, have formed fingerprints on the finger of both hands. This is the starting point of our research. [1] [5] [16] [18]

Considering this scientific fact, which is crucial for our patent and this device, this research and this project realization will provide a qualitative jump in gynecology, midwifery and nursing in every hospital in the entire world. We need to provide 100% accuracy and guarantee the maternity over a newborn baby anywhere in the world. We achieve this by placing one of the baby's fingers on the smaller field of our device simultaneously with the mother's finger on the larger field of the device and initiating the scans. At this point, the device also generates a unique reference for this pair of scanned data-minutiae (mother and baby). Later, after a few days, when parents leave the ward, the next step is to check the fingerprint on the same device, which should confirm the baby's identity. This procedure will provide new quality and it will prevent any possible misplacements or baby theft. [1] [2] [6]

Figure 3 presents the final results of the empirical test results, made of a new born The results presented here are an absolute novelty in fingerprint biometrics.



Attempt / Scanner type	Optical	Capacitive	Preasure	Thermal
Finger 1	10	7	3	2
Finger 2	10	6	2	2
Finger 3	10	6	2	1
Finger 4	10	5	1	0
Finger 5	9	4	0	0
Percentage of success	98.00%	56.00%	16.00%	10.00%
Total	49	28	8	5

Figure 3

Results of research and experiment enrolled per each finger and each scanner type

When it comes to other biometrics, such as iris recognition, they did not prove to be as useful, since they are unstable on babies. The reason lies in the fact that iris and eye pigmentation is not formed until 4 years of age, and it keeps changing in shape and in color. Therefore, it cannot be used for this purpose. [5] [11]

Other body parts, especially palm or other limbs, cannot be used because they rapidly change due to the normal process of growing up. This is the reason why this excellent, scientific fact concerning fetus fingerprints, their prenatal formation, by the end of the 7th month to be precise, in the pregnant mother, and their minutiae construction remaining unchanged, is marvelous and useful. [1] [6]

Mothers-to-be, as well as, medical staff in maternity wards, in health care facilities, have many different concerns, during the birthing process. A study that was carried out in Australia and New Zealand, by the Woman and Birth Journal, in 2009 included 17 different workshops, with more than 700 midwives, over a period of 5 years, points out that women had 144 various fears at the moment of giving birth. Taking this fact into consideration, our device can prevent a great deal of those tremendous fears and perhaps eliminate them in a total number of n=43. [6]

We have to mention that all the data obtained, from the mother and baby, during the process of fingerprint scanning, together with the unique ID reference encryption and storage in the device memory and/or on a server in the encrypted forms never leave the device in vulnerable state, or available to the public. The data is only to authorized nurses, doctors and midwifes in the maternity ward. [5] [6]

4.2 Improvements in The New Information System

The new IS combines device for biometric identification of newborn babies based on the fingerprint scan will enable the following:

- Improvement and evidence of maternity for every newborn baby
- Exclusion of any possible replacement or identity theft of newborn babies
- Safety for each parent couple
- Portability due to the small dimensions and low weight; it is ambient friendly and does not pollute environment
- A good ratio of price/quality
- A wide range of applications and use

In order to understand better the functionality and application of the device, as well as its practical realization, there are a few pictures that illustrate the device and algorithms, long with the cross-section of the patented device.

Figure 4 illustrates the device for biometric identification of maternity in a completely new view with digital display, a switch and two fields for fingerprint scanning.

Figure 4 contains the following remarks:

B Body of the device

I Power switch with two positions (on/off)

D Device display for displaying all the details, such as a unique ID reference generated during the process of scanning

S Set button for starting the scanning process and reading parameters obtained by fingerprint scanning

R Reset button

R1 Command button that saves and stores data

S1 Field for scanning the fingerprint of the baby's finger, smaller than the field for scanning the mother's fingerprint

S2 Field for scanning the fingerprint of the mother's finger, larger field

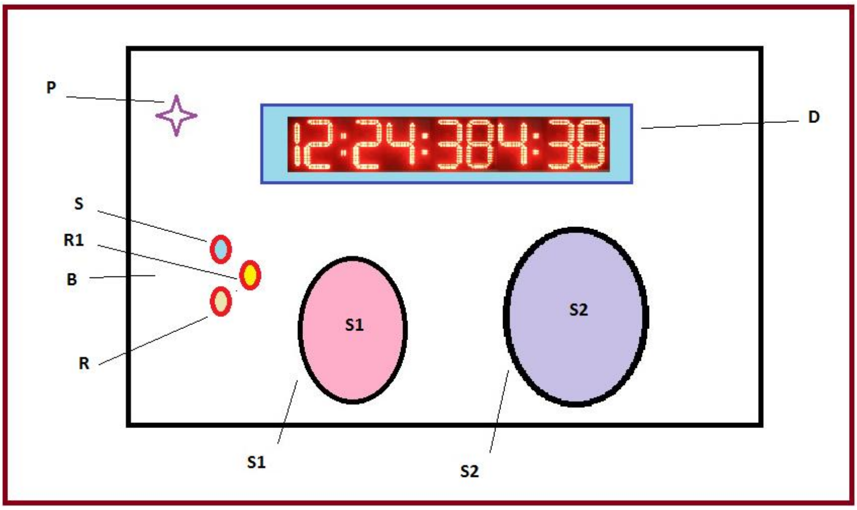


Figure 4
Detail description of device functionality

5 Acquiring Software Algorithm in Pseudo Code

The algorithm, in pseudo code, is listed here, in order to illustrate the logic and all possible software features that the device possesses, to explain in details how the entire system is supposed to accomplish its purpose of providing a completely new quality implementing information technology. [11]

```
----------------------------------------------------------------------------
START
BEGIN LOOP 1 TO 3
FIELD-1 SCANN
IF F1 OK
  THEN GOTO GENERATE UNIQE ID
                  ELSE IF LOOP < 3 GOTO END
BEGIN LOOP 1 TO 3
FIELD-2 SCANN
IF F2 OK
  THEN GOTO GENERATE UNIQE ID
                  ELSE IF LOOP < 3 GOTO END
GENERATE UNIQE ID REFERENCE
GENERATE PIN CODE
ENCRYPT DATA
GENERATE HASH VALUE
STORE AND SAVE DATA
DISPLAY SUCCESS MESSAGE
FINISH
----------------------------------------------------------------------------
```

Figure 5 represents the device algorithm, that is main software part of our innovative device and it can provide the process functionality with the deduction process. After the scanning process, it generates unique ID, encrypts those data and generates a hash value providing information about a successful scanning event.

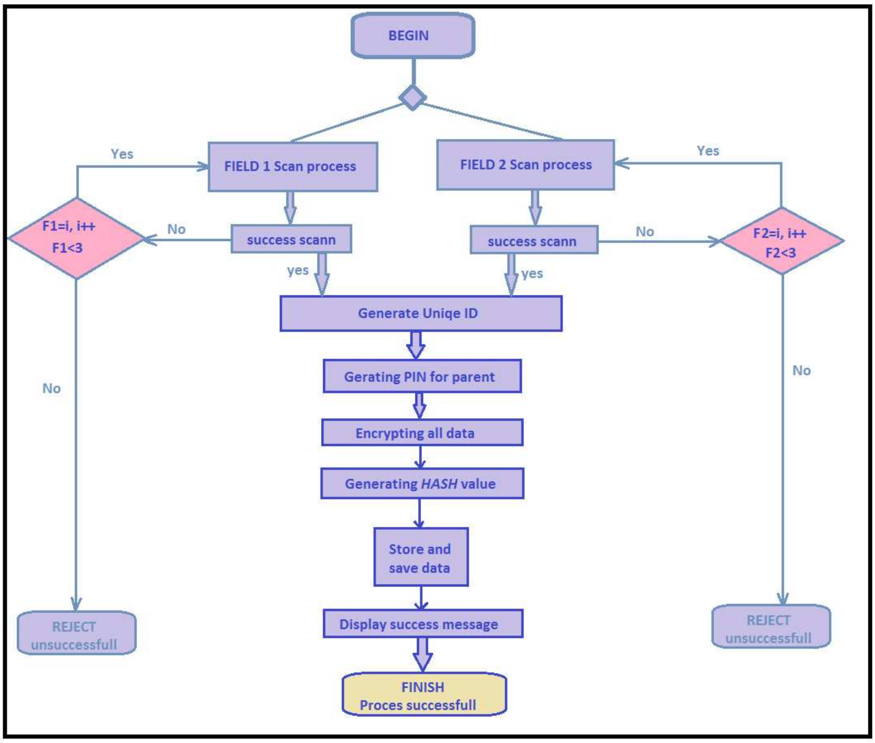


Figure 5
Device software algorithm for data acquisition

6 Verification Software Algorithm in Pseudo Code

This algorithm in pseudo code illustrates the logic and all possible software features that the device possesses for the verification of fingerprint minutiae and explains how the system uses cancelable biometrics to enable security and private data. [11]

Figure 6 represents the essence of this device algorithm which is a part of our innovated device and which provide the process functionality to verify stored data. It has a double check to provide total security of its functionalities. It compares encrypted data and also check the hash value of a stored data.

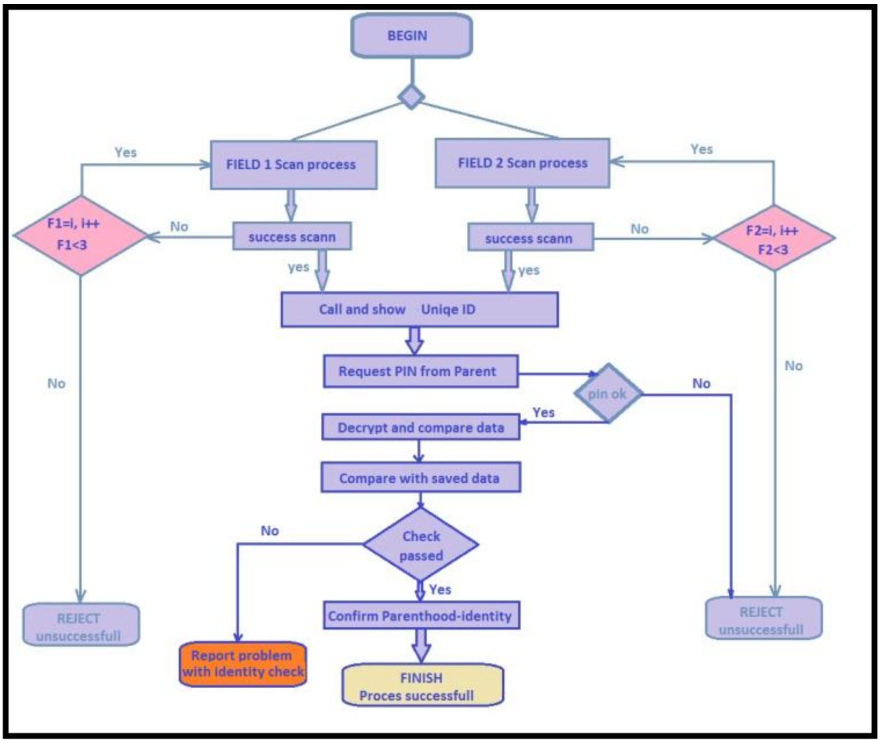


Figure 6

Device software algorithm for data verification

7 Possibilities of Further Development

The Information System with the device and novel method of implementing the new device, with biometric fingerprint identification of maternity and scanning each newborn baby can be used as a brand new model, in the public health care systems. Considering further development of similar fingerprint biometrics system in everyday care, hospitals may solve various problems, regarding moving small children and their continuous monitoring. It is necessary to provide further development and research based on the patented device, innovation and qualitative research that were performed during this study.

Software algorithms given herein, will improve the device functionality and enable it to work faster. They can be used as a part of much larger health care system, regarding young children and their pediatric care; they can also provide basic data concerning possible allergies and specific health states for each child.

Combinations may allow for great improvements in that part of the health care system at a global level.

Conclusion

This novel device combines the objectives of three parts: the patented device, the information system and a safe maternity ward approach. It can improve the level of public health in the Republic of Serbia. The system is modular, it can be updated and, most importantly, it can be the basis for future developments in biometric systems. The device can be applied in a number of countries, to fight the organized criminals and help to prevent the theft or misplacement of newborn babies, especially in territories with a low IT infrastructure and technological development.

Every kind of biometry is eager to minimize both FAR[1] and FRR[2] in order to be much more accurate and secure. This device has accomplished this, since it combines two scanned data and its accuracy grows exponentially. In the modern IoT (Internet of things), the majority of countries try to provide a completely new quality of health care service, help the staff in maternity wards, make the process of giving birth much easier and more relaxed, both for mothers-to-be, gynecologists, midwifes and nurses.

Conflict of Interest

None.

Acknowledgement

We want to thank every single person and baby involved into this research with purpose to improve IT in public health system.

References

[1] Anil K. Jain-*Michigan State University,* USA, Patric Flynn-*University of Notre Dame, USA,* ARUN A. ROSS-*West Virginia University, USA* (2008)*:* Handbook of Biometrics – Sringer, USA

[2] Komlen Lalović, Nemanja Maček, Milan Milosavljević, Mladen Veinović, Igor Franc, Jelena Lalović, Ivan Tot - Biometric Verification of Maternity and Identity Switch Prevention in Maternity Wards, Acta Polytechnica Hungarica, Volume 13, Issue Number 13, 2016 DOI: 10.12700/APH.13.5.2016.5.4

[3] NIST, *A Survey of Access Control Methods*

[4] Nemanja Maček, Borislav Đorđević, Jelena Gavrilović, Komlen Lalović - An Approach to Robust Biometric Key Generation System Design, Acta

[1] FAR – False Accept Rate

[2] FRR – False Reject Rate

Polytechnica Hungarica, Volume 12, Issue Number 8, 2015, DOI: 10.12700/APH.12.8.2015.8.3

[5] Komlen Lalović, Milan Milosavljević, Ivan Tot, Nemanja Maček: Device for Biometric Verification of Maternity, Serbian Journal of Electrical Engineering-Vol. 12, No. 3, October 2015, DOI: 10.2298/SJEE1503293L

[6] Hannah Grace Dahlen, Shea Caplice: "What do midwives fear?", Published Online: July 24, 2014 – Elsevier, *Women and Birth, Journal of Australian College of Midwives*

[7] Komlen Lalović, Ivan Tot, Svetlana Andjelić - How to Guarantee Baby Identity based on Fingerprint Biometry, Bisec 2017 - International conference in Security ICT, October 18th-Belgrade, Serbia

[8] Komlen Lalović, Jasmina Nikolić, Ivan Tot, Žana Lalović - Software Algorithm of Device for biometric identification of Parenthood, BISEC 2016 - International conference in Security ICT, October 15th-Belgrade, Serbia

[9] Keith Moore, T. V. N. Peraud, Mark Torchia: Before We Are Born, Elsevier UK, Saunders, ISBN: 9780323313377, 2014, 9th Edition.

[10] NIST publishes compression guidance for fingerprint, Journal Elsevier - Biometric Technology Today, Volume 2014 Issue 4, April 2014, Pages 12

[11] Komlen Lalović, Patent Overview: Device for Fingerprint Identity Guarantee - Military Technical Courier, 2018, Vol. 66, Issue 2, http://dx.doi.org/10.5937/vojtehg66-15868

[12] Chouaib Moujahdia, George Bebisb, Sanaa Ghouzalic, Mohammed Rziza: Fingerprint shell: Secure representation of fingerprint template,and Pattern Recognition Letters, Volume 45, 1 August 2014, pp. 189-196

[13] Chungkeun, L., Hang, S. S., Jongchul, P., Myoungho, L. The optimal attachment position for a fingertip photoplethysmographic sensor with low DC. IEEE Sens. J. 2012;12:1253-1254

[14] Elgendi, M. On the analysis of fingertip photoplethysmogram signals. Curr. Cardiol. Rev. 2012;8:14-25

[15] Martti Juhola, Youming Zhang, Jyrki Rasku: Biometric verification of a subject through eye movements, Computers in Biology and Medicine, Vol. 43, Issue 1, pp. 42-50, Published in issue: January 01, 2013

[16] Jan Evangelista Purkynje (1787-1869): First to describe fingerprints, Andrzej Grzybowski, Krzysztof Pietrzak, Clinics in Dermatology, Vol. 33, Issue 1, pp. 117-121, Published in issue: January, 2015

[17] Esperanza Gutiérrez-Redomero, Noemí Rivaldería, Concepción Alonso-Rodríguez, Ángeles Sánchez-Andrés: Assessment of the methodology for

estimating ridge density in fingerprints and its forensic application, May 2014, Volume 54, Issue 3, pp. 199-207

[18] Kimberly Kaplan-Sandquist, Marc A. LeBeau, Mark L. Miller: Chemical analysis of pharmaceuticals and explosives in fingermarks using matrix-assisted laser desorption ionization/time-of-flight mass spectrometry, Forensic Science International, Vol. 235, pp. 68-77

[19] Lynsey Nicholson, Kevin Farrugia, David Bremner, Dennis Gentles: A preliminary investigation into the acquisition of fingerprints on food Sarah Ferguson, Science and Justice, Vol. 53, Issue 1, pp. 67-72

[20] https://brandongaille.com/20-babies-switched-at-birth-statistics

[21] https://patents.google.com/patent/WO2016036267A1/fi

Determinants of Cost Efficiency: Evidence from Banking Sectors in EU Countries

Jaroslav Belas[1], Kristina Kocisova[2], Beata Gavurova[3]

[1] Center for Applied Economic Research, Zlin, Faculty of Management and Economics, Tomas Bata University in Zlín, Czech Republic.
E-mail: belas@utb.cz

[2] Faculty of Economics, Technical University of Košice. Slovakia.
E-mail: kristina.kocisova@tuke.sk

[3] Center for Applied Economic Research, Zlin, Faculty of Management and Economics, Tomas Bata University in Zlín, Czech Republic.
E-mail: gavurova@utb.cz - Corresponding author

Abstract: The purpose of this study was to examine the cost efficiency of banking sectors within the European Union (EU) counties during the period 2008-2017 and to find out which banking sector specific variables and macroeconomic variables influenced cost efficiency. We compared cost efficiency estimated by the traditional model of Data Envelopment Analysis presented by [15] and new cost efficiency model under different unit prices presented by [37] which aimed to make the reader aware of the pros and cons of both methods. Our second stage of analysis included estimation of the regression model in order to find out the determinants of banking sectors´ cost efficiency. Panel data multiple regression was applied to find out the relationship between the depended variable (cost efficiency) and independent variables (banking sector specific variables and macroeconomic variables). Our results showed that cost efficiency was mainly explained by the capitalisation, profitability, loan risk, market structure, conditions of the economy and development of inflation.

Keywords: Data Envelopment Analysis; Cost efficiency; Banking sector; European Union countries

1 Introduction

Research problem. Under the condition of the European Union (EU) countries commercial banks as principal financial intermediaries play an important role in capital allocation. The role is to provide financial intermediation and economic acceleration by converting deposits into productive investment. The role is not only to convert deposits mainly into the loans, but this transformation should be done with minimal cost. Therefore, it is crucial to study cost efficiency by comparing the transformation process in different banking sectors with different labour costs, interest expenses and other types of costs. In the literature, there are three basic methods through which to measure cost efficiency: the ratio analysis, the parametric approach and the non-parametric approach. The parametric and non-parametric approaches differ primarily in the underlying assumptions applied in estimating cost efficient frontiers. The most commonly employed parametric procedure is the Stochastic Frontier Approach (SFA) as it allows for the effect of statistical noise to be separated from the effect of inefficiency, thereby resulting in a stochastic frontier. However, this approach requires a specific functional form that presupposes the shape of the cost efficiency frontier and assumes a specific probability distribution for the efficiency level. Additionally, if the assumptions are incorrectly specified, the estimated cost efficiency will contain errors. The non-parametric approach, commonly referred to as Data Envelopment Analysis (DEA), avoids this type of specification error because it does not require a priori assumptions about the analytical form of the cost function or an assumed probability distribution for efficiency. However, it has one major drawback in that it does not allow for random errors (e.g. measurement errors, good or bad luck) in the optimisation problem and all deviations from the cost efficiency frontier are marked as inefficiency. As both parametric and non-parametric approaches have their own merits and limitations and as the correct level of cost efficiency is unknown, the choice of a suitable efficiency estimation procedure has been quite controversial [13]. However, in the banking area, some researchers prefer to use parametric method ([40], [17]), while some studies mainly used the non-parametric approach ([20], [25], [19]). We can also find some studies comparing the results of cost efficiency estimated by both methods simultaneously ([41], [24]). Most of the mentioned studies apply the traditional cost frontier to measure efficiency. In modern literature, we can also find some studies dealing with the application of a new cost efficiency function [13], but only a few studies are dealing with the application of this method in the condition of Slovak banking ([32], [42]).

Aim and motivation. Technology and cost are the wheels that drive modern enterprises; some enterprises have an advantage regarding technology and others in cost. Hence, the management is eager to how and to what extent their resources are being effectively and efficiently utilised, compared to other similar enterprises in the same or similar field [10]. Regarding this subject, there are two different situations: one with common unit prices and costs for all Decision-Making Units (DMUs) and the other with different prices and costs from Decision-Making Unit to Decision-Making Unit. Cost efficiency evaluates the ability to produce current outputs at minimal cost. The concept of cost efficiency was first introduced by [15] and then developed by [16] by using linear programming technologies. In this model, it was assumed that input prices are the same across all Decision-Making Units. However, the prevailing price and cost assumption is not always valid in actual business, and it is demonstrated that efficiency measures based on this assumption can be misleading. So we decided to present a new cost efficiency related model introduced by [37] and compare results obtained by traditional and new cost model. Therefore, in the first stage of the analysis, we applied the DEA model under the condition of variable return to scale to compare traditional and new cost model and then used the more appropriate model to examine cost efficiency within the European Union countries during the period 2008-2017. In the second stage, we aim to examine the determinants of cost efficiency and to find out the relationship between efficiency and banking sectors specific variables and macroeconomic variables. The analysis is realised in a sample of 28 banking sectors based on the data available at the web page of the European Central Bank and Eurostat. The structure of the paper is the following. Section 2 presents a review of the literature on bank efficiency; Section 3 explains the methodology and data used in this paper; while Section 4 presents the results of the analysis. The last section concludes the paper with a summary of key findings.

2 Literature Review

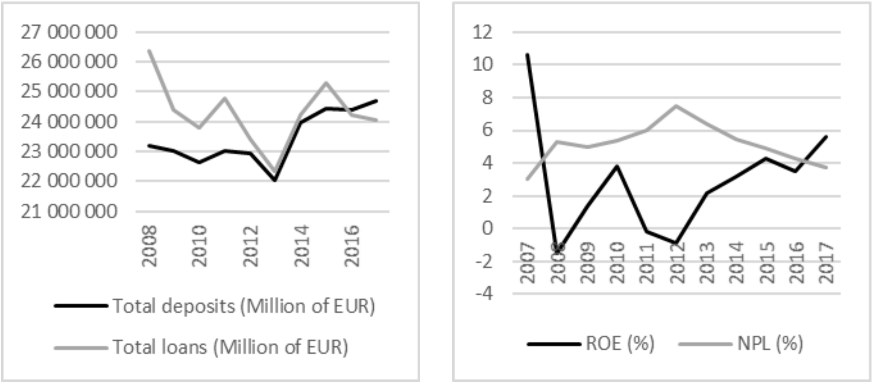
The analysis of determinants of pricing policy at the bank level in the Czech Republic was examined in the study of [22]. One of the main aims of the study was to evaluate cost efficiency. The authors used DEA and SFA. They pointed to the fact, that cost efficiency is one of the most critical factors of bank pricing policy. The study identified the crisis impact on the behaviour of the banking sector and consequently, on a higher tendency to avoid risk, while the level of flexible response to any changes in interest rate decreased.

During the period 2005-2011 the differences in cost efficiency in six countries of Central and Eastern Europe was examined by [30]. The research was based on the fact that the cost efficiency of banks is inevitable for their stability. The results showed that the macroeconomic stability of a country significantly supported efficiency. The bank with a higher risk profile was considered as more inefficient. Similarly, the bank with lower liquidity, lower level of solvency and higher credit risk were considered more inefficient than a bank that was more conscious in taking risks. Until 2008, the potential for efficiency increase was evident in the banks of all analysed banking systems. The efficiency decrease was registered in Poland, Romania, Russia and Hungary since 2009. However, the banks in Bulgaria and the Czech Republic noted efficiency stagnation. During the period 2002-2010 cost efficiency of Indonesian banks was researched by [3]. The tendency was decreasing during the examined period. The following factors influenced the cost efficiency of Indonesian banks: bank size, profitability and capital. The positive trend of efficiency is related to the lessons to be learnt by Indonesian banks from the previous crisis during 1997-1998. An adequate environment for management created by the Government of Indonesia resulted in positive performance and cost efficiency measured during the realisation of this study. Also, the authors positively assessed implementation of various programs that focus on the capital increase in the banks, improvement of competition and efficiency that resulted in a higher bank's stability to the financial crisis's impacts during 2007-2008. The most significant determinants of the cost efficiency were the following: inflation rates, the growth rate of the GDP, as well as the unemployment rate. The study provided recommendations for the management of banks in order to more effectively manage banks themselves, increase banks' sizes regarding assets and capital, manage liquidity and risk, and also provide credits for small business to maintain its credit portfolio. The cost efficiency in the Western European countries was examined by [28]. Authors emphasised numerous changes in the regulatory and competitive environment of banks, including elimination or total removal of trade and entry barriers in these European markets. Also, the authors focused on the need to use the best technologies to create new financial products. The cost efficiency of the European banks was considered crucial. The lack of management skills was regarded as the primary source of inefficiency. Similarly, the technology gap ratio (TGR) that may be interconnected with environmental variables typical of a particular country was a subject of the study. TGR and metafrontier cost efficiency showed a gradual upward trend during 1996-2000. However, the downward trend followed after this period, especially after the subprime mortgage crisis in 2007-2010. The results showed that the competitive banking market influenced higher cost efficiency and reduction of the technology gap in the member states of the EU during 1996-2000. However, international economic integration was accompanied by an influence of higher risks. The global recession and also the financial crisis affected the evolution of cost efficiency.

The influence of the financial crisis on the banking efficiency in the Euro Area countries was explained by [4]. The result indicated a gradual process of efficiency convergence among the banks of core and peripheral countries until 2008. The financial crisis had an influence on the structure of banking performance during the period 2009-2012. In many cases, the efficiency improved after the crisis that is also related to actual processes of risk management in the banks, processes of lowering costs, and a wide range of monitoring and regulatory processes that had been set in the post-crisis period. The authors highlighted the fact that positive evaluations may be influenced by the misleading interpretation of analytical results about efficiency and may appeal to a necessity of multi-dimensional assessment of efficiency aspects. The impact of financial partnership on efficiency was analysed by [31]. The research focused on Malaysia's and Islamic banks, their efficiency during 1996-2012, as well as determinants that influenced it. The authors applied SFA. The banks with allocated financial partnership reached higher efficiency than other banks. Similarly, banks with lower capital risk and a higher rate of financial partnership tended to be more efficient. It is necessary to cautiously interpret these results, especially during the period of financial crisis. Governmental authorities agreed that the financial partnership of the banks might represent an efficient strategy in increasing their efficiencies. It is important to evaluate the capital risk of the banks and the rate of their financial partnerships in research and selection of suitable banks for these types of partnerships. The study supported an idea that partnership may be used as a strategy for improving efficiency, especially due to low capital risk. The competent authorities need to examine the level of partnership as well as capital risk in monitoring banks that offer to finance by a partnership [34]. The DEA method to research banking sector efficiency was preferred by [5]. The main aim was to analyse the efficiency in Turkey and subsequent comparison of participation and conventional banks' efficiencies. The result indicated that the efficiency of a participation bank was higher than the efficiency of a conventional bank. Both groups of Turkish banks had higher technical efficiency than allocative efficiency. The technical efficiency had a higher impact on cost efficiency. In conclusion, higher allocative efficiency and effective use of resources may increase the cost efficiency in the Turkish banking system. Efficiency and loans represented a positive relationship in forming efficiency determinants. On the other hand, efficiency and expenses, capital, deposit, non-performing loans, bank size, GDP growth and inflation represent a negative relationship especially in the case of conventional banks. Both participation and conventional banks and their efficiency may be influenced by the same determinants differently. The determinants of banking performance about banking indicators, such as ROA, ROE and cost efficiency was examined by [11]. In 2005-2009, 12 banks participated in this research. The result indicated an increase of ROE by 0.06% and improvement of efficiency by 0.09% by increasing an income diversity of 5%. The authors deduced that those variables which are connected to government intervention had a negative impact on banks' performance in conventional banking model. The market share, indicators of population solvency and net credit and total net assets had the most significant influence on ROE. Variables in the DEA model were examined by [36] in order to improve the methodology of efficiency measurement. The idea emerged from methodological procedures' diversity in measuring banking sector performance. The authors suggested that the banks' loans and deposits should be used as key variables in the DEA model in order to measure the efficiency of the Lithuanian banks. The authors used the input-oriented DEA model and under the assumption of variable returns to scale. They concluded that the securities should not be used as a DEA model's outcome because this variable was not evident as statistically significant in a sample of the Latvian banks. The following inputs were used: deposits from customers, total administrative expense, balances due to credit institutions, equity, interest expense, fees and commission expense, staff expense. The outputs were as follows: loans, securities, net interest margin, operating profit, interest income, fees and commission income. However, the authors did not take into consideration the fact that some variables may be both input and output. In conclusion, the authors stated that it is essential to use such variables that reflect business specificities of all financial market participants in terms of variables' selection. The determinants of bank efficiency by DEA models during 2006-2011 was examined by [1]. The conclusion was that cost efficiency was positively affected by market concentration and demand density, while inversely related to branching. Also, these results were robust to any sample restriction anchored to the distribution of efficiency. Sensitivity analysis highlighted a significant source of cost inefficiency that was related to the risk in local markets. The comparative analysis, where the meta-regression analysis was used to compare the results from 120 papers published during 2000-2014 was prepared by [2]. The conclusions were following: banking efficiency is lower when parametric methods were used; value-added approach with intermediation method increased banking efficiency, while hybrid approach lowered it; quality of studies, number of observations and variables determined the level of efficiency; sign and magnitude were different between parametric and non-parametric studies. The Nerlovian Revenue Inefficiency model was applied in the study of [18]. This model enabled to differentiate between technical and allocative inefficiency. As a consequence of this fact, the authors also examined if the inefficiency of banks was caused by technical or allocative inefficiency. The authors suggested researching the influence of the banks' geographical locations on their total efficiency. The efficiency of the European banks before the crisis, during the crisis and also in the post-crisis period from a technical and allocative efficiency point of view was examined by [39]. The gradual improvement of technical and allocative efficiency was evident, while in a majority of countries in the post-crisis period, the trend of efficiency decrease was notified. The authors used Bayesian dynamic modelling and panel data to analyse the commercial banks that operate in the 'old' member states of the EU (15) during 2005-2012. In 2008-2012, there was a decrease in efficiency in the group of large banks. Also, an influence of the lack of banking capital as a consequence of the financial crisis that caused the stagnation of banking credit activities was presented. The authors recommend research of differences between short-term and long-term banking efficiency within the use of dynamic models. The results of [23] are linked to the study mentioned above. They examined 74 Chinese commercial banks during 2006-2013. The findings showed that banks allocate roughly 59% and 61% of labour and capital, respectively, to collect deposits in the first stage and that the average technical efficiency scores in both production stages were respectively 68% and 84%. The results also support findings that joint-stock banks were the most technically efficient, while larger commercial banks, including the big four state-owned banks, were the least technically efficient. Similar research ambitions were obvious by [13] who examined the consistency of efficiency evaluation results that were obtained by using SFA and DEA methods. Authors used panel data of the Chinese banks during 1994-2007. The results moderated consistency between parametric and non-parametric methods in efficiency rankings, identification of best and worst practice banks, the stability of efficiency over time, and the correlation between frontier efficiency and accounting-based performance measures. The only limitation would be a fact that the Chinese banking sector may be subject to important technological and regulatory processes from a long-term point of view that was not a part of this analysis. Authors appeal to an inevitability of multiple methods' use at the same time in order to evaluate banking efficiency, and a combination of these methods enable cross-check of the results. Thus, it is possible to reach a more plausible evaluation of the banking sector efficiency.

The primary role of commercial banks is to realise the financial intermediation, i.e. channell of resources from lenders to borrowers with the view to putting the existing assets to best economic use. In European Union countries, a universal banking model is prevalent. A well-developed financial infrastructure helps citizens save for the future, helps companies borrow to invest and expand, and facilitates the trade of goods and services. The scale of bank´s operations in European Union countries in 2008 was striking: banks managed an equivalent of 144% of EU-27 GDP as loans extended to households and enterprises and held an equivalent of 135% GDP in deposits. During the next year, we can see an increase in deposits and a decrease in loans. In 2017 the total deposit in the EU countries rose to 24.7 trillion EUR, which represents an increase of 6.52% compared to 2008 (Figure 1). This growth was driven by an increase in deposits from households and non-financial corporations. The share of deposit over total assets increased in 2017 to 53.4% from 51.3% in 2016, in line with the rising trend since 2007 when the share of deposit over total assets was 47.3%. That reveals a shift towards greater deposit dependency as a source of funding. The total value of loans outstanding from the EU decreased by 8.67% in 2017 compared to 2008. It was influenced by development between 2008 and 2010 when the total loans decreased by 9.77% due to restriction in loan policy as a consequence of the global financial crisis. After this year the loans started to increase, and the peak was reached in 2015 when the value of total loans in the EU countries was 25.3 trillion EUR (Figure 1).

Figure 1
Development of EU banking



Source: Prepared by the authors

The downward trend in the number of EU credit institutions, which started in 2008, continued in all years untill 2017. While in 2008 the number of credit institutions was 8525, in 2017 the number of credit institutions drops to 6250. This trend includes factors such as mergers in the banking sector to enhance profitability. The downward trend can also be seen in case of a number of employees in credit institutions. By end-2017 banks in EU countries employed about 2.71 million people, compared to 3.28 million people in end-2008. The global financial crisis has led to a substantial increase in non-performing loans (NPLs) in banks' balance sheets. This trend has been increasing since 2008 leading to a maximum NPL ratio of 7.5% in EU countries in 2012 (Figure 1).

However, the NPL ratio trajectories show a significant decline across the EU countries, which can be attributed to increased bank lending activities in recent years. From 2017, the NPL ratio for the EU was only 3.7% suggesting that NPLs are no longer a specific problem of European banks. European banks have continued to build a strong capital position and strengthen their balance sheets. Capital continued to growth, with Tier 1 ratio in EU banks was 13.8% in 2017. All banks have met the liquidity coverage ratio above the minimum. Also, the leverage and Net stable funding ratio shortfalls continued to decrease. Given that the ECB maintains its ultra-low interest rates, profitability remains a key challenge facing EU banks. The return on equity (ROE), a key indicator for assessing the attractiveness of banking sector for investors is slowly recovering. The ROE of EU banks was 5.6% in 2017, which represents only half of the values before the outbreak of the financial crisis, but is the highest since 2007 (Figure 1). The same tendencies can also be seen in the case of ROA [14].

3 Methodology and Data

Data Envelopment Analysis (DEA) was first developed by [8] under the constant return to scale assumption and provides a measure of technical efficiency. Following [15] and [16], a sequence of linear programmes was applied to construct cost efficiency frontiers, and from these, measures of traditional cost efficiency were calculated. The traditional cost efficiency model assumes that the unit cost of inputs is identical among Decision-Making Units. According to the [32], to be cost-efficient, the Decision-Making Unit must be both technically efficient (adopting the best practice technology) and allocative efficient (selecting the optimal mix of inputs to minimise the costs for a given output).

We define $\mathbf{y}_o$ as the $s \times 1$ vector of the *o*-th production unit´s *s* outputs $(r = 1,...,s)$, $\mathbf{x}_o$ is the $m \times 1$ vector of its *m* inputs $(i = 1,...,m)$, **Y** is the $s \times n$ matrix of outputs (*n* denotes the number of DMUs, $(j = 1,...,n)$), and **X** is the $m \times n$ matrix of inputs. Let us consider we have prices associated with inputs. Let **c** = $(c_1,...,c_m)$ be the standard unit input-price or unit-cost vector. Then the cost efficiency γ^* of DMU_o is defined as the ratio between minimal the cost and the actual cost:

$$\gamma^* = \frac{cx^*}{cx_o} \tag{1}$$

Where x_o^* is an optimal solution of the constant return to scale cost minimisation DEA model defined in the following terms:

Cost $$cx^{*} = \min_{x,\lambda} cx \tag{2}$$

Subject to $$x \geq X\lambda \tag{3}$$

$$y_o \leq Y\lambda \tag{4}$$

$$\lambda \geq 0 \tag{5}$$

The solution to this optimisation problem is known to be the point x^{*} where the isocost line is tangent to the isoquant. This point represents the cost minimising vector of input quantities for the evaluated production unit, given the vector of input prices $\mathbf{c}_o$ and output levels $\mathbf{y}_o$. The isoquant represents all possible combinations of inputs amount (x_1, x_2) that are needed to produce the same amount of a single output. The point *x* is a point in the interior of the production possibility set representing the activity of a Decision-Making Unit which produces this same amount of output but with a more significant amount of both inputs. To evaluate the performance of this production unit we can use the common Farrell measure of radial efficiency. The result is the measurement of technical efficiency which can be calculated as the ratio between the distance from 0 to $\tilde{x}$ and distance from 0 to *x*. If the information about the input prices is available, we can also define the isocost line whose slope is given by the ratio of input prices. Isocost line shows all combinations of inputs which cost the same total amount. The relative distance of $\hat{x}$ and $\tilde{x}$ refers to allocative efficiency which can bring minimal cost but is connected with the loss of technical efficiency [6].

In traditional cost efficiency DEA models, we assume that input prices are the same across all decision-making units. However, real markets do not necessarily function under perfect competition, and unit input prices might not be identical across all Decision-Making Units. Thus, as pointed out by [37] the traditional DEA cost efficiency model does not take account of the fact that costs can be reduced by reducing the input factor prices. For example, if two production units have the same inputs and outputs while the unit input prices for one DMU are twice those of the other DMU, then the total costs of the DMU with the higher unit input prices will be higher than those of the DMU with the lower unit input prices. However, under the traditional DEA model, the cost function is homogeneous of degree one in input prices, and the scaling factor cancels out in the cost efficiency ratio, and thus, the two DMU will be assigned the same measure of cost efficiency irrespective of the fact that they have significantly different input prices. It represents a severe drawback for assessing relative efficiency levels under the traditional DEA model and is caused by the peculiar structure of the DEA model which exclusively focuses on the technical efficiency of two DMU and cannot take account of variations in unit input prices between the DMUs. Therefore, in order to avoid this shortcoming, [37] proposed a new scheme for evaluating cost efficiency under which the production technology is homogeneous of degree one in the total costs as distinct from being homogeneous of degree one in the input prices under the traditional DEA model. It means that under the new DEA model DMUs with different input prices will return different measures of cost efficiency [13].

The new cost efficiency model is based on the definition of another cost-based production possibility set P_c as [10]:

$$P_c = \{(\bar{x}, y) | \bar{x} \geq \bar{X}\lambda, y \leq Y\lambda, \lambda \geq 0\} \tag{6}$$

Where $\bar{X} = (\bar{x}_1,...,\bar{x}_n)$ with $\bar{x}_j = (c_{1j}x_{1j},...,c_{mj}x_{mj})^T$ where $(j = 1,...,n)$. Here we assume that the matrices X and C are non-negative, and elements of $\bar{x}_{ij} = (c_{ij}x_{ij})(\forall(i, j))$, where $(i = 1,...,m)$ and $(j = 1,...,n)$, are denominated in similar units in monetary terms (e.g. euro). The new cost efficiency $\bar{\gamma}^*$ is defined as [10]:

$$\bar{\gamma}^* = \frac{e\bar{x}_o^*}{e\bar{x}_o} \tag{7}$$

Where $e \in R^m$ is a row vector with elements being equal to 1, and $\bar{x}_o^*$ is the optimal solution of the linear programmes given below:

New Cost $$e\bar{x}_o^* = \min_{\bar{x},\lambda} e\bar{x} \tag{8}$$

Subject to $$\bar{x} \geq \bar{X}\lambda \tag{9}$$

$$y_o \leq Y\lambda \tag{10}$$

$$\lambda \geq 0 \tag{11}$$

In the new cost efficiency model, the optimal input mix $\bar{x}_o^*$ that produces the output y_o can be found independently of the DMU´s current unit price c_o, whereas in the traditional cost efficiency model is keeping the unit cost of DMU *j* fixed at c_o we search for optimal input mix x^* for producing output y_o. These are fundamental differences between the two models. Using traditional cost efficiency model we can fail to recognise the existence of other cheaper input mixes. In our research, we focused on the evaluation of the cost efficiency of EU banking sectors. Banking institutions within the banking sector usually operate under the condition of imperfect competition, financial constraints, regulatory requirements and other factors that do not allow them to operate at their optimal size. For this reason, we used DEA models under the conditions of variable return to scale which minimises the impact of mentioned restrictions. We evaluated the relative efficiency of 28 banking systems during 2008-2017 based on the data available at the website of the European Central Bank (ECB). The term "relative" efficiency refers to the achieved efficiency of the evaluated banking system within the group of evaluated banking systems and of the criteria used (input and output variables according to the applied approach). We used the intermediation approach to evaluate the cost efficiency of the banking sectors. This approach views the bank as an intermediary of financial services and assumes that banks collect funds (deposits and purchases funds) with the assistance of labour and capital and transform them into loans and other assets. For each banking sector in the sample, it was necessary to select inputs, outputs and input prices. Total deposits and output variables, as well as selected types of costs, are measured in thousands of EUR. We consider two inputs, namely, total deposits (x_1), and the number of employees (x_2). Each of these inputs generates costs, referred to total interest expenses, and staff costs. Therefore, we can easily calculate prices for each input as a ratio of the particular cost to the selected input. The price of deposits (c_1) can be calculated as the ratio of total interest expenses to total deposits, and the price of labour (c_2) as the ratio of staff costs to the number of employees. On the output side, we consider two types of outputs: total loans (y_1) and other earning assets (y_2), which refer to non-lending activities. We provide descriptive statistics of all input, output variables and input prices in selected years used to calculate efficiencies in *Table 1*. The calculations of cost efficiency were done using the R software [33]. In the process of calculation of cost efficiency all input, output variables, and input prices were put together in one dataset. The reason for putting data together is that we would like to eliminate the change of efficiency affected by the change due to the technological progress, which could lead to the shift of the efficiency frontier. Within the second stage of the analysis, in order to examine the internal (banking sector specific variables) and external (macroeconomic) factors that affect the cost efficiency of banking sectors in EU countries, the following model has been developed:

$$CE_{it} = \sum_{j=1}^{J} \beta_j X_{j,it} + \sum_{l=1}^{L} \delta_l Y_{l,it} + \varepsilon_{it} \tag{12}$$

CE_{it} is the cost efficiency of the banking sector i at time t, with $i = 1,\ldots,N$; and $t = 1,\ldots,T$; $X_{j,it}$ is the banking sector specific variables of bank i at time t, with $j = 1,\ldots,J$; and $Y_{j,it}$ is the macroeconomic variables with $l = 1,\ldots,L$; and ε_{it} is the disturbance. To examine the determinants of cost efficiency, we selected the independent variables, which has been used in most studies on bank efficiency. We can divide the independent variables into two groups: the banking sector specific variables and macroeconomic variables. As a banking sector, specific variables were used: total equity over the total assets (capitalisation), net interest margin, total loans to total assets, and cost to income ratio. We used the ratio of total equity to total assets (ETA) to measure the capital strength of the banking sector. In general, we assume that a higher capital ratio indicates higher safe in the banking sector. To determine a relationship in case of this variable is not entirely clear. One point of view is that capital ratio is expected to have a positive sign since we assumed that banks are predicted to be rewarded with additional revenues for holding the optimal amount of capital ([27], [20]). The second point of view says that capital ratio is expected to have a negative sign since it is assumed that banks which hold the higher value of capital cannot provide these funds in the form of loans and this way reduces the value of potential interest income ([38], [26]). We used the net interest margin (NIM) as the profitability indicator. We can expect that more cost-efficient banking sectors can earn a higher profit, which should lead to a positive relationship between NIM and cost efficiency. The existence of a positive relationship was described for example in the work of [32] and [3]. We used the share of total loans over the total assets (TLTA) as the indicator of credit risk. As the loans are the primary item on the bank´s balance sheet, we can expect that increasing share of loans on the total assets indicate a higher probability of clients´ default and this way higher risk of bank failure. The higher share of problematic loans can lead to additional cost, and therefore we can assume a negative relationship between the TLTA and cost efficiency. The existence of a negative relationship was described, for example, in the works of [21] and [35]. The cost to income ratio (CI) as the indicator of operating efficiency represents the share of operating costs to operating income. Decreasing value of this indicator suggests that banks use their resources rationally and effectively. Therefore, we expect a negative relationship between CI and cost efficiency.

Table 1

Descriptive statistics on variables used for efficiency measurement in selected years

Variable		2008	2011	2014	2017
Total deposits (thousands of EUR)	Minimum	22529538	16204830	18391566	20990278
	Maximum	4825242292	4365091099	5165173780	4964894562
	Average	828314461	823053108	856281966	882280472
	St.dev.	1247055577	1254404587	1390576096	1431605287
Number of employees	Minimum	3872	4026	4426	4920
	Maximum	685550	663800	649900	597319
	Average	117192	110695	101976	96812
	St.dev.	170551	162459	153096	143242
Price of deposits	Minimum	0.0291	0.0148	0.0034	0.0026
	Maximum	0.2433	0.0808	0.0462	0.0351
	Average	0.0616	0.0331	0.0189	0.0118
	St.dev.	0.0385	0.0144	0.0088	0.0075
Price of labour (thousands of EUR)	Minimum	108.54	109.22	121.98	133.83
	Maximum	1606.20	1576.59	1697.62	1937.94
	Average	679.00	705.00	704.71	742.91
	St.dev.	433.27	466.09	487.17	496.59
Total loans (thousands of EUR)	Minimum	25494424	15165481	16146218	14612609
	Maximum	5109275497	4729262760	4687794919	4286872201
	Average	941490575	885183481	865485292	859829418
	St.dev.	1443001244	1339990410	1346364169	1333638597
Other earnings assets (thousands of EUR)	Minimum	1499987	988552	1340534	645256
	Maximum	3879104144	5874885874	5364815194	3117402948
	Average	500810288	554052432	505773886	339879781
	St.dev.	1007846235	1243601727	1133071099	705709326
Cost to income ratio	Minimum	0.4050	0.3124	0.3695	0.4064
	Maximum	1.8618	0.7213	0.7256	0.7402
	Average	0.6300	0.5624	0.5650	0.5665
	St.dev.	0.2673	0.0934	0.0876	0.0788

Total equity to total assets	Minimum	0.0293	-0.0059	0.0411	0.0528
	Maximum	0.1402	0.1945	0.1511	0.1508
	Average	0.0629	0.0738	0.0855	0.0912
	St.dev.	0.0266	0.0405	0.0289	0.0276
Net interest margin	Minimum	0.0079	0.0058	0.0067	0.0046
	Maximum	0.0473	0.0477	0.0411	0.0337
	Average	0.0210	0.0217	0.0208	0.0190
	St.dev.	0.0107	0.0116	0.0094	0.0082
Total loans to total assets	Minimum	0.4720	0.3573	0.3850	0.4302
	Maximum	0.9634	0.8612	0.7663	0.7648
	Average	0.6970	0.6615	0.6441	0.6493
	St.dev.	0.1193	0.1260	0.0852	0.0794
Herfindahl-Hirschman index	Minimum	0.0191	0.0317	0.0300	0.0250
	Maximum	0.3490	0.3880	0.3630	0.2419
	Average	0.1136	0.1112	0.1150	0.1145
	St.dev.	0.0775	0.0758	0.0747	0.0622
Index of a gross domestic product	Minimum	102.40	89.40	80.80	81.40
	Maximum	126.90	132.50	141.10	165.20
	Average	111.73	109.26	112.54	124.34
	St.dev.	6.72	8.86	13.19	19.56
The harmonised index of consumer prices	Minimum	78.33	92.43	98.84	99.45
	Maximum	99.50	102.36	101.57	104.48
	Average	89.75	95.72	100.09	101.90
	St.dev.	3.95	2.20	0.62	1.35

Source: Prepared by the authors

In addition to the banking sector, specific variables the analysis included a set of macroeconomic variables like an indicator of market structure, real gross domestic product, and inflation. The market structure in the banking industry is usually measured by the Herfindahl-Hirschman index (HHI). The increasing value of HHI indicates that the level of competition in the banking sector decreases and the market power is concentrated in the hand of the biggest banks on the market. In our study, we expect a positive relationship between the bank concentration ratio and cost efficiency, which is in line with the traditional structure-conduct-performance paradigm. In a highly concentrated market, enterprises have higher market power which allows them to set prices above marginal costs and achieve higher efficiency. Higher concentration reduces competition by fostering collusive behaviour among firms, whether more concentrated market improves market performance as a whole. Index of gross domestic product (GDP) reflects the conditions of the economy. We assume that the growing economy will provide a growing demand for banking services and lower risk; therefore, we expect a positive relation with cost efficiency. According to [12] the effect of inflation on efficiency depends on whether wages and other operating expenses increase faster than inflation. Many studies ([7], [29]) have found a positive relationship between inflation and cost efficiency. However, if inflation is not anticipated and banks do not adjust their interest rates correctly, there is the possibility that costs may increase faster than revenues and hence affect bank efficiency negatively. Inflation is measured by the Harmonised index of consumer prices (HICP). We present the descriptive statistics of all banking sector specific variables and macroeconomic variables in selected years in *Table 1*. In order to examine the determinants of EU banking sectors´ cost efficiency, we applied the regression analysis for panel data.

Model (1) was estimated through a pooling regression taking each banking sector´s cost efficiency (CE) as the dependent variable. The opportunity to use a panel structure of the data frame was tested with the Chow test. If the p-value of the Chow test is lower than 0.05 at 95% significance level, then it is suitable to use a panel structure of the model. A precondition for the use of a linear model is stationarity of time series. In the literature, there are several tests of stationarity of time series. To verify the stationarity in the case of our sample, we used the augmented Dickey-Fuller test (ADF test). If the data are stationary, it allows us to analyse the relationship between variables through a linear model. Primary and most used method for estimating the parameters of a linear model (regression coefficients) is the ordinary least squares method (OLS). The normality of residues distributions was tested by Lilliefors (Kolmogorov-Smirnov) normality test. The presence of the heteroscedasticity by the Goldfeld-Quandt test. To verify the correlation between the independent variables was used VIF test.

4 Results and Discussion

At first, we want to demonstrate the difference between traditional and new cost efficiency model with a simple example involving 28 banking sectors in 2017 with each using two inputs (x_1, x_2) to produce two outputs (y_1, y_2) along with input costs (c_1, c_2). The data and the resulting measurement are exhibited in *Table 2*. For the banking sector in Netherland, the traditional cost model gives the efficiency score $\gamma^* = 1$. The traditional cost model assumes that the unit cost of inputs is identical among, so do not take into account the actual prices of production units.

Table 2

Comparison of traditional and new cost efficiency

	x_1	x_2	c_1	c_2	y_1	y_2	$\bar{x}_1$	$\bar{x}_2$	e_1	e_2	Cost γ^*	Cost$\overline{\gamma^*}$
Austria	683563615	71927	0.0097	897.97	668211550	159978945	6638108	64588090	1	1	0.4475	0.4880
Belgium	736069787	53002	0.0159	1270.92	663830236	191132030	11694628	67361120	1	1	0.5842	0.4368
Bulgaria	43339685	30070	0.0046	133.83	31444887	7197650	199209	4024320	1	1	0.1547	0.5004
Cyprus	60280326	10632	0.0151	528.24	43345469	4814871	913225	5616260	1	1	0.4415	0.4171
Czech Republic	212269639	41566	0.0056	392.57	201357569	30748224	1183177	16317680	1	1	0.2939	0.6185
Germany	4643112339	597319	0.0126	757.27	4151780258	1828168231	58549722	452333000	1	1	0.4763	0.4176
Denmark	277901421	42240	0.0237	770.58	635571192	159863483	6591714	32549240	1	1	0.8472	0.8452
Estonia	20990278	4920	0.0030	385.08	18857486	645256	63361	1894600	1	1	0.8057	0.7499
Spain	2573541252	183016	0.0153	1591.53	2298517928	684184706	39425041	291274830	1	1	0.6363	0.3578
Finland	288022396	20999	0.0085	819.08	273029377	47404295	2441404	17199900	1	1	0.6636	0.7381
France	4013514129	398516	0.0209	1431.96	4218882008	1594188332	84004911	570659000	1	1	0.6994	0.3311
United Kingdom	4964894562	353299	0.0104	1348.42	4286872201	3117402948	51721425	476394970	1	1	0.8287	0.5233
Greece	212619376	41707	0.0103	550.81	177600789	30243587	2196832	22972700	1	1	0.2750	0.3816

Croatia	51008131	20434	0.0113	303.50	41854113	8465001	575900	6201750	1	1	0.2535	0.3906
Hungary	100176994	38877	0.0066	307.08	71914285	34218858	656250	11938500	1	1	0.1681	0.3325
Ireland	275382985	26891	0.0106	1131.33	262388042	95891662	2913747	30422600	1	1	0.5166	0.4185
Italy	1853534489	281928	0.0103	943.47	1893333371	547181303	19089514	265989930	1	1	0.3420	0.3422
Lithuania	24320392	8922	0.0030	187.73	19175587	1510252	73834	1674930	1	1	0.4655	0.8490
Luxembourg	622719094	26149	0.0113	1451.17	477274000	137463504	7062961	37946750	1	1	0.7943	0.5547
Latvia	24126954	8492	0.0058	352.39	14612609	4996593	141073	2992480	1	1	0.4890	0.4979
Malta	40027646	4924	0.0188	436.36	24568413	17566753	752613	2148630	1	1	0.8492	0.8245
Netherlands	1550227495	75215	0.0351	1937.94	1825197654	325947293	54435385	145761970	1	1	1.0000	0.4699
Poland	310320217	168800	0.0139	187.03	290719709	106852515	4300291	31570810	1	1	0.1530	0.4294
Portugal	306306111	46238	0.0114	687.94	241657251	85870011	3488864	31809000	1	1	0.3080	0.3652
Romania	80598771	55044	0.0059	204.32	55812592	19855043	475572	11246680	1	1	0.1091	0.2869
Sweden	637987857	70877	0.0239	1078.36	1101745032	254319209	15258667	76430780	1	1	0.7906	0.6214
Slovenia	35453871	9844	0.0037	434.23	26490120	9294822	131151	4274560	1	1	0.4057	0.4428
Slovakia	61543416	18879	0.0026	280.36	59179989	11228503	159229	5292980	1	1	0.2815	0.6484

Source: Prepared by the authors

The new scheme devised as in [37] distinguishes banking sectors by according them different cost efficiency scores. This is due to the difference in their unit costs. We can see the drop in Netherland banking sector from one $\left(\gamma^{*}_{Netherland}\right)$ to 0.4699 $\left(\bar{\gamma}^{*}_{Netherland}\right)$. Its higher cost structure explains this drop in banking sector performance. We can see that banking sector in Netherland uses 1550227495 thousand of EUR of input 1 (total deposits) with a price of 0.0351 EUR per one unit of deposits and 75215 persons of input 2 (number of employees) with a price of 1937,94 thousand of EUR per one employee. When we look at the unit cost in different banking sectors, we can see that unit cost was the highest. Therefore the banking sector in the Netherlands could not be considered as cost-efficient. It indicates that all banking sectors that use the same amount of inputs to produce the same amount of outputs but take into account different unit prices then the total costs of the banking sectors are different. Therefore we could not consider them the same cost-efficient. Therefore, we decided to analyse the cost efficiency of the banking sectors in EU countries under the new scheme described by [37]. Following the described methodology we evaluate the new cost efficiency of banking sectors within the EU countries during 2008-2017. As it was mentioned above, the intermediation approach was applied. According to the intermediation approach the input and output variables, and their prices, for each banking sector were defined. *Table 3* shows the development of new cost efficiency in individual EU banking sectors and average values for the whole EU banking sector during 2008-2017. We observed no dramatic changes in the average new cost efficiency during the analysed period, but we can see notable differences among the observed countries. *Table 3* shows the results of an average new cost efficiency obtained relative to the whole sample during the analysed period. The minimum average value was reached in 2008, the maximum average value in 2014. Results showed that the average new cost efficiency increased from 39.88% in 2008 to 53.31% in 2014 and then decreased to 51% in 2017. The average new cost efficiency at the beginning of the analysed period was 39.88% indicating that on average, banking sectors could save 62.12% of their costs by using the inputs in optimal combination while maintaining the given input prices. On average the European banking sector did not use the minimum amount of inputs for producing the given outputs, and the proportion of inputs did not guarantee the minimum possible costs. At the end of the analysed period, the average new cost efficiency was 51%, indicating potential cost-saving equal to 49%. The results of analysis per country, indicate that the new cost efficiency ranged from 14.45% (in Belgium in 2008) to 100%. The highest scores were recorded in countries like Germany (2011), Estonia (2014), the United Kingdom (2011, 2014 and 2015), Ireland (2011), and Malta (2014). The lowest scores were observed in countries like Belgium (2008, and 2009), Hungary (2012 and 2013), and Romania (2010, 2011, 2014, 2015, 2016 and 2017). Improvement in new cost efficiency during the analysed period can be seen in Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Italy, Lithuania, Luxembourg, Latvia, Malta, Netherlands, Poland, Portugal, Romania, Sweden, Slovenia and Slovakia. The decline in new cost efficiency can be seen in Germany, Spain, France, the United Kingdom, and Ireland. The most significant decrease between the years 2008 and 2017 occurred in Germany, where the new cost efficiency decreased from 86.30% to 41.769%. On the other hand, the highest increase was recorded in Belgium, where the new cost efficiency increased from 14.45% to 43.68%. The result of the analysis can suggest different banking behaviour for specific countries. Therefore, in the second stage, the regression analysis with a set of banking sector specific variables and macroeconomic variables will be done.

Table 3

New cost efficiency of the EU banking sectors, 2008-2017

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Austria	0.3624	0.4146	0.4360	0.4340	0.4296	0.4330	0.4680	0.4680	0.4493	0.4880
Belgium	0.1445	0.1788	0.3036	0.2871	0.3133	0.3670	0.3768	0.3978	0.4228	0.4368
Bulgaria	0.4176	0.4226	0.4342	0.4341	0.4290	0.4468	0.4576	0.4410	0.4772	0.5004
Cyprus	0.3998	0.4001	0.3923	0.3593	0.3634	0.3570	0.4564	0.4809	0.3911	0.4171
Czech Republic	0.3361	0.3488	0.3595	0.3553	0.3538	0.3955	0.4198	0.4329	0.4819	0.6185
Germany	0.8630	0.6763	0.6613	1	0.7642	0.6596	0.7016	0.6902	0.4991	0.4176
Denmark	0.4133	0.4331	0.5171	0.4861	0.5055	0.5264	0.6358	0.6497	0.6701	0.8452
Estonia	0.4887	0.5882	0.6435	0.5847	0.7591	0.7721	1	0.9566	0.9491	0.7499
Spain	0.3693	0.4059	0.4178	0.3837	0.3732	0.3478	0.3530	0.3770	0.3680	0.3578
Finland	0.4790	0.4776	0.6615	0.8440	0.7171	0.6642	0.8445	0.7129	0.7446	0.7381
France	0.3693	0.2985	0.2989	0.3066	0.3028	0.3075	0.3194	0.3171	0.3337	0.3311
United Kingdom	0.6292	0.6972	0.7791	1	0.8332	0.6594	1	1	0.6028	0.5233
Greece	0.2820	0.3071	0.3192	0.2856	0.2916	0.2797	0.3232	0.3428	0.3731	0.3816
Croatia	0.3916	0.3882	0.3848	0.3814	0.3779	0.4061	0.3943	0.3876	0.3885	0.3906
Hungary	0.2357	0.2298	0.2280	0.2387	0.1954	0.1964	0.3324	0.3242	0.2869	0.3325
Ireland	0.6124	0.7004	0.8124	1	0.8322	0.9568	0.4928	0.4838	0.4133	0.4185
Italy	0.2657	0.2892	0.2995	0.2915	0.2956	0.2952	0.3020	0.3137	0.3035	0.3422
Lithuania	0.5208	0.5671	0.6109	0.6751	0.7215	0.7318	0.7103	0.7625	0.8115	0.8490
Luxembourg	0.3489	0.5304	0.7173	0.6210	0.6716	0.7326	0.6848	0.6383	0.6215	0.5547
Latvia	0.4448	0.5049	0.5637	0.5976	0.6172	0.6130	0.6121	0.6143	0.5063	0.4979
Malta	0.6044	0.7792	0.8806	0.8626	0.9112	0.9978	1	0.8362	0.8377	0.8245
Netherlands	0.2990	0.3529	0.3793	0.3732	0.3922	0.4141	0.4133	0.4454	0.4569	0.4699
Poland	0.2311	0.2374	0.2475	0.2711	0.2584	0.2801	0.4060	0.4089	0.4592	0.4294

Portugal	0.3054	0.3561	0.3655	0.3408	0.3496	0.3463	0.3646	0.3365	0.3829	0.3652
Romania	0.1918	0.1896	0.2136	0.2171	0.2152	0.2248	0.2394	0.2512	0.2743	0.2869
Sweden	0.3390	0.4356	0.4493	0.4705	0.5226	0.5908	0.6501	0.6681	0.6406	0.6214
Slovenia	0.3788	0.4224	0.4310	0.4060	0.3980	0.3751	0.5181	0.5231	0.4455	0.4428
Slovakia	0.4439	0.4270	0.4495	0.4279	0.4281	0.4467	0.4509	0.4735	0.6048	0.6484
Minimum	0.1445	0.1788	0.2136	0.2171	0.1954	0.1964	0.2394	0.2512	0.2743	0.2869
Maximum	0.8630	0.7792	0.8806	1.0000	0.9112	0.9978	1.0000	1.0000	0.9491	0.8490
Average	0.3988	0.4307	0.4735	0.4977	0.4865	0.4937	0.5331	0.5262	0.5070	0.5100
St.dev.	0.1511	0.1564	0.1837	0.2410	0.2094	0.2124	0.2200	0.1980	0.1736	0.1658

Source: Prepared by the authors

To explain the variability in new cost efficiencies, we regressed the new cost efficiencies (CE) on the set of relevant banking sector specific variables and macroeconomic variables. The testing of the model was implemented in program R. The opportunity to use a panel structure of the data frame was tested with the Chow test. The proposed model (12) was tested for statistical significance of the model (F-statistics). The normality of residues distributions was tested by Lilliefors (Kolmogorov-Smirnov) normality test. The presence of the heteroscedasticity by the Goldfeld-Quandt test, and the multicollinearity by the VIF test.

Table 4

Determinants of new cost efficiency

	Coefficient	t-statistics	p-value
Cost to income ratio (CI)	-0.02061	-0.5284	0.5976703
Total equity to total assets (ETA)	1.596	4.3704	0.000018 ***
Net interest margin (NIM)	-10.197	-9.1985	< 2.2e-16 ***
Total loans to total assets (TLTA)	-0.30635	-3.4010	0.000772 ***
HHI index (HHI)	0.66848	4.9918	0.000001 ***
GDP index (GDP)	0.0024755	3.3602	0.00089 ***
Inflation (HICP)	0.0044088	4.2337	0.000031***
Sample size	Balanced Panel: n=28, T=10, N=280		
R^2 (Adjusted R^2)	0.37556 (0.36184)		
F-statistics	F-statistics: 23.4277 on 7 and 273 DF, p-value: < 2.22e-16		
Chow Test	F = 5.6295, df1 = 637, df2 = 1700, p-value < 2.2e-16		
ADF test	stationary		
Lilliefors normality test	D = 0.05637, p-value = 0.03161		
Goldfeld-Quandt test	GQ = 0.69277, df1 = 133, df2 = 133, p-value = 0.9824		
VIF test	CI (1.1009); ETA (1.8609); NIM (1.9271); TLTA (1.3187); HHI (1.1386); GDP (1.2903); HICP (1.1234)		

'***' 0.01 '**' 0.05 '*' 0.1

Source: Prepared by the authors

Table 4 reports the regression results for our models. As can be seen, there was no problem with heteroscedasticity, and multicollinearity and the residues were normally distributed. Six variables were identified as the statistically significant: capitalisation, profitability, loan risk, market structure, conditions of the economy and inflation. The capitalisation measured as the ratio of total equity and total assets had positive and significant, similar to the findings of [20], [27], [32], or [31], and [3], who assumed that banks are predicted to be rewarded with additional incomes to maintaining the optimal amount of capital. It signalises that better-capitalised banking sectors were safer compared to those with lower capitalisation and therefore better-capitalised banking sectors might face lower costs of funding due to lower prospective bankruptcy costs. In concrete terms, an increase of the capitalisation by one percentage point led to an average increase of the new cost efficiency of 1.596 percentage point. This significant positive development could also be affected by the implementation of additional capital buffers in line with Basel III in the post-crisis period, wherein the period of new cost efficiency increase also increase the level of capitalisation [9]. The indicator of profitability, net interest margin, was negative and significant. The results in not in line with findings of [32] and [3] who expected that more efficient banking sectors could earn a higher profit, and therefore they expected a positive relationship between NIM and new cost efficiency. In our study, the relationship was marked as negative. It should be influenced by the policy of low interest rates of the European Central Bank (ECB). This low-interest rates of ECB passed through to the interest rates of commercial banks for loan and deposit product. The commercial banks, as well as whole banking sectors, must face this situation which eliminates the level of net interest margin. The commercial banks tried to replace the shortage of interest income by non-interest income, which is evident by the increase of non-interest incomes on gross incomes in the post-crisis period. Therefore, the commercial banks were able to increase their cost efficiency also in the time when the net interest margin decreased. The regression coefficient can be interpreted as follows, a decrease of net interest margin by one percentage point led to an average increase of the new cost efficiency of 10.197 percentage points. The ratio of total loans to total assets, as the indicator of loan risk, was significant and had a negative impact on new cost efficiency, similar to findings of [21], [35], and [30]. According to [32], this finding might be a result of holding riskier loans or having poor credit management. The impact of the Herfindahl-Hirschman index, as an indicator of the market, was positive and significant. It is in line with our assumption and also with the structure-conduct-performance paradigm. In a highly concentrated market, banks had higher market power which allowed them to set prices above marginal costs and achieve higher efficiency. Higher concentration reduced competition by fostering collusive behaviour among banks, whether more concentrated market improves cost efficiency as a whole, which is also in line with the finding of [1] and [2]. The condition of the economy described by the index of the Gross domestic product had a significant and positive impact on new cost efficiency. This finding is similar to findings of [32], [30], [5] and [3]. Also, inflation had a positive and significant impact on new cost efficiency as was also found out by [7] and [29]. The coefficient of operating efficiency, cost to income ratio, was negative but not significant. The negative sign of cost to income ratio confirmed our expectation. The more efficient banking sector (banking sectors with lower cost to income ratio) were also more cost-efficient. This result clearly shows that efficient cost management was a prerequisite to improve the overall cost efficiency of the banking sectors in EU countries. To check the robustness of our results, we decide to exclude not significant variables from the regression model and test the model again. After the exclusion of cost to income ratio, the results of new testing made all variables as significant. The impact of individual parameters has not changed, and the overall adjusted R square slightly increases (0.3635). Within the new model, there was no problem with heteroscedasticity, and multicollinearity and the residues were normally distributed.

Conclusions

The study aimed to find out which banking sector specific variables and macroeconomic variables influenced new cost efficiency of banking sectors in European Union countries during the period 2008-2017. In the first stage of our analysis, we compare the traditional cost model proposed by [15] and the new cost efficiency model under different unit prices presented by [37]. We have found out that in traditional cost model the banking sectors with different unit cost could be considered as efficient. However, under the new cost efficiency model, where information about unit cost is taking into account, the banking sector where the unit cost was higher could not be considered as cost-efficient anymore. It indicates that all banking sectors that use the same amount of inputs to produce the same amount of outputs but take into account different unit prices then the total costs of the banking sectors are different. Therefore, we could not consider them the same cost-efficient. Therefore, we decided to analyse the cost efficiency of the banking sectors in EU countries under the new scheme described by [37]. The results of our cost efficiency analysis indicated no dramatic changes in the average new cost efficiency during the analysed period, but we can see notable differences among the observed countries. Results showed that the average new cost efficiency increased from 39.88% in 2008 to 53.31% in 2014 and then decreased to 51% in 2017. The results of analysis per country, indicate that the new cost efficiency ranged from 14.45% (in Belgium in 2008) to 100%. The highest scores were recorded in countries like Germany, Estonia, the United Kingdom, Ireland, and Malta. The lowest scores were observed in countries like Belgium, Hungary, and Romania. The result of the analysis can suggest different banking behaviour for specific countries. Therefore, in the second stage, the regression analysis with a set of banking sector specific variables and macroeconomic variables will be done. We found that the statistically significant variables were: capitalisation, profitability, loan risk, market structure, conditions of the economy and inflation. These results had some political implications, as the capitalisation, loan risk and marked structure can be regulated. In banking sectors where the level of new cost efficiency was low, one way how to improve the efficiency could be the improve the level of capitalisation. The regulation authority can implement strict rules for capital regulation, which should lead to an improvement in cost efficiency. The results also pointed to the fact that more risky activities decline the level of new cost efficiency. Therefore, the regulation authorities in countries with higher level of loan risk can implemented additional measure to eliminate the level of loan risk (e.g. tightening credit requirements, reducing the loan to value ratio, limit on the indicator of the ability to repay the loan, limit on the indicator of total indebtedness to overall income of household), which should lead to increase in new cost efficiency.

References

[1] Aiello, F., & Bonanno, G. (2016a) Efficiency in banking: a meta-regression analysis. *International Review of Applied Economics,* Vol. *30*, No. 1, pp. 112-149

[2] Aiello, F., & Bonanno, G. (2016b) Looking at the determinants of efficiency in banking: evidence from Italian mutual-cooperatives. *International Review of Applied Economics,* Vol. *30*, No. 4, pp. 507-526

[3] Anwar, M. (2018) Cost efficiency performance of Indonesian banks over the recovery period: A stochastic frontier analysis. *The Social Science Journal*

[4] Asimakopoulos, G., Chortareas, G., & Xanthopoulos, M. (2018) The eurozone financial crisis and bank efficiency asymmetries: Peripheral versus core economies. *The Journal of Economic Asymmetries*, Vol. *18*

[5] Batir, T. E., Volkman, D. A., & Gungor, B. (2017) Determinants of bank efficiency in Turkey: Participation banks versus conventional banks. *Borsa Istanbul Review*, Vol. *17,* No. 2, pp. 86-96

[6] Bogetoft, P., & Otto, L. (2010) *Benchmarking with DEA, SFA, and R*. Springer, London

[7] Bourke, P. (1989) Concentration and other determinants of bank profitability in Europe, North America and Australia. *Journal of Banking & Finance*, Vol. *13*, No. 1, pp. 65-79

[8] Charnes, A., Cooper, W. W., & Rhodes, E. (1978) Measuring the efficiency of decision-making units. *European journal of operational research*, Vol. *2*, No. 6, pp. 429-444

[9] Cipovová, E., & Belas, J. (2012) Assessment of Credit Risk Approaches in Relation with Competitiveness Increase of the Banking Sector. *Journal of Competitiveness*, Vol. *4*, No. 2, pp. 69-84

[10] Cooper, W. W., Seiford, L. M., & Zhu, J. (2007) *Data envelopment analysis*. Springer, Boston, MA

[11] Daly, S., & Frikha, M. (2016) Banks and economic growth in developing countries: What about Islamic banks?. *Cogent Economics & Finance*, Vol. *4*, No. 1

[12] Dietrich, A., & Wanzenried, G. (2011) Determinants of bank profitability before and during the crisis: Evidence from Switzerland. J*ournal of*

International Financial Markets, Institutions and Money, Vol. *21*, No. 3, pp. 307-327

[13] Dong, Y., Hamilton, R., & Tippett, M. (2014) Cost efficiency of the Chinese banking sector: a comparison of stochastic frontier analysis and data envelopment analysis. *Economic Modelling*, Vol. *36*, No. 1, pp. 298-308

[14] European Banking Federation (2018) *Banking in Europe: EBF Facts and Figures 2018*. Available at, https://www.ebf.eu/facts-and-figures/, referred on 20/02/2019

[15] Farrell, M. J. (1957) The measurement of productive efficiency. *Journal of the Royal Statistical Society*, Vol. *120*, No. 3, pp. 253-290

[16] Färe, R., Grosskopf, S., & Lovell, C. K. (1985) *The measurement of the efficiency of production*. Springer, Boston, MA

[17] Fries, S., & Taci, A. (2005) Cost efficiency of banks in transition: Evidence from 289 banks in 15 post-communist countries. *Journal of Banking & Finance*, Vol. *29*, No. 1, pp. 55-81

[18] Fukuyama, H., & Matousek, R. (2017) Modelling bank performance: A network DEA approach. *European Journal of Operational Research,* Vol. *259*, No. 2, pp. 721-732

[19] Gavurová, B., Belás, J., Kocisova, K., Dapkus, R., & Bartkute, R. (2017) Revenue and cost efficiency of banking sectors in the European Union countries: Do they depend on size, location or crisis period?. *Transformation in Business & Economics*, Vol. *16*, No. 2, pp. 124-146

[20] Grigorian, D. A., & Manole, V. (2006) Determinants of commercial bank performance in transition: an application of Data Envelopment Analysis. *Comparative Economic Studies*, Vol. *48*, No. 3, pp. 497-522

[21] Hassan, M. K., & Bashir, A. M. (2003) Determinants of Islamic Banking Profitability. *Paper presented at 10th ERF Annual Conference*, Morocco

[22] Havranek, T., Irsova, Z., & Lesanovska, J. (2016) Bank efficiency and interest rate pass-through: Evidence from Czech loan products. *Economic Modelling*, Vol. *54*, pp. 153-169

[23] Huang, T.-H., Lin, C.-I., & Chen, K.-C. (2017) Evaluating efficiencies of Chinese commercial banks in the context of stochastic multistage technologies. *Pacific-Basin Finance Journal*, Vol. *41*, No. 2, pp. 93-110

[24] Iršová, Z. (2009) Measuring bank efficiency. Charles University in Prague, Faculty of Social Sciences, Institute of Economic Studies, Master Thesis

[25] Jimborean, R., & Brack, E. (2010) The cost-efficiency of French banks

[26] Kočišová, K. (2014) Profitability Determinants and the Impact of Global Financial Crisis. *5th Central European Conference in Regional Science*, 396-406

[27] Kosmidou, K., Pasiouras, F., Doumpos, M., & Zopounidis, C. (2006) Assessing Performance Factors in the UK Banking Sector: A Multicriteria Methodology. *Central European Journal of Operations Research*, Vol. *14*, No. 1, pp. 25-44

[28] Lee, C. C., & Huang, T. H. (2017) Cost efficiency and the technological gap in Western European banks: A stochastic meta-frontier analysis. *International Review of Economics & Finance*, Vol. *48*, No. 3, pp. 161-178

[29] Molyneux, P., & Thornton, J. (1992) Determinants of European bank profitability: A note. *Journal of Banking & Finance*, Vol. *16*, No. 6, pp. 1173-1178

[30] Niţoi, M., & Spulbar, C. (2015) An Examination of Banks’ Cost Efficiency in Central and Eastern Europe. *Procedia Economics and Finance*, Vol. *22*, pp. 544-551

[31] Othman, N., Abdul-Majid, M., & Abdul-Rahman, A. (2017) Partnership financing and bank efficiency. *Pacific-Basin Finance Journal*, Vol. *46*, pp. 1-13

[32] Pančurová, D., & Lyócsa, S. (2013) Determinants of commercial banks' efficiency: evidence from 11 CEE Countries. *Finance a Uver*, Vol. *63*, No. 2, pp. 152

[33] R core team. (2013) *R: a language and environment for statistical computing*. Available at, http://www.r-project.org/, referred on 18/02/2017

[34] Rahman, A., Rozsa, Z., & Cepel, M. (2018) Trade Credit and Bank Finance – Evidence from the Visegrad Group. *Journal of Competitiveness*, Vol. *10*, No. 3, pp. 132-148

[35] Rumler, F., & Waschiczek, W. (2012) Have Changes in the Financial Structure Affected Bank Profitability? Evidence for Austria. Oesterreichische Nationalbank, *Working Paper*, No. 180

[36] Titko, J., Stankevičienė, J., & Lāce, N. (2014) Measuring bank efficiency: DEA application. *Technological and Economic Development of Economy*, Vol. *20*, pp. 739-757

[37] Tone, K. (2002) A strange case of the cost and allocative efficiencies in DEA. *Journal of the Operational Research Society*, Vol. *53*, No.11, pp. 1225-1231

[38] Tregenna, F. (2009) The Fat Years: The Structure and Profitability of the US Banking Sector in the Pre-crisis Period. *Cambridge Journal of Economics*, Vol. *33*, No. 4, pp. 609-632

[39] Tsionas, E. G., Assaf, A. G., & Matousek, R. (2015) Dynamic technical and allocative efficiencies in European banking. *Journal of Banking & Finance,* Vol. *52*, pp. 130-139

[40] Weill, L. (2003) Banking efficiency in transition economies. *Economics of Transition*, Vol. *11*, No. 3, pp. 569-592

[41] Weill, L. (2004) Measuring cost efficiency in European banking: A comparison of frontier techniques. *Journal of Productivity Analysis*, Vol. *21*, No. 2, pp. 133-152

[42] Zimková, E. (2015) Cost efficiency of Slovak commercial banks under the standpoint of the intermediation approach. *Conference Proceedings 18th Applications of Mathematics and Statistics in Economics*

A Comprehensive Overview of Digital Signal Processing Methods for Voltage Disturbance Detection and Analysis in Modern Distribution Grids with Distributed Generation

Aleksandar M. Stanisavljević, Vladimir A. Katić, Boris P. Dumnić, Bane P. Popadić

University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, acas@uns.ac.rs, katav@uns.ac.rs, dumnic@uns.ac.rs, banep@uns.ac.rs

Abstract: The rapid trends towards smart grids and implementation of distributed generation (DG) and renewable energy sources bring new challenges in power quality domain. Modern distribution grids have a higher amount of voltage disturbances due to DGs power converters, nonlinear loads and system faults. The on-going research on development of new, faster and more reliable techniques for detection and analysis of voltage variations in order to prevent malfunction of equipment or to support gird and enhanced its operation, is at present very important topic. The paper presents a comprehensive overview of voltage disturbances detection and analysis methods, which use different digital signal processing techniques for use in modern distribution grids. Comprehensive, critical literature review encompassed wide range of methods, from standard, well-known ones over digital signal processing (DSP) ones to the advanced, hybrid methods. Simulation and laboratory evaluation of methods applied as part of grid-tie inverter control is presented. Advantages and disadvantages are underlined and critical evaluation of selected methods is presented. The main criteria for evaluation of methods are the speed of detection, a reliability of methods, analysis capability and computational complexity (i.e. cost of application).

Keywords: Power quality disturbances; fault analysis; artificial intelligence; signal processing; smart grids

1 Introduction

Modern concept of smart grid implies multilayer structure around power system with wide application of digital technologies and encompasses integration of energy network with digital communication network, wide-area measurements, powerful computer data processing, management and large data bases. In energy layer, it enables two way energy flows due to connections of distributed generation (DG), renewable energy sources (RES), electric vehicles with energy recovery feature, fast energy storage, high efficient and sophistically controlled industrial and domestic loads and other devices making distribution network active one. Most of them are connected to the grid with some type of power electronics (PE) systems. It could be grid-tied inverter (DC/AC converter), as in case of PV systems, or grid-tied rectifier (AC/DC converter) in cases of industrial or domestic loads, or their combination, AC/DC/AC (back-to-back) converter, as in cases of some types of wind generators or some other PE converter.

All these PE systems are sensitive to voltage disturbance (VD) in the grid. The major disturbances are large power variation, either on load side or on generation side (in case of renewable generation) leading to voltage variation and unbalance. Another set of voltage disturbances result from different type of faults (short-circuits) resulting in voltage interruptions, voltage dips (sags), voltage swells or other. These disturbances affect proper operation of different loads, especially sensitive ones, cause load tripping, overheating and might produce significant economic and production losses [1]. The generation units are affected, also, especially in cases of voltage dips. On the other hand, power electronic devices having non-linear characteristics induce additional distortion on voltage waveform (harmonics, flickers, etc.). There are also other sources of VDs, like overvoltage due to lightning strikes, impulses due to capacitor bank switching, etc.

In this paper, focus will be on VD, especially on voltage dips and their interaction with DGs. In such cases PE devices are subject to high over current stresses, errors in synchronization circuits (PLL), increase of current distortion and other effects, which may result in their tripping. However, according to recent grid codes the PE based generation units need to stay connected to the grid during the voltage dips (for a defined period of time) and support the grid by supplying some amount of reactive and active power or only reactive power, depending on the voltage dip depth [2, 3].

The first step in reducing the effects of VDs, especially of voltage dips, is fast and reliable detection. The control system (as a part of PV system connected to the LV or MV grid) should switch from the normal operation mode to grid fault operation mode as soon as possible[4]. In that case, behaviour of the whole control system of the grid-tie converter or similar device may be swiftly adopted to low voltage ride-through (LVRT) requirements. Also, there are different applications of voltage dips detection and analysis (VDDaA) methods in Dynamic Voltage Restorers (DVRs) [1, 5-7], Series and Shunt controllers based on voltage-source converters, Unified Power Quality Conditioning Systems (UPQCSs) [1, 6, 8, 9], microprocessor relay protection [10], DGs control algorithms [3, 11-13], PQ monitoring algorithms [14, 15] and FACTS. For all these systems, it is desirable that VD is detected with the shortest delay that is achievable.

In modern power systems a large number of voltage disturbances data, which may be recorded makes analysis very complex [16]. Many researchers have applied some type of the digital signal processing (DSP) based methods [17, 18] for

VDDaA. They are using a large collected scientific experience from other fields, like telerobotics [19], numeric estimation [20] or nanostructure analysis [21] to name a few. The existing paper reviews [14, 22-29] present mainly wide-range overview of the technical literature based on comparison of results given in these papers. They are obtained in different conditions and for different PE systems. From these references, it can be concluded that proposed algorithms for different voltage dips detection and analysis (VDDaA) are tested by computer simulations, only. Also, it can be observed that the main advantage of new methods is in their ability to detect and analyse multiple disturbances and to successfully classify them even in noisy conditions [22, 25, 30, 31].

In this paper focus is on application of VDDaA methods in distribution grids, with special emphasis on characteristics that are important for applications in such conditions. The paper’s aim is to present a comprehensive and critical literature overview for VDDaA methods. Detailed classification of these methods is given. Based on reviewed literature, comparison of selected VDDaA methods is presented. Also, for compared methods advantages and disadvantages are highlighted. The comparison is made according to the three main criteria: speed of detection, analysis capability and complexity and cost of implementation.

The main contribution of the paper is that comparison and evaluation of VDDaA methods are done under the same conditions and performed by experimental testing in laboratory using both grid emulator generated voltage dips and voltage dips measured in real grid. The comparison is done in the case of application of all these methods for control of a grid-tied inverter using three mentioned criteria and by evaluating each result with specific unique grade (from 1 to 10). In this case the optimal method may be selected with more reliability than in previous reported researches.

The practical value of this overview is that it may be a relevant source for insight in potential and features of a broad spectrum of VDDaA methods. Also, the best ones can be used as part of grid-connected converters control, in LVRT support algorithms, PQ monitoring devices or for other applications. By using of the selected optimal method significant improvement in the control algorithm of these PE devices is possible, i.e. control engineers will have possibility to achieve better performances and capabilities of the control systems.

The paper is organized as follows. Theoretical background is given in the second section and contains brief description of the PQ standards and basics on voltage disturbance and analysis algorithms. In the third section a comprehensive critical overview of VDDaA methods with classification is presented. In the fourth section, the results of comparison of previously reported methods and the ones achieved by laboratory testing using real measurement data and grid emulation are presented and described. The conclusions, future scope, acknowledgements and references are given in final part of the paper.

2 Theoretical Background

This section describes a theoretical background on important power quality standards and gives details of the VDDaA algorithm.

2.1 Voltage Dip

Voltage signal should be less than 90% of the RMS nominal voltage value to consider it a voltage dip and perform detection. The common detection method is the RMS. This method is standard one, and for many years it has been used in practice [1]. Voltage dips can be classified in different ways, for example, using voltage amplitude and phase angle variation, ABC classification (7 types of dips) [1], or using amplitude time change and measuring duration of dips.

2.2 Power Quality Standards

Harmonics in power systems attracted a lot of attention and large effort is made in order to achieve accurate estimation and reliable mitigation of them. Many standards, guidelines and recommendations are published, including IEEE 519-2014, EN 50160 and several IEC 61000 standards (6100-4-30) [32].

Also, other VDs are addressed in several other IEEE standards. In the IEEE 1159-2009 the classification and definition of VD are presented. According to it voltage dips are defined as a decrease of 0.1-0.9 p.u. in the voltage magnitude at system frequency with the duration of half cycle to 1 min [33]. The IEEE Std. 1564-2014 identifies, describes and defines appropriate voltage dip indices, as well as characteristics of electrical power systems [34].

2.3 Voltage Disturbance Detection and Analysis

Normal duration for voltage dip detection that a standard VD algorithm requires is 1 to 2 grid cycles. Such a reaction time may not be always appropriate, as modern grids have new types of PE equipment and grid requirements are upgraded.

VD analysis is a complex task which can be divided in several stages. The first stage of VD analysis is measurement. Depending on the application, type of device and equipment, measurement usually includes some sort of transformers. Measurement can also include sampling, analogue anti-aliasing filtering, down-sampling, or other signal preparation steps. The next step, after measurement, is transforming voltage signal values from analogue to digital (A/D conversion). This paper assumes that A/D conversion is done without any errors and with sufficient sampling rate according to the Niquist-Shennons theorem [32]. A simplified algorithm of typical VDDaA scheme (based on algorithm described in [32]) is shown in Fig. 1. Digital waveform data are then pre-processed with a transformation or feature extraction, as the voltage waveform cannot be directly used to detect VDs (and any other PQ disturbance) [32]. Also, it is not suitable for voltage magnitude analysis. For example, for voltage dips detection the one-cycle RMS voltage should be compared with a 90% value of magnitude every half cycle, according to IEEE 1564-2014 [34] and IEC 61000-4-30 [35]. Once the disturbance is detected, this information can be used in devices to aid LVRT, or in inverter control for choosing proper power profiles (if it is recognized as voltage dip) [4]. After detection, the so-called single-event indices (also known as single-event characteristics) which typically include duration and some kinds of magnitude are obtained. Besides these, actual analysis differs for different types of events, and signal can be further analysed and other indices can be calculated and stored or further processed. Some of these indices are depth of a dip, phase shift, voltage dip energy (Evs), voltage dip severity (Se), system index (a parameter indicating the voltage or current quality), harmonics, type of fault, estimated distance of fault, harmonics, etc. These data can be used for diagnostics, for calculating additional fault parameters and causes of disturbance, improvement of control, in PQ classification algorithms, etc.

3 Overview and Classification of Voltage Disturbances Detection and Analysis Methods

A large number of papers that present new methods and algorithms for detection and analysis of voltage disturbances are published in the last two decades. Mostly, they use some type of the digital signal processing (DSP) algorithm to extract features and further analyse them, to obtain detection or classification of the disturbances, to estimate there's a characteristic, to calculate distance of the fault, etc. General classification of the voltage disturbance detection and analysis methods is presented in Fig. 2. It can be seen that the DSP methods for VDDaA can be divided in the three large categories: Standard DSP methods, DSP based methods and DSP and AI based methods.

Both voltage dips and voltage variations use the RMS of voltage as their basic measurement quantity [32, 34, 35]. Because of that, the RMS method is the most commonly used method for detection and segmentation. According to IEC 61000-4-30 [35] for the detection of voltage dips, the one-cycle RMS voltage value is

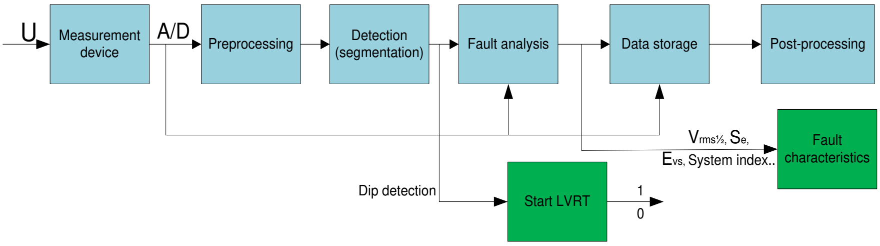


Figure 1
A general scheme of voltage dips detection and analysis methods

compared with a threshold every half cycle, also, in IEEE Std. 1564-2014 [34] for voltage dip characteristic voltage, depth of dip, etc. In addition, several different variations of the RMS method exist and they can be characterized as advanced RMS calculation methods. Because RMS methods are well known, they will be only briefly addressed.

The second group of DSP VDDaA methods is DSP methods based on transforms (or just DSP based methods). Algorithms in this group use mathematical transforms (usually harmonic estimation) to obtain voltage disturbance indices in time, frequency, or other domain. Based on transformed signals, they detect and further analyse disturbances. This is probably the largest group, which is further developed in several different directions.

The third group, most up-to-date, covers methods that utilize some form of artificial intelligence (AI). The AI is used in order to improve performances of detection and analysis. Comparing a feature of voltage, e.g. RMS with the threshold (0.9 p.u.) is replaced with complex pattern recognition and learning models. Usually, some form of neural networks (NN) or Fuzzy logic (FL) is used to improve detection and analysis of disturbances. For pre-processing or segmentation these methods use some of the DSP methods, e.g. Wavelet transform (WT), FFT, Hilbert-Huang transform (HHT), Short Time Fourier transform (STFT), S-transform (ST), etc. These methods are not the topics of this paper, because they are still in developing and methods are not common in applications that are addressed.

3.1 Standard DSP Methods

The voltage waveform cannot be directly used to detect or classify events. Because of that, simple and the most common methods are based on the direct extraction of the voltage magnitude RMS from the voltage waveform. Also, very frequent approach is to calculate fundamental-voltage magnitude sequences (the approximated RMS) and to detect and analyse disturbances on the basis of that.

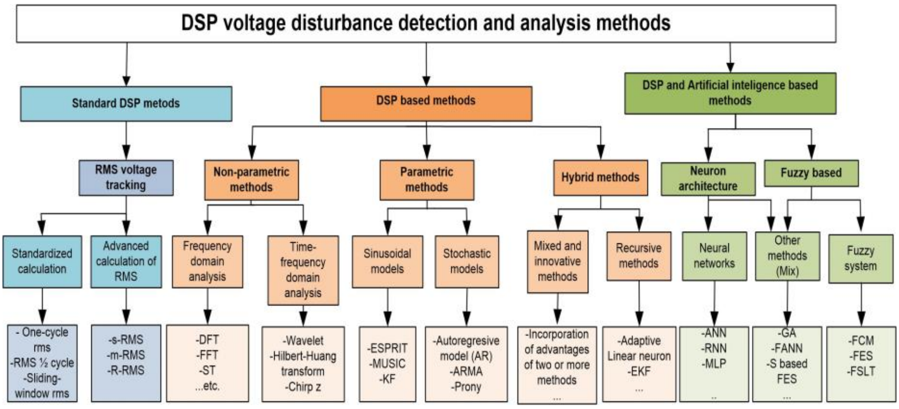

DSP voltage disturbance detection and analysis methods
Standard DSP metods
DSP based methods
DSP and Artificial inteligence based methods
RMS voltage tracking
Non-parametric methods
Parametric methods
Hybrid methods
Neuron architecture
Fuzzy based
Standardized calculation
Advanced calculation of RMS
Frequency domain analysis
Time-frequency domain analysis
Sinusoidal models
Stochastic models
Mixed and innovative methods
Recursive methods
Neural networks
Other methods (Mix)
Fuzzy system
- One-cycle rms -RMS ½ cycle -Sliding-window rms
-s-RMS -m-RMS -R-RMS
-DFT -FFT -ST ...etc.
-Wavelet -Hilbert-Huang transform -Chirp z
-ESPRIT -MUSIC -KF
-Autoregresive model (AR) -ARMA -Prony
-Incorporation of advantages of two or more methods ...
-Adaptive Linear neuron -EKF ...
-ANN -RNN -MLP ..
-GA -FANN -S based FES ...
-FCM -FES -FSLT

Figure 2

Classification of voltage disturbance detection and analysis methods

An important parameter for the RMS is block (buffer) size of a data sequence that is analysed. RMS obtained from using a half-cycle window has a higher time resolution, but with more fluctuation compared with RMS obtained from a one-cycle window [32]. Voltage RMS magnitude is usually obtained from discrete signal using (1).

$$V_{rms} = \sqrt{{}^{1}/_{2} * \sum_{n=1}^{N} v_i^2} \tag{1}$$

where *N* is a number of samples (buffer size), *n* is the *n*th sample of the data and v_i is digital voltage signal.

In IEEE [34] and IEC [35] standards, $V_{rms1}/2$ is defined as a value of RMS voltage measured during one cycle and results are updated each half cycle (RMS ½ cycle). In [36] different ways in which RMS can be calculated are presented, using fix window of different durations (s-RMS), moving average technique (m-RMS) or infinite impulse response (recursive moving average, r-RMS). If the RMS is continuously calculated over a windowed signal, using past samples from an input, it is called a moving average finite impulse response (FIR) filtering, and it is abbreviated as s-RMS [36].

Beside delay in detection, limitation of estimation of magnitude and duration (especially for short duration faults), as well as inability to calculate phase-angle information nor the point-on-wave when fault starts are drawbacks of these methods [37].

3.2 DSP-based Methods

DSP based methods for VDDaA include methods that use various types of transformations and can be further divided into three sub-categories as non-parametric methods (NPM), parametric methods (PM) and hybrid methods (HM).

3.2.1 Non-Parametric Methods

The NPMs have low computational complexity. They calculate harmonics with algorithms that are applied directly on discretized voltage waveform [38]. Also, they are well-known, easy to use and implementation costs of these methods are low.

In literature, two subgroups of the NPMs can be found: Frequency domain analysis and Time-frequency domain analysis [38]. Transformation from time domain to the frequency domain is usually done with FFT. The FFT is a way of calculation of the DFT that can be defined as in (2).

$$H(m) = \sum_{n=0}^{N-1} x(n) * e^{-j(2\pi/N)kn} \tag{2}$$

where *H(m)* is calculated harmonic, *n* is the *n*th sample of the data, *m* is frequency index and *N* is number of samples (buffer size).

However, FFT has many known problems: leakage effect, sensitivity to frequency deviation, etc. [32, 39]. Many researchers proposed new solutions in order to improve FFT and to solve these well-known problems. Newly proposed algorithms try to improve FFT using synchronization [40, 41], windowing [42, 43], interpolation [44, 45] or using different sampling techniques [46], etc. Besides these algorithms, new NPMs emerged. Many of them are developed for PQ event detection and analysis and become widely used, like WT [47, 48] and HHT [39, 49]. Also, advanced successors of the FT and FFT, like S-transform or STFT, show very good results in different applications (detection of faults in modern grids [50] or in power quality analysis [51, 52]).

The WT is one of the most commonly used methods for harmonic analysis in PQ associated applications. The WT estimates a local representation of signal in a time domain and in a frequency domain, and this is usually consider as time-frequency representation. The discrete WT can be calculated as shown in (3).

$$\mathrm{F}(i,j) = L^{-j/2} * \sum_{i=0}^{N} \mathrm{f(n)} * \omega * \left(\frac{n-i}{L^j}\right) \tag{3}$$

where v_i is digital voltage signal, F_{ij} is matrix that consists of decomposed v_i values, j is the level of the decomposition, i is band index, L is dilatation translation parameter (for Dyadric wavelets it is equal to 2), N is number of samples, ω is complex conjugate and n is nth sample of the data [32]. It shows very good results as tool for analysing fast-changing signals, like VDDaA [53]. Mostly, the highest frequency band is used for detection of voltage disturbances [32]. In [54] method for VDDaA, with WT used as a tool for detection and extraction of useful information from disturbance is presented. Probabilistic NN (PNN) is used for detection of patterns and classification. A main disadvantage of wavelets is that the centre frequencies of the sub band filters are difficult to be set in the harmonic frequencies, making them less attractive to harmonic-related disturbance analysis [32]. Also, detection using wavelets is prone to noise and signal deterioration [55-57].

S-transform (ST) is modified version of WT that is well-known, mainly for application in PQ analysis and classification algorithms [58]. In [59] comparative study for wavelet and ST for PQ disturbance detection, analysis and is landing detection is presented. It is concluded that S-transform is better than wavelet for detection and localization of PQ events based on simulations and experimental results. Mathematical model of ST (continuous integration formulation) can be written as in (4):

$$\mathrm{S}(\tau,j) = \int_{-\infty}^{\infty} \mathrm{x(t)w(t} - \tau, j)\, dt \tag{4}$$

where $\mathrm{w[t} - \tau, j]$ is a scaled replica of the fundamental mother wavelet, as defined for WT, t in this case is dilation that determines the width of the wavelet and resolution of transformation. Further, if for multiplication is used function S:

$$S = e^{i2\pi ft} \tag{5}$$

And for mother wavelet is used function w:

$$w(t,f) = \frac{f}{\sqrt{2\pi}} e^{\frac{-t^2 f^2}{2}} e^{-2\pi i f t} \tag{6}$$

The final form of ST combined (4-6) can be written as [60]:

$$\mathrm{S}(\tau,f) = \int_{-\infty}^{\infty} \mathrm{x(t)} \frac{f}{\sqrt{2\pi}} e^{\frac{-t^2 f^2}{2}} e^{-2\pi i f t}\, dt \tag{7}$$

In [61] the analysis of voltage disturbance with WT and STFT methods is discussed. From studies and examples presented in this paper, advantages and disadvantages of WT and STFT are described. Both methods are able to detect the transient of disturbance. STFT is better for time-frequency analysis of disturbances, while WT presents better results for detecting events. Both methods are very similar and they showed similar results. STFT, as alternative to FFT, differs from FFT because it uses a window function w[n-m], and this window translate in time by m samples. STFT can be defined as a sum, as presented in (8).

$$\mathrm{Fs}_{\mathrm{i,j}} = \sum_{n=0}^{N} \mathrm{x[n]} e^{-\frac{2j\pi k n}{N}} \mathrm{w[n-m]} \tag{8}$$

where w[n-m] is window function and x[n] is n$^{\text{th}}$ digital sample of voltage signal.

Research which also compares PQ analysis capabilities of WT and STFT is presented in [52]. Conclusion of this research is similar to the conclusion presented in [61], i.e. STFT is more suitable for disturbance signal analysis, while WT obtained better results for detection of disturbances. In [62], different PQ VDDaA methods are presented and compared. Between RMS, STFT and high pass filter, STFT showed the best results. In [37], a comparative study of RMS, DFT, EKF and WT for detection and analysis of voltage disturbances is presented. In this paper, it is concluded that STFT and RMS methods in all tested cases have delay in detection, EKF shows good results and WT shows the best results in the detection and analysis. However, WT must be used with other method in order to differentiate voltage disturbances from frequency disturbances. In [63] comparison of KF, WT and FFT for voltage dip parameters estimation is addressed. The methods are tested with different signals, including signal with noise, phase angle jump, etc. In this paper, it is concluded that WT is prone to noise and other disturbances with higher frequency components, and that KF and FFT performances are acceptable and satisfy mitigation requirements. Also, it is concluded that the RMS shows the worst results in comparison. In [64] two methods for voltage dip detection are tested as part of grid-tie inverter system. Reduced FFT (RFFT) method shows better results in comparison with FFT, both in speed of detection and in complexity.

HHT is signal analysis method, which consists of two-part transformation, the empirical mode decomposition (EMD) and Hilbert transformation. The HHT of the signal in time domain calculates also real valued time domain signal $\overline{x(t)}$. This two values can form analytical signal: $z(t) = x(t) + j\overline{x(t)}$, where x(t) is original signal. Transformation can be written as [22]:

$$\overline{x(t)} = \int_{-\infty}^{\infty} \frac{x(\tau)}{\pi(t-\tau)} \tag{9}$$

The amplitude signal and instantaneous phase angle $\theta(t)$ and frequency f_0 can be written as (10-12):

$$A(t) = [x^2(t) + \overline{x^2(t)}]^{1/2} \tag{10}$$

$$\theta(t) = \frac{1}{\tan(\overline{x(t)}/x(t))} \tag{11}$$

$$f_0 = \frac{1}{2\pi t}\frac{1}{\tan(\overline{x(t)}/x(t))} \tag{12}$$

HHT is often used as part of algorithms for PQ detection and classification. In [65] application of HHT in wind power systems for voltage dips detection is presented. It is shown that HHT can successfully detect a dip with good detection times, very accurately, but only voltage dips are examined, in simulations, and further examination of this method as stand-alone is needed. In [66] method based on HHT and Symbolic Aggregate appro Ximation (SAX) is proposed for analysis and identification of sudden changes in waveform. The method is tested for general sudden changes and non-stationary signals, to identify frequency amplitude and phase angle. Tests for any type of real PQ disturbance for detection, identification or analysis are not performed in the paper. In [67] HHT method is used for detection, analysis and classification, only with addition of fuzzy rules in classification part of method. Both detection and analysis of single and multiple disturbances are tested. It is stated that HHT can extract from disturbance signal instantaneous amplitude, frequency and phase. Also, many features of disturbances can be calculated from this data set.

From presented literature review, it can be concluded that results of specific methods depend on their application. However, some methods present better results in most of the applications, while others always underperform. Methods based on WT, FFT, HHT and ST in most cases show at least good results, while RMS usually shows the worst results. HHT and ST are mainly used as part of PQ classification algorithms.

3.2.2 Parametric Methods

PMs are the second most important group of DSP based methods. This class of methods use model of signal to perform analysis. Appropriate model is chosen based on knowledge about signals properties. If the model has good matching with the signal, this type of method can achieve high accuracy [32]. Otherwise, if signal is not properly modelled, PM methods can induce significant error.

In [68] AutoRegressive model (AR) is applied for VDDaA. In this paper, it is shown that AR models can be used for detection transitions and potential for tracking time behaviour of dominant frequency, and that this method can be used for event analysing, but that further studies are needed. In [69] detection of voltage disturbances in noisy signals is addressed using AR model in combination with sequential generalized local likelihood ratio detector. In presented simulations, superior performance of proposed method is observed.

In [70] performances of Adaptive Linear Neurone (ADALINE) based method is compared with RMS, WT and HPF for detection of voltage dips. ADALINE is an adaptive filter that is usually used for extraction of waveform features from signals and for reducing noise. In this paper, it is concluded that problem of ADALINE method, as well as AR and ARMA methods, is classification of disturbances. Also, problem for these methods is determination of threshold value that is used for detection. For WT, it is concluded that WT is suitable for the detection of PQ disturbances, but analysis of disturbance is sensitive to noise.

Kalman filter is a method that shows good results in voltage disturbance detection and analysis. This method has good accuracy in amplitude estimation, phase and frequency estimation for application in analysis of disturbance [71]. Method that uses three KF for detection of voltage events and to estimate single-event characteristics is presented in [72]. Results of method using real-grid measurements, applied in real-time environment shows that method is suitable for detection of voltage disturbances, with much faster detection in comparison with RMS ½. Results for precision and reliability of method are not presented.

The Estimation of signal parameters via rotation invariance technique (ESPRIT) and the Multiple Signal Classification (MUSIC) method can be applied for stationary signals analysis [32]. These methods can be further upgraded to work with sliding-window processing methods or as block-based processing methods and can be used to analyse non-stationary signal, but this requires further research.

It can be concluded that PM methods are suitable mainly for analysis of disturbances. Also, these methods can be good choice for offline processing where a delay is required or for improving reliability of classification [38].

3.2.3 Hybrid Methods

HM are mainly methods that cannot be classified as previously addressed groups and do not have implemented some of the AI algorithms. Current classification of DSP (or just signal processing) methods known in literature [32, 38] is further upgraded in [73]. HM can be divided in two sub-groups: Mixed and innovative methods and Recursive methods.

A new method for detection and analysis of VDs which is a combination of WT and sliding-window is presented in [74]. WT is used for detection and good results, even for noisy signals, are obtained. However, in this paper, accuracy and reliability of method, as well as the exact time delay of detection are not summarized. Another example of combining methods to achieve better results is presented in [75]. This method is proposed for harmonics estimation in power systems, and shows good results in online tracking of dynamic changes that can be very useful for voltage disturbance analysis. The method combines Least Square (LS) with ADALINE algorithm, to decompose analysis into a linear and a nonlinear part. It shows better performance in comparison with EKF method for tracking harmonics in normal and noisy conditions. Method is not tested for detection of PQ disturbances, but can be very useful for PQ analysis.

Methods that are based on different use of well-known methods are presented in [76, 77]. Most of the VDDaA methods that are based on WT apply detail coefficient of the highest frequency band for disturbance detection. In [77] method based on improved WT is proposed. This method utilizes two different mother wavelets (db2 and db8). Comparison of proposed method with EPLL and FFT is presented. Very good detection times are obtained. However, despite hybrid structure, high frequency noise can deteriorate abilities of proposed algorithm.

ADALINE is an adaptive filter that can be used in extracting signals from noisy environments, in model identification and in linearization of nonlinear problems [78]. In ADALINE is used with AI methods for VDDaA [78], as part of control algorithm of Shunt active power filter [79] and for dynamic phasor estimation [80] and promising results are obtained. In [70] comparison of RMS, ADALINE, AR, ARMA, HPF and WT are presented. ADALINE and RMS detection do not have required precision. WT is suitable for detection of PQ events and reduction of noise enchased performances. However, much higher complexity of AR and ARMA is not justified with only slight improvements in results.

WT is combined with KF to achieve better performances in [81]. Fuzzy-expert system (FES) is used only for classification. Accuracy over 90% is achieved. The method has the ability to detect and successfully classify different disturbances with relatively low computational complexity. Method that overcomes some known problems of the KF, extended KF (EKF) is applied for detection and classification of voltage disturbances in [82]. EKF method showed good accuracy, but requires all input data for modelling to be known. In [83] hybrid method that includes EKF and ST for detection and analysis of short duration disturbances is addressed. Based on simulation and laboratory research, it is concluded that ST alone can detect and localize disturbances, while KF can successfully extract important parameters of fault. Combined, these two methods show good results in both detection and analysis of disturbances. In [84] the design principles of EKF are presented, together with experimental results and implementation. Based on experimental results of extracting voltage disturbance parameters during transient, it is concluded that estimation includes error and that distortion is present in extracted signal. It is stated that cost of implementation is high because algorithm is highly iterative and needs a fast microprocessor for calculation. However, today's micro-processors can support calculation of EKF with ease.

4 Comparative Study

4.1 Comparison-based on Overview of Research Papers

Comparative study of VDDaA is carried out on the basis of critical overview of a large number of findings and conclusions presented in previously published papers. The results of comparative study are presented in Tables 1 and 2. Ten different and the most frequently applied DSP methods have been taken into consideration: RMS, s-RMS, FFT, WT, KF, STFT, ST, HHT and EKF. The methods are commented and rated according to three here defined criteria: 1. Speed of detection (SoD), 2. Analysis capability (AnC), and 3. Computational demands/cost of implementation (CDi). SoDis time delay between occurrence of disturbance in the grid and its successful detection with tested method. AnC examines method's potential to precisely extract and calculate parameters of a fault that are needed for successful characterization or classification of a VD, and to successfully detect disturbance. CDi is the parameter of a method that defined its complexity, i.e.it can be considered an amount of microprocessor power (time) that must be reserved for implementation of some method, in some hardware unit (e.g. grid-tie inverter control unit or PQ monitoring device).

The SoD, AnC and CDi are rated with numbers from 1 to 10, where 1 is the worse and 10 is the best, based on results that are presented in literature. As an averaged value, a parameter named averaged Total result (TR_a) is introduced and defined with (13). Further on, the three presented criteria are weighted, according to their importance and presented as another new parameter, the weighted Total result (TR_w). In this paper, the SoD and AnC are weighted with coefficients of 0.4, while CDi is weighted with 0.2, like it is shown in (14).

$$TR_a = (SoD + AnC + CDi)/3 \tag{13}$$

$$TR_w = 0.4 * (SoD + AnC) + 0.2 * CDi \tag{14}$$

Table 1 shows advantages and disadvantages of all addressed methods according to the reports in available literature. The methods are not rated. Table 2 presents results of comparison of above mentioned methods according to three here defined criteria and averaged and weighted TR are given.

It is important to notice that some researchers use hybrid methods, which typically contain several DSP methods in combination, while others separately address and test each of them. From these results useful information may be obtained, both about each DSP method and of a whole hybrid algorithm.

The Table 2 shows that WT and HHT methods in the most cases achieve the best overall result. Methods that utilized EKF and ST and STFT follow them as the second best. After these three groups, other popular DSP methods are ranked from place 4 to 10. Standard DSP methods, based on RMS, are ranked with the lowest overall result.

Table 1
Advantages and disadvantages of DSP methods

	Advantages	**Disadvantages**
RMS [36][37][63]	Very simple, standard solution.	Underperforms in comparison with any other method.
s-RMS [36][26]	Improved version of RMS.	Better results than RMS, overall underperforms.
FFT [37][50][51] [63]	Well known. Standard solution for harmonics analysis.	Have problems with analysis of transients.
WT [37][61][47, 48] [53–57]	Very fast SoD. Better for analysis of transients that FFT.	Low reliability, prone to noises. Noise (harmonics) in signal can deteriorate performances significantly.
KF [63][90][72] [81]	Good amplitude and frequency estimation capability even in noisy condition, acceptable SoD and AnC.	More complicated than FFT and similar results of SoD.
STFT [51][52][61] [62]	Good harmonics estimation, useful for voltage disturbance analysis (better than WT), good detection abilities.	Induces a significant delay in detection. Limited performance for analysis of short duration disturbances.
ST [30][58][59] [83][91]	Works better in noisy conditions than other FT based methods.	Results in real-time environment are not good. Because it is based on WT, due to harmonics estimation has error.
HHT [49][65][66] [92]	Good results in noisy conditions, very good AC. Good time-frequency estimation. More adaptive that WT. Low sensitivity to noise.	Short disturbances transients are difficult to detect and analyse with HHT. Should be further tested with real grid disturbances.
EKF [37][82][83] [84]	Simple, fast SoD. Shows good results both in detection and analysis.	Results for SoD and AnC are good, but for AC much better solutions are proposed. Also, WT have faster SoD.

Table 2
Comparison of DSP methods from literature

	RMS	s-RMS	FFT	WT	KF	STFT	ST	HHT	EKF
SoD (1-10)	2	3	4	10	7	7	8	8	8
AnC (1-10)	2	2	4	5	5	6	6	7	6
CDi (1-10)	10	9	8	5	7	7	5	5	6
TRa (1-10)	4.67	4.7	5.33	6.67	6.4	6.67	6.3	6.67	6.67
TRw(1-10)	**3.6**	**3.8**	**4.8**	**7**	**6.2**	**6.6**	**6.6**	**7**	**6.8**

4.2 Comparison based on Real Grid Measurements and Laboratory Evaluation

Based on authors previous research [12, 13, 50, 64, 85–88], comprehensive testing with real grid measurements and with grid-emulator in laboratory were done in order to further evaluate presented methods in the same conditions. Out of 680 recordings in real grids, 127 contain some type of voltage dips or interruptions or other disturbances. From these 127 faults, 10 were selected for testing. In selected sample of 10 faults, various types of dips and interruptions are present. Some of them are very interesting, like multiple disturbances and multi-level faults with developing and changing types.

Rating using AnC criteria is based on ability to detect all disturbances in multiple-events (ME), and on ability to extract key features from all disturbances. Estimated key features must enable proper recognition and classification of each stage of ME. The MOV is magnitude of voltage, which represents minimum value of voltage RMS (calculated with RMS ½-cycle) during disturbance, according to [34]. Results of detection time and AnC grade for ten recorded signals of voltage disturbance (dips) based on real grid measurements are presented in Table 3.

The RMS and the s-RMS can obtain only single-event characteristics (duration and magnitude). Because of that, in terms of analyzing they are usually graded with 4 (AnC). The RMS based methods successfully detect start and the end (if it is recorded) of every disturbance, and obtain magnitude. With average detection time of 19.05 ms and median of 16.7 ms, the RMS ½-cycle is the slowest. Estimated magnitude contains less variation in comparison to magnitudes obtained with other methods (s-RMS, FFT, KF and EKF). The s-RMS with average detection time of 12.51 ms and median of only 7.15 ms is much better and it does not lag considerably in comparison to more complex methods.

The FFT successfully detected all tested events, and obtained enough information from voltage signal from the most of disturbances, so multiple events can be successful classified. Some information are not extracted precisely, like phase angle in some cases. With average detection time of 11.77 ms, median of 6.31 ms and considerably good feature extraction, the FFT presents a method that is in the middle of the list by performance. The AnC grade is 7 and reliability is 100%.

The WT detected six out of ten tested faults with reliability of 70%. Such result may be explained by speed of voltage dip amplitude change. The WT cannot detect slowly developing disturbances that have low transient changes despite that signal has low noise level and even using energy of wavelets. But, for more severe disturbances, the WT performs remarkably well, with average detection time of only 4.22 ms and median of 4ms, which makes it the fastest method. Also, the WT enables successful classification of a disturbance, even if it is complex one. Because of low reliability AnC grade is 6, but SoD grade is 10.

The KF and EKF are applied in a similar way, using fundamental harmonic for a model. The EKF is more complex and better in dynamic state estimation, as it is modified version of linear KF. The EKF´s average detection time is 8.08 ms with median of 4.76 ms. Only for one shallow dip, the EKF underperform with 28.1 ms. The KF average detection time is 11.74 ms with median 6.1 ms. Both methods have reliability of 100%, with AnC grade of 7.8.SoD grades are 6 and 8 for KF and EKF, respectively.

All voltage dips detection methods are tested in laboratory conditions, as well, using voltage dips which have been generated by a grid-emulator. The detection methods were applied as part of grid-tie inverter control. Primary task was to observe the methods' behaviour in real-time systems, measure computational complexity in real-time environment, and compare methods from viewpoint of ease-of-use. Fig. 3 presents overall look of such laboratory setup. It consists of advanced hardware in the field of electrical drives and of the control units based on highly modular dSpace control hardware and modified industrial converters [89]. The system is paired with AC grid emulator GE 15-AC and connected using Yd transformer to the supply. Computational complexity is measured on dSpace, which utilizes DS1006 processor board (AMD Opteron™ processor). System is set to works at a PWM frequency of 6.4 kHz and generates a synchronized software interrupt with a 3.2 kHz frequency.

Table 3

Real grid method testing – detection times, reliability and analysis capability

Description of disturbances / No.	MOV [%]	Detection time [ms] / AnC [1-10]					
		RMS	s-RMS	FFT	WT	KF	EKF
#1. Type G, five cycles, develops into Type A	**79**	22.6/4	19.7/4	19/7	$4.4^{e}/8$	19.5/7	10.3/7
#2. Non-fault interruption	**5.5**	35.3/4	6.29/5	4.6/10	7.34/10	4.4/10	3.77/10
#3. Type C, 15 cycles	**87**	39.7/4	40.4/4	39/7	/	39.9/7	28.1/7
#4. Balanced dip with unbalanced recovery	**48**	9.5/5	7.35/5	6.3/8	$3.5^{e1}/9$	6.1/8	4.1/8
#5. Remarkable multiple event *	**59**	7.2/4	6.6/4	5.9/8	$3^{e1}/9$	5.7/8	3.6/8
#6. Type D dip	**56**	9.19/4	4.15/4	3.7/8	/	3.72/8	3.26/8
#7. Type F, 15 into type A	**84**	28.7/4	21.8/4	21/8	/	21.5/8	13.4/8
#8. Three-phase fault	**47**	16.2/4	6.95/4	6.3/8	$3.1^{e1}/10$	6.11/8	5.7/8
#9. Unbalanced dip (Type C).	**67**	4.9/4	4.3/4	3.6/9	4/8	3.6/9	3.2/9
#10. Single-phase fault with over-voltages (multiple events)	**0.47**	17.2/5	7.53/5	6.91/8	$4.17^{e1}/9$	6.9/8	5.43/8
Average detection time		**19.05**	**12.51**	**11.77**	**4.22**	**11.74**	**8.08**
Reliability [%]		**100**	**100**	**100**	**70**	**100**	**100**
Mean of AnC [1-10]		**4.2**	**4.3**	**7**	**6**	**7.8**	**7.8**

^{e}Energy of wavelet is used for detection, fault cannot be detected with detail coefficients;^{e1}Energy of wavelet is used for detection, but fault can be detected with detail coefficient/detection is slower; *Starts as type C, slow recovery one phase up, two phases down, repeat of the first event

The testing showed that all methods were successfully applied. Results of digital processor computation times are presented in Table 4. Standard RMS method is one with the lowest execution time, following with the s-RMS, FFT, KF, EKF and WT, with delays of 8%, 57%, 61%, 71%, and 146%, respectively.

In Fig. 4 graphical presentation of DSP VDDaA method results are shown. In Fig. 5 examples of signal processing with 3 tested methods are given. Signal of disturbance fault #10 from Table 3 is analysed with different algorithms: RMS, FFT and WT. Outputs of these algorithms are presented. Fig. 5a presents voltage signals recorded in real grid, Fig. 5b shows time representation obtained with the

RMS, Fig. 5c shows frequency representation obtained with the FFT and Fig. 5d shows time-frequency representation derived using the WT.

Based on SoD, AnC and CDi presented in Tables 3 and 4, a comparison of all tested methods is given in Table 5. Also, TR_a and TR_w are calculated. Ratings for SoD were presented in a way that the best method obtained rating 10, as shorter time presents better result. It can be seen that the WT based method that analyzes voltage signals is the best in term of detection speed, but has lower CDi and reliability problems.EKF shows the best overall results of 7.72, while WT, KF and FFT follows (7.4, 7.12 and 6.8 respectively). RMS based methods underperforms.

Conclusion

A comprehensive and critical review on methods for detection and analysis of voltage disturbances, based on DSP methods is presented. The major advantages and disadvantages are outlined, as well as the comparison of the wide range of methods in the DSP domain.

Based on comprehensive laboratory and real grid measurement signals testing, it can be concluded that EKF and WT have the best overall grades. Also, the FFT and KF can be distinguished as the ones with high detection capabilities. On the other hand, the RMS based method underperforms. However, it should be noted that each of these methods has its own advantages and drawbacks, and selection should be done based on specific application and priorities. Based on presented comprehensive literature review, it can be concluded that the DSP techniques can be successfully used for VDDaA in modern distribution grids.

Signals with significant amount of noise are challenge even for advanced methods, and detection and analysis methods can underperform due to noise in signals. Also, complex multiple disturbances, or very distant disturbances that cause shallow dips may be challenging, as well.

Table 4

Microprocessor execution time of a voltage detection method

	RMS	s-RMS	FFT	WT	KF	EKF
Laboratory execution time on dSpace [μs]	7.1	7.67	11.2	17.5	11.5	12.2

Table 5

Comparison of DSP methods based on comprehensive evaluation

	RMS	s-RMS	FFT	WT	KF	EKF
SoD (1-10)	3	6	6	10	6	8
AnC (1-10)	4.2	4.3	7	6	7.8	7.8
CDi (1-10)	10	10	8	5	8	7
TR_a (1-10)	**5.73**	**6.77**	**7**	**7**	**7.27**	**7.6**
TR_w(1-10)	**4.88**	**6.12**	**6.8**	**7.4**	**7.12**	**7.72**

Future scope

Despite large amount research results in the field of VDDaA, several challenges remain. There is still space to find better method in terms of the detection and analysis performances, and to optimize it in term of computational complexity using artificial intelligence techniques. Also, improvements and additional research in finding a method that has the ability to provide good results in noisy conditions and in analysing events with multiple disturbances are needed.

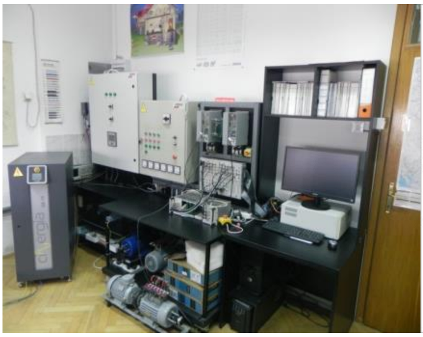


Fig. 3 Outlook of the laboratory setup.

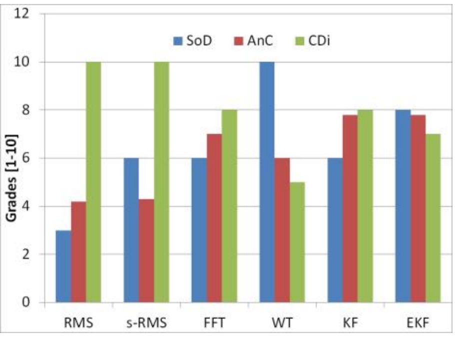


Fig. 4 Grafic presentation of DSP methods results

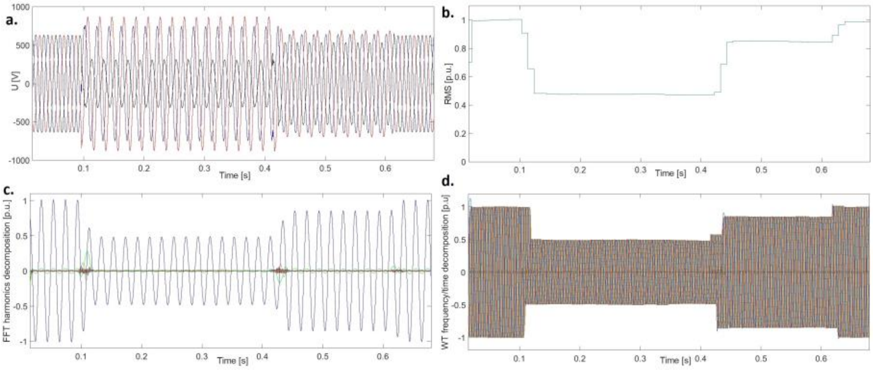


Fig. 5a Voltage signal for fault #10 (Table 3) processing with: b. RMS, c. FFT harmonics decomposition, d. WT frequency/time decomposition

Acknowledgement

This work was supported by the Republic of Serbia, Ministry of Education, Science and Technological Development, project No. III 042004 entitled “Smart Electricity Distribution Grids Based on Distribution Management System and Distributed Generation“.

The authors would like to express their acknowledgement to Prof. Math Bollen who has provided in-field measurement results.

References

[1] Bollen, M. H.: Understanding Power Quality Problems, IEEE, 1999

[2] Bae, Y. et al.: Implemental control strategy for grid stabilization of grid-connected PV system based on German grid code in symmetrical low-to-medium voltage network, IEEE Tran. En. Conv., 28(3), 2013, pp. 619-631

[3] Yang, Y. et al.: Benchmarking of Grid Fault Modes in Single-Phase Grid-Connected Photovoltaic Systems, IEEE Trans. Ind. Appl., 49 (5), 2013, pp. 2167-2176

[4] Yang, Y., Blaabjerg, F.: Low-voltage ride-through capability of a single-stage single-phase photovoltaic system connected to the low-voltage grid, Int. J. Photoenergy, 2013, pp. 1-9

[5] Bhavaraju, V. B., Enjeti, P.: A Fast Active Power Filter to Correct Line Voltage Sags, IEEE Trans. Ind. Electron., 41 (3), 1994, pp. 333-338

[6] Akagi, H.: New trends in active filters for power conditioning, IEEE Trans. Ind. Appl., 32 (6), 1996, pp. 1312-1322

[7] Nielsen, J. G., Blaabjerg, F.: A detailed comparison of system topologies for dynamic voltage restorers, IEEE Tran. Ind. Appl., 41(5), 2005, pp. 1272-1280

[8] Rufer, A. et al.: Power quality compensation using universal power quality conditioning system, IEEE Power Eng. Rev., 20 (12), 2000, pp. 58-60

[9] Kwan, K. H. et al.: An output regulation-based unified power quality conditioner with Kalman filters, IEEE Trans. Ind. Electron., 59 (11), 2012, pp. 4248-4262

[10] Zhao, W. et al.: Microgrid relay protection scheme based on harmonic footprint of short-circuit fault, Chinese J. Elec. Eng., 4(4), 2018, pp. 64-70

[11] Afshari, E. et al.: Control Strategy for Three-Phase Grid-Connected PV Inverters Enabling Current Limitation under Unbalanced Faults, IEEE Trans. Ind. Electron., 64 (11), 2017, pp. 8908-8918

[12] Katic, V. A., Stanisavljevic, A. M.: Smart Detection of Voltage Dips Using Voltage Harmonics Footprint, IEEE Trans. Ind. Appl., 54 (5), 2018, pp. 5331-5342

[13] Katic, V. A., Stanisavljevic, A. M.: Novel voltage dip detection algorithm using harmonics in the dip’s transient stage, Proc. IECON 2017 - 43rd Annu. Conf. IEEE Ind. Electron. Soc., 2017, pp. 351-356

[14] Granados-Lieberman, D. et al.: Techniques and methodologies for power quality analysis and disturbances classification in power systems: a review, IET Gener. Transm. Distrib., 5 (4), 2011, p. 519

[15] Abdelsalam, A. A. et al.: Classification of power system disturbances using linear Kalman filter and fuzzy-expert system, Int. J. Electr. Power Energy Syst., 43 (1), 2012, pp. 688-695

[16] Borges, F. A. S. et al.: Feature Extraction and Power Quality Disturbances Classification Using Smart Meters Signals, IEEE Trans. Ind. Informatics, 12 (2), 2016, pp. 824-833

[17] Blackledge, J.: Digital Signal Processing (2ndEd.), Horwood Publ., 2006

[18] Antoniou, A.: Digital Signal Processing: Signals, Systems and Filters, McGraw-Hill, 2016

[19] Haidegger, T. et al.: Simulation and control for telerobots in space medicine, Acta Astronaut., 81 (1), 2012, pp. 390-402

[20] Spall, J. C.: Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation, IEEE Trans. Automat. Contr., 37 (3), 1992, pp. 332-341

[21] Ürmös, A. et al.: Application of Self-Organizing Maps for Technological Support of Droplet Epitaxy, Acta Polytechnica Hungarica, 14 (4), 2017, pp. 207-224

[22] Khokhar, S. et al.: A comprehensive overview on signal processing and artificial intelligence techniques applications in classification of power quality disturbances, Renew. Sust. Energy Rev., 51, 2015, pp. 1650-1663

[23] Hosseini, S. A. et al.: An overview of microgrid protection methods and the factors involved, Renew. Sust. Energy Rev., 64, 2016, pp. 174-186

[24] Saini, M. K., Kapoor, R.: Classification of power quality events - A review, Int. J. Electr. Power Energy Syst., 43 (1), 2012, pp. 11-19

[25] Mahela, O. P. et al.: A critical review of detection and classification of power quality events, Renew. Sust. Energy Rev., 41, 2015, pp. 495-505

[26] Gururajapathy, S. S. et al.: Fault location and detection techniques in power distribution systems with distributed generation: A review, Renew. Sust. Energy Rev., 74, 2017, pp. 949-958

[27] Khokhar, S. et al.: Automatic Classification of Power Quality Disturbances : A Review, IEEE Student Conf. Res. Dev., 2013, pp. 16-17

[28] Barros, J. et al.: Review of signal processing techniques for detection of transient disturbances in voltage supply systems, IEEE Instrum. Meas. Technol. Conf., 2013, pp. 450-455

[29] Prakash M. O., Gafoor S. A.: Topological aspects of power quality improvement techniques: A comprehensive overview, Renew. Sust. Energy Rev., 58, 2016, pp. 1129-1142

[30] Mishra, S. et al.: Detection and Classification of Power Quality Disturbances Using S-Transform and Probabilistic Neural Network, IEEE Trans. Power Deliv., 23 (1), 2008, pp. 280-287

[31] Hooshmand, R., Enshaee, A.: Detection and classification of single and combined power quality disturbances using fuzzy systems oriented by particle swarm optimization algorithm, Electr. Power Syst. Res., 80 (12), 2010, pp. 1552-1561

[32] Bollen, M. H. J., Gu, I. Y. H.: Signal processing of power quality disturbances, New York: Press, Series on Power Eng., 2006

[33] Institute of Electrical and Electronics Engineers: 1159-2009 - IEEE Recommended Practice for Monitoring Electric Power Quality, 2009

[34] Institute of Electrical and Electronics Engineers: IEEE Std 1564 - Guide for Voltage Sag Indices, 2014

[35] International Electrotechnical Commission: Electromagnetic compatibility (EMC): IEC 61000-4-30 Edition 3.0 2015-02, 2015

[36] Albu, M., Heydt, G. T.: On the use of RMS values in power quality assessment, IEEE Trans. Power Deliv., 18 (4), 2003, pp. 1586-1587

[37] Perez, E., Barros, J.: Voltage Event Detection and Characterization Methods: A Comparative Study, IEEE/PES Transm. Distrib. Conf. Exp. Lat. Am., 2006, pp. 1-6

[38] Jain, S. K., Singh, S. N.: Harmonics estimation in emerging power system: Key issues and challenges, Electr. Power Syst. Res., 81 (9), 2011, pp. 1754-1766

[39] Huang, N. E. et al.: The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non- Stationary Time Series Analysis, Proc. R. Soc. London A, 454 (1971), 1998, pp. 903-995

[40] Zhao, F., Yang, R.: Power-quality disturbance recognition using S-transform, IEEE Trans. Power Deliv., 22 (2), 2007, pp. 944-950

[41] Aiello, M. et al.: Synchronization techniques for power quality instruments. IEEE Trans. Instrum. Meas., 56 (5), 2007, pp. 1511-1519

[42] Belega, D., Petri, D.: Frequency estimation by two- or three-point interpolated Fourier algorithms based on cosine windows, Signal Processing, 117, 2015, pp. 115-125

[43] Chintakindi, S. R. et al.: Improved Hanning window based interpolated FFT for power harmonic analysis, IEEE TENCON, 2016, pp. 1-5

[44] Wen, H. et al.: Harmonic Estimation Using Symmetrical Interpolation FFT Based on Triangular Self-Convolution Window, IEEE Trans. Ind. Informatics, 11 (1), 2015, pp. 16-26

[45] Belega, D. et al.: Iterative sine-wave frequency estimation by generalized Fourier interpolation algorithms, 11th Int. Symp. Electron. Telecommun. ISETC 2014 - Conf. Proc., 2014, pp. 1-4

[46] Van Der Byl, A., Inggs, M. R.: Recursive sliding discrete Fourier transform with oversampled data, Digit. Sig. Proc. A Rev. J., 25(1), 2014, pp. 275-279

[47] Thirumala, K. et al.: Estimation of single-phase and three-phase power-quality indices using empirical wavelet transform, IEEE Trans. Power Deliv., 30 (1), 2015, pp. 445-454

[48] Poisson, O. et al.: Detection and measurement of power quality disturbances using wavelet transform, IEEE Trans. Power Deliv., 15 (3), 2000, pp. 1039-1044

[49] Yang, L. et al.: Disturbance source identification of voltage sags based on Hilbert-Huang transform, APPEEC, 2010, pp.1-4

[50] Stanisavljevic, A. M. et al.: Voltage dips detection in a system with grid-tie inverter, Proc. of 18th EPE 2016 ECCE Europe, 2016, pp. 1-10

[51] Ingale, R.: Harmonic Analysis Using FFT and STFT, Int. J. Signal Process. Image Process. Pattern Recognit., 7 (4), 2014, pp. 345-362

[52] Jurado, F., Saenz, J. R.: Comparison between discrete STFT and wavelets for the analysis of power quality events, Electr. Power Syst. Res., 62 (3), 2002, pp. 183-190

[53] Santoso, S. et al.: Power quality assessment via wavelet transform analysis, IEEE Trans. Power Deliv., 11 (2), 1996, pp. 924-930

[54] Lin, C.-H., Tsao, M.-C.: Power quality detection with classification enhancible wavelet-probabilistic network in a power system, IEE Proc. - Gener. Transm. Distrib., 152 (6), 2005, pp. 969-976

[55] Barros, J. et al.: Applications of wavelets in electric power quality: Voltage events, Electr. Power Syst. Res., 88, 2012, pp. 130-136

[56] Kezunovic, M., Liao, Y.: A novel software implementation concept for power quality study, IEEE Trans. Power Deliv., 17 (2), 2002, pp. 544-549

[57] Perez, E., Barros, J.: A proposal for on-line detection and classification of voltage events in power systems, IEEE Trans. Power Deliv., 23 (4), 2008, pp. 2132-2138

[58] He, S. et al.: A real-time power quality disturbances classification using hybrid method based on s-transform and dynamics, IEEE Trans. Instrum. Meas., 62 (9), 2013, pp. 2465-2475

[59] Ray, P. K. et al.: Islanding and Power Quality Disturbance Detection in Grid-Connected Hybrid Power System Using Wavelet and S-Transform, IEEE Trans. Smart Grid, 3 (3), 2012, pp. 1082-1094

[60] Dash, P. et al.: Power quality analysis using s-transform, IEEE Trans. Power Deliv., 18 (2), 2003, pp. 406-411

[61] Gu, Y., Bollen, M. H. J.: Time-frequency and time-scale domain analysis of voltage disturbances, IEEE Tran. Power Deliv., 15(4), 2000, pp.1279-1284

[62] Ingale, R., Tawade, L.: Detection and Comparison of Power Quality Disturbances using Different Techniques, Int. J. Comput. Appl., 75 (18), 2013, pp. 48-53

[63] Amarís, H. et al.: Computation of voltage sag initiation with Fourier based algorithm, Kalman filter and Wavelets, 2009 IEEE Bucharest PowerTechInnov. Ideas TowardElectr. Grid Futur., 2009, pp. 1-6

[64] Stanisavljevic, A. M. et al.: Wavelet transform for voltage dips detection in a microgrid with distributed generation, Proc. of 19th EPE'17 ECCE Europe, 2017, pp. 1-10

[65] Li, Y. et al.: Study on Voltage Sag Detection of Wind Power System Based on HHT, Energy and Power Engineering, 05 (04), 2013, pp. 922-926

[66] Afroni, M. J. et al.: Analysis of nonstationary power-quality waveforms using iterative Hilbert Huang transform and sax algorithm, IEEE Trans. Power Deliv., 28 (4), 2013, pp. 2134-2144

[67] Das, D. et al.: Hilbert huang transform with fuzzy rules for feature selection and classification of power quality disturbances, Proc. of 4th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Comput. Electron., 2017, pp. 439-445

[68] Gu, I. Y. H. et al.: The use of time-varying AR models for the characterization of voltage disturbances, Proc. of IEEE Power Eng. Soc. Conf., 2000, pp. 2943-2948

[69] Li, S., Wang, X.: Cooperative Change Detection for Voltage Quality Monitoring in Smart Grids, IEEE Trans. Inf. Forensics Secur., 11 (1), 2016, pp. 86-99

[70] Chang, G. W., Cheng-I Chen: Performance evaluation of voltage sag detection methods, Proc. of IEEE PES General Meeting, 2010, pp. 1-6

[71] Moreno Saiz, V. M., Barros Guadalupe, J.: Application of Kalman filtering for continuous real-time tracking of power system harmonics, IEE Proc. - Gener. Transm. Distrib., 144 (1), 1997, p. 13

[72] Barros, J. Perez, E.: Automatic Detection and Analysis of Voltage Events in Power Systems, IEEE Trans. Inst. Meas., 55 (5), 2006, pp. 1487-1493

[73] Stanisavljevic, A. M. et al.: Overview of voltage dips detection analysis methods. Proc. of19thIntern. Symposium on Power Elec., 2017, pp. 1-6

[74] De Apráiz, M. et al.: A real-time method for time-frequency detection of transient disturbances in voltage supply systems, Electr. Power Syst. Res., 108, 2014, pp. 103-112

[75] Joorabian, M. et al.: Harmonic estimation in a power system using a novel hybrid Least Squares-Adaline algorithm. Electr. Power Syst. Res., 79 (1), 2009, pp. 107-116

[76] Costa, F. B. et al.: Assessment of Voltage Sag Indices Based on Scaling and Wavelet Coefficient Energy Analysis,IEEE Trans. on Power Delivery, 28 (1), 2013, pp. 336-346

[77] Latran, M. B., Teke, A.: A novel wavelet transform based voltage sag/swell detection algorithm, Int. J. Elec. Power Ener. Syst., 71, 2015, pp.131-139

[78] Valtierra-Rodriguez, M. et al.: Detection and classification of single and combined power quality disturbances using neural networks, IEEE Trans. Ind. Electron., 61 (5), 2014, pp. 2473-2482

[79] Martinek, R. et al.: An Efficient Control Method of Shunt Active Power Filter Using ADALINE, IFAC-PapersOnLine, 49 (25), 2016, pp. 352-357

[80] Nanda, S., Dash, P. K.: A Gauss-Newton ADALINE for dynamic phasor estimation of power signals and its FPGA implementation, IEEE Trans. Instrum. Meas., 67 (1), 2018, pp. 45-56

[81] Abdelsalam, A. A. et al.: Characterization of power quality disturbances using hybrid technique of linear Kalman filter and fuzzy-expert system, Electr. Power Syst. Res., 83 (1), 2012, pp. 41-50

[82] Ghahremani, E., Kamwa, I.: Dynamic state estimation in power system by applying the Extended Kalman filter with unknown inputs to phasor measurements, IEEE Trans. Power Syst., 26 (4), 2011, pp. 2556-2566

[83] Dash, P. K., Chilukuri, M. V.: Hybrid S-transform and Kalman filtering approach for detection and measurement of short duration disturbances in power networks, IEEE Trans.Instrum. Meas., 53 (2), 2004, pp. 588-596

[84] Routray, A. et al.: A novel Kalman filter for frequency estimation of distorted signals in power systems, IEEE Trans. Instrum. Meas., 51 (3), 2002, pp. 469-479

[85] Katic, V. A. et al.: Comparison of voltage dips detection techniques in microgrids with high level of distributed generation, Proc. of 17th IEEE Intern. Conf. on Smart Technologies, EUROCON, 2017, pp. 1-6

[86] Stanisavljević. A. M. et al.: Voltage dips detection using Kalman filter in a microgrid with high level of distributed generation, Proc. of 20th EPE 2018 ECCE Europe, 2018, pp. 1-10

[87] Stanisavljević, A. M. et al.: Reduced FFT algorithm for network voltage disturbances detection, Proc. of Int. Sym. Ind. Elect. INDEL, 2016, pp. 1-6

[88] Stanisavljevic, A. M. et al.: Voltage dips detection in a microgrid with distributed generation for grid-tie inverter protection purposes. Proc. of 19th EPE 2017 ECCE Europe, 2017, pp. 1-10

[89] Dumnic, B. et al.: Advanced laboratory setup for control of electrical drives as an educational and developmental tool, Proc. of 15th IEEE Intern. Conf. EUROCON, 2013, pp.903-909

[90] Moreno Saiz, V. M., Barros Guadalupe, J.: Application of Kalman filtering for continuous real-time tracking of power system harmonics, IEE Proc. - Gener. Transm. Distrib., 144 (1), 1997, pp. 13-21

[91] Mohanty, S. R. et al.: Classification of disturbances in hybrid DG system using modular PNN and SVM, Int. J. Electr. Power Energy Syst., 44 (1), 2013, pp. 764-777

[92] Das, D. et al.: Hilbert Huang transform with fuzzy rules for feature selection and classification of power quality disturbances, Proc. of 4th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Comput. Electron., 2017, pp. 439-445

Innovation in Healthcare Performance among Private Brand's Healthcare Services in Small and Medium-sized Enterprises (SMEs)

Rohaizat Baharun[1], Tock Jing Mi[2], Dalia Streimikiene[3], Abbas Mardani[4], Jawaria Shakeel[5, 6], Vitalii Nitsenko[7]

[1] Azman Hashim International Business School, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia, E-mail: m-rohaizat@utm.my
[2] Azman Hashim International Business School, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
[3] Vilnius University, Kaunas Faculty, Muitines 8, Kaunas, LT-42280, Lithuania, E-mail: dalia.streimikiene@khf.vu.lt
[4] Azman Hashim International Business School, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia, E-mail: mabbas3@live.utm.my
[5] Azman Hashim International Business School, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia, E-mail: sjawaria@graduate.utm.my
[6] COMSATS University Islamabad, 5400, Lahore Campus, Pakistan, E-mail: jawaria@cuilahore.edu.pk
[7] Department of Accounting, Analysis and Audit, Odessa I. I. Mechnikov National University, Odessa, Ukraine

Abstract: Small and medium-sized enterprises (SMEs) in Malaysia are rapidly expanding their businesses; they use international diversification as an imperative strategic route to attain growth. Due to the great potential of SME in a developing market and the importance of branding that induces the performance of a company, there is a need for more research to explore branding dimensions. This research is mainly aimed at empirically examining the interrelated relationships that exist among various SME branding constructs (i.e., brand orientation, brand trust, brand equity, innovation, SME performance, marketing, and financial performance) and testing whether the proposed SME branding dimensions model efficiently helps us to understand the role the SME branding plays in growth and success of a firm. The research adopts a combination of qualitative and quantitative approaches, known as mixed model method. A sample containing 67 items of data was collected from a healthcare center, involving the owner, health specialists, and customers. This study aims to contribute to new marketing knowledge on the area of healthcare branding in SMEs, and identify the branding dimensions that affect the financial and marketing performance of small and medium sized enterprises. The results of this study found that a good brand orientation is the most important component that contributes to the healthcare performance.

Keywords: Brand equity; Brand trust; Innovation; Healthcare; SME's Performance

1 Introduction

Brand management has been mentioned extensively in the marketing literature since decades ago and its strategic importance to a company has been well recognized [15, 39, 47]. Past studies have suggested that companies that direct their managerial activities and practices in the direction of the expansion, procurement, and leveraging of branded products and services are in a better position to improve their performance [33, 57]. Though, previous researchers have linked several branding determinants with the company performance in SMEs [11, 22, 25, 52, 64, 69]. Those studies are usually focused on SME in the manufacturing field, neglecting the application of branding in services industry, e.g., the healthcare services. Only a small number of researchers in the field of branding have focused on smaller units such as dental, clinics, and maternity centers. Additionally, according to Barbis [9], the literature available on the healthcare topic is very limited. From a larger perspective, brand literature has mainly focused on international large-scale industries only. Hence. Neglecting the enterprises of small and medium size [5, 14, 25]. Another limitation is that the majority of companies studied have been from the western and eastern developed countries only (see Inter-brand). Thus, plenty of work is required for a neglected country like Malaysia. The SMEs branding studies in Malaysia suffers from a lack of consensus, since there are several different streams that are contradictory to each other and have little, or nothing, that links branding, SME, and performance together [6].

One of the most powerful and important asset that each company needs to have is brand equity [1]. Based on the study conducted by Piaralal and Mei [58], building brand equity in healthcare sector is not an easy task. However, it should be delivered consistently within the center because it ensures quality assurance; this is what most customers seek when it comes to wellness services. On the other side, Schindehutte et al. [62] stated that the illustration of innovation is about reconfiguring, realigning, and renewing the marketing activities within a planned progression and development in spite of dramatic transformation. Innovation is one of the driving forces in defining effective strategy for SMEs. Innovativeness helps companies to see the significance of implementing branding [61] as an essential instrument to be well adopted to innovative services that meet consumer demands; this is because a strong brand gives credibility and security [1]. However, the stricter national policies on healthcare branding have put additional pressures on privately owned centers such as clinics and pharmacies (see The Medicine (Advertisement and Sale) Act 1956 and Malaysian Health Promotion Board Act 2006). For example, according to The Medicine (Advertisement and Sale) Act 1956, Section 3 to Section 4A, all advertisement related to medicines, diseases, skills, and services are prohibited. No one person is allowed to be involved in publishing any advertisement referring to any medicine, an appliance, or a remedy except those published by the Federal or State Government. The major gap to the marketing literature is limited branding studies on small and medium sized enterprises, which is considered comparatively new in Malaysia's [6] healthcare centers. In fact, Malaysia generally lacks research on branding, which is due to the managers' and owners' ignorance of branding practice. As literature shows, most of the studies in this field are conducted by Western scholars [14, 25, 42, 52].

This study is aimed at providing empirical evidence in the related area by developing a model to examine the link between brand trust and brand orientation as independent variables and performance of the company as a dependent variable with intervening role of brand equity and the moderating role of innovation in the context of small and medium sized enterprises in Malaysia. The present study addresses the above argument with an effort to increase the understanding of healthcare branding context among SME in Malaysia. With regard to the literature, this study makes two important contributions:

(1) To new marketing knowledge on the area of healthcare branding in SMEs. This research demonstrates how ideas about branding are translated and communicated from the perspective of SME healthcare owners.

(2) To identification of the branding dimensions that contribute to the financial and marketing performance of small and medium sized enterprises (more specifically, healthcare centers). To date, this research will be a pioneering study in Malaysia in measuring the healthcare branding and performance in SME.

2 Literature Review

2.1 Hypotheses Development

To further understand the theories related to the framework, we can use major theories such as theory of Resource-Based View (RBV) and Diffusion of Innovation (process innovation) to explain the construct validity of SME Branding for this research. The resource-based view argues that firms possess resources that help to gain competitive advantage, which leads to a superior long-term performance [10, 35]. Many researchers have argued that differences may occur in many forms of resources such as innovations and patents [19]. However, Klein [41] stated that such differences may also formed by subjective judgments that imagined by entrepreneurs. In various marketing fields, RBV is mainly used to compel the structure it suggests in the integration of diverse resources in a way to make clear the differential, synergistic effects on performance and the contingencies that are interrelated [27]. The same is applied to this study's framework that involves multiple dimensions structure such as branding effects on the organization performance [29, 30]. Three main variables such as brand trust, brand orientation, and brand equity selected for this study were found effective resources in building brands internally. On the other hand, RBV in marketing strategy indicates that performance is measured considering several indicators, e.g., market share [34], profitability [68], and return on investments [51]. To the other side, Medical and technological innovation adoption in healthcare differs when made by owners or individuals. Once an owner decides to use a device or piece of technology, he or she must consider the impact not only on the patient and the practice, but also on the performance of the company. The value created from the innovation adoption must be evaluated. An example of a great diffusion of innovation is the adoption of X-ray. Seelor and Mair [63] proposed a framework of adopting innovations which entered an organization through diffusion process in order to guide evaluation of factors influencing organization capacity in continuous innovation for social sector organizations such as the healthcare centers. In short, the diffusion of innovations theory is a useful systemic framework to describe how well small and medium sized enterprises implement and capture innovation culture in the strategic management. If SMEs contribute to productivity in developing market offerings, their competencies can result in an economic dynamism. These offerings can lead to durable and beneficial market positions, which can bring about greater financial performance for SMEs.

2.1.1 Linking Brand Equity and Brand Trust

To achieve customer loyalty in the context of brand building, one of the most important components, which needs to be taken well into account, is the concept of trust [7, 17, 25]. There are two-dimensional ideas of trust, which are commonly found in the management and marketing literature [23, 24, 28, 54]. On the other side, brand equity also creates value for both the customer and the company [8]. In addition, its incremental utility and value is endowed to a product or service by the brand name [40, 49, 69, 71]. Attributes such as provocativeness, risk-taking and innovation portray an entrepreneurial mindset [20, 43]; they are able to identify the opportunities in market and exploit them through combination or recombination of the resources obtainable by the owner's venture [21, 37, 65].Moreover, according to Mohamed and Daud [55], no study has examined the firms' values such as brand trust and brand equity in one particular construct. Therefore, there is a significant gap that should be filled in order to gain knowledge about trust and equity relationship.

H1: Brand trust affects brand equity in SME

2.1.2 Linking Brand Orientation Affects Brand Equity

Literature characterize brand orientation by brand dominance of incorporate strategic thinking and a relatively consistent, constant, consumer-relevant branding strategy that can be plainly distinguished from competition [11, 33]. Mzungu et al., [48] conducted research to measure the first stage of safeguarding the brand equity. That is to adopt brand orientation. It is one of the significant ingredients to manage the brand strategically for all types of companies even SMEs. Within the first stage, there are three key propositions adopted for building brand orientation. First, creating a brand orientation mind set which helps the organization to create a sustainable competitive advantage for the brand. The second step involves clearly defining the brand in terms of its purpose, vision, values, competencies, and aspirations. The last step of building brand orientation mind set is communicating the brand because defining the brand without communicating it within the organization will open to multiple interpretations at its various touch points. The literature has shown that brand orientation has a powerful impact on brand equity [12, 48, 61] and also influences the performance of a company [11, 32]. Brand orientation is suitable to be tested as part of this research because of its robust and dynamic interaction with brand management and it is rarely being examined within healthcare industry specifically. Thus, the following hypothesis is needed to be proposed.

H2: Brand orientation affects brand equity in SME

2.1.3 Link between Brand Equity and SME Performance

Brand equity has been a link between customer and firm in the past [64]. However, to numerous managers and researchers, it has been attractive to measure the return of intangible assets, e.g., brand equity [64] to the company. Realizing the immense standing of brand equity in a firm performance, there is a need to execute brand equity to increase the value of a company. Berthon et al. [14] strongly agree that owners or managers are required to monitor brand equity. In a healthcare point of view, it is important to employ available healthcare marketing resources and programs in an improved way in order to gain a greater influence within the community. Understanding brand equity is a critical starting point for planning marketing strategy and tracking progress toward goals [31]. The brand management research is primarily aimed at exploring the actual value of these intangible assets and applying that information concretely to the improvement of the firm's standing and perception. Thus, the third and fourth hypotheses are in two aspects:

H3a: Brand equity affects SME financial performance

H3b: Brand equity affects SME marketing performance

2.1.4 Brand Equity Mediates the Relationship between Brand Trust, Brand Orientation, and SME Performance

In a study conducted by Yoo et al. [69], the framework was conceptualized based on the extension of [2]model. Three main propositions on brand equity were derived. First, brand equity creates value for both the customer and the firm. Second, the value for the customer enhances value for the firm, and finally, brand equity consists of multiple dimensions. Yoo et al. [69]extended the model by placing into a separate construct, brand equity dimensions, and value for the customer and the firm. Brand equity acted as a mediator between brand assets and its consequences. The theory proves that brand equity can be created, maintained, and expanded by strengthening its dimensions. Yoo et al. [69]emphasized on brand equity linkages and noted a very imperative future research issue, namely the interaction effects and consequences of brand equity. Although there are some studies suggesting that brand equity acts as intermediate variable [4, 64], none has measured branding in such an approach amongst small and medium sized healthcare centers. It provides directions for owners of healthcare centers in terms of creating and enhancing brand trust and brand orientation through brand equity in a way to lead to sustainable marketing and financial performance. This has resulted in the fourth hypothesis as follows.

H4a: Brand equity mediates the relationship between brand trust and financial performance

H4b: Brand equity mediates the relationship between brand trust and marketing performance

H4c: Brand equity mediates the relationship between brand orientation and financial performance

H4d: Brand equity mediates the relationship between brand trust and marketing performance

2.1.5 Innovation Moderates the Relationship between Brand Equity and SME Performance

Literature consists of numerous studies carried out on performance of SMEs plus their financial and marketing issues and their drivers [50, 56, 59] because it is significant to evaluate in a different manner in a way to be adapted with customer's constantly evolving needs and preferences [36, 46, 66]. The main contribution of the study conducted by Merrilees et al., [52] to the social sciences is the evaluation of auxiliary SME capabilities as determinants of marketing performance. The notion explaining the marketing performance lies in two main marketing variables: branding and innovation. In our study, innovation is included in the framework to moderate the relationship between branding and SME performance (marketing and financial). Moreover, findings of Li et al. [45] showed that innovation moderated the relationship between market orientation and performance. In addition, innovation for this study's framework has been built further as a moderator by relating it to the business performance in the presence of a branding plan. Only one study has been conducted so far adopting this approach [52] but it was constructed in a developed country raising a query if the model works for an emerging market. Accordingly, the conceptual framework in this study was modified to examine how the relative contribution of mechanism works in a developing market. Therefore, the fifth hypothesis was constructed as follow, *H5a: Innovation moderates the relationship between brand equity and financial performance*

H5b: Innovation moderates the relationship between brand equity and marketing performance

2.2 Development of the Brand Dimension Model

The conceptual framework is extended in two ways. First, we are setting brand equity in a joint construct illustrated in Figure 1: brand equity with its dimensions of brand trust and brand orientation and the influence of brand equity on the performance of the healthcare center. Secondly, innovation has positive impact on SME performance in several studies (e.g. [44] and [60]) and supported by theory of diffusion of innovation.

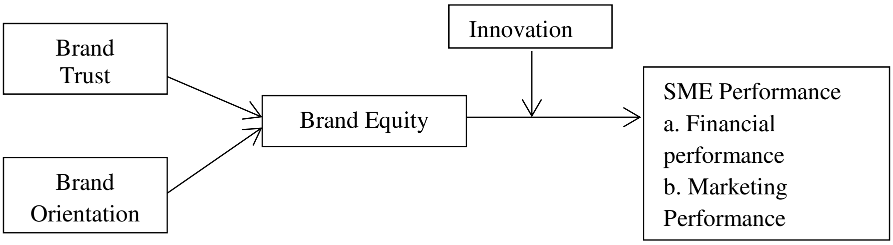


Figure 1
Conceptual Framework of SME Branding Dimension

3 Research Method

3.1 Variables and Measurement

This study requires measurement of brand trust, brand orientation, brand equity, and SME performance. We made use of a Likert scale to score all measures; the options were ranged between 1 (‘do not agree at all’) and 5 (‘strongly agree’). In this scale, higher scores showed a higher level of the construct in question. Pilot testing for both qualitative and quantitative data was done in order to avoid vagueness or confusing questions. Validity of qualitative and quantitative data was examined through the reflective measurement model assessment. The key criteria for this study were indicator reliability, composite reliability, and convergent validity. Furthermore, discriminant validity was achieved. Every reflective construct had to share more variance with its own indicators compared to other constructs in the path model [26]. The constructs were deemed appropriate for PLS-SEM analyses in case all these requirements criteria were met.

3.2 Sampling

The samples from this study are collected from healthcare services provided in Johor Bahru, Malaysia, mainly from the developed area around Iskandar Malaysia Township or Nusajaya. Johor Bahru is strategically located near many other medical hubs such as Singapore and Indonesia. The health care development in Iskandar, Malaysia aims to capture patients who are from around these regions and seeking quality and cost-effective healthcare services. It is aimed at becoming the next medical destination. Therefore, Iskandar Malaysia in Johor Bahruis known as a billion-dollar industry projected to grow. It is the economic potential that has led the Malaysian government to consider the healthcare sector as one of the country's 12 National Key Economic Areas (ETP Annual Report, 2014).

The sampling procedure for both qualitative and quantitative parts was done using non-probability sampling. The reasons for choosing non-probability sampling were (i) first, this research cannot meet the criteria of probability sampling; most of the experts, doctors, and pharmacist in the health industry decline to cooperate. (ii) Second, the procedure of selecting respondents to be included in the sample is much easier, quicker, and cheaper.

3.2.1 Qualitative Sampling

For qualitative part of the study, a snow ball sampling procedure was done. As respondents from health industry are hard to reach, the potential subjects were selected based on recommendation and identification by the initial subject who also meets the criteria of the research. In this study, one of the branding experts from SME Corporation was chosen followed by the Chairman of Malaysia Medical Association. The determination of sample size follows Becker, Bryman et al. [13] theory where the observation stops when no new theoretical insights are being gleaned from the data.

3.2.2 Quantitative Sampling

The sampling of the SME entrepreneurs, owners, or managers was initialized using probability sampling based on the lists provided by Syarikat-Syarikat Suruhanjaya Malaysia (SSM) and followed by a quota sampling method that is a non-probability sampling method. There are four town councils administered by the local authorities in Johor Bahru, the largest population in Johor state, as explained in Figure 2. About 400 questionnaires were distributed to potential respondents of the four main Johor Bahru regions namely: City council of Johor Bahru, City council of Central Johor Bahru, City council of Pasir Gudang and Nusajaya. Initially, 200 questionnaires were distributed in the first phase to all health care-related providers from clinics, pharmacies, dental clinics, and maternity centers in Johor Bahru. A total of 57 completed questionnaires were received. Of that number, the study then extended again to gather even more respondents. In the second phase, 200 questionnaires were distributed, and we could manage to gather 67 completed questionnaires (Figure 2).

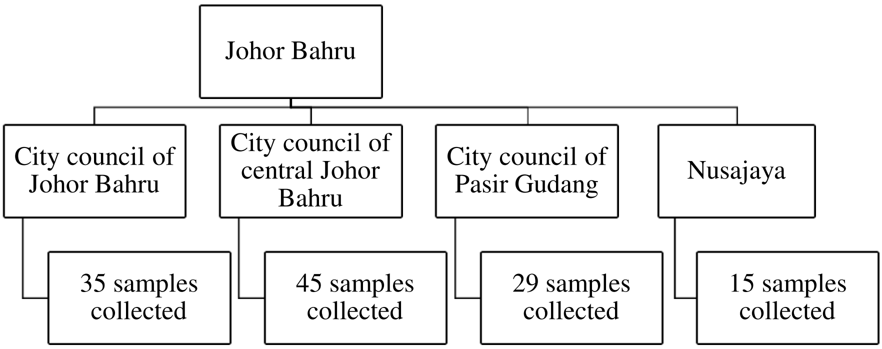


Figure 2

Cluster – Systematic (Proportions) Sampling Area Distribution and Total Samples Collected

4 Data Analysis

In-depth interviews were conducted as the preliminary stage for qualitative research followed by a distribution of survey questionnaires as the quantitative part of the study. All collected qualitative data were tabulated, coded, retrieved, summarized, drawn, and verified. To calculate, Burn and Bush [16] formula was adopted, assuming that there was a great expected variability (50%) and for ±10 percent accuracy at the 95 percent level of confidence, and sufficient sample size for data collection was at least 96. Descriptive analysis was done to describe the basic features of the data in the study. Bootstrapping, blindfolding, CTA-PLS, analysis of moderating effect, and multi-group analysis were tested and analyzed. Consequently, a comprehensive evaluation was verified using reflective and formative measurement. Finally, PLS-SEM analyses for moderating and mediating and importance-performance analyses were described as closing.

5 Results

Five interviewees were selected from different nature of businesses such as SMEs, Health center and pharmacy. The respondents were selected carefully by taking into account their experience in the industry. The main focus of the interview was mainly to answer the aspects of branding dimension and the performance of a company.

4.1 Analysis of Qualitative Data

The analysis of all branding dimensions were analyzed in the interview and divided into categories: general branding categories, brand trust, brand orientation, brand equity, innovation financial and marketing performance. The first part of the interview on general branding inquiries showed that most of the respondents realized the importance of branding for company performance. Trust is evaluated among the healthcare owners based on few indications of a good brand trust; reliability, credibility, competitive advantage, partnering with reputable associates, customer recognition, values and keeping promises. Brand dimension was claimed to be understood by most of the respondents. Nevertheless, when probed further, only two out of the five respondents were able to describe the mechanism of brand dimension. Three out of five participants were clear about the identity of the health care centers. Most of the owners were still unclear of their organization image. Even though four out of five respondents stated that they recognized brand as a valuable asset and strategic resource for development; only a few actually understood the brand values. Three respondents, which included managers and owners, stated that the development of brand in the health care center is the responsibility of every employee and there is in fact a good communication in regard to branding within the organization. However, throughout the interviews, only one owner stated that the health care center uses all marketing activities to develop a brand. Surprisingly, none of the respondents specified that an active and effective management is essential for achieving competitive advantage. From the insights gained from the interview regarding brand trust, four out of five respondents indicated that reliability, competitive advantage, and credibility were important for brand trust. Furthermore, innovation is fundamental for health care providers. Four out of five respondents stated that new ideas and new services must be constantly introduced to the company to keep updated with the competitors. Lastly, the health center performance is based on two factors: financial performance and marketing performance. All of the interviewees believed that financial performance is measured by how profitable the business is. Moreover, the return of investment and how well the company reaches financial goals are also two main measurements used to quantify the financial performance.

4.2 Analysis of Quantitative Data

This section describes the results of descriptive analysis of demographic variable. The data showed that most of the respondents were male (75.8%), while female owners were only 24.2% of the total respondents. Most of the respondents were Chinese with a total of 49.2 percent and the rest of them were Indian (30.6%) and Malay (20.2%). Most of the small and medium sized enterprises were run by the owners (80.6%) themselves and only few were supervised by a manager (19.4%). Most of the owners of the small and medium sized enterprises were not given financial assistance for brand building activities. Only 12.9 percent was funded before the business started.

According to the demographic data, 41.1% of the health centers claimed that branding needs to be considered before setting up the organization. Almost half of the respondents thought that branding is a strategic point of view for company management in which the rewards of having a strong brand is become advantage to them. Most of the companies have an annual average growth between 0 to 20%. Despite having no financial assistance, these health centers were able to grow in business performance annually. Surprisingly, about 26.6% of the health centers have experienced no growth.

4.3 Measurement Model

The reflective measurement model was evaluated in terms of both validity and reliability. For each construct in this research, the reflective scale items were considered to be sufficient and appropriate to represent the construct domain. The overall measurement model and the indicator loadings of the final set of items were used for independent and dependent measurements. From the measures, eight indicators showed factor loading below 0.7. In case of the six reflectively-measured constructs, the composite reliabilities were ranged between 0.913 and 0.924, which is much greater than the minimum requirement of 0.70. Each latent variable AVE (Average Variance Extracted) was checked regarding the convergent validity. All AVE values were shown greater than the threshold of 0.5, hence indicating convergent validity for all constructs. As obviously shown by the criterion, in case of the reflective constructs, all AVE values were greater than the squared inter construct correlation indicating that the discriminant validity was well established.

4.4 Structural Model

A highly recommended approach is PLS-SEM path weighting since it is capable of providing the greatest R^2 value for endogenous latent variables and also it can be generally applied to all estimations and specifications of the PLS path model. However, before interpreting the path coefficients, a test was conducted on the structural model regarding its collinearity. It was of a high importance since the path coefficients estimation was on the basis of the ordinary least squares regressions [53]. VIF values of the analyses found were in ranged between 1.062 and 1.568; this ensured that the structural model results were not negatively affected by collinearity. The calculation process of the PLS results was iterated for 300 times, which is large enough for data analysis [26]. When checking the PLS-SEM result, the algorithm needs to be checked to be terminated because of the stop criterion. This value should be sufficiently small, i.e., 10^-5. In the structural model analysis, the last step is about the relevance and significance of the structural model relationships. As demonstrated by the results obtained from the bootstrapping procedure, seven out of eight structural relationships were proved significant at ($p \leq 0.05$), see Figure 3. The inner model suggested that brand orientation had the strongest effect on brand equity with path coefficient of 0.679 in this research. Statistics showed that there was a significant relationship between brand equity and brand orientation. Surprisingly, the path coefficient between brand trust and brand equity is not significant at only 0.075. This is because the standardized path coefficient is lower than 0.1. The hypothesized path relationship between brand trust with financial performance (0.251) and marketing performance (0.239) were rather weak but significant. To conclude, brand orientation is a strong predictor for brand equity, while brand trust is a weak predictor for brand equity. The results showed that all relationships are significant with t value ≥ 1.96 except for one relationship where brand trust does not affect brand equity. At 90% level of confidence, t value is 1.066 (see Table 1).

Table 1
Path Significance in Bootstrapping

Hypothes is	Paths	Path Co.	SD	T Statistics	P Values	Result
H1	Brand Trust -> Brand Equity	0.075	0.071	1.066***	0.293	rejected
H2	Brand Orientation -> Brand Equity	0.579	0.062	9.413**	0.000	Accepted
H3a	Brand Equity -> Financial Performance	0.545	0.094	5.800**	0.000	Accepted
H3b	Brand Equity -> Marketing Performance	0.679	0.075	9.001**	0.000	Accepted

4.5 Mediation Analysis

After determination of the valid path coefficients, mediation analysis was examined. The analysis was based on the study of Hair et al. [26] where three main ideas listed by F. Hair Jr, Sarstedt, Hopkins and G. Kuppelwieser [26] were fulfilled. To satisfy the mediation effect rules of thumb, brand trust must have a relationship with brand equity. Looking at Table 1, we can see that the path coefficient of brand trust and brand equity is not significant. Therefore, brand equity does not mediate the relationship between brand trust with financial and marketing performance. In the final step, the strength of mediation was tested using variance accounted for (VAF). This final analysis step resulted in VAF values that were smaller than 1 as stated in Table 2). Based on findings of Hair et al. [63], this value indicates that the relationship is mediated.

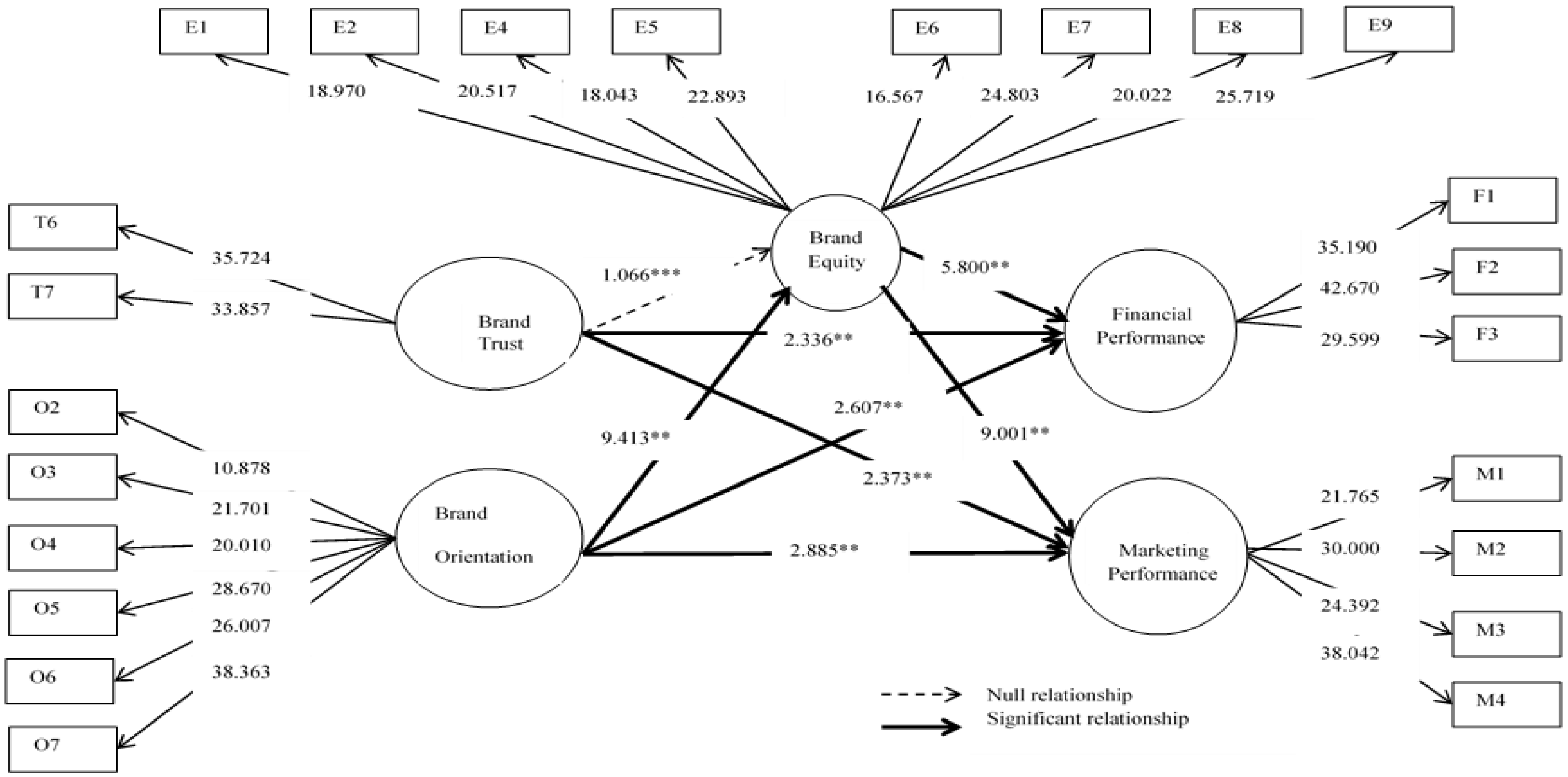


Figure 3

Bootstrap result

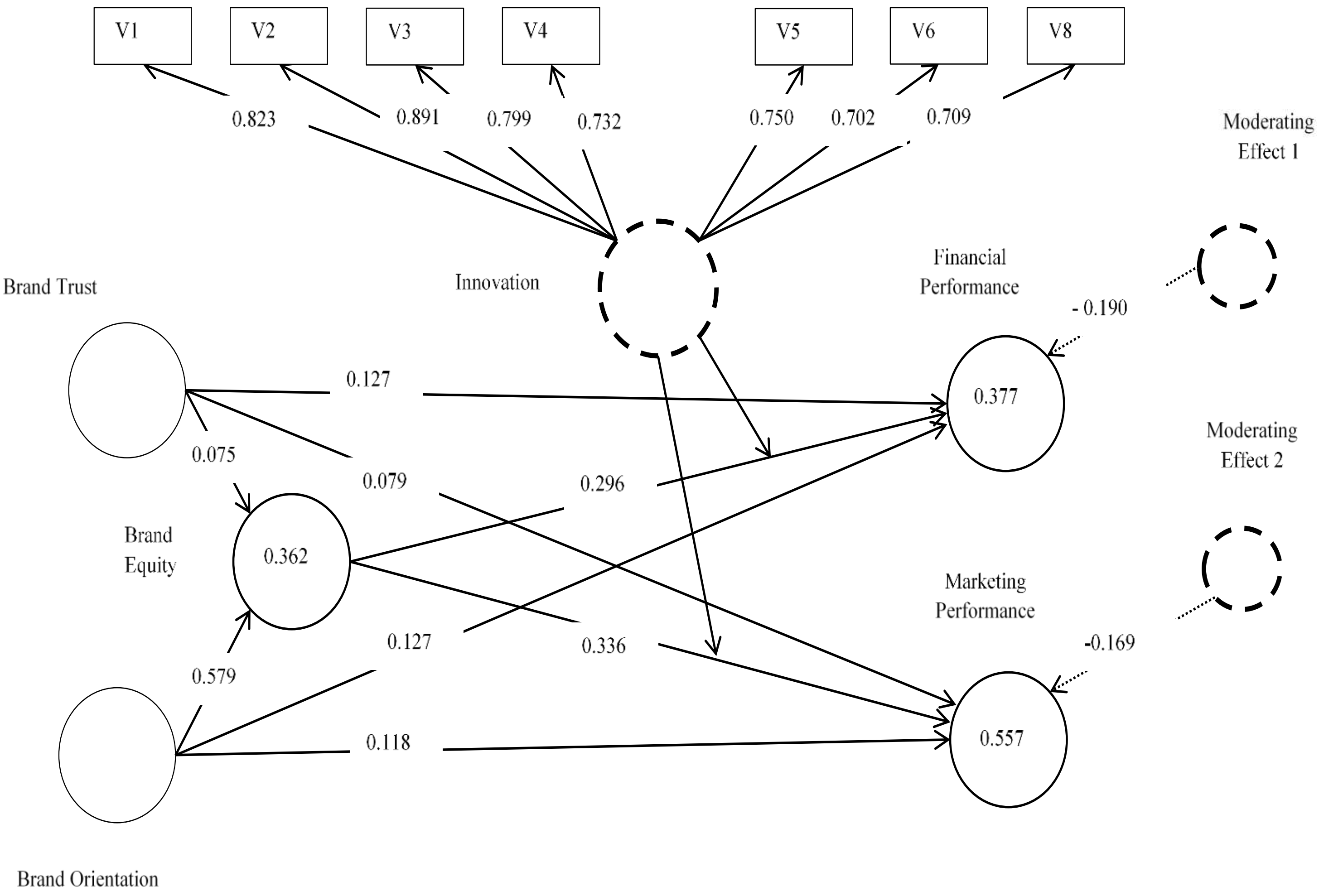


Figure 4

Moderator Model in SMART PLS

Brand orientation and marketing performance indicated the highest VAF value of 0.610, which showed that about 61.0% of the total effect of brand orientation on marketing performance was explained by indirect effect. Brand orientation and financial performance achieved a VAF value of 0.569. Thus, the two relationships were mediated by brand equity (see Table 2). These relationships showed a partial mediation effect. According to Hair *et al.*, (2014), a situation in which the VAF is larger than 20% and less than 80% can be characterized as partial mediation. Anything higher than 80% is considered full mediation.

Table 2

Mediation Value of Direct Effect, Indirect Effect, Total Effect and Variance Accounted For (VAF)

	Direct effect ***(c)***	**Indirect effect** ***(axb)***	**Total effect** ***(c)*** **+** ***(axb)***	**VAF**
brand orientation → brand equity (*a*) brand equity → marketing performance *(b)* brand orientation → marketing performance *(c)*	0.251	0.393	0.644	0.610
brand orientation → brand equity (*a*) brand equity → financial performance *(b)* brand orientation → financial performance *(c)*	0.239	0.316	0.555	0.569

4.6 Moderation Analysis

To conduct the significance test, bootstrapping procedure with 500 bootstrap samples using no sign changes option was used. The analysis yields a *t* value of 0.717 for path linking the interaction term and financial performance. Similarly, for marketing performance in Figure 4, interaction effect was at *t* value = 0.794. According to Hair *et al.*, (2014), there is no significant moderating effect of innovation on the relationship between brand equity and financial performance and marketing performance where the analysis yields less than *t* = 1.96 (see Table 3).

Table 3

Summary of moderating effect

	Predictive Value R^2	**Effect Size** f^2	**Relationship**
Moderating Effect 1 (Financial Performance)	0.023	0.001	Non-significant
Moderating Effect 2 (Marketing Performance)	0.010	0.000	Non-significant

6 Discussions

Multiple studies have shown that brand trust affects brand equity [22, 55]. However, the results of this study showed no significant relationship between brand trust and brand equity as shown in figure 4. Ultimate justification can be the fact that trust is not evaluated by one person alone because trust comes to exist at the highest level between the consumer and the brand consumed, where the emotional investment is made between the two parties. A health center does not typically move forward to the identity or consistency level in establishing a trusted relationship with a certain brand until the brand effectively proves its capability of living up to expectations. Respondents to the qualitative part of the research reflected the importance of brand orientation in the health care business. According to literature, recognizing brand as a valuable asset and strategic resource has to be continuously developed and protected in the best possible way [32]. The outcomes were also consistent with Gromark and Melin [32] findings where indicating that with a proper brand orientation, a company will benefit in terms of profits. Therefore, the development of healthcare brand is not the responsibility of only a small group within the company, but everyone in the company is responsible [12, 32].

Results showed that brand equity was not a mediator between brand trust and any one the marketing or financial performance; thus, hypotheses 4a and 4b were rejected. The reason behind this is that findings showed no significant relationship between brand trust and brand equity opposing to Ballester and Aleman [22] study indicating the significance of brand trust in the development of brand equity. As mentioned earlier, brand trust cannot determine brand equity of a health care center because the trust relationship is between a health care provider and the customer rather than the firm equity with the customer.

From the moderating analysis, innovation is not a moderator for brand equity and SME performance, therefore, hypothesis H5a and H5b were rejected. It is surprisingly different from the literature arguing innovation influences brand equity, hence affecting the performance of a company [18, 45]. The main cause behind this might be the degree of innovation in the respective health center. According to one study conducted by Zhang et al. [70], the degree of innovation has a positive effect on both brand equity and value equity, which eventually impacts a company's performance. Thus, the reason a health care center launch innovative products or services must be not only boosting the sales through enhancing the customer value. Additionally, it also improves the brand image, which affects the organization performance.

The SME business success measurement using financial and marketing performance was supported by Hooley et al. [34]; Vorhies and Morgan [68]; Merrilees et al. [52]. All items loadings were found to be above 0.8; therefore, items were reliable for interpretation. The results of marketing and financial performance, similar to performance measurement and evaluation systems, are important to both owners and managers. Moreover, brand equity was found to have relationship with financial performance. Hypothesis 3a was accepted and also supported by past studies where financial performance was reported to be greatly affected by brand equity [3, 18, 67, 69]. The wisdom of brand equity concept even for services industry like health care is found to prevail. The results imply that the health care providers need to build and safeguard brand equity to have a better performance. They must design appropriate marketing activities to have their brand internally and externally known by customers and employees, hence increasing the sales revenue.

Supported by Merrilees et al., (2011), the performance measurement falls into two categories: financial and marketing performance. These two measurements have shown different significance level toward brand equity. According to findings, brand equity showed a higher significant value to marketing performance compared to financial performance. Hypothesis 3b was accepted; it follows Keller [38] definition where brand equity is defined in terms of the marketing effects uniquely attributable to the brand.

Conclusion

This research has highlighted the need for a greater appreciation of the importance and relevance of brand orientation, brand equity, and RBV and how they can affect the business performance of small firms. In addition, the model was tested using an advance technique via partial least square of structural equation modeling (SEM) and resulted in a strong empirical support. It is proposed that the new SME branding dimension is appropriate to conceptualize the SME branding situation in Johor Bahru, Malaysia. The small sample size was the main issue for this study. This research was limited to a number of private health care centers selected from the four-main district of Johor Bahru and responses were collected from one owner of each center. The limitation of questionnaire was found where there was no way of checking misinterpretations and unintelligible replies by the respondents. Generally, this study had low response rates, due to uncooperative respondents. Future research should focus on implementing that the current model be tested in different regions or countries and different economic status to measure the branding efforts, and the influences in the business performance. The economic effects may portray differently especially for each group, respectively region in North America, Europe and Asia. Moreover, future studies could focus on empirically examining the framework with a large sample size.

References

[1] Aaker, D. A. and Jacobson, R.: The value relevance of brand attitude in high-technology markets, Journal of marketing research, 38 (2001) No. 4, pp. 485-493

[2] Aaker, D. A.: Managing Brand Equity, New York, Maxweel Macmillan-Canada, Inc, (1991)

[3] Aaker, D. A.: Managing brand equity, simon and schuster, (1991)

[4] Aaker, D.: Innovation: Brand it or lose it, California Management Review, 50 (2007) No. 1, pp. 8-24

[5] Aaker, D.: Innovation: Brand it or lose it. California Management Review, 50 (2007) No. 1, pp. 8-24

[6] Abimbola, T. and Kocak, A.: Brand, organization identity and reputation: SMEs as expressive organizations: A resources-based perspective, Qualitative Market Research: An International Journal, 10 (2007) No. 4, pp. 416-430

[7] Ahmad, F. S. and Baharun, R.: A crucial role of entrepreneur in B2B branding: A case from Malaysia, Faculty of Management and Human Resource Development (2010)

[8] Akbar, M. M. and Parvez, N.: Impact of service quality, trust, and customer satisfaction on customers loyalty, ABAC Journal, 29 (2009)

[9] Baldauf, A., Cravens, K. S. and Binder, G.: Performance consequences of brand equity management: evidence from organizations in the value chain, Journal of product & brand management, 12 (2003) No. 4, pp. 220-236

[10] Barbis, D.: Brand model creation for a small healthcare service, (2012)

[11] Barney, J.: Firm resources and sustained competitive advantage, Journal of management, 17 (1991) No. 1, pp. 99-120

[12] Baumgarth, C. and Schmidt, M.: How strong is the business-to-business brand in the workforce? An empirically-tested model of 'internal brand equity'in a business-to-business setting, Industrial Marketing Management, 39 (2010) No. 8, pp. 1250-1260

[13] Baumgarth, C.: Brand orientation of museums: Model and empirical results, International Journal of Arts Management (2009) pp. 30-45

[14] Becker, S., Bryman, A. and Ferguson, H.: Understanding research for social policy and social work: themes, methods and approaches, Policy Press (2012)

[15] Berthon, P., Ewing, M. T. and Napoli, J.: Brand management in small to medium- sized enterprises, Journal of Small Business Management, 46 (2008) No. 2, pp. 27-45

[16] Berthon, P., Hulbert, J. M. and Pitt, L. F.: Brand management prognostications, MIT Sloan Management Review, 40 (1999) No. 2, p. 5

[17] Burns, A. C. and Bush, R. F.: Prentice Hall, Upper Saddle River, New Jersey (2003)

[18] Chaudhuri, A. and Holbrook, M. B.: The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty, Journal of marketing, 65 (2001) No. 2, pp. 81-93

[19] Cheng, C. C. and Krumwiede, D.: The effects of market orientation and service innovation on service industry performance: An empirical study, Operations Management Research, 3 (2010) No. 3-4, pp. 161-171

[20] Chuang, L.-M. and Chao, S.-T.: A Cross-National Comparison of Entrepreneurial Opportunity Recognition: Application of a Self-Organizing Map with a Resource-Based view, INTERNATIONAL RESEARCH JOURNAL OF FINANCE AND ECONOMICS (2013) No. 108, p. 8

[21] Covin, J. G. and Slevin, D. P.: Strategic management of small firms in hostile and benign environments, Strategic management journal, 10 (1989) No. 1, pp. 75-87

[22] Davidsson, P., Delmar, F. and Wiklund, J.:Entrepreneurship as growth: Growth as entrepreneurship, In Entrepreneurship and the Growth of Firms, Edward Elgar Publishing, 2006 , pp. 21-38

[23] Delgado-Ballester, E. and Luis Munuera-Alemán, J.: Does brand trust matter to brand equity?, Journal of product & brand management, 14 (2005) No. 3, pp. 187-196

[24] Delgado-Ballester, E., Munuera-Aleman, J. L. and Yague-Guillen, M. J.: Development and validation of a brand trust scale, International Journal of Market Research, 45 (2003) No. 1, pp. 35-56

[25] Doney, P. M. and Cannon, J. P.: An examination of the nature of trust in buyer-seller relationships, the Journal of Marketing (1997) pp. 35-51

[26] Eggers, F., O'Dwyer, M., Kraus, S., Vallaster, C. and Güldenberg, S.: The impact of brand authenticity on brand trust and SME growth: A CEO perspective, Journal of World Business, 48 (2013) No. 3, pp. 340-348

[27] F. Hair Jr, J., Sarstedt, M., Hopkins, L. and G. Kuppelwieser, V. Partial least squares structural equation modeling (PLS-SEM) An emerging tool in business research, European Business Review, 26 (2014) No. 2, pp. 106-121

[28] Fang, E., Palmatier, R. W. and Grewal, R.: Effects of customer and innovation asset configuration strategies on firm performance, Journal of Marketing Research, 48 (2011) No. 3, pp. 587-602

[29] Ganesan, S.: Determinants of long-term orientation in buyer-seller relationships, the Journal of Marketing (1994) pp. 1-19

[30] Gavurová, B., Kováč, V. and Fedačko, J.: Regional disparities in medical equipment distribution in the Slovak Republic–a platform for a health policy regulatory mechanism. Health economics review, 7 (2017) No. 1, p. 39

[31] Gavurová, B., Kováč, V. and Šoltés, M.: Medical Equipment and Economic Determinants of Its Structure and Regulation in the Slovak Republic,in Encyclopedia of Information Science and Technology, Fourth Edition, (2018) pp. 5841-5852

[32] Gombeski Jr, W. R., Martin, B. and Britt, J.: Marketing-Stimulated Word-of-Mouth: A Channel for Growing Demand, Health marketing quarterly, 32 (2015) No. 3, pp. 289-296

[33] Gromark, J. and Melin, F.: The underlying dimensions of brand orientation and its impact on financial performance, Journal of Brand Management, 18 (2011) No. 6, pp. 394-410

[34] Hankinson, G.: Location branding: A study of the branding practices of 12 English cities, Journal of Brand Management, 9 (2001) No. 2, pp. 127-142

[35] Hooley, G. J., Greenley, G. E., Cadogan, J. W. and Fahy, J.: The performance impact of marketing resources, Journal of business research, 58 (2005) No. 1, pp. 18-27

[36] Hoopes, D. G., Madsen, T. L. and Walker, G.: Guest editors' introduction to the special issue: why is there a resource- based view? Toward a theory of competitive heterogeneity, Strategic Management Journal, 24

[37] Hult, G. T. M., Hurley, R. F. and Knight, G. A.: Innovativeness: Its antecedents and impact on business performance. Industrial marketing management, 33 (2004) No. 5, pp. 429-438

[38] Ireland, R. D., Hitt, M. A., Camp, S. M. and Sexton, D. L.: Integrating entrepreneurship and strategic management actions to create firm wealth, Academy of Management Perspectives, 15 (2001) No. 1, pp. 49-63

[39] Keller, K. L.: Brand synthesis: The multidimensionality of brand knowledge, Journal of consumer research, 29 (2003) No. 4, pp. 595-600

[40] Keller, K. L.: Conceptualizing, measuring, and managing customer-based brand equity, the Journal of Marketing (1993) pp. 1-22

[41] Keller, K. L.: Strategic brand management, Upper Saddle River, NJ, Prentice Hall, 1998

[42] Klein, P. G.: Opportunity discovery, entrepreneurial action, and economic organization, Strategic Entrepreneurship Journal, 2 (2008) No. 3, pp. 175-190

[43] Krake, F. B.: Successful brand management in SMEs: a new theory and practical hints, Journal of Product & Brand Management, 14 (2005) No. 4, pp. 228-238

[44] Kraus, S.: The role of entrepreneurial orientation in service firms: empirical evidence from Austria, The Service Industries Journal 33 (2013) pp. 427-444

[45] Laforet, S.: In Organizational Culture, Business-to-Business Relationships, and Interfirm Networks, Emerald Group Publishing Limited (2010) pp. 341-362

[46] Li, Y., Zhao, Y., Tan, J. and Liu, Y.: Moderating effects of entrepreneurial orientation on market orientation- performance linkage: Evidence from

Chinese small firms, Journal of small business management, 46 (2008) No. 1, pp. 113-133

[47] Lin, C.-H., Peng, C.-H. and Kao, D. T.: The innovativeness effect of market orientation and learning orientation on business performance, International journal of manpower, 29 (2008) No. 8, pp. 752-772

[48] Low, G. S. and Fullerton, R. A.: Brands, brand management, and the brand manager system: A critical-historical evaluation, Journal of marketing research (1994) pp. 173-190

[49] Marinova, S., Cui, J., Marinov, M. and Shiu, E.: Customers relationship and brand equity: A study of bank retailing in China, WBC, poznau, 9 (2011) pp. 6-9

[50] Mavondo, F. and Farrell, M.: Cultural orientation: its relationship with market orientation, innovation and organisational performance, Management Decision, 41 (2003) No. 3, pp. 241-249

[51] Menguc, B. and Auh, S.: Creating a firm-level dynamic capability through capitalizing on market orientation and innovativeness, Journal of the academy of marketing science, 34 (2006) No. 1, pp. 63-73

[52] Merrilees, B., Rundle-Thiele, S. and Lye, A.: Marketing capabilities: Antecedents and implications for B2B SME performance, Industrial Marketing Management, 40 (2011) No. 3, pp. 368-375

[53] Mooi, E. and Sarstedt, M.: A concise guide to market research, chapter 9:"Cluster analysis", Berlin: Springer-Verlag, 10 (2011) pp. 978-973

[54] Morgan, R. M. and Hunt, S. D.: The commitment-trust theory of relationship marketing, The journal of marketing (1994) pp. 20-38

[55] M'zungu, S. D., Merrilees, B. and Miller, D.: Brand management to protect brand equity: A conceptual model, Journal of Brand management, 17 (2010) No. 8, pp. 605-617

[56] Naina Mohamed, R. and Mohd Daud, N.: The impact of religious sensitivity on brand trust, equity and values of fast food industry in Malaysia, Business Strategy Series, 13 (2012) No. 1, pp. 21-30

[57] Nasution, H. N., Mavondo, F. T., Matanda, M. J. and Ndubisi, N. O.: Entrepreneurship: Its relationship with market orientation and learning orientation and as antecedents to innovation and customer value, Industrial marketing management, 40 (2011) No. 3, pp. 336-345

[58] Noble, C. H., Sinha, R. K. and Kumar, A.: Market orientation and alternative strategic orientations: A longitudinal assessment of performance implications, Journal of marketing, 66 (2002) No. 4, pp. 25-39

[59] Piaralal, S. and Mei, T. M.: Determinants of Brand Equity in Private Healthcare Facilities in Klang Valley, Malaysia, American Journal of Economics, 5 (2015) No. 2, pp. 177-182

[60] Rajaguru, R. and Jekanyika Matanda, M.: Influence of inter-organisational integration on business performance: The mediating role of organisational-level supply chain functions, Journal of Enterprise Information Management, 22 (2009) No. 4, pp. 456-467

[61] Rosenbusch, N., Brinckmann, J. and Bausch, A.: Is innovation always beneficial? A meta-analysis of the relationship between innovation and performance in SMEs, Journal of business Venturing, 26 (2011) No. 4, pp. 441-457

[62] Santos-Vijande, M. L., del Río-Lanza, A. B., Suárez-Álvarez, L. and Díaz-Martín, A. M.: The brand management system and service firm competitiveness, Journal of Business Research, 66 (2013) No. 2, pp. 148-157

[63] Schindehutte, M., Morris, M. and Allen, J.: Beyond achievement: Entrepreneurship as extreme experience. Small Business Economics, 27 (2006) No. 4-5, pp. 349-368

[64] Seelos, C. and Mair, J.: Innovation is not the Holy Grail, Stanf Soc Innov Rev, 10 (2012) No. 4, pp. 44-49

[65] Seggie, S. H., Kim, D. and Cavusgil, S. T.: Do supply chain IT alignment and supply chain interfirm system integration impact upon brand equity and firm performance?, Journal of business research, 59 (2006) No. 8, pp. 887-895

[66] Sexton, D. L. and Smilor, R. W.: Entrepreneurship 2000, Kaplan Publishing, (1997)

[67] Theoharakis, V. and Hooley, G.: Customer orientation and innovativeness: Differing roles in New and Old Europe, International Journal of Research in Marketing, 25 (2008) No. 1, pp. 69-79

[68] Vazquez, R., Del Rio, A. B. & Iglesias, V.: Consumer-based brand equity: development and validation of a measurement instrument, Journal of Marketing management, 18 (2002) No. 1-2, pp. 27-48

[69] Vorhies, D. W. and Morgan, N. A.: Benchmarking marketing capabilities for sustainable competitive advantage, Journal of marketing, 69 (2005) No. 1, pp. 80-94

[70] Yoo, B., Donthu, N. and Lee, S.: An examination of selected marketing mix elements and brand equity, Journal of the academy of marketing science, 28 (2000) No. 2, pp. 195-211

[71] Zhang, H., Ko, E. and Lee, E.: Moderating Effects of Nationality and Product Category on the Relationship between Innovation and Customer Equity in K orea and China, Journal of Product Innovation Management, 30 (2013) No. 1, pp. 110-122

[72] Zheng, W., Yang, B. and McLean, G. N.: Linking organizational culture, structure, strategy, and organizational effectiveness: Mediating role of knowledge management, Journal of Business research 63 (2010) No. 7, pp. 763-771

Visemes of Chinese Shaanxi Xi’an Dialect Talking Head

Lu Zhao[1], László Czap[*2]

Institute of Automation and Infocommunication, University of Miskolc
H-3515 Miskolc-Egyetemváros, Hungary
e-mail: [1] qgezhao@uni-miskolc.hu, [2] czap@uni-miskolc.hu

Abstract: Animated 3D articulation models – called talking heads – can be utilized, for instance, in speech assistant systems for children who are hard-of-hearing or when teaching learners of a second language. In this study, the objective is to identify articulation features and a dynamic system for visual representation of speech sounds for a Shaanxi Xi’an dialect talking head. In the first phase of the study, a phonetic alphabet of the dialect (northwest China) is formed following the official Romanization system used for Mandarin (Standard Chinese). After relating the phonemes of the dialect to those of Mandarin, we introduce the SAMPA code developed for the dialect, in addition to the correspondent regularities for whole syllable pronunciation. Secondly, we display the classification of static visemes (phonemes represented in visual form) for the dialect and describe an experiment carried out to articulatory movements of the tongue (features of timing and position) in dialect speech utterances recorded at different tempos. Finally, we discuss the results of an analysis of the images based on spatial-temporal tracking of the tongue movement contour. For definition of each uttered viseme the visual information obtained is classified and then used to create the dynamic viseme system of the tongue for a talking head using the Shaanxi Xi’an dialect of Chinese.

Keywords: Shaanxi Xi’an dialect; talking head; tongue contour tracking; dynamic viseme system; speech assistant system

1 Introduction

In this paper we identify the fundamental aspects needed for forming a talking head for the Shaanxi Xi’an dialect of Chinese [1]. The structure of the talking head system is illustrated in Figure 1. Using the X-SAMPA code created here, visemes are created for the consonants and vowels of the dialect. The viseme library also contains a dominance model (see Section 4.2) for the talking head [2]. Viseme classification is aided by the X-SAMPA code of consonants (C) and vowels (V) and identification of the regularities of their usage in whole-syllable pronunciation. We categorized the static visemes of the Shaanxi Xi’an dialect using the classification method for static visemes of Mandarin (Standard Chinese). Then an experiment was conducted focusing on the timing and position features of articulatory movements of the tongue in VCV and CVC utterances in the Chinese Shaanxi Xi'an dialect.

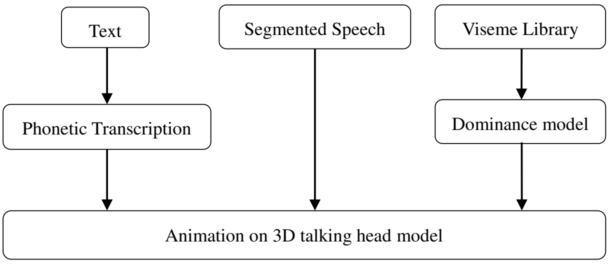


Figure 1
Structure of Shaanxi Xi'an dialect talking head system

Mandarin is a common focus of basic linguistic research in the fields of speech synthesis and speech recognition technology. However, other minority languages and a variety of Chinese dialects are present in modern-day multi-ethnic China. There are various studies relevant to the current topic. For instance, one study [3] has looked at phonetic conversion from Mandarin to the Min dialect of Taiwanese, along with mixed speech synthesis in Chinese with English. Another investigation of speech synthesis involving dialects focused on Tibetan, using a computer readable SAMPA scheme for conversion of text [4]. The Lanzhou, Liaocheng, Shenyang, and Tianjin dialects of Chinese have also been represented in speech synthesis [5].

The talking head being developed will form the foundation of a system to support deaf and hard-of-hearing children in learning to produce speech. It is possible to make the digital face transparent so that tongue placement and movement are clearly displayed – an advantage over a human speaker.

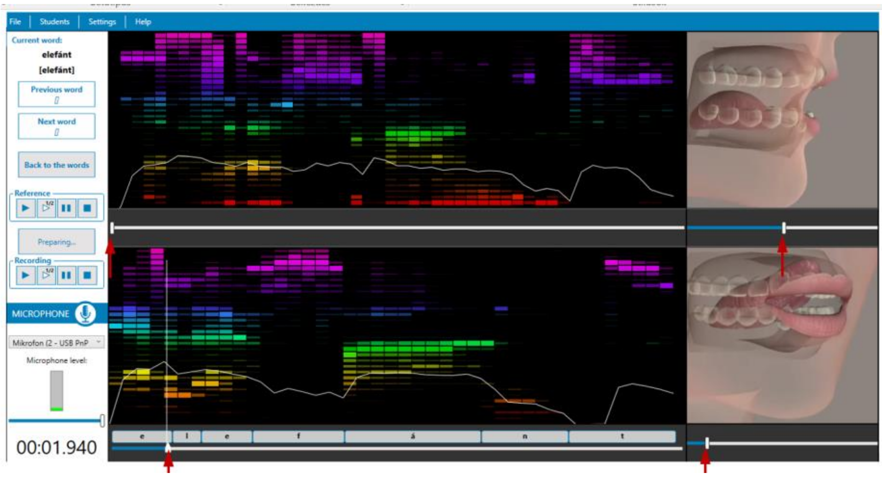
Xi'an, located in northwest China, was a capital during 13 dynasties of ancient China, and remains important today. Its dialect, Shaanxi Xi'an (also called Qin), has been spoken for over 3,000 years and has 8 million speakers today. It is the representative dialect of a large and influential region [6], making it well worth researching. Shaanxi Xi'an displays differences from Mandarin in its vocabulary, grammar, and most particularly in its articulation.

In order to create visual speech synthesis of the dialect, a transcription system is needed to label its phonetic information [7]. Based on this, it is possible to convert to SAMPA – Speech Assessment Methods Phonetic Alphabet. This is a machine-readable phonetic alphabet developed within the ESPRIT project that was first used with languages in the European Community, but has expanded to a variety of languages worldwide [8] [9]. This prompted the development of X-SAMPA (Extended-SAMPA), which covers all of the International Phonetic Alphabet (IPA) characters and remaps them into 7-bit ASCII, meaning that computer-readable phonetic transcriptions can be generated for any language [10]11]. X-SAMPA analysis of dialect phonemes and comparison with Hungarian and Pinyin ones enables us to derive certain viseme features from these existing systems, meaning that the main tasks are to distinguish unique features and identify regularities in articulation compared to those languages.

The speech assistant (SA) system being developed will highly rely on visual modality, especially on the visual representation of tongue movement, which is hardly observable in real conversation. As human speech perception involves both visual and auditory modalities, it is clearly multimodal, and the conditions of speech determine which modality has more effect [12]. Various studies have examined the development of normally hearing children in comparison with deaf or blind children and have found that insufficient exposure to stimuli in both modes has a substantial effect on the ability to perceive and produce speech that speech [13]. The audiovisual mode is more effective in transmitting articulatory features than any form of unimodal communication [14]. It has been proven by a number of clinical and laboratory investigations, that combined auditory-visual perception yields better results than perception through one mode alone, and this has been found true for normal-hearing and hearing-impaired children and adults alike [15]. People understand speech better when they can see articulators like the lips, jaw, tongue tip and teeth, as well as, the face. Thus, visual speech is an important aspect of speech perception, especially for deaf or hard-of-hearing people but also for normally hearing people in noisy surrounding [16]. Visual speech is studied and utilized in the fields of speech recognition [17], speech processing [18] [19], audio-visual speech synthesis [20], virtual talking head animation [21] and lip or tongue synchronization [22].

During the visual perception of speech, it is not the sight of the movement of lips and the face alone that matters; it has been found that the motion of the tongue, even though it is partially hidden, also conveys articulatory information that lip reading alone cannot access [23]. Development of infocommunication systems encourages speech researchers to deal with the speech of hard of hearing people and to study its physiological and acoustic characteristics. Computers can reveal features unseen before. In view of this, there are some pioneer Hungarian applications, such as the "Magic Box" (Varázsdoboz) package, developed by a team led by Klára Vicsi, which offers a tool for correcting pronunciation using spectrograms [24].

The rapidly developing capabilities in computing, 3D modeling and animation have contributed to the visualization tools that can be utilized, as audiovisual talking heads can display a human-like face while also making internal articulators visible. A talking head labeled Baldi was used for computer-assisted pronunciation training (CAPT) by Massaro and Cohen, who employed it as a tool in speech therapy and second language learning [25]. They went on to compare the effectiveness of instruction in phonetic contrasts between languages through illustrations of the processes taking place in the oral cavity along with an external view of Baldi's face [26]. Badin et al. attempted to use MRI and CT data to configure 3D tongue positions and forms. Their corpus consisted of sustained articulations from a single subject speaking French. With this, they developed a linear articulatory tongue model [27] that was later built into an audiovisual talking head that was able to display the normally hidden articulators (tongue, velum) during articulation [28]. A 3D talking head was proposed by Fagel et al. as a tool for speech therapy; it was capable of making a large variety of synthesized utterances for visualizing articulatory movements inside the oral cavity [29]. Wik and Engwall described how intra-oral articulations displayed in animation were able to contribute to the perception of speech [30]. A synthetic talking head using computer animation to illustrate the facial motions of lips, the jaw and the tongue with was utilized in training in speech perception and production by Beskow et al. [31]. An audio-visual representation of speech processes is also given by talking head developed at the University of Miskolc in Hungary, which is intended primarily to act as an aid in teaching speech to hard-of-hearing children [32] [33].



Current word:
elefánt
[elefánt]

00:01.940

Figure 2

Sample image of Hungarian speech assistant system with transparent face talking head and bar chart

On the right-hand side of Figure 2 the transparent talking head is shown from two different views. In either or both windows the head can be displayed from 45° and 90° angle views, enabling comparison of the articulation in two separate phases of the same word or sentence. In the Hungarian speech assistant system, beside the visualization of the lips and tongue movements, an additional speech sound visualization technique helps in developing speech production. In the center of Figure 2 the bar chart represents the visualized reference sound (bottom) and the recorded sound of the trainee (top). The pointer of the bar chart and that of the Talking Head can be moved in parallel from one picture frame to another, thus, the special educational needs (SEN) teacher can associate each sound graph with its articulation position. The Speech Assistant system proved to be a beneficial aid in individual speech therapy of hard of hearing pupils. The pedagogically planned methodology makes the speech therapy complete.

In the SA the bar chart and the talking head visually represent the speech signal and articulation, realizing sensor bridging between the modalities [34] [35] [36]. "The sensory information is transformed to an appropriate sensory modality in a way that the user can process it effectively" [37]. In both hearing-impaired and normal- hearing people acoustic and visual signals are integrated by the brain. The degree of sensor sharing of modalities depends on the grade of hearing impairment. The more severe the hearing loss, the more the subjects rely on the visual modality. For profoundly deaf people the visual representation of speech can be considered as a sensory substitution [38].

An expert system has been designed that aims at extending the Speech Assistant to ease the work of the SEN teacher for the hearing impaired as well as assist those practicing on their own [39]. Through automatic assessment of articulation, the system recommends the next word for practicing that can be most easily uttered built upon the sounds and sound connections already pronounced correctly. Thus, a scheme for individual development can be planned that takes linguistic, acoustic and phonetic knowledge and regularities into consideration. In this way the SA can adapt the order of speech items to the cognitive abilities and the current speech production level of the trainee.

Tongue movement measurement technologies have progressed in a number of stages. An x-ray microbeam system to investigate the effect of rate of speech on tongue-tip and lower-lip velocity profiles when uttering stop consonants was introduced by Adams et al [40]. A non-invasive NMR tagging technique representing tissue as discrete deforming elements was used by Napadow et al. for quantifying the degree of local tongue deformation [41]. Systems involving electromagnetic tracking utilize transmitters placed around the head and sensor coils positioned in the mid-sagittal plane and fixed to a variety of points on the jaw, lips and tongue [42]-[45]. Tongue positions can be revealed by ultrasound imaging, providing 2D images of the tongue surface contour [46]. Tongue dorsum movements were monitored during C-V sequences with varying speech rates with a computerized pulsed-ultrasound system [47]. Stone introduces methods for extracting, displaying and analyzing ultrasound image contours [48]. Ultrasound technology, despite some disadvantages, is a practical way of obtaining an image sequence for tongue motion. This non-invasive technology offers real-time capture rates, is relatively affordable, and can be easily incorporated into experimental setups. Other available methods such as slow motion recording, high cost (MRI), or radiation exposure (X-ray) have major drawbacks [49].

The following section deals with identifying different articulation of consonants and vowels between Mandarin and the Shaanxi Xi'an dialect and extension of the Pinyin Scheme to form a phonetic alphabet. The third section focuses on analysis of the CV phonetic features of the Shaanxi dialect and its phoneme changes compared to Mandarin. In the fourth section we report on an experiment studying articulatory movements of the tongue in terms of timing and position when a Chinese Shaanxi Xi’an dialect speaker is making VCV and CVC utterances of at varied tempos. Methods for modeling the tongue movements and creating the dynamic viseme system are detailed. In addition, we provide an example showing the result of tracing the tongue contour obtained by ultrasound technology for the phonemes of the dialect based on the algorithm we developed in MATLAB. Finally, we summarize our progress towards achieving a talking head for Chinese Shaanxi Xi’an dialect.

2 Phonetic Alphabet of Shaanxi Xi’an Dialect

2.1 Pinyin Scheme for Mandarin Chinese

The Pinyin scheme, the official Romanization system for Mandarin Chinese, contains basic phonemes (56), consonants (23) and simple vowels (36). This leads to 413 potential CV combinations plus special cases. When the four tones of Mandarin are considered, there are about 1,600 unique syllables. Pinyin is illustrated in Table 1 [50] [51].

2.2 Shaanxi Xi’an Dialect and Its Phonetic Alphabet

The Shaanxi Xi’an dialect consists of 26 consonants and 40 simple vowels. Its phonemes represented in IPA can be found in [52] [53] [54]. Its five unique consonants are presented in Table 2: ‘pf’ and ‘pf^{h}’ are both labiodental plosive fricative consonants, ‘v’ is a voiced labiodental fricative consonant, ‘ŋ’ is a velar nasal initial, and ‘ɳ’ is a retroflex nasal consonant [55] [56].

For establishing the relationship between phonemes and SAMPA code, first we need to identify the articulation of Shaanxi Xi’an dialect in IPA, then transcribe it into X-SAMPA [57] [58] [59]. Results are given in Tables 2 and 3.

Table 3 shows both the simple and compound vowels of the Shaanxi Xi’an dialect, which has 13 unique vowels compared to Mandarin.

X-SAMPA code analysis has shown that 5 vowels and 13 consonants are similar to Hungarian phonemes. The dialect talking head visemes of these phonemes can be easily derived from their Hungarian counterparts.

Table 1[1]

Romanized phonetic alphabet of Mandarin

23 consonants					
Type	Unaspirated	Aspirated	Nasal	Voiceless fricative	Voiced fricative
Bilabial	b	P	m		
Labiodental				f	
Alveolar	d	T	n		L
Velar	g	K		h	
Palatal	j	Q		x	
Dental sibilant	z	C		s	
Retroflex	zh	Ch		sh	R
	w, y				

36 vowels		
6 simple vowels		a, o, e, i, u, ü
14 compound vowels		ai, ao, ei, ia, iao, ie, iou, ou, ua, uai, üe, uei, uo, er
16 nasal vowels	8 front nasals	an, en, ian, in, uan, üan, uen, ün
	8 back nasals	ang, eng, iang, ing, iong, ong, uang, ueng

Table 2

X-SAMPA mapping for consonants of Shaanxi Xi’an dialect

Character	IPA	X-SAMPA
追	Pf	pf
吹	pfʰ	pv
味	V	v
爱	ŋ	N
女	ɳ	n`

Table 3

X-SAMPA mapping for vowels of Shaanxi Xi’an dialect

Character	IPA	X-SAMPA	Character	IPA	X-SAMPA
哀	æ	{	恩	ẽ	e~
岩	iæ	i{	因	iẽ	ie~
歪	uæ	u{	温	uẽ	ue~
安	æ̃	{~	晕	yẽ	ye~
烟	iæ̃	i{~	核	ɯ	M
弯	uæ̃	u{~	药	yo	Yo
冤	yæ̃	y{~			

[1] Note on pronunciation: The letters ‘y’ and ‘w’ can mark a new syllable: the syllable ‘wu’ is pronounced as the Pinyin ‘u’ , ‘yi’ as the Pinyin ‘i’ and ‘yu’ as the Pinyin ‘ü’

3 Phonemic Differences between Mandarin and the Shaanxi Xi'an Dialect

Pronunciation is the main difference between the dialect and Mandarin, especially with consonants, although variation is also found in vowels. Though quite complex, the variation follows some rules [60].

3.1 The Correspondence of some Consonants between Shaanxi Xi'an Dialect and Mandarin [61]

Table 4 shows the different consonants used in Shaanxi Xi'an dialect and in Mandarin when articulating the same Chinese character.

Table 4

Correspondence of some consonants in Mandarin and dialect

Mandarin	Dialect	Examples (Mandarin/ Shaanxi Xi'an dialect)
n	l	拿 ná/la; 奈 nài/lai; 弄 nòng/lòng; 暖 nuān/luan
ch	sh	尝 chāng/shǎng; 盛 chéng/sheng; 晨 chén/shen;
t	q	踢 tī/ qi; 调 tiáo/qiao; 田 tián/qian; 贴 tiē/qie
d	j	滴 dī/ji; 跌 diē/jie; 掉 diào/jiao; 丢 diū/jiu
k	f	哭 kū/fu; 苦 kǔ/fu; 酷 kù/fu
j	z	俊 jùn/zun; 炯 jiǒng/ziong; 精 jīng/zing
q	c	全 quān/cuan; 群 qún/cun；晴 qíng/cing
x	s	选 xuān/suan; 讯 xùn/sun; 削 xūe/suo; 行 xíng/sing

The syllables beginning with the consonants 'n' and 'l' basically have the same articulation as Mandarin, but in situations such as '农(nóng)', 'n' is articulated as 'l'. The consonants 'ch', 't', 'd', 'k' in Mandarin also have corresponding consonants in dialect. It can also be seen that when a syllable starts with the Mandarin consonants 'j', 'q', and 'x', in dialect these are articulated as 'g', 'k', and 'h', respectively.

A considerable number of non-aspirated consonants ('b', 'd', 'g', 'j', 'z') in Mandarin are always replaced by the aspirated initials ('p', 't', 'k', 'q', 'c', respectively) in Shaanxi Xi'an dialect. We present a series of examples in Table 5 to demonstrate this phenomenon.

Table 5

Correspondence between non-aspirated and aspirated consonants

Character	鼻	柜	旧	知	早	国
Mandarin	bí	guì	Jiù	zhī	zǎo	guó
Dialect	pi	kui	Qiu	chi	cao	gui

In most parts of the Xi’an area, when the phonemes ‘zh’, ‘ch’, or ‘sh’ are used at the beginning of the syllables they will be articulated as either ‘zh’, ‘ch’, ‘sh’ or as ‘z’, ‘c’, ‘s’ depending on the following phonemes. When ‘zh’, ‘ch’, ‘sh’ are followed by finials such as ‘a’, ‘ai’, ‘an’, ‘en’ they are articulated ‘z’, ‘c’, ‘s’; otherwise, they are pronounced ‘zh’, ‘ch’, ‘sh’. Some examples are listed to express this rule in Table 6.

Table 6

Correspondence among the phonemes ‘zh’, ‘ch’, ‘sh ’and ‘z’, ‘c’, ‘s’

Character	暂	知	产	潮	省	陕
Mandarin	zhǎn	zhī	Chǎn	chǎo	shěng	shǎn
Dialect	zan	zi	Can	cao	seng	san

In addition, there is a special pronunciation for syllables such as ‘zhu’, ‘chu’, ‘shu’, and ‘ru’, which most Xi’an dialect speakers articulate differently. ［pf］, ［pfʰ］ and ［f］ are fricatives, voiceless and unaspirated, ［v］ is fricative, voiced (note that the use of square brackets indicates IPA). Table 7 gives some examples to explain this articulation of phonemes.

Table 7

Complex reading of syllables ‘zhu’, ‘chu’, ‘shu’, ‘ru’

Character	猪	追	入	吹
Mandarin	zhū̄	Zhuī	rù	chuī
Dialect	［pfu］	［pfui］	[vu]	［pfʰei］

3.2 Vowel Features of Dialect Compared with Mandarin [62]

When the consonants ‘d’, ‘t’, ‘n’, ‘l’, ‘z’, ‘c’, ‘s’ and ‘zh’, ‘ch’, ‘sh’ appear in the front of the vowel ‘u’, ‘u’ will be changed to ‘ou’. This phenomenon is extremely common, and is widely perceived as an accent feature of the Shaanxi people. Table 8 presents examples of this and other vowel changes to show the corresponding phonemes between the two languages.

Table 8

Correspondance between vowels

Mandarin	Dialect	Examples
u	ou	读 dú-dou; 路 lù-loù; 足 zú-zoú; 醋 cù-cou; 数 shù-soù
e	i	液 yè-yi
ie	i	携 xié-xī
u	i	婿 xù-xi
uo	u	措 cuò-cù

It is very common to articulate ‘an’, ‘ian’, ‘uan’, ‘üan’ as ‘a’ or ‘ai’ ([æ]), which is a nasalization tone in Xi’an dialect. Table 9 presents examples.

Table 9

Special vowel changes

Character	三	端	捐	电
Mandarin	sān	Duān	jüān	diàn
Dialect	sain	Duain	jüai	die

4 Modeling the Tongue Movement of Chinese Shaanxi Xi'an Dialect Speech

This section describes the process used to gather data and model motion of the tongue for use in the talking head. The process begins with a static viseme classification of the dialect based on the method used to classify Mandarin static visemes [12] [63]. Then we introduce the ultrasound system used to obtain the speech materials (VCV and CVC sequences at different tempos). The visual information thus obtained is used to define the uttered viseme (the visual representation of a phoneme) for the dynamic viseme system for the tongue. This dynamic viseme system forms the fundamentals of a talking head – in this case, an animated articulation model for Shaanxi Xi’an dialect speech [64] [65].

4.1 Static Viseme Classification

We established a set of static visemes reflecting characteristics of the Shaanxi Xi’an dialect and their phonetic composition. When the dialect and Mandarin share a phoneme, we refer to the classification of Mandarin static visemes; when not, we provide dialect-specific phonemes. The viseme system in Standard Chinese is carried out by means of statistical analysis [12]. The viseme classification of the 26 consonants of Chinese Shaanxi Xi’an dialect speech is shown in Table 10.

Table 10[2]

Static consonant visemes classification for Shaanxi Xi’an dialect

Consonant	b,p,m	d,t,n	l	g,k,h	j,q,x	zh,ch sh,r	z,c ,s	f,v	[pf], [pf^h]	[ɳ]	[ŋ]
开口呼	爸	大	拉	哈	机	沙	杂	发	追	女	爱
合口呼		毒	路	姑	句	书	组		吹		

[2] ‘v’, ‘pf’, ‘pf^h’, ‘ɳ’, ‘ŋ’ are phonemes illustrated by IPA and others are a Romanized expression of phonemes.

The 40 vowels of the dialect [66] are classified into Static vowel viseme groups in Table 11. Fifteen basic static vowel visemes are classified; compound vowels are a combination of visemes for single vowels, as illustrated in Table 11.

Table 11[3]

Static vowel visemes classification and examples for Shaanxi Xi’an dialect

Simple Vowel	a, ang	啊，昂	er	儿
	æ, æ̃	哀，安	i	衣
	ao	奥	u	乌
	o	喔	ü	迂
	ou	欧	-i(-i front)	是
	e,eng	鹅，鞥	-i(-i back)	失
	ei, ẽ	诶，恩	ɯ	核
	yo	药		
Compound Vowel	ia=i+a;ie=i+e; iẽ =i+ẽ; ing=i+eng; iao=i+ao; iou=i+ou			
	iæ=i+æ; iæ̃ =i+ æ̃; iang=i+ang; ua=u+a; uo=u+o; uæ =u+ æ			
	uei=u+ei; uæ̃, uæ̃ =u+ æ̃; uẽ =u+ ẽ; uang=u+ang; ueng,ong=u+eng			
	yæ̃=y+ æ̃; üe=ü+e; yẽ=y+ẽ; iong=i+ong			

4.2 Dominance Classification Concept

While some parameters reach their target values during pronunciation, others do not, especially during fast speech. Grouping was performed according to the dominance of each feature determining tongue position and lip share and the articulation features of each speech sound were placed in a dominance class [66]. This differs from the standard approach, which classifies only the dominance of the phonemes. Four dominant grades emerge from the parametric model features:

- stable — co-articulation has no effect (e.g. tongue position of alveolar plosives, lip shapes of bilabials),

- dominant — co-articulation has only a slight effect (e.g. lip shapes of vowels),

- flexible —neighboring sounds affect the feature (e.g. tongue positions of vowels),

- uncertain — the neighborhood defines the feature (e.g. tongue position of bilabials, lip shapes of ‘h’ and ‘r’).

[3] ‘æ’, ‘iæ’, ‘uæ’, ‘æ̃’, ‘iæ̃’, ‘uæ̃’, ‘yæ̃’, ‘ẽ’, ‘iẽ’, ‘uẽ’, ‘yẽ’, ‘ɯ’, ‘yo’ are phonemes illustrate by IPA and others are a Romanized expression of phonemes.

4.3 Tongue Movement Contour Tracking

In this study we record a small-scale visual speech database using combinations of the consonants and vowels of Chinese Shaanxi Xi’an dialect. The tongue movement contour is tracked through processing of the ultrasound image in the speech database, while the viseme system for Chinese Shaanxi Xi’an dialect determined through dynamic analysis.

4.3.1 Subjects and Speech Material

The subject was one adult female – the first author of this paper – who is a native speaker of Chinese Shaanxi Xi’an dialect.

Two structures were investigated, VCV and CVC (‘V’ indicates vowel while ‘C’ indicates consonant), covering all phonemes involved in our experiment (the 26 consonants and 40 vowels of Shaanxi Xi’an dialect). In the VCV structure, ‘e’ and ‘a’ (eCe and aCa) are used to compare the different dominance features of the same consonant. Similar tongue positions mean high dominance, while different values mean low dominance. In the CVC structure, ‘k’ and ‘t’ are the two phonemes used (kVk and tVt) to compare the different dominance features of the same vowel. These phonemes were chosen for the database because of the rear tongue articulating the phonemes ‘a’ and ‘k’ and front position when articulating ‘e’ and ‘t’.

4.3.2 Tongue Movement Recording Method

In order to follow the motion of the tongue we use the 'Micro' ultrasound system (Articulate Instruments Ltd.). a speech recording instrument with a 2–4 MHz/64 element 20 mm radius convex ultrasound transducer at roughly 82 fps. The angle of view was 90°; there were 842 pixels in each of the 64 scanlines in the raw data. An ultrasound stabilization headset (Articulate Instruments Ltd.) fixed the transducer during recording. An Audio-Technica ATR 3350 omnidirectional condenser microphone was set approximately 20 cm from the lips when recording the speech samples.

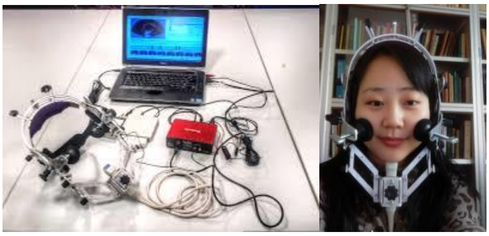


Figure 3

Left: ‘Micro' Ultrasound System. Right: Probe Stabilization Headset installation

A photograph of this instrumentation is presented on the left side of Figure 3, while the Probe Stabilization Headset is shown in place on the right side. The headset was individually fitted with the main goal of obtaining an image appropriate for phonetic analysis [67]. The headset fixes the subject’s head while capturing the images and also fixes the ultrasound transducer under the chin [49].

4.4 Tongue Movement Contour Tracking

Problems revealed in biomedical image analysis such as user fatigue, user bias, and difficulty in reproducing results also may occur when manually tracking tongue contours [68]. We thus developed an algorithm to extract and track 2D tongue surface contours from ultrasound sequences in the 'Micro' ultrasound system. Traditionally, visemes are defined as a set of static mouth shapes that represent clusters of contrastive phonemes [69]. However, the movement of phoneme pronunciation is less a static state and more a dynamic process. Here we present the concept of the dynamic viseme, representing the entire process of organ motion during articulation of a given phoneme. Similarly to the co-articulation model of Cohen [70], our dynamic viseme model blends dominance and parameter values.

4.4.1 Dominance Classification for the Shaanxi Xi’an Dialect

Two central frames of the same consonant or vowel in the audio speech spectrum of a paired structure are selected after manual segmentation in Praat. JPG images of the ultrasound frames are analyzed. Then we trace the tongue feature points of the targeted phoneme in the paired structure and compare the tongue contour of the same phonemes in both structures in MATLAB using the algorithm to trace tongue contour. In Figure 4, (b) and (c) show the tongue contour comparison in the frame belonging to the burst of ‘t’ and ‘p’ in the two structures.

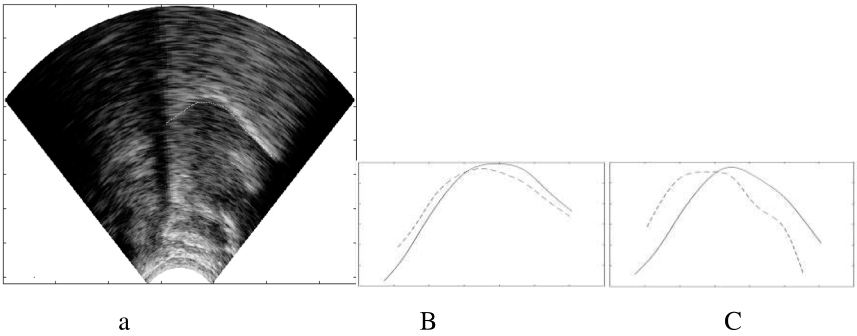


Figure 4

a: Sample ultrasound image with tongue contour tracking; b: Tongue contours of ‘t’ in ‘ete’ (—) and ‘ata’ (- -); c: Tongue contours of ‘p’ in ‘epe’ (—) and ‘apa’ (- -)

The continuous curve shows the tongue contour in that frame for ‘t’ in the structure ‘ete’ and ‘p’ in ‘epe’, while the dashed line shows the tongue contour of ‘t’ in the structure ‘ata’ and ‘p’ in ‘apa’.

The dominance feature of the invisible tongue tip of ‘t’ is classified as stable, while the tongue position of ‘p’ is uncertain, approaching that of the neighboring sounds. In our future research, we plan to focus on a sequence of frames in order for more accurate classification of the dominance grade of viseme features.

In Figure 4, the complete tongue contour that can be seen in the ultrasound image is shown after automatic contour tracking, the uneven curve being smoothed with discrete cosine transformation filtering. The description of the smoothed tongue contour makes it possible to draw further conclusions on the basis of the selected feature points of the curve. Four feature points were selected at 20, 40, 60 and 80% of the arc of the smoothed curve [71]. In Figure 5 (a), the positions of the feature points of the sound ’p’ in VCV words ’apa’ and ’epe’ can be seen for the three image frames before burst (altogether 36 ms). Figure 5 (b) shows the position of the feature points of the sound ’sh’ in words ’asha’ and ’eshe’ for the whole range of the sound. The uncertain character of ’p’ and the dominant character of ’sh’ can be seen very well. (Similarly to Figure 4, on the left hand side, the back and on right hand side, the front of the tongue can be seen. The numbers of rows increase from top to bottom, as it is usual in the representation of images.)

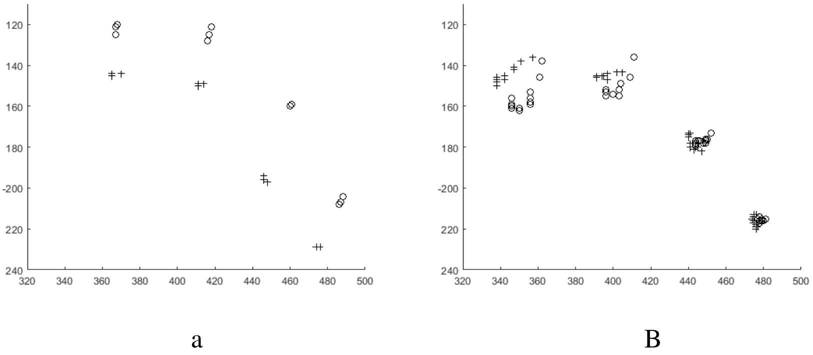


Figure 5

a: positions of the four feature point of sounds ’p’ and b: ’sh’ in the environment of ’e’ (o) and ’a’ (+)

The movement of the tongue during speech can be described with the changes in the coordinates of the feature points. Figure 6 shows the vertical positions of the two front feature points while VCV words ’ama’ and ’ala’ are being uttered. This representation not only shows the uncertain character of ’m’ and the dominant character of the vertical position of the front part of the tongue in case of ’l’ but also makes possible the investigation of the interpolation between key frames.

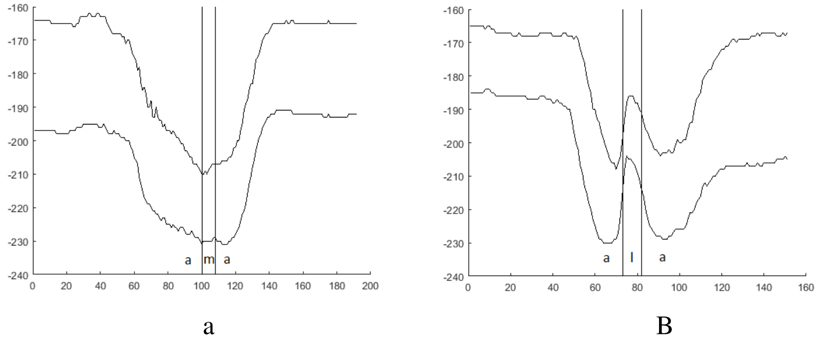


Figure 6
Vertical position of the first two feature points of the tongue while the words a: ’ama’ and b: ’ala’ are being uttered

Dominance is analyzed for all the features involved in the animation. The bilabial sounds in the previous examples are, e.g. stable as regards the shape of the lips but are of uncertain character as regards the position of the tongue.

This method will be used to determine the dominance grade for all viseme features of the dialect, so that we can create a dynamic viseme system using the tongue contour with a dominance model.

5 Conclusion and Future Work

This paper presents a method for the phonetic transcription of the Shaanxi Xi’an dialect of Chinese and the conversion of its basic phomes into a computer readable phonetic alphabet. Transcription was based on the phonetic alphabet of the dialect, mapping the phonemes shared with Mandarin supplemented by several phonemes unique to Shaanxi Xi’an. The purpose is to obtain the fundamental data needed to create a talking head for the Shaanxi Xi’an dialect. The classification method for Mandarin static visemes was applied to static viseme classification of Chinese Shaanxi Xi’an dialect speech. We studied both the timing and position properties of articulatory movements of the tongue in Chinese Shaanxi Xi’an dialect speech utterances spoken at different tempos by one native speaker of the dialect. She read randomized lists of VCV utterances containing the vowels /e/ or /a/ and CVC utterances containing the consonants /k/ or /t/ in all possible combinations of the dialect’s 26 consonants and 41 vowels. The 'Micro' ultrasound system recorded the utterances and the Assistant Advanced software formed JPG images and MP4 videos. We developed an algorithm to automatically track spatial-temporal tongue movement contours from the ultrasound images. The visual information is classified by dominance and other features to define the uttered viseme and will form the basis of a dynamic viseme system of tongue motion for the Shaanxi Xi'an dialect. Similar classification of lip shape features is in progress through analysis of the video recordings. The interpolation between articulation features is being refined with the analysis of the ultrasound image (position of the tongue) and video (shape of the lips) made during the continuous reading of a long text. The standard deviation of the feature examined well combines the essence of the analyses shown in Figures 4-6: the greater the deviation, the lower the dominance. Our animation process, elaborated for the Hungarian language, accomplishes the screening of the features according to dominance class.

The long-term objective is to create a dynamic articulation model that can be applied to animate articulation for Shaanxi Xi'an dialect speech within a 3D virtual talking head. This is intended for use in a speech assistant system for hard-of-hearing children and second language learners.

Acknowledgement

The described study was carried out as part of the EFOP-3.6.1-16-00011 "Younger and Renewing University – Innovative Knowledge City – institutional development of the University of Miskolc aiming at intelligent specialization" project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

References

[1] Czap L, Mátyás J: Hungarian talking head. Proceedings of Forum Acusticum 4th European Congress on Acoustics. Budapest, Hungary, 2005, pp. 2655-2658

[2] Czap L, Mátyás J: Virtual speaker [J] Infocommunications Journal Selected Papers, 2005, Vol. 60, 6, pp. 2-5

[3] Lyu R. Y: A bi-lingual Mandarin/Taiwanese (Min-nan), Large Vocabulary, Continuous speech recognition system based on the Tong-yong phonetic alphabet (TYPA) [C] Sixth International Conference on Spoken Language Processing, 2000

[4] Liu B, Yang H, Gan Z: Grapheme-to-phoneme conversion of Tibetan with SAMPA [J] Jisuanji Gongcheng yu Yingyong (Computer Engineering and Applications) 2011, 47(35): 117-121 (In Chinese)

[5] Guo W, Yang H, Song J: Research on Text Analysis for Dialect Speech Synthesis [J] Computer Engineering, 2015 (In Chinese)

[6] Wurm S A, Li R, Baumann T: Language Atlas of China [M] Australian Academy of the Humanities; Longman Group (Far East) 1987

[7] Lu Tuanhua: Comparison of Phonetic Features and Pronunciation of Mandarin in Xi'an Dialect [J] Journal of Test Sciences, 2010, 9: 23-24 (In Chinese)

[8] Arora K K, Arora S, Singla S R: SAMPA for Hindi and Punjabi based on their Acoustic and Phonetic Characteristics [C]//Proc. International Oriental COCOSDA 2007 Conference (Hanoi, Vietnam. 2007: 17-22

[9] Kabir H, Saleem A M: Speech assessment methods phonetic alphabet (SAMPA): Analysis of Urdu [J] CRULP Annual Student Report published in Akhbar-e-Urdu, 2002

[10] Tseng C, Chou F: Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan [J] Journal of the Acoustical Society of Japan (E) 1999, 20(3): 215-223

[11] Zu Y, Chen Y, Zhang Y: A super phonetic system and multi-dialect Chinese speech corpus for speech recognition [C] International Symposium on Chinese Spoken Language Processing, 2002

[12] Wang A, Bao H, Chen J: Primary research on the viseme system in standard Chinese [J] Proceedings of the International Symposium of Chinese spoken language Processing, 2000

[13] Bailly G, Badin P: Seeing tongue movements from outside [C]//Seventh International Conference on Spoken Language Processing. 2002

[14] Robert-Ribes J, Schwartz J L, Lallouache T: Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise [J] The Journal of the Acoustical Society of America, 1998, 103(6): 3677-3689

[15] Erber N P: Auditory-visual perception of speech [J] Journal of Speech and Hearing Disorders, 1975, 40(4): 481-492

[16] Czap L, Pinter J M: Multimodality in a Speech Aid System [J] Journal on Human Machine Interaction, 2014, 1: 64-71

[17] Werda S, Mahdi W, Hamadou A B: Lip localization and viseme classification for visual speech recognition [J] arXiv preprint arXiv:1301.4558, 2013

[18] Massaro D W, Beskow J, Cohen M M: Picture my voice: Audio to visual speech synthesis using artificial neural networks [C]//AVSP'99-International Conference on Auditory-Visual Speech Processing, 1999

[19] Czap L: On the Audiovisual Asynchrony of Speech. Proceedings of Auditory-Visual Speech Processing (AVSP) 2011, Volterra, Italy, International Speech Communication Association (ISCA) pp. 137-140

[20] Železný M, Krňoul Z, Císař P: Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis [J] Signal Processing, 2006, 86(12): 3657-3673

[21] Pintér J M, Czap L: Improving Performance of Talking Heads by Expressing Emotions. 3nd CogInfoCom Conference, Košice, Slovakia, IEEE, pp. 523-526

[22] Zorić G, Pandžić I S: Real-time language independent lip synchronization method using a genetic algorithm [J] Signal Processing, 2006, 86(12): 3644-3656

[23] Montgomery D: Do dyslexics have difficulty accessing articulatory information? [J] Psychological Research, 1981, 43(2): 235-243

[24] Vicsi, K., Hacki, T.: 'CoKo' - Computerised audiovisual feedback speech-tutoring system for children with articulation disorders and impaired hearing. (In German: 'CoKo' - Computergestützter Sprechkorrektor mit audiovisueller Selbstkontrolle für artikulationsgestörte und hörbehinderte Kinder.) Sprache-Stimme-Gehör 20, 141-149, 1996

[25] Massaro D W, Cohen M M: Visible speech and its potential value for speech training for hearing-impaired perceivers [C]//STiLL-Speech Technology in Language Learning, 1998

[26] Massaro D W, Light J: Read my tongue movements: bimodal learning to perceive and produce non-native speech [C] Eighth European Conference on Speech Communication and Technology, 2003, CD-ROM 4 pp

[27] Badin P, Bailly G, Reveret L: Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images [J] Journal of Phonetics, 2002, 30(3): 533-553

[28] Badin P, Elisei F, Bailly G: An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data [J] Articulated Motion and Deformable Objects, 2008: 132-143

[29] Fagel S, Madany K: A 3-D virtual head as a tool for speech therapy for children [C]//Ninth Annual Conference of the International Speech Communication Association. 2008

[30] Wik P, Engwall O: Can visualization of internal articulators support speech perception? [C]//INTERSPEECH. 2008: 2627-2630

[31] Beskow J, Engwall O, Granström B: Visualization of speech and audio for hearing impaired persons [J] Technology and Disability, 2008, 20(2): 97-107

[32] Czap L, Pintér J M, Baksa-Varga E: Features and Results of a Speech Improvement Experiment on Hard of Hearing Children Speech Communication 2019 106 pp. 7-20, 14 p.

[33] Czap, L.: Speech Assistant System. INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association,

Singapore, International Speech Communication Association (ISCA) pp. 1486-1487

[34] Baranyi, P., Csapo, Á., Sallai, Gy.: Cognitive Infocommunications, Springer, 2015

[35] Baranyi P., Csapó Á.: Definition and synergies of Cognitive Infocommunications, Acta Polytechnica Hungarica, 9 (1) pp. 67-83, 2012

[36] CogInfoCom - Cognitive Infocommunications www.coginfocom.hu, accessed: 29.05.2018

[37] Sallai, G.: The Cradle of the Cognitive Infocommunications, Acta Polytechnica Hungarica 9 (1) pp. 171-181

[38] D. W. Massaro, M. M. Cohen: Integration of visual and auditory information in speech perception, Journal of Experimental Psychology Human Perception & Performance November 1983, pp. 753-771

[39] Kovács, S., Tóth, Á., Czap, L. "Fuzzy model based user adaptive framework for consonant articulation and pronunciation therapy in Hungarian hearing-impaired education." CogInfoCom 2014: Proceedings. 2014, pp. 361-366

[40] Adams S G, Weismer G, Kent R D: Speaking rate and speech movement velocity profiles [J] Journal of Speech, Language, and Hearing Research, 1993, 36(1): 41-54

[41] Napadow V J, Chen Q, Wedeen V J: Intramural mechanics of the human tongue in association with physiological deformations [J] Journal of Biomechanics, 1999, 32(1): 1-12

[42] Dromey C, Nissen S, Nohr P: Measuring tongue movements during speech: Adaptation of a magnetic jaw-tracking system [J] Speech Communication, 2006, 48(5): 463-473

[43] Perkell J S, Cohen M H, Svirsky M A: Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements [J] The Journal of the Acoustical Society of America, 1992, 92(6): 3078-3096

[44] Schönle P W, Gräbe K, Wenig P: Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract [J] Brain and Language, 1987, 31 (1): 26-35

[45] Dromey C, Nissen S, Nohr P: Measuring tongue movements during speech: Adaptation of a magnetic jaw-tracking system [J] Speech Communication, 2006, 48(5): 463-473

[46] Lundberg A J, Stone M: Three-dimensional tongue surface reconstruction: Practical considerations for ultrasound data [J] The Journal of the Acoustical Society of America, 1999, 106(5): 2858-2867

[47] Ostry D J, Munhall K G: Control of rate and duration of speech movements [J] The Journal of the Acoustical Society of America, 1985, 77(2): 640-648

[48] Stone M: A guide to analysing tongue motion from ultrasound images [J]. Clinical linguistics & phonetics, 2005, 19(6-7): 455-501

[49] Akgul Y S, Kambhamettu C, Stone M: Automatic extraction and tracking of the tongue contours [J] IEEE Transactions on Medical Imaging, 1999, 18(10): 1035-1045

[50] Zein P.: Mandarin Chinese Phonetics. http://www.zein.se/patrick/chinen8p.html, accessed: 11.07.2017

[51] Zhou Youguang: Basic knowledge of Hanyu Pinyin Schedule [M] Language Publishing House, 1995 (In Chinese)

[52] Sun Lixin: Xi'an dialect research [M] Xi'an publishing house, 2007 (In Chinese)

[53] Kang Jizhen: An Experimental Study of Phonetics in Xi'an Dialect [C] Northwest University, 2015 (In Chinese)

[54] Guo Weitong: Analysis of Acoustic Features and Modeling of Prosody in Xi'an Dialect [C] Northwest Normal University, 2009 (In Chinese)

[55] Yuan Jiahua: Outline of Chinese Dialects [M] Text Reform Press, 1983 (In Chinese)

[56] Chinese Dialect Vocabularies [M] Text Reform Press, 1989 (In Chinese)

[57] Jialu Z: SAMPA_SC for standard Chinese (Putonghua) [J] Acta Acustica, 2009, 34:82-86 (In Chinese)

[58] SAMPA: Computer Readable Phonetic Alphabet. http://www.phon.ucl.ac.uk/home/sampa/home.htm; accessed: 11.07.2017

[59] Wells J: Computer-coding the IPA: a proposed extension of SAMPA. http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm, accessed: 11.07.2017

[60] Zhao L: The correspondent regularities between Mandarin and Shaanxi Dialect [J] Journal of Baoji University of Arts & Sciences, 2008, Vol. 28, 1 (In Chinese)

[61] Yang Jinfeng: Corresponding speech sound in west Shannxi Dialect and Mandarin. [J] Journal of Xianyang Teachers' College. 2003 Vol. 18, 5 (In Chinese)

[62] Wang Y: Research on the types of phonetic changes in Xi'an Dialect. [J] Journal of Yanan University (Social Science Edition) 1995, Vol. 17, 2 (In Chinese)

[63] Wu Z, Zhang S, Cai L: Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar[C]//INTERSPEECH. 2006 (4): 1802-1805

[64] Zhao H, Tang C: Visual speech synthesis based on Chinese dynamic visemes[C] Information and Automation, 2008. ICIA 2008, International Conference on Information and Automation, IEEE, 2008: 139-143

[65] Sztahó D, Kiss G, Czap L, Vicsi K: A computer-assisted prosody pronunciation teaching system[C]//WOCCI. 2014: 45-49

[66] Czap L, Zhao L: Phonetic Aspects of Chinese Shaanxi Xi'an Dialect. 8th International Conference on Cognitive InfoCommunications: CogInfoCom. Debrecen, Hungary, Piscataway: IEEE Computer Society, 2017, pp. 51-56

[67] Scobbie J M, Wrench A A, van der Linden M: Head-Probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement[C]//Proceedings of the 8th International Seminar on Speech Production. 2008: 373-376

[68] Xu K., Csapó T. G., Roussel P., Denby B.: A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization, Journal of the Acoustical Society of America. 2016 Vol. 139, 5 pp. 154-160

[69] Taylor S L, Mahler M, Theobald B J: Dynamic units of visual speech[C]//Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation. Eurographics Association, 2012: 275-284

[70] Aghaahmadi M, Dehshibi M M, Bastanfard A: Clustering Persian viseme using phoneme subspace for developing visual speech application [J] Multimedia Tools and Applications, 2013, 65(3): 521-541

[71] Zhao L., Czap L: A nyelvkontúr automatikus követése ultrahangos felvételeken (Automatic tongue contour tracking on ultrasound images, In Hungarian) Beszédkutatás 2019, Vol. 27 : 1 pp. 331-343

An Approach to Evaluating the Quality of Websites Based on the Weighted Sum Preferred Levels of Performances Method

Darjan Karabasevic[1], Dragisa Stanujkic[2], Mladjan Maksimovic[3], Gabrijela Popovic[4], Oliver Momcilovic[5]

[1, 3] Faculty of Applied Management, Economics and Finance, University Business Academy in Novi Sad, Jevrejska 24, Belgrade, Serbia, E-mail: darjan.karabasevic@mef.edu.rs; mladjan.maksimovic@mef.edu.rs

[2] Technical Faculty in Bor, University of Belgrade, Vojske Jugoslavije 12, 19210 Bor, Serbia, E-mail: dstanujkic@tfbor.bg.ac.rs

[4] Faculty of Management in Zajecar, Megatrend University Belgrade, Park šuma “Kraljevica” bb, 19000 Zaječar, Serbia, E-mail: gabrijela.popovic@fmz.edu.rs

[5] Faculty of Information Technology and Engineering, University „Union – Nikola Tesla“, Jurija Gagarina 149a (TC Piramida), Staro sajmište 29, 11070 Belgrade, Serbia, E-mail: oliver.momcilovic@fiti.edu.rs

Abstract: Measuring the quality of a website is important for companies in order to maintain their competitiveness. This manuscript intends to present a new multiple criteria decision-making approach devoted to the evaluation of the quality of websites from the viewpoint of their visitors. The proposed approach uses gaps between expectations and perceptions similar to the well-known SERVQUAL methodology. The proposed approach is also based on the use of the Weighted Sum Preferred Levels of Performances (WS PLP) approach and the proven smaller set of criteria, which enables the forming of less complex questionnaires, and as such, it should enable us to more easily collect the real attitudes of surveyed website visitors. The usability and efficiency of the proposed approach are considered in the case study devoted to the evaluation of the websites of three telecommunication companies in Serbia.

Keywords: MCDM; WS PLP approach; Company website; Website quality evaluation

1 Introduction

The emergence of new technologies, such as the Internet, has caused change in the manner companies do their business. In the last decades, the Internet has undoubtedly become the fastest-growing communications medium; accordingly, many companies have adopted and have been taking the advantages the Internet offers. The Internet has allowed companies of any size to be easily accessible in the cyberspace, in accordance with which small and large size companies are able to create websites in order to present their respective corporate missions, products or services to the world [1]. So, modern companies perform the largest portion of their communications with their customers over the Internet, usually in order to promote their own products and services. Therefore, Cebi [2] emphasizes the fact that the Internet is an effective tool for companies to reach their customers via their own websites.

Similarly to any commercial, the website helps companies to inform, persuade and remind their customers about and of, respectively, the company and its products and services. In his study, Ibrahim [3] has confirmed the fact that websites are the main bearers of the marketing activities of a company, which are actively used today. However, bearing in mind the fact that there is increasing competition, the quality of companies' websites is of great importance and, undoubtedly, it is critical for a company's achieving of goals [4]. For that reason, the quality of the website has become an important tool for the acquisition of new customers, as has been confirmed in numerous studies, only to name some of them: Stanujkic et al. [5], Al-Manasra et al. [6], Bai et al. [7], Lin [8], Kim and Stoel [9] and so on.

Therefore, measuring the quality of a website is important for companies in order to maintain their competitiveness. The SERVQUAL model, proposed by Parasuraman et al. [10], was often used for measuring a service quality. Somewhat later, based on SERVQUAL, other models for the evaluation of a service quality have been proposed, amongst which: WebQual [11; 12], E-S-QUAL, E-RecS-QUAL [13] and so forth. Measuring the quality of a website mainly based on the WebQual has been the subject matter of numerous studies conducted by numerous researchers, such as: Loiacono et al. [14], Barnes and Vidgen [15-16], Shchiglik and Barnes [17], Park et al. [18], Park and Baek [19] and so on.

Quality is an attribute of a service that gives an insight how well it fulfils the customers' needs. Evaluating the service, in this case the website, quality is very complex and acquiring task. According to the previous mentioned models proposed for solving such a problem, different evaluation dimensions are emphasized which additionally complicates already complicated and complex issue. Different lists of evaluation criteria are proposed which often are not mutually compatible. If decision maker (hereinafter marked as DM) gives the priority to the certain set of evaluation criteria, neglecting the others, the decision would not be representative because it would not be based on the whole group of the involved criteria. In order to overcome the problems in decision-making process related to the appreciating of all evaluation criteria the Multiple Criteria Decision Making (MCDM) methods are introduced.

MCDM is one of the significant branches of operational research and it deals with problems which we are faced with when deciding in the presence of a large number of, usually conflicting, criteria. Therefore, Keeney and Raiffa [20] suggest the five key principles that should be considered when formulating criteria: completeness, the operational ability, decomposability, non-redundancy and the minimum size. It is stated by Beynon [21] that the DM's ability to make preference judgments about a number of different decision alternatives is the basis for making a decision. Further, Korhonen [22] points out the fact that the solving of a multi-criteria decision-making problem requires that the DM should make choice of the "reasonable" alternative out of a set of available alternatives the most consistent with his/her preferences. Until today, many different MCDM methods are proposed as well as its appropriate extensions that enable DM to incorporate the vagueness and uncertainty into the decision-making process [23-26].

In this manuscript the WS PLP method [27] is proposed as a convenient tool that could be used for the website quality evaluation from two aspects, i.e. customers and companies. Customers could apply the WS PLP method in the process of the evaluation and selection of the websites of different kind. By applying the aforementioned method, companies could investigate how consumers evaluate their websites and what is their position related to their main competitors. The main goal is dichotomous: (1) first is to facilitate the decision-making process and enable DM (in this case customer) to choose website in accordance with his/her needs and requirements; (2) second is to enable companies to, by using the same tool on the group of examinees, estimate the quality of theirs websites, their rang according to competitors and to define what dimensions should be improved.

For the need of this paper, the evaluation process is performed by using the set of the evaluation criteria retrieved from the Webby Awards (http://webbyawards.com/judging-criteria/). The website's quality of the three telecommunication companies that operate in the Republic of Serbia are assessed by DMs involved in the process of the evaluation because of obtaining the more reliable results as possible. In order to show the possibilities of the proposed method, the manuscript is structured in the following manner: Section 2 presents Literature Review and Section 3 demonstrates the WS PLP approach. In Section 4, the framework for evaluating the website quality is presented and a case study is given in Section 5. Ultimately, the conclusions are given at the end of the manuscript.

2 Literature Review

Website quality is topic that occupies the scientific attention because it has a significant impact on the business results and success of an organization. For example, Bai et al. [7] examined how the website quality have influence on the consumer satisfaction and willingness to shop online, and Jones and Kim [28] explore does website quality have impact on the young female consumers in the US to buy clothes online. The appropriate websites also could contribute to the hotel business by developing good relationships with their clients [29]. Certain attributes are very important when the website quality is in question because they contribute to the consumer satisfaction in a greater degree [30]. The consumer satisfaction is crucial dimension that is connected with website quality and question that arises is how to measure the quality of webiste as well as the consumer satisfaction that is connected to it?

Websites and its applications represent a kind of service and the tool proposed for the measuring the service quality is SERVQUAL, as previously stated. According to Parasuraman et al. [10] service quality is defined as a gap between consumer expectations and consumers' perceptions and five dimensions crucial for the measuring of service quality are identified and they are: tangibility, reliability, responsiveness, assurance and empathy. Further, the new appropriate models for the estimating of website quality was proposed such as: WebQual model [11-12], Web Quality (WQM) model [31] and WebQual TM quality evaluation model [14]. In various cases, when certain type of websites are evaluated that fullfils concrete customer need, different model is used. The mentioned models are very convinient for applying by practitioners who want to identify what performance of their website is good as those that need improvement so they could be in accordance with customer desire. But, on other hand there are consumers who also wants to choose the website according to their needs, but the use of mentioned models is not quite appropriate for them. Also, there is a need for the simpler questioners that companies could use for the investigating the opinions of their customers. In that case, the MCDM methods seems like appropriate solution that could help in resolving the mentioned issues.

In the past decades, the rapid development of operational research has caused the creation of many MCDM methods, such as: the Weighted Sum (WS) method [32-33], the ELECTRE method [34], the AHP method [35-36], the TOPSIS method [37], the PROMETHEE method [38], the COPRAS method [39], and the VIKOR method [40]. Recently, the new generation of the MCDM methods is proposed such as: the MUSA method [41], the MULTIMOORA method [42], the ARAS method [43], the SWARA method [44], the FARE method [45], the WASPAS method [46], the KEMIRA method [47] and the EDAS method [48].

In a number of studies, MCDM methods have been successfully used for the purpose of the evaluation of the quality of websites. Burmaoglu and Kazancoglu [49] have applied the hybrid MCDM method AHP and VIKOR in a fuzzy environment for the evaluation of the e-government website. Akincilar and Dagdeviren [50] have proposed a hybrid multi-criteria decision-making model based on the AHP and PROMETHEE in order to evaluate the websites of the hotels. Ecer [51] uses the AHP and COPRAS-G to conduct the evaluation of the quality of websites in the banking industry. Stanujkic et al. [5] provide an approach to the measuring of the website quality in the rural tourism industry based on Atanassov’s intuitionistic fuzzy sets. Jain et al. [52] uses the TOPSIS method for the evaluation of e-learning websites; also, Jain et al. [53] use weighted distance-based approximation for the selection and ranking of e-learning websites. Kang et al. [54] uses the fuzzy hierarchical TOPSIS based on the ES-QUAL model for the evaluation of e-commerce websites. Stanujkic and Karabasevic [55] uses extension of the WASPAS method with intuitionistic fuzzy numbers for the evaluation of websites.

In this manuscript, the application of the WS PLP method introduced by Stanujkic and Zavadskas [27] is proposed. This method represents the improvement of the WS method. The mentioned method is very applicable and easy to use and until now it is successfully applied for the solving of various types of the decision-making problems, such as: the ranking of transportation zones [56], the estimation of technologies for the power supply [57], the evaluation of the sustainability of transport noise [58] and performing a decision-making process in a fuzzy environment [59-60]. The WS PLP method is used in the field of human resources management [61-63] and for the estimation of the froth flotation reagents [64]. The WS PLP method could, also, enable the facilitating of the evaluation process in the field of the website quality performed by the customers by utilizing its advantage of incorporating the DM’s point of view in a better way.

3 The WS PLP Approach

Based on the WS method, proposed by Churchman and Ackoff [32] and MacCrimmon [33], Stanujkic and Zavadskas [27] proposed the WS PLP approach. The normalization procedure introduced by Stanujkic et al. [65] that incorporates the DM`s preferences for preferred performance ratings (*ppr*) initiated the idea for the forming of new decision-making method. The standpoint of the DM related to the value of criteria is incorporated in every MCDM method in certain degree, but in this case, through the *ppr* values, the DM`s concretely and exactly express the desired values of the considered criteria. This specified set of *ppr* values directly affect evaluation of alternatives by transforming them into the group of acceptable alternatives among which the selection should be performed.

In the case of website quality estimation, DMs usually exactly know what features should have the certain website. The *ppr* values, as a part of the WS PLP method, enable making of an adequate decision that includes expectations and requirements of the DM in greater extent. Exactly this *ppr* values lead to the similarity of the WS PLP method with the SERVQUAL model because this values represents the expectations of the DM, while the given estimations of the alternatives represent their perceptions. Also, there is a possibility to make a choice between the alternative that better match with the given *ppr* values and the alternative that has the best overall performance rating. Compensation coefficient that is a part of the procedure of the WS PLP method, gives the opportunity to the DM to choose between the mentioned options. Besides, the DM could easily define whether an alternative is better ranked because of only or several criteria that has good ratings and, in that way, the possibility of selection of the alternative along with neglecting of other requirements is avoided.

Evaluation of the quality of website given in this manuscript is performed by using the WS PLP method, whose computational procedure can accurately be presented as follows [27]:

$$S_i' = \sum_{j=1}^{n} w_j \, r_{ij} - \gamma \, c_i , \tag{1}$$

where S_i' denotes the adjusted overall performance rating of the alternative *i*; $S_i' \in [-1,\ 1]$, *w_j* is the weight of the criterion *j*, r_{ij} is the normalized performance rating of the alternative *i* with respect to the criterion *j*, *n* is the number of the criteria, c_i is the compensation coefficient; $c_i > 0$, γ is the coefficient; $\gamma = [0,1]$.

The normalized performance ratings in the WS PLP approach should be calculated as follows:

$$r_{ij} = \frac{x_{ij} - x_{0j}}{x_j^{+} - x_j^{-}} , \tag{2}$$

where: x_{ij} denotes performance rating of alternative *i* in relation to the criterion *j*, x_{0j} denotes the preferred performance rating of the criterion *j*, x_j^{+} and x_j^{-} denotes the best and the worst ratings of criterion *j*, respectively, and they are calculated as follows:

$$x_j^{+} = \begin{cases} \max\limits_i x_{ij} \middle| j \in \Omega_{\max} \\ \min\limits_i x_{ij} \middle| j \in \Omega_{\max} \end{cases} , \text{ and} \tag{3}$$

$$x_j^{-} = \begin{cases} \min\limits_i x_{ij} \middle| j \in \Omega_{\max} \\ \max\limits_i x_{ij} \middle| j \in \Omega_{\max} \end{cases} . \tag{4}$$

The preferred performance rating x_{0j} of the evaluation criterion *j* should be set on the basis of the preferences made by the decision maker/respondent. If he/she does not have any preferences in relation to any criterion, such a criterion should be determined as follows:

$$x_{oj} = \begin{cases} \max\limits_i x_{ij} \middle| j \in \Omega_{\max} \\ \min\limits_i x_{ij} \middle| j \in \Omega_{\min} \end{cases} . \tag{5}$$

In the proposed approach, the alternatives whose S_i' is greater than or equal to zero form a set of the most appropriate alternatives, from which one should be selected.

The part $-\gamma c_i$ of Eq. (1) can be used to reduce the number of the alternatives or to fine tune the values of S_i' in the set of the most appropriate alternatives. However, its use is not mandatory.

The meaning and the usage of $-\gamma c_i$ part of Eq. (1) are explained in detail in Stanujkic and Zavadskas [27].

4 The Framework for Evaluating the Website Quality

Different authors have identified different phases in the multiple criteria evaluation process. Many of them have highlighted some as those important for the further consideration of the proposed approach, such as [66-68]:

- the selection of relevant evaluation criteria,
- the determination of the criteria weights, and
- the aggregation and selection phase.

In the following text every phase is further elaborated.

4.1 The Selection of Relevant Evaluation Criteria

The choice of the appropriate set of the evaluation selection criteria is very important for the successful solving of each MCDM problem. The evaluation of the website quality performed by Kaya [69] is based on the four groups of criteria which are as follows: information quality, service quality, system quality and vendor quality. Every of the mentioned criteria group consists of two additional criteria that better explain the considered aspect of website quality. Kaya and Kahraman [70] in their manuscript based the evaluation of the e-banking websites on the two dimensions: (1) customer service quality criteria that involve: product quality, reliability, responsiveness, competence and access; and (2) online systems quality criteria that involves: information content, ease of use and security. Akincilar and Dagdeviren [50] evaluated the quality of the hotel websites by using the following criteria that is further elaborated in the certain number of sub-criteria: customer oriented, technology oriented, marketing oriented, security oriented and other factors.

The common denominator for all previous approaches is that the evaluation process is based on the greater number of criteria and sub-criteria. Using a larger number of criteria usually leads to the formation of more precise models, on the one hand, whereas on the other, a larger number of criteria can, however, be less desirable if some data should be collected through a survey. In contrast to the said, a smaller number of criteria can be much more efficient when some data should be collected through a survey, on the one hand, whereas on the other, the usage of more complex criteria is required sometimes, which can also lead to the forming of less accurate models. There is also a set of rules that a set of evaluation criteria should satisfy, such as: completeness, the operational ability, decomposability, non-redundancy and the minimum size [71].

Therefore, in this approach a proven set of evaluation criteria, which has been adopted from the Webby Awards[1], is proposed for evaluating the quality of the website:

- **Content (*C*)** - The content is the information provided on the website. It is not just the text, but also the music, the sound, the animation or a video – anything that communicates the body of knowledge of the website.
- **Structure and Navigation (*S*)** - The structure and navigation refer to the framework of the website, the organization of the content, the prioritization of the information and the method in which the website is navigated by the visitor. Websites with the good structure and navigation are consistent, intuitive and transparent.
- **Visual Design (*V*)** - Visual design is the appearance of the website. It is more than just a nicely designed homepage and it does not have to be the cutting edge or anything trendy. A good visual design is of a high quality, appropriate and relevant for the audience and the message it supports. It communicates a visual experience and may even take your breath away.
- **Functionality (*F*)** - Functionality is the use of technology on the website. Good functionality means that the website works well. It loads quickly, has live links and the different kinds of the new technology applied to it are functional and relevant for the intended audience.
- **Innovation (*I*) –** Innovation is the idea that is completely new and contributes to the better functioning or visual design of some website.
- **Overall Experience (*O*)** - Demonstrating the fact that websites are frequently more or less than the sum of their parts, the overall experience encompasses the content, visual design, functionality, interactivity and the structure and navigation, on the one hand, also including the intangibles that make the visitor stay on it or leave it, on the other.

In this case, the criterion Innovation is omitted from the evaluation procedure. The reason for excluding of mentioned criterion from the assessment of the websites` quality is twofold. Firstly, examination of the research studies shows that evaluation of the quality of websites mainly relies on the criteria that roughly could be categorized as follows: content, navigation, visual appeal, multimedia

[1] http://webbyawards.com/judging-criteria/

and ease of use [72-74]. Secondly, the Innovation criterion is omitted from the assessment because its meaning could be confusing to the respondents. Besides, the Overall Experience criterion is used for checking the reliability of the collected data and the evaluation results.

The given set of criteria is reliable because it is proved to be useful for the estimation of the website quality. Besides, its application facilitates decision-making process, which is liberated from the large number of criteria that complicates the procedure.

4.2 The Determination of Criteria Weights

The determination of the significance of criteria is of great importance in multiple criteria evaluation models, which is why a number of methods for their determination have been proposed, such as: the Analytic Hierarchy Process (AHP), developed by Saaty [35-36], the Step-wise Weight Assessment Ratio Analysis (SWARA) technique, developed by Kersuliene et al. [44], and the pivot pairwise relative criteria importance assessment method for determining the weights of criteria (PIPRECIA) [75].

In this approach, the preferred ratings obtained from the respondents are used for the determining of the significance of the evaluation criteria, i.e. the weights of the criteria, as follows:

$$w_j = \frac{x_{0j}}{\sum_{l=1}^{n} x_{0l}} , \tag{6}$$

where x_{0j} denotes the preferred performance rating of the criterion j and x_{0l} denotes the preferred performance ratings.

4.3 The Aggregation and Selection Phase

In this approach, the WS PLP approach is chosen for aggregating the ratings collected during the survey for each one of the respondents separately. This means that the *K* ranking orders will be formed in the case of a survey that includes *K* respondents.

There are several ways for the evaluation of the alternatives. The first approach is the theory of dominance [42] based on the number of the occurrences of some alternative in the first position. The number of the occurrences in the other positions, the second, and the third and so on, can also be significant for a more precise evaluation.

The Weighted Averaging (WA) operator, proposed by Harsanyi [76], can be used as an alternative way for the transformation of the *K* ranking list into the resulting ranking list.

5 A Case Study

In order to verify the proposed framework, a limited research related to the quality of the websites of the three telecommunication companies in Serbia was conducted.

The e-mail survey was carried out, and e-mails were sent to more than 80 pre-selected email addresses. The positive feedback response was obtained from 51 respondents, out of which 45 surveys were selected as properly completed. Respondents appraised before mentioned websites against the given set of criteria using the marks 1 to 5 (1 as the worst and 5 as the best mark).

In order to demonstrate not only the efficiency, but also the simplicity of the use of the proposed framework, the evaluation performed on the basis of the five randomly selected respondents is presented below.

The ratings obtained from the selected respondents are shown in Tables 1 to 5.

Table 1

The ratings and the preferred ratings obtained from the first of the five selected respondents

Criteria	C_1	C_2	C_3	C_4	C_5	Overall
ppr	3	4	5	2	1	
A_1	2	4	5	3	2	4
A_2	2	3	3	4	2	1
A_3	3	4	5	2	1	4

Table 2

The ratings and the preferred ratings obtained from the second of the five selected respondents

Criteria	C_1	C_2	C_3	C_4	C_5	Overall
ppr	3	5	5	4	1	
A_1	2	4	5	3	2	4
A_2	2	3	4	2	2	2
A_3	3	4	5	2	1	4

Table 3

The ratings and the preferred ratings obtained from the third of the five selected respondents

Criteria	C_1	C_2	C_3	C_4	C_5	Overall
ppr	3	4	5	2	1	
A_1	2	4	5	3	2	2
A_2	2	3	3	4	2	1
A_3	3	4	5	2	1	4

Table 4
The ratings and the preferred ratings obtained from the fourth of the five selected respondents

Criteria	C_1	C_2	C_3	C_4	C_5	Overall
ppr	3	4	3	3	1	
A_1	2	4	5	3	2	3
A_2	2	3	3	4	2	1
A_3	3	4	1	4	1	2

Table 5
The ratings and the preferred ratings obtained from the fifth of the five selected respondents

Criteria	C_1	C_2	C_3	C_4	C_5	Overall
ppr	3	4	5	2	1	
A_1	2	4	5	3	2	4
A_2	2	3	3	4	2	1
A_3	3	4	5	2	1	3

The normalized ratings and the weights of the criteria obtained on the basis of the ratings and the preferences of the first respondent by using Eq. (2) and Eq. (6) are shown in Table 6.

Table 6
The normalized ratings and the weighting of the criteria obtained on the basis of the ratings and the preferences of the first of the five selected respondents

Criteria	C_1	C_2	C_3	C_4	C_5
w_j	0.20	0.27	0.33	0.13	0.07
A_1	-1	0	0	1	1
A_2	-1	-1	-2	2	1
A_3	0	0	0	0	0

The weighted normalized ratings and the overall ratings obtained by using Eq. (1), as well as the ranking order obtained on the responses of the first respondent are accounted for in Table 7.

Table 7
The computational details obtained on the basis of the responses received from the first of the five selected respondents

Criteria	C_1	C_2	C_3	C_4	C_5	S_i	Rank
A_1	0.00	0.00	0.00	0.07	0.07	0.13	1
A_2	-0.20	-0.27	-0.33	0.13	0.07	-0.60	3
A_3	0.00	0.00	0.00	0.00	0.00	0.00	2

According to Table 7, the website of the telecommunication company labelled as A_1 is the best-ranked and the website of the company labelled as A_3 is the second best-placed.

The Overall parameter applied in the conducted survey is used for the purpose of determining the consistency of the responses obtained through the survey.

The results of the ranking obtained on the basis of the responses received from the first of the selected respondents are given in Table 8.

Table 8
The ranking of the websites on the basis of the Overall parameter obtained from the first of the five selected respondents

Alternatives	Overall	Rank
A_1	4	1
A_2	1	3
A_3	4	1

According to Tables 7 and 8, the ranking order obtained on the basis of the five evaluation criteria and the ranking order obtained on the basis of the Overall parameter are similar to one another, which confirms the consistency of the responses received from the first of the selected respondents.

The ranking orders of the considered websites obtained from the responses given by the five selected respondents are shown in Table 9.

Table 9
The ranking orders obtained on the basis of the responses received from the five selected respondents

Alternatives	DM_1	DM_2	DM_3	DM_4	DM_5	I	Rank
A_1	1	1	2	2	1	3	1
A_2	3	2	3	3	3	0	3
A_3	2	3	1	1	2	2	2

According to Table 9, the website A_1 is the most appropriate, based on the responses obtained from the five selected respondents because it has the greatest number of appearances in the first position.

The website A_3 is the runner up, with three appearances in the first position and two appearances in the second position, which is indicative of the fact that this particular website could surpass the website A_1, based on certain adjustments to that website.

The ranking order obtained on the basis of appearance in the first position is also confirmed by using the WA operator, as Table 10 shows.

Table 10
The ranking of the websites on the basis of the WA operator and the ratings obtained from the five selected respondents

Alternatives	S_i	Rank
A_1	-0.14	1
A_2	-0.57	3
A_3	-0.27	2

The value -0.14 of the overall rating of the website A_1 indicates the fact that its overall quality is below the respondents' expectations. By analyzing their responses, the website A_1 was found to mainly fail in relation the Content criterion, as is shown in Table 11.

Table 11
The weighted normalized ratings of the website A_1, obtained from the selected respondents

Criteria	C_1	C_2	C_3	C_4	C_5
DM_1	0.00	0.00	0.00	0.07	0.07
DM_2	-0.17	-0.28	0.00	-0.22	0.02
DM_3	-0.20	0.00	0.00	0.07	0.07
DM_4	-0.21	0.00	0.11	0.00	0.07
DM_5	-0.20	0.00	0.00	0.07	0.07

The ranking orders of the considered websites obtained on the basis of the responses received from all the surveyed respondents are accounted for in Table 12.

Table 12
The ranking orders obtained on the basis of all the respondents' responses

Alternatives	I	Rank	%
A_1	30	1	67%
A_2	6	3	13%
A_3	9	2	20%

According to Table 12, the website A_1 is still the first-placed, with 30 appearances in the first position, which is 67% if expressed in percentage.

The dominance of the website A_1 could also be confirmed by using the WA operator, as well as by the ranking based on the Overall parameter, as is shown in Table 13 and Table 14.

Table 13
The ranking orders of the websites on the basis of the WA operator

Alternatives	S_i	Rank
A_1	-0.35	1
A_2	-0.72	3
A_3	-0.64	2

Table 14
The ranking orders obtained on the basis of the Overall parameter

Alternatives	I	Rank	%
A_1	10	48%	1
A_2	5	24%	3
A_3	6	29%	2

Additionally, in order to verify the reliability of the proposed approach, sensitivity analysis was conducted with the comparison of the ranking results of the WS PLP method with 5 other well-known MCDM methods (TOPSIS, VIKOR, ARAS, MULTIMOORA and WASPAS). The obtained results are shown in Figure 1.

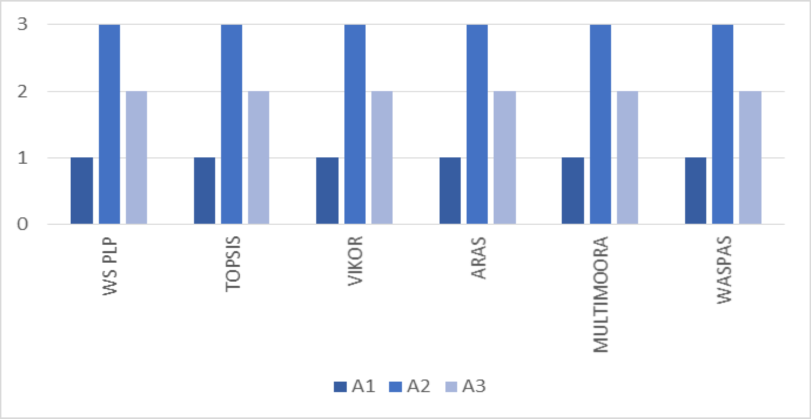


Figure 1

Results of sensitivity analysis in comparison of ranks with other MCDM methods

The comparison results of the conducted sensitivity analysis indicate that proposed approach have the same ranking results as well as other methods, which confirms that the proposed approach is reliable and adequate when it comes to website evaluation.

Conclusion

A new approach to the evaluation and ranking of websites from the viewpoint of their respective visitors is the subject matter of consideration in this manuscript. The main goal of the paper is proposing the WS PLP method as a suitable technique for the website quality evaluation that could be used by the customers or the companies. The capability of the WS PLP method application in the mentioned case is dual. This method could help customers to select the website according to their needs, but also, companies could evaluate the quality of their websites, and determine its position in relation to the main competitors as well providing the information regarding the main aspects that should be improved.

The approach proposed herein also has significant similarities to the proven approaches of the determination of customers' satisfaction, such as SERVQUAL or models similar to that one. The main similarity to the SERVQUAL model is related to the fact that the proposed approach also uses gaps between expectations and perceptions. By introducing the *ppr* values the DM's could better to express their expectations and perform the comparisons with the given performances of the alternatives. Also, the DM is in the position to choose whether he/she wants to give priority to the alternative that is better matching with his/her expectations or to the one that is the best of all.

The before mentioned authors such as Kaya [69], Kaya and Kahraman [70] and Akincilar and Dagdeviren [50] have based their evaluation of the website quality on the greater number of criteria and sub-criteria that leads to the more complex procedure and that is very complicated for the application in the practice by ordinary users. Approach proposed in this paper, which is based on the significantly smaller number of criteria, enables forming of the simpler questionnaires that could be more appropriate when preferences and ratings are collected through conducting surveys with ordinary respondents, i.e. those unprepared in advice for surveying.

The principal disadvantage of this approach is the use of the crisp numbers. Despite that, the proposed approach proved to be useful when it comes to solving problems of websites evaluation.

In the end, the usability of the proposed approach is tested and verified in the case study on the evaluation of the websites of the telecommunication companies. Also, additional verification of the proposed approach is demonstrated in the conducted sensitivity analysis. The obtained result has confirmed that the proposed approach is proven to be reliable and adequate for solving problems of website quality evaluation.

References

[1] Taylor, M. and Kent, M. L., (2004) Congressional web sites and their potential for public dialogue. *Atlantic Journal of Communication*, 12(2), pp. 59-76

[2] Cebi, S., (2013) Determining importance degrees of website design parameters based on interactions and types of websites. *Decision Support Systems*, 54(2), pp. 1030-1043

[3] Ibrahim, M., (2015) Evaluating hotel websites as a marketing communication channel A dialogic perspective. *Information Development*, 32(3), pp. 1-10

[4] Lee, Y. and Kozar, K. A., (2006) Investigating the effect of website quality on e-business success: An analytic hierarchy process (AHP) approach. *Decision support systems*, 42(3), pp. 1383-1401

[5] Stanujkic, D. Kazimieras Zavadskas, E. and Tamošaitienė, J., (2015) An approach to measuring website quality in the rural tourism industry based on Atanassov intuitionistic fuzzy sets. *Ekonomie a Management*, 18(4), pp. 184-199

[6] Al-Manasra, E. Khair, M. Zaid, S. A. and TaherQutaishat, F., (2013) Investigating the Impact of Website Quality on Consumers’ Satisfaction in Jordanian Telecommunication Sector. *Arab Economic and Business Journal*, 8(1), pp. 31-37

[7] Bai, B. Law, R. and Wen, I., (2008) The impact of website quality on customer satisfaction and purchase intentions: Evidence from Chinese online visitors. *International Journal of Hospitality Management*, 27(3), pp. 391-402

[8] Lin, H. F. (2007) The impact of website quality dimensions on customer satisfaction in the B2C e-commerce context. *Total Quality Management and Business Excellence*, 18(4), pp. 363-378

[9] Kim, S., and Stoel, L., (2004) Apparel retailers: website quality dimensions and satisfaction. *Journal of Retailing and Consumer Services*, 11(2), pp, 109-117

[10] Parasuraman, A. Zeithaml, V. A. and Berry, L. L., (1985) A conceptual model of service quality and its implications for future research. *Journal of Marketing*, 49, pp. 41-50

[11] Barnes, S. J. and Vidgen, R., (2001) An evaluation of cyber-bookshops: the WebQual method. *International Journal of Electronic Commerce*, 6(1), pp. 11-30

[12] Barnes, S. and Vidgen, R., (2000) *WebQual: an exploration of website quality*. ECIS 2000 Proceedings, 74

[13] Parasuraman, A. Zeithaml, V. A. and Malhotra, A., (2005) ES-QUAL a multiple-item scale for assessing electronic service quality. *Journal of service research*, 7(3), pp. 213-233

[14] Loiacono, E. T. Watson, R. T. and Goodhue, D. L., (2002) WebQual: A measure of website quality. *Marketing theory and applications*, 13(3), pp. 432-438

[15] Barnes, S. J. and Vidgen, R. T., (2003) *Assessing the quality of a cross-national e-government Web site: a study of the forum on strategic management knowledge exchange*. In System Sciences, 2003, Proceedings of the 36th Annual Hawaii International Conference on (pp. 10-pp) IEEE

[16] Barnes, S. J. and Vidgen, R., (2003) Measuring Website Quality Improvements: A Case Study of the Forum on Strategic Management Knowledge Exchange. *Industrial Management and Data System*, 103(5), pp. 297-309

[17] Shchiglik, C. and Barnes, S. J., (2004) Evaluating website quality in the airline industry. *Journal of Computer Information Systems*, 44(3), pp. 17-25

[18] Park, Y. A. Gretzel, U. and Sirakaya-Turk, E., (2007) Measuring web site quality for online travel agencies. *Journal of Travel & Tourism Marketing*, 23(1), pp. 15-30

[19] Park, H. and Baek, S., (2007) *Measuring service quality of online bookstores with WebQual*. In International Conference on Human-Computer Interaction, Springer, Berlin Heidelberg, pp. 95-103

[20] Keeney, R. and Raiffa, H., (1976) *Decision with multiple objective: Preference and value tradeoffs*. Wiley, New York

[21] Beynon, M. J., (2006) The role of the DS/AHP in identifying inter-group alliances and majority rule within group decision making. *Group decision and negotiation*, 15(1), pp. 21-42

[22] Korhonen, P., (2005) *Interactive methods. In Multiple criteria decision analysis: state of the art surveys*. Springer, New York

[23] Karabasevic, D., Popovic, G., Stanujkic, D., Maksimovic, M. and Sava, C., (2019) An approach for hotel type selection based on the single-valued intuitionistic fuzzy numbers. *International Review* (1-2), pp. 9-16

[24] Nunić, Z., (2018) Evaluation and selection of Manufacturer PVC carpentry using FUCOM-MABAC model. *Operational Research in Engineering Sciences: Theory and Applications*, 1(1), pp. 13-28

[25] Pamučar, D., Lukovac, V., Božanić, D. and Komazec, N., (2018) Multi-criteria FUCOM-MAIRCA model for the evaluation of level crossings: case study in the Republic of Serbia. *Operational Research in Engineering Sciences: Theory and Applications*, 1(1), pp. 108-129

[26] Vesković, S., Stević, Ž., Stojić, G., Vasiljević, M., and Milinković, S., (2018) Evaluation of the railway management model by using a new integrated model DELPHI-SWARA-MABAC. *Decision Making: Applications in Management and Engineering*, 1(2), pp. 34-50

[27] Stanujkic, D. and Zavadskas, E. K., (2015) A modified weighted sum method based on the decision-maker's preferred levels of performances. *Studies in Informatics and Control*, 24(4), pp. 61-470

[28] Jones, C. and Kim, S., (2010) Influences of retail brand trust, off-line patronage, clothing involvement and website quality on online apparel shopping intention. *International Journal of Consumer Studies*, 34(6), pp. 627-637

[29] Wang, L. Law, R. Guillet, B. D. Hung, K. and Fong, D. K. C., (2015) Impact of hotel website quality on online booking intentions: eTrust as a mediator. *International Journal of Hospitality Management*, 47, pp. 108-115

[30] Yang, K. Li, X. Kim, H. and Kim, Y. H., (2015) Social shopping website quality attributes increasing consumer participation, positive eWOM, and co-shopping: The reciprocating role of participation. *Journal of Retailing and Consumer Services*, 24, pp. 1-9

[31] Calero, C. Ruiz, J. and Piattini, M., (2005) Classifying Web Metrics Using the Web Quality Model. *Online Information Review*, 29(3), pp. 227-248

[32] Churchman, C. W. and Ackoff, R. L., (1954) An approximate measure of value. *Journal of the Operations Research Society of America*, 2(2), pp. 172-187

[33] MacCrimon, K. R., (1968) *Decision Marking Among Multiple-Attribute Alternatives: A Survey and Consolidated Approach*. RAND memorandum, RM-4823-ARPA

[34] Roy, B., (1968) Classement et choix en présence de points de vue multiples. Revue française d'automatique, d'informatique et de recherche opérationnelle. *Recherche opérationnelle*, 2(1), pp. 57-75

[35] Saaty, T. L., (1977) A scaling method for priorities in hierarchical structures. *Journal of mathematical psychology,* 15(3), pp. 234-281

[36] Saaty, T. L., (1980) *The Analytic Hierarchy Process*. McGraw Hill Company, New York

[37] Hwang, C. L. and Yoon, K., (1981) *Multiple Attribute Decision Making Methods and Applications*. Springer-Verlag, Heidelberg

[38] Brans, J. P. and Vincke, P., (1985) Note—A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making). *Management science*, 31(6), pp. 647-656

[39] Zavadskas, E. K. Kaklauskas, A. and Sarka V., (1994) The New Method of Multicriteria Complex Proportional Assessment of Projects. *Technological and Economic Development of Economy*, 1(3), pp. 131-139

[40] Opricovic, S., (1998) *Multicriteria optimization of civil engineering systems*. Faculty of Civil Engineering, Belgrade

[41] Grigoroudis, E. and Siskos, Y., (2002) Preference disaggregation for measuring and analysing customer satisfaction: The MUSA method. *European Journal of Operational Research,* 143(1), pp. 148-170

[42] Brauers, W. K. M., and Zavadskas, E. K., (2010) Project management by MULTIMOORA as an instrument for transition economies. *Technological and Economic Development of Economy*, 16(1), pp. 5-24

[43] Zavadskas, E. K. and Turskis, Z., (2010) A New Additive Ratio Assessment (ARAS) Method in Multicriteria Decision-Making. *Technological and Economic Development of Economy*, 16(2), pp. 159-172

[44] Kersuliene, V. Zavadskas, E. K. and Turskis, Z., (2010) Selection of rational dispute resolution method by applying new step-wise weight assessment ratio analysis (SWARA). *Journal of Business Economics and Management*, 11(2), pp. 243-258

[45] Ginevičius, R., (2011) A new determining method for the criteria weights in multicriteria evaluation. *International Journal of Information Technology & Decision Making*, 10(6), pp. 1067-1095

[46] Zavadskas, E. K. Turskis, Z. Antucheviciene, J. and Zakarevicius, A., (2012) Optimization of Weighted Aggregated Sum Product Assessment. *Elektronika ir elektrotechnika*, 122(6), pp. 3-6

[47] Krylovas, A. Zavadskas, E. K. Kosareva, N. and Dadelo, S., (2014) New KEMIRA method for determining criteria priority and weights in solving MCDM problem. *International Journal of Information Technology & Decision Making*, 13(6), pp. 1119-1133

[48] Ghorabaee, M. K. Zavadskas, E. K. Olfat, L. and Turskis, Z., (2015) Multi-Criteria Inventory Classification Using a New Method of Evaluation Based on Distance from Average Solution (EDAS). *Informatica*, 26(3), pp. 435-451

[49] Burmaoglu, S. and Kazancoglu, Y., (2012) E-government website evaluation with hybrid MCDM method in fuzzy environment. *International Journal of Applied Decision Sciences*, 5(2), pp. 163-181

[50] Akincilar, A. and Dagdeviren, M., (2014) A hybrid multi-criteria decision making model to evaluate hotel websites. *International Journal of Hospitality Management*, 36, pp. 263-271

[51] Ecer, F., (2014) A hybrid banking websites quality evaluation model using AHP and COPRAS-G: A Turkey case. *Technological and Economic Development of Economy*, 20(4), pp. 758-782

[52] Jain, D. Garg, R. and Bansal, A., (2015) A Parameterized Selection and Evaluation of E-Learning Websites Using TOPSIS Method. *International Journal of Research & Development*, 22(3), pp. 12-26

[53] Jain, D. Garg, R. Bansal, A. and Saini, K. K., (2016) Selection and ranking of E-learning websites using weighted distance-based approximation. *Journal of Computers in Education*, 3(2) pp. 193-207

[54] Kang, D. Jang, W. and Park, Y., (2016) Evaluation of e-commerce websites using fuzzy hierarchical TOPSIS based on ES-QUAL. *Applied Soft Computing*, 42, pp. 53-65

[55] Stanujkić, D. and Karabašević, D., (2018) An extension of the WASPAS method for decision-making problems with intuitionistic fuzzy numbers: a case of website evaluation. *Operational Research in Engineering Sciences: Theory and Applications*, 1(1), pp. 29-39

[56] Jakimavičius, M. Burinskiene, M., (2009) A GIS and multi-criteria based analysis and ranking of transportation zones of Vilnius city. *Technological and Economic Development of Economy* 15(1), pp. 39-48

[57] Shakouri, H. Nabaee, M. and Aliakbarisani, S., (2014) A quantitative discussion on the assessment of power supply technologies: DEA (data envelopment analysis) and SAW (simple additive weighting) as complementary methods for the "Grammar". *Energy*, 64, pp. 640-647

[58] Oltean-Dumbrava, C. Watts, G. Miah, A., (2016) Towards a more sustainable surface transport infrastructure: A case study of applying multi criteria analysis techniques to assess the sustainability of transport noise reducing devices. *Journal of Cleaner Production*, 112, pp. 2922-2934

[59] Chen, T. Y., (2012) Comparative analysis of SAW and TOPSIS based on interval-valued fuzzy sets: discussions on score functions and weight constraints. *Expert Systems with Applications*, 39(2), pp. 1848-1861

[60] Wang, P. Zhu, Z. and Wang, Y., (2016) A novel hybrid MCDM model combining the SAW, TOPSIS and GRA methods based on experimental design. *Information Sciences,* 345, pp. 27-45

[61] Karabašević, D. Stanujkić, D. Đorđević, B. and Stanujkić, A., (2018) The weighted sum preferred levels of performances approach to solving problems in human resources management. *Serbian Journal of Management*, 13(1), pp. 145-156

[62] Stanujkic, D. Karabasevic, D. and Zavadskas, E. K., (2017) A New Approach For Selecting Alternatives Based On The Adapted Weighted Sum And The Swara Methods: A Case Of Personnel Selection. *Economic Computation & Economic Cybernetics Studies & Research*, 51(3) pp. 39-56

[63] Vujić, D. Stanujkić, D. Urošević, S. and Karabašević, D., (2016) An approach to leader selection in the mining industry based on the use of weighted sum preferred levels of the performances method. *Mining and Metallurgy Engineering Bor*, 4, pp. 53-62

[64] Stanujkić, D. Milanović, D. Magdalinović, S. and Jovanović, I., (2017) An approach to the evaluation of froth flotation reagents based on the use of the SWARA and WS-PLP methods. *Mining and Metallurgy Engineering Bor*, 3-4, pp. 103-110

[65] Stanujkic, D. Magdalinovic, N. and Jovanovic, R., (2013) A multi-attribute decision making model based on distance from decision maker's preferences, *Informatica* 24(1), pp. 103-118

[66] Čupić, M. Tummala, R. and Suknović, M., (2003) *Decision-making–Formal approach*. Faculty of Organizational Sciences, Belgrade (In Serbian)

[67] Schoenfeld, A. H., (2010) *How we think: A theory of goal-oriented decision making and its educational applications*. Routledge, New York

[68] Stanujkić, D. Đorđević, B. and Đorđević, M., (2013) Comparative analysis of some prominent MCDM methods: A case of ranking Serbian banks. *Serbian Journal of Management*, 8(2), pp. 213-241

[69] Kaya, T., (2010) Multi-attribute evaluation of website quality in E-business using an integrated fuzzy AHPTOPSIS methodology. *International Journal of Computational Intelligence Systems*, 3(3), pp. 301-314

[70] Kaya, T. and Kahraman, C., (2011) A fuzzy approach to e-banking website quality assessment based on an integrated AHP-ELECTRE method. *Technological and Economic Development of Economy*, 17(2), pp. 313-334

[71] Keeney, R. L. and Gregory, R. S., (2005) *Selecting attributes to measure the achievement of objectives*. Operations Research, 53(1), pp. 1-11

[72] Moustakis, V., Litos, C., Dalivigas, A. and Tsironis, L., (2004) Website Quality Assessment Criteria. *ICIQ*, pp. 59-73

[73] Kincl, T. and Štrach, P., (2012) Measuring website quality: asymmetric effect of user satisfaction. *Behaviour & Information Technology*, 31(7), pp. 647-657

[74] Rocha, Á., (2012) Framework for a global quality evaluation of a website. *Online Information Review*, 36(3), pp. 374-382

[75] Stanujkic, D. Zavadskas, E. K. Karabasevic, D. Smarandache, F. and Turskis, Z., (2017) The use of the pivot pairwise relative criteria importance assessment method for determining the weights of criteria. *Romanian Journal for Economic Forecasting*, 20(4), pp. 116-133

[76] Harsanyi, J. C., (1995) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4), pp. 309-321

On-Board Diagnostic-based Positioning as an Additional Information Source of Driver Assistant Systems

Tibor Busznyák[1], Gergő Pálfi[2], István Lakatos[3]

[1] Széchenyi István University, Egyetem tér 1, H-9026 Győr, Hungary
busznyak.tibor@sze.hu

[2] Vehicle Engineering MSc, Egyetem tér 1, H-9026 Győr, Hungary
drlakatosi@ga.sze.hu

[3] Széchenyi István University Egyetem tér 1, H-9026 Győr, Hungary
lakatos@sze.hu

Abstract: Nowadays the role of driver assistant systems is becoming more and more pertinent in the vehicle industry. The increasing automation level presents opportunity to involve the driver in any kind of network such as network among vehicles (V2X) and into vehicle and infrastructure connections. The essence of automation in case of the vehicle industry is to facilitate for the vehicle to provide services based on multi-information sources to help the driver because in modern transport systems the most cardinal criteria are environment friendly, plannable, cost efficient and safety travel. Thus, the stability of data transport plays the key role in case of connected vehicles. The following sections will present an opportunity in a detailed way to determine the vehicle's position relying on the vehicle's On-Board Diagnostic system based on the developed algorithm and will show how this method can help the driver assistant system's reliable operation as well.

Keywords: vehicle; RTK-GPS; OBD; satellite; V2X

1 Introduction

Innovation can be chosen for the password of the 21st Century vehicle industry. Increasingly severe environmental standards, safety and traffic optimization criteria encourage the automotive industry to devise more and more complex solutions.

Smart city conceptions have come to the forefront, the importance of sustainable traffic has become fundamental. Online communication systems, consumption optimization, autonomous transportation systems and transport conceptions realization are very complex and challenging tasks, which require cooperation from more sectors such as the infocommunication industry, urban development and automotive industry [11] [12] [13]. Communication flow among the components of traffic determines the basics of automated transport [16] [17] [18]. Based on these facts it can be seen the stability of traffic data has a key-role in this complex system.

As the quality of the data is determined by the content of the information, there is a growing demand for the components involved in the transport as well. The essence of automation is that the vehicle (along with the service infrastructure) can provide a quality service based on the multiple information sources, as nowadays the main priorities are environmentally conscious, planned and safe travel.

The current traffic trend in our world is determined by the growing number of vehicles. Huge traffic can be expected near peak times in big cities. In this case, the spaces become narrower for the driver and the roads become more and more crowded. These conditions also test the drivers, as their proper handling requires a lot of routine and concentration. Driving assistant systems offer a great and effective solution to this [5] [8]. Automation provides the opportunity to involve the vehicle in a network to implement a vehicle-to-infrastructure (V2X-Vehicle to Everything) connection [9] [10].

The network brings together all the vehicles that are involved in this way of transport, providing information on the routes to be crawled, thus congestion and accidents are avoidable. Setting up such a system can greatly contribute to secure, predictable traffic. Vehicle-to-Everything (V2X) refers to the connection between the vehicle and its surrounding units. The surrounding infrastructure in collaboration with the Intelligent Transportation System (ITS) can create more optimally operating traffic [7].

Therefore, mapping potential databases that can provide information as the basis for these systems is an important task.

2 Concerning GPS, the First Basic System

GPS (Global Positioning System or american satellite system), or GNSS (Global Navigational Satellite System) based positioning is a central question of today. It is a world wide global source of information. Developments, new horizons, new possibilities are known, from military uses to weather forecasts [4]. All regions are building their own satellite systems, adding together the different systems, thus resulting in GNSS services [6]. Advanced satellite systems:

- GPS – USA,
- GLONASS – Russian,

- Galileo – EU,
- Beidou – Chinese.

We have to mention the QZSS (Quasi-Zenith Satellite System) from Japan and the Indian IRNSS (Indian Regional Navigation Satellite System) system as well. In Hungary, both of the latest are rare or not available. The most advanced is the American satellite system. In Hungary, the American and the Russian systems are frequented. For visibility, an observer point in Győr, Hungary:

- Latitude: N 47°41’2”; Longitude: E 17°38’6”

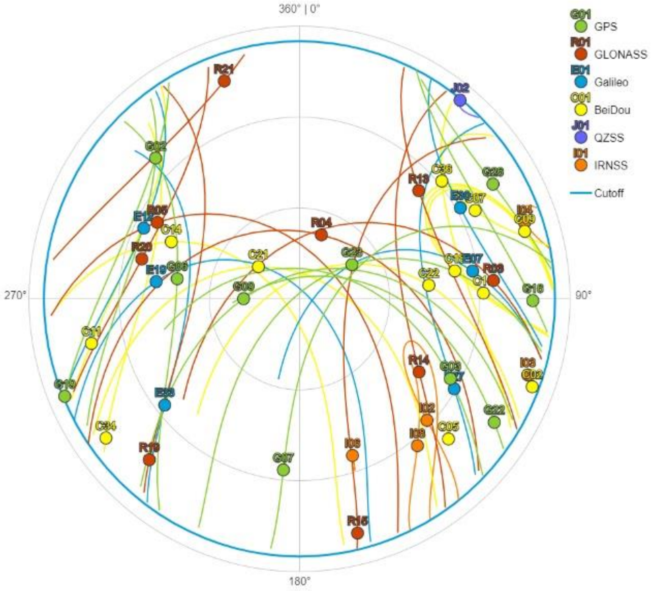


Figure 1

Sky plot of different satellite systems, observer point: Győr, Hungary

(source: www.gnssplanningonline.com, opened: 2019.03.17)

Figure 1 shows the availability of the different satellite systems. Observer point is in the city of Győr, Hungary. QZSS and IRNSS systems are not or rarely available, J02 is on the first qadrant sector with the only one member of QZSS. The IRNSS satellites are downside of the second quadrant sector assuming the specific ’form 8’. Online planning of the survey is very helpful. We can estimate the best interval for measurement. The boundary conditions of our survey will be presented later, including requirements of the number of satellites.

3 RTK – Real Time Kinematic Survey

RTK is a positioning method. It provides geodesic accuracy geographic points not just with fix point surveys, but with continuous topographic survey too. During the survey moving is possible. Geodesic accuracy is 3-4 cm horizontal, 5-6 cm vertical precision. RTK survey needs 5 satellites access at the same time, and connection with the base. Base is a pre-recorded point, it helps increase accuracy during survey Figure 2 shows the constellation of satellites. It was mounted on an atomic clock. Its delay is 1 secundum in 300 000-3 000 000 year. It depends on its regulation. In this case it is possible to measure time accurately. The satellite now has an accurate time stamp, communicates with the observer on Earth. This observer has a time stamp too and the difference between the two states gives the parameter 'elapsed time'. In a given medium, radio wave velocity is known, 300.000 km/s. The covered distance can be calculated. This is the radius of the sphere around the given satellite. 2 satellites and their spheres define a circle. Cutting a circle with another sphere gives 2 points, two further satellites increase the accuracy of the survey. To sum it all up, the RTK method works with 5 different satellites, and correction data from the base.

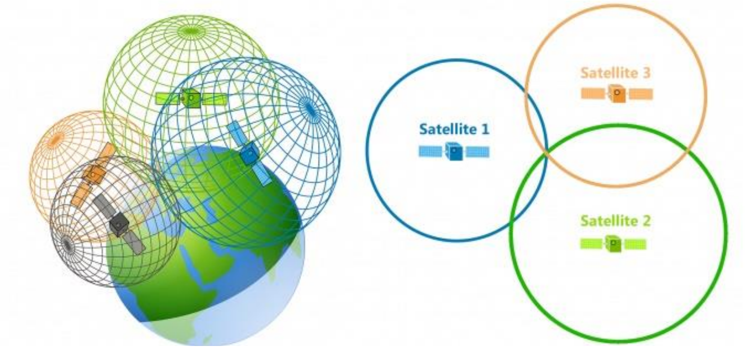


Figure 2

RTK method

(source: https://gisgeography.com/trilateration-triangulation-gps/, opened: 2019.03.17.)

Possible use of navigation and RTK method is very varied, for example drone control systems. These 'non piloted' or unmanned aerial vehicles (UAV) are qualified to detect non or hardly accessible surfaces and for aerial photoshooting using GPS navigation signals [14]. It is highly frequented in remote sensing and photogrammetry, in agricultural sciences for example vegetable covering of a given territory [3] [21]. NDVI, normalised difference vegetable index supplemented by RTK signal make it possible to analyze agronomic measurements [15]. RTK is capable of being used in the case of industrial facilities or monuments analyzis [19].

4 Foundation of OBD-based Positioning

There are several ways to get the databases, depending on whether you start from the vehicle or from the infrastructure. If a vehicle is almost always within an intelligent infrastructure, the number and complexity of the in-built systems can be greatly reduced as the information is provided by the surrounding infrastructure instead of the vehicle's sensors. This is best suited for vehicles that travel on specific road sections or areas, such as public transport units. It is easier to build the infrastructure around these vehicles mostly because of the predefined routes. On the other hand, we consider the vehicle as an independent unit. Without infrastructure, relying on vehicle's sensors and sensor systems the vehicle can move virtually anywhere without the need for a deployed ITS server. The idea of simplifying the vehicle's own sensor system to rely on the OBD (On-Board Diagnostics) arises - which is basically a part of the vehicle - to perform auxiliary system tasks closely related to ITS.

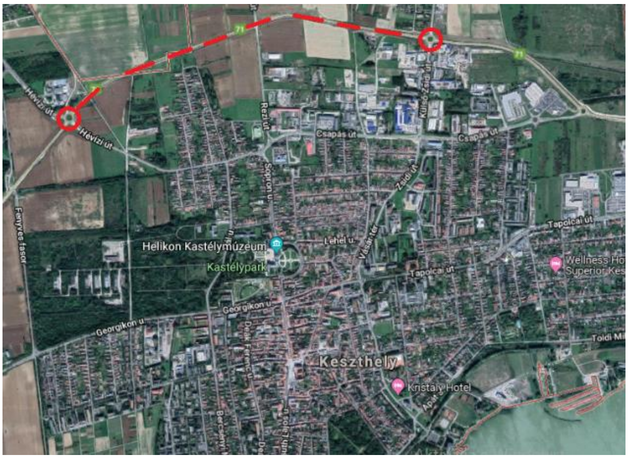


Figure 3
Route of the survey
(source of the original map: www.googlemaps.com)

Experiments - carried out in previous years [1] [2] - resulted in the mapping of digital tools that can be used in the investigation of the vehicle from the V2X point of view, focusing on the issue of fuel consumption. Figure 3 shows the map of the survey [20].

A method for defining the relationship between the fuel consumption (airflow rate) provided by the OBD system and the terrain produced by the RTK-GPS was successfully identified.

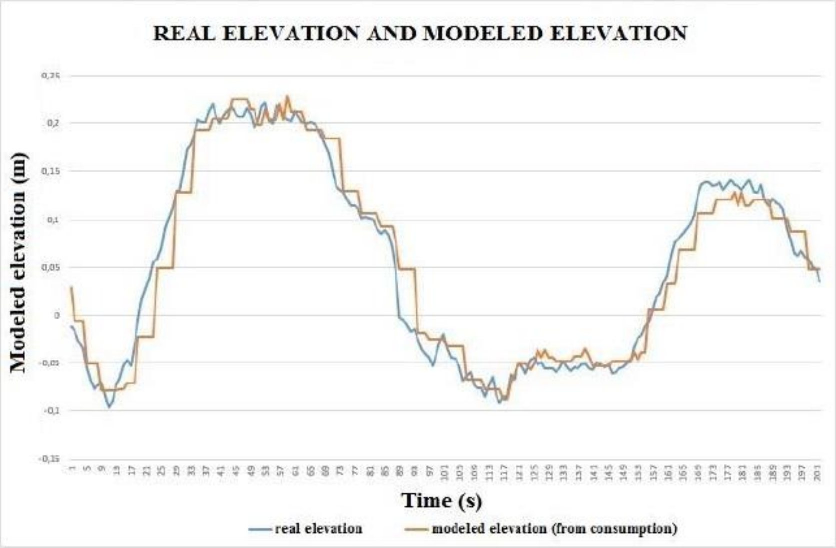


Figure 4

Real RTK elevation and modelled elevation trends

The determination coefficients given by the statistical analyzes at 3 different constant speeds are shown in the Table 1. It presents determination coefficients to describe the correlation between fuel consumption in case of different constant speeds and the elevation.

Table 1

Determination of coefficients at 3 different, standard speeds

Velocity [km/h]	Determination of coefficient
30	0,9549
40	0,9160
50	0,8370

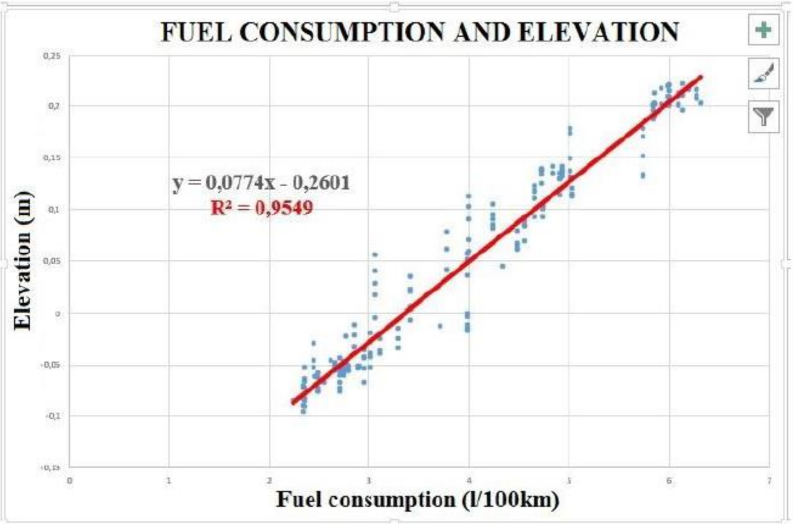


Figure 5

Connection between the measured elevation with RTK GPS and the fuel consumption provided by the vehicles OBD system in case of 30 km/h constant speed

These data show that we can deduce the fuel consumption of the vehicle on the basis of the known elevation conditions at high confederation levels (see Figure 4 and Figure 5).

5 Shortcomings of the High Precision GPS Infrastructure

Nowadays some automated vehicle functions use GPS based services. There are several shortcomings of the high precision GPS measurements in real traffic situations. A boundary condition system for GPS survey can be set up:

- Adequate number of GNSS satellites;
- Ensuring seamless connection with GNSS satellites;
- Base station providing clarification data;
- Ensure a seamless connection with a base station.

We can eliminate a high security risk if we can replace GPS with other alternatives. Such alternatives may be the combined use of the altitude database of routes and the availability of OBD data which is available for each vehicle individually to replace online connections.

6 Concept for Positioning Information Provided by On-Board Diagnostics

The close correlation between the above-mentioned relief and fuel consumption is based on the thought that the actual position of the vehicle can be determined based on the available reference database (altitude database of the route to be crawled) and fuel consumption data provided by On-Board diagnostics. The foundation of it is the monitoring of fuel consumption based on OBD data, a pre-measured and scaled database of the road section to be traversed by the vehicle and an algorithm that establishes a relationship between height coordinates and vehicle fuel consumption. The developed algorithm establishes a link between the elevation data points (MAP) and the OBD's output fuel consumption data points (Fig. 6), with its help the fuel consumption data point sampled from the current OBD can be retrieved on the reference database.

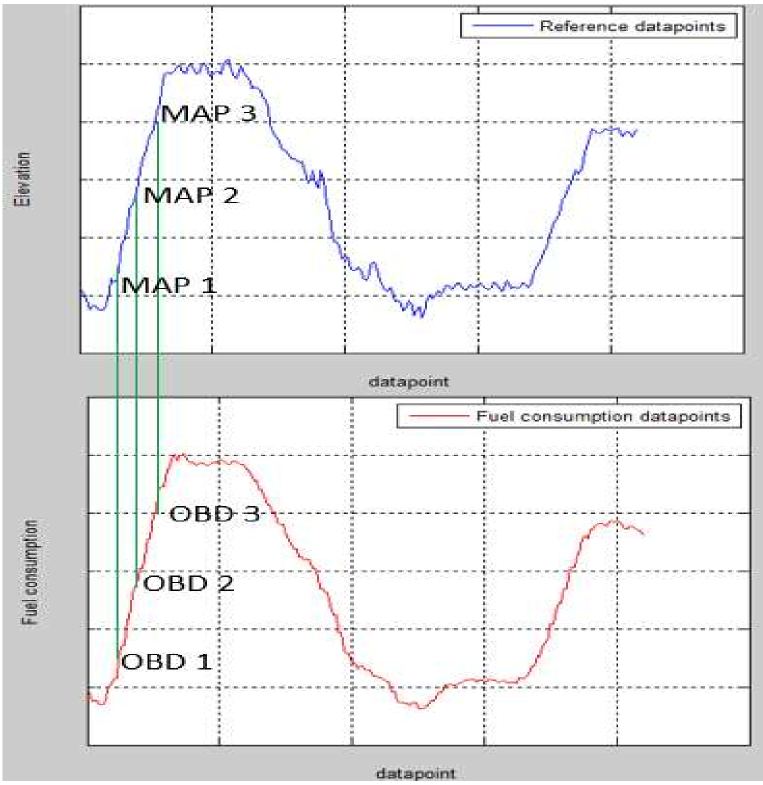


Figure 6
The matching of the two independent databases – fuel consumption based on the incoming OBD data and the altitude database measured by the GPS

As the first step, the algorithm examines the sampling of the fuel consumption data point from the OBD (n). These data points are compared by the algorithm with the reference database (MAPj).

The LS matrix includes square differences between fuel consumption sampled from OBD and MAP database data. The first element of the LS matrix summarizes the square deviations of fuel consumption data and reference data from the OBD based on the available OBD data points (4).

The LS matrices other than the first rows continue to examine the currently available OBD data points but extend the investigation to all available elements of the map database (4) by increasing the map database index with one from row to row in the LS matrix until the maximal index of the map database data is reached in the LS matrix's rows.

From the calculated squared deviations minimal value (contained by the LS matrix) is searched in case of each OBD data points. This minimal value will be contained by the S matrix (5). Selecting the minimum value and its location from the LS matrix (i.e. the row of the LS matrix containing the minimum value) and fixing the index of this row we can determine where the sum of the square deviations was the smallest from which the highest index value used in the map database (MAP) during the current calculation can be retrieved.

This value will coincide with the value of the instantaneous consumption data, so the altitude for the current consumption can be selected on the map. Height data have X and Y coordinates from the map database, thus the investigation can be extended to three dimensions.

When the available fuel consumption associated with the maximal MAP database index is reached, the OBD database (n) increasing is resulted.

$$OBD = \begin{pmatrix} OBD_1 \\ \vdots \\ OBD_n \end{pmatrix} \tag{1}$$

$$MAP = \begin{pmatrix} MAP_1 \\ \vdots \\ MAP_j \end{pmatrix} \tag{2}$$

$$LS_N^n = \begin{pmatrix} LS_1^n \\ \vdots \\ LS_N^n \end{pmatrix} \tag{3}$$

$$\begin{pmatrix} Ls_1 \\ \vdots \\ Ls_N \end{pmatrix} = \left\{ \begin{matrix} \sum_{\substack{i=1 \\ J=1}}^{n} \left(OBD_i - MAP_j\right)^2, If\ N = 1 \\ \sum^{n}{}_{\substack{i=1 \\ j=i-1+N}} \left(OBD_i - MAP_j\right)^2, If\ \ N \neq 1 \end{matrix} \right\} \tag{4}$$

Where:

- OBD_i = The 'i.' sampled OBD data point;
- MAPj = The 'j.' element of reference database;
- n = NUmber of OBD data points sampled to the actual time;
- LS = Matrix of the sum of the squared deviations;
- LS_N = Current row number of the LS column matrix;
- N = Counts rows of the LS matrix.

$$S = LS_N|_{min} \tag{5}$$

Where:

- S = Minimum of the sum of the squared deviations.

7 Test Environment of the Developed Algorithm

For the test of the developed algorithm a Matlab program was created (Figure 7).

During the test of the algorithm and the evaluation of the measurement data we considered the data point in case of 30 km/h constant speed - measured by RTK GPS - as the reference database. Due to the low sampling frequency of the OBD data logger device in case of 40 and 50 km/h constant speed measurement less fuel consumption data are available than in case of the reference database. Thus, linear interpolation can be used in case of this higher speed measurements to extend the interval of the fuel consumption.

Based on the investigations beside 30 km/h the sum of stored square differences data from the LS matrix with selecting the least value in the data series the biggest index of the connected reference database can be selected. With this selected index from the reference database the actual position of the vehicle can be determined. Based on the selected indexes the convergence curve can be presented (Figure 8).

In case of the 30 km/h constant speed measurement from the 5th datapoint of the fuel consumption data coming from the vehicle's OBD system begins to converge the data to the indexes of the reference map database indexes, thus from the 5th point of the fuel consumption the vehicle's position is definable.

Based on the first incoming OBD consumption data, the algorithm is activated. The above mentioned least squares method determines which of the points in the map database is the one that best matches the given consumption data. It has been previously determined that the correlation between fuel consumption and altitude change is higher than 95% in case of 30 km /h constant speed. However, as shown in Figure 9, based on the first (OBD data 1) consumption data, the algorithm is not positioned in the correct location.

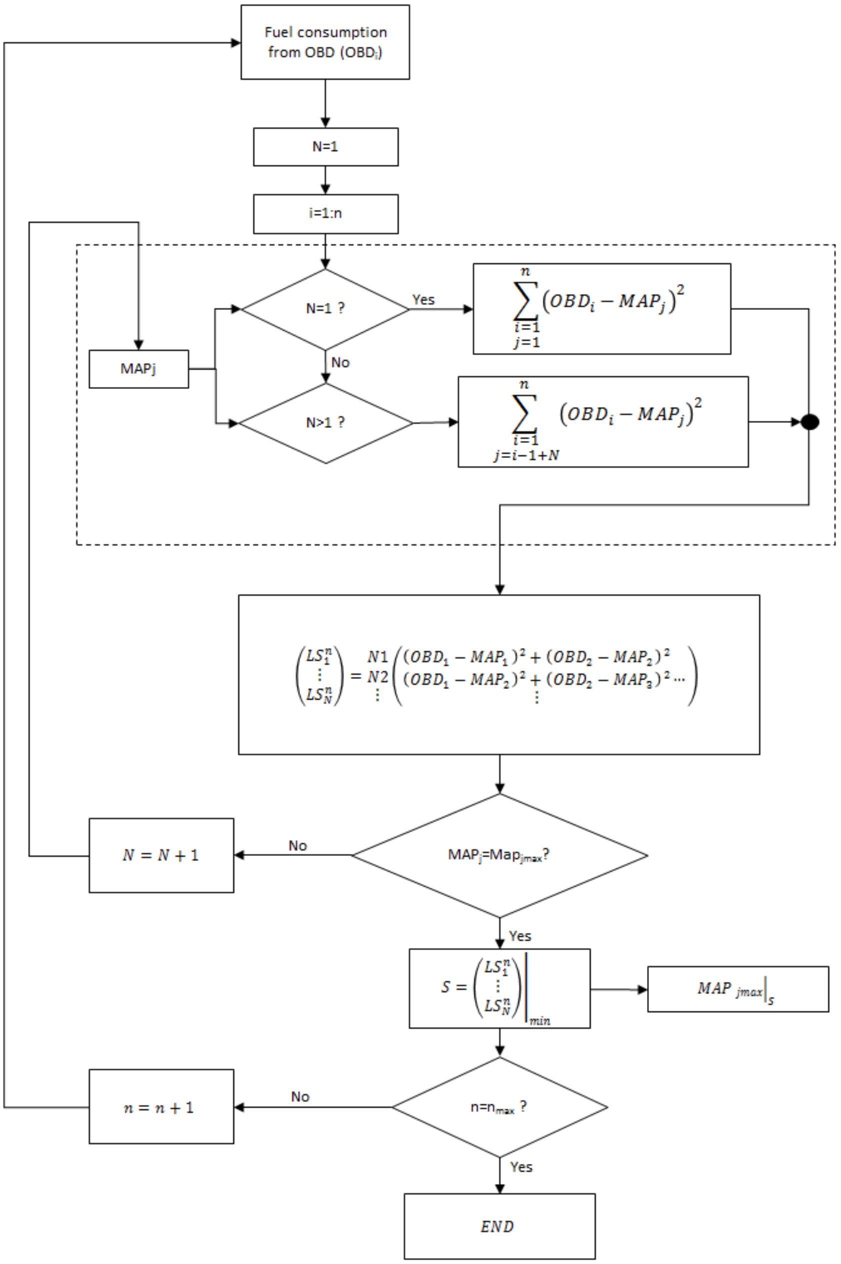

Fuel consumption from OBD (OBD_i)
N=1
i=1:n
N=1 ?
Yes
No
MAPj
N>1 ?
$\sum_{\substack{i=1\\j=1}}^{n}(OBD_i - MAP_j)^2$
$\sum_{\substack{i=1\\j=i-1+N}}^{n}(OBD_i - MAP_j)^2$
$\begin{pmatrix} LS_1^n \\ \vdots \\ LS_N^n \end{pmatrix} = \begin{matrix} N1 \\ N2 \\ \vdots \end{matrix}\begin{pmatrix} (OBD_1 - MAP_1)^2 + (OBD_2 - MAP_2)^2 \\ (OBD_1 - MAP_2)^2 + (OBD_2 - MAP_3)^2 \cdots \\ \vdots \end{pmatrix}$
$MAP_j=Map_{jmax}$?
No
$N = N + 1$
Yes
$S = \left.\begin{pmatrix} LS_1^n \\ \vdots \\ LS_N^n \end{pmatrix}\right|_{min}$
$MAP\ _{jmax}\big|_S$
$n=n_{max}$?
No
$n = n + 1$
Yes
END

Figure 7

Flow chart of the developed algorithm

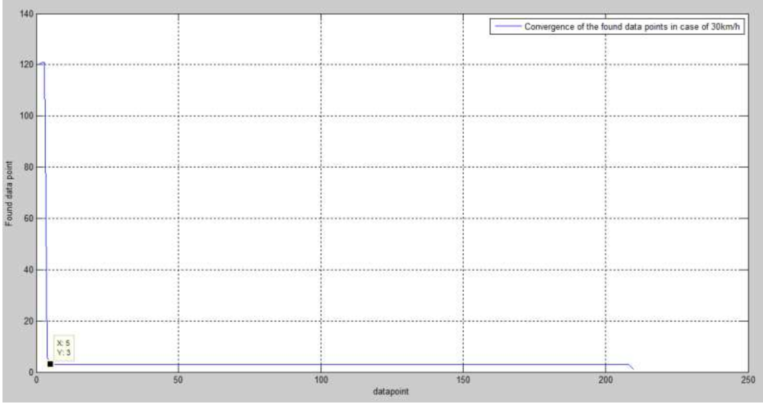


Figure 8

Convergence curve of the found data point indexes

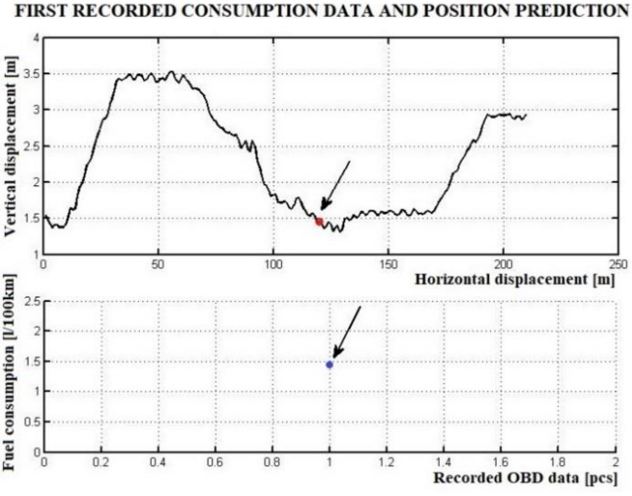


Figure 9

Identification of the first consumption data

During the operation of the algorithm consumption data arrive in time series. The recorded data follow each other at fixed intervals.

When analyzing multiple points, the algorithm steps over the map database. Suppose that we try to estimate the position based on 3 received consumption data. In this case, the algorithm overlaps the map database this means that we create data blocks with 3 points. From the 1, 2, and 3 consumption data for the first time the 1, 2 and 3 map data are analyzed after the 2, 3 and 4 data points and so on. Figure 10 shows that the estimated point is not yet good after 3 consumption data points.

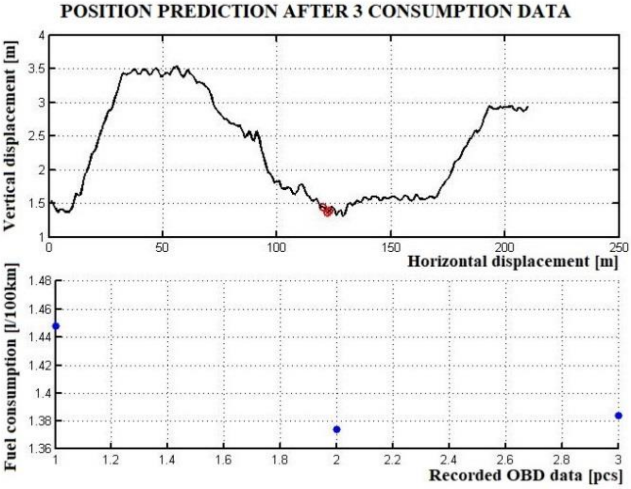


Figure 10

Identification of the third consumption data

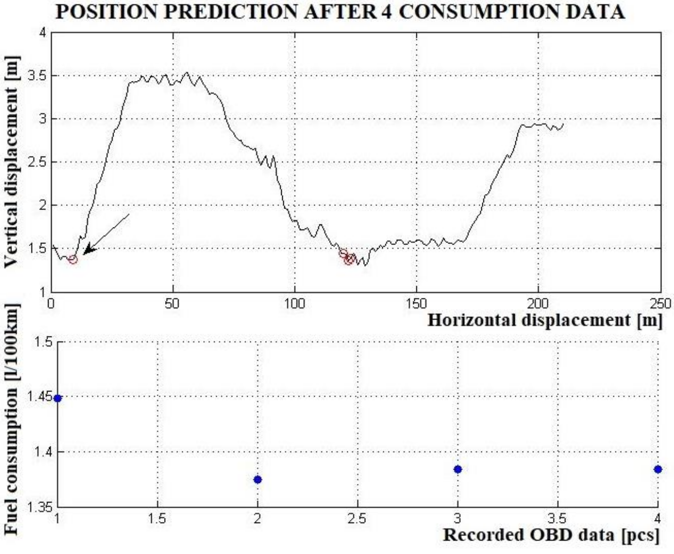


Figure 11

Identification of the fourth consumption data

Estimation based on the fourth incoming OBD consumption data brings the first adequate result. The algorithm then places the point to the beginning of the cycle to the point 4 of the map, as illustrated on Figure 11.

The algorithm then works smoothly. Building on each other guarantees the stability of the process. Since there is more data in the analysis in each iteration it is constantly built up and does not ignore the points that have already been identified. This is the result of the continuously overlapping method with the given number of data in the reference database (Figure 12).

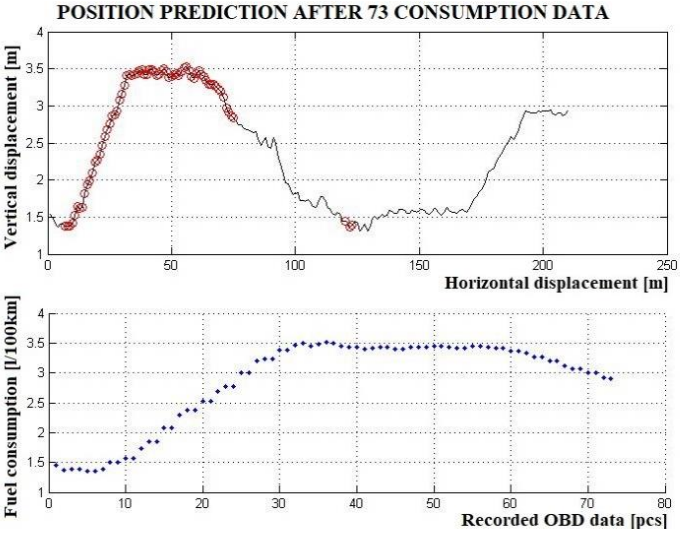


Figure 12

Illustration of the working of the algorithm with several points

This method allows us to decide where the vehicle is located on the map database based on 4 incoming OBD data in case of 30 km/h constant speed. The above shown problems with the high precision RTK - GPS measurement can be eliminated by using predefined map databases as a reference.

It is not necessary to get GPS signal from the vehicle in all times or the communication with the server (AGPS - Assisted GPS) as the OBD data source may be capable to supplement the local information gap on the basis of the presented principles. This method is able to replace and help the positioning GPS component at an appropriate level.

8 Possible Extension to Electric Vehicles

Feasibility of the measurement and compliance of the map data base and the power consumption of the electric motor should be investigated in case of electric vehicles as well.

The main concept in case of electric vehicle’s measurement is to measure the power consumption of the electric motor. This data can be available via the on board diagnostic system of the electric vehicle.

Previous studies were done with special regard to the electric vehicles power consumption measurement in city driving and in freeway driving. The investigations have revealed that the grade of the road has a significant impact on the electric vehicle's consumption [22].

If the power consumption data from the electric vehicle's OBD system and the map database are available, the statistical analyses can be done and the coefficient of determination is definable.

From the elevation database the rising resistance can be predictable and optimal operating status of self-driving vehicle can be adjustable in case of vehicles with internal combustion engines and electric vehicles as well. Hence, the fuel and power consumption and the emission can be reduced. The rising resistance can be calculated from the previously measured road database and the optimal road can be chosen with the algorithm developed for this purpose.

This algorithm can also take into consideration the length of the roads, the permitted speeds and calculates the available optimal operating states of the vehicle in the chosen road which leads to energy saving and contributes to reducing the environment pollution.

Conclusions

In this article we have shown a method to define vehicle position without constant GPS connection.

We carried out measurements with 30 km/h constant speed in order to have the reference road database. During the measurements the fuel consumption and the vertical displacement were recorded the former from the vehicle's OBD system the latter based on the RTK GPS measurement system. According to our measurement the determination of coefficient value is 95,49% in case of 30 km/h constant speed.

We have found that the vehicle's position can be determined from the fuel consumption and the available map database, which database contains the vertical displacement ($R^2 = 0,9549$).

We presented an algorithm which investigates and compares the input data (fuel consumption) from the OBD system with the available reference database, so the position of the vehicle can be defined with it. The developed algorithm investigates the sum of the square differences of the fuel consumption and the available map data and selects the smallest value from it. The smallest value from the calculations and the associated maximal index of the MAP database results the actual position of the vehicle.

The algorithm was implemented into Matlab environment and the results show that the position of the vehicle can be determined exactly after 4 datapoints of fuel consumption, these four datapoints mean an interval of 10-15 seconds.

The result of this algorithm development constructs the basics of vehicle positioning without GPS, which can be very useful part of the Intelligent Transportation System if the proper GPS coverage is not existent and when keeping functionality and safety are the main priority.

Possible steps for further methods:

- Development opportunity in this topic is the mapping without RTK-GPS such as 3-axis accelerometer;
- The presented algorithm can be extended to 40 and 50 km/h standard speeds;
- Evaluation can be done at various speeds and various routes as well;
- Investigation of the application fields of the developed algorithm for driver assistant systems with the special regarding of the predictive cruise control systems;
- Analyzis of electric vehicles.

Acknowledgement

The research presented in this paper was carried out as part of the EFOP-3.6.2-16-2017-00016 project in the framework of the New Széchenyi Plan. The completion of this project is funded by the European Union and co-financed by the European Social Fund.

References

[1] Busznyák, T., Lakatos, I. (2017) "Automotive Engineering possibilities in combining Global Postioning and Vehicle Diagnostic" *5th International Scientific Conference on Advances in Mechanical Engineering (ISCAME 2017),* University of Debrecen Faculty of Engineering, pp. 84-89, ISBN 978-963-473-304-1

[2] Busznyák, T., Lakatos, I. (2018) "Digitális eszközrendszerek a gépjárművekben, mint az autonomizálódó közlekedés fejlesztésének információforrásai" *IFFK 2018: XII. Innováció és fenntartható felszíni közlekedés,* 2018.08.29-2018.08.31, Budapest, Hungary, MMA, Paper: 13, ISBN 978-963-88875-3-5

[3] Colomina, I., Molina, P. (2014) "Unmanned aerial systems for photogrammetry and remote sensing: A review" *ISPRS Journal of Photogrammetry and Remote Sensing,* Volume 92, pp. 79-97, doi: https://doi.org/10.1016/j.isprsjprs.2014.02.013

[4] Dabove, P., Manzino, M. M., Gogoi, N. (2018) "Assessment of positioning performances in Italy from GPS, BDS and GLONASS constellations" *Geodesy and Geodynamics*, Volume 9, Issue 6, pp. 439-448, doi: https://doi.org/10.1016/j.geog.2018.06.009

[5] Derbel, O.; Peter, T.; Zebiri, H.; Mourllion, B.; Basset, M. (2013) Modified intelligent driver model for driver safety and traffic stability improvement *IFAC Proceedings Volumes* 46(21): 744-749, https://doi.org/10.3182/20130904-4-JP-2042.00132

[6] Han, H., Wang, J., Du, M. (2017) "GPS/BDS/INS tightly coupled integration accuracy improvement using an improved adaptive interacting multiple model with classified measurement update" *Chinese Journal of Aeronautics,* Volume 30, Issue 3, pp. 556-566, doi: https://doi.org/10.1016/j.cja.2017.12.011

[7] Iordanopoulos, P., Mitsakis, E. and Chalkiadakis, C. (2018) "Prerequisites for Further Deploying ITS Systems: The Case of Greece" *Periodica Polytechnica Transportation Engineering*, 46(2), pp. 108-115, doi: https://doi.org/10.3311/PPtr.11174.

[8] Omae, M., Fujioka, T., Hashimoto, N., Shimizu, H. (2006) "The application of RTK-GPS and Steer-by-wire technology to the automatic of vehicles and an evaluation driver behavior" *IATSS Research,* Volume 30, Issue 2, pp. 29-38, doi: https://doi.org/10.1016/S0386-1112(14)60167-9

[9] Péter, T., Bokor, J. (2010.2) "Modeling road traffic networks for control" *Annual international conference on network technologies communications: NTC 2010.* Thaiföld, 2010.11.30-2010.11.30. pp. 18-22, Paper 21, ISBN:978-981-08-7654-8

[10] Péter, T., Bokor, J. (2011) "New road traffic networks models for control" *GSTF International Journal on Computing*, Vol. 1, Number 2, pp. 227-232, DOI: 10.5176_2010-2283_1.2.65 February 2011

[11] Pokorádi, L. (2018) "Graph model-based analysis of technical systems" *IOP CONFERENCE SERIES: MATERIALS SCIENCE AND ENGINEERING 393: 1*, Paper: 012007, 8 p.

[12] Pokorádi, L. (2018) "Methodology of Advanced Graph Model-based Vehicle Systems Analysis" In: Szakál, Anikó (szerk.) *IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI 2018*) Budapest, IEEE Hungary Section (2018) pp. 325-328, 4 p.

[13] Pokorádi, L., Lázár-Fülep, T. (2017) "Jármű-irányítási rendszerek megbízhatósági és kockázatelemzési modellezése – Egy kutatási projekt beharangozója" In: Péter, Tamás (szerk.) *IFFK 2017: XI. Innováció és fenntartható felszíni közlekedés* Budapest, Magyarország: Magyar Mérnökakadémia (MMA) (2017) pp. 181-186, 6 p.

[14] Rabah, M., Basiouny, M., Ghanem, E., Elhadary, A. (2018) "Using RTK and VRS in direct geo-referencing of the UAV imagery" *NRIAG Journal of Astronomy and Geophysics,* Volume 7, Issoue 2, pp. 220-226, doi: https://doi.org/10.1016/j.nrjag.2018.05.003

[15] Ramli, M., F., Aburas, M., M., Abdullah, S., H., Ash'aari, Z., H. (2015) "Measuring Land Cover Change in Seremban, Malaysia Using NDVI Index" *Procedia Enviromental Sciences,* Volume 30, pp. 238-243, doi: https://doi.org/10.1016/j.proenv.2015.10.043

[16] Rudas, I., J., Haidegger, T., Takács, Á., Drexler, D., A., Galambos, P. (2018) "Assessment and Standardization of Autonomous Vehicles" *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)* 21-23 June, Las Palmas de Gran Canaria, Spain, pp. 185-192, ISSN: 1543-9259, DOI: 10.1109/INES.2018.8523899

[17] Rudas, I., J., Horváth, L. (2018) "Information Content Driven Model for Virtual Engineering Space" *ACTA POLYTECHNICA HUNGARICA 15: 2,* pp. 7-32, 26 p.

[18] Rudas, I., J., Haidegger, T., Takacs, Á., Bosl D., (2018) "Highly Automated Vehicles and Self-Driving Cars" *IEEE ROBOTICS & AUTOMATION MAGAZINE 25: 4*, pp. 106-112, 7 p.

[19] Tapete, D., Morelli, S., Fanti, R., Casagli, N. (2015) "Localising deformation along the elevation of linear structures: An experiment with space-borne InSAR and RTK GPS on the Roman Aqueducts in Rome, Italy", *Applied Geography,* 58, pp. 65-83, doi: https://doi.org/10.1016/j.apgeog.2015.01.009

[20] Xia, J., Sun, Q., Foster, J., Falkmer, T., Lee, H. (2017) „Pursuing Precise Vehicle Movement Trajectory in Urban Residential Area Using Multi-GNSS RTK Tracking" *Transportation Research Procedia,* Volume 25, pp. 2356-2372, doi: https://doi.org/10.1016/j.trpro.2017.05.255

[21] Xu, H. (2012) "Application of GPS-RTK Technology in the Land Change Survey" *Procedia Engineering,* 29, pp. 3454-3459, doi: https://doi.org/10.1016/j.proeng.2012.01.511

[22] Wu, X., Freese, D., Cabrera, A., Kitch, W., A. (2015) „Electric vehicles' energy consumption and estimation" *Transportation Research Part D 34,* pp. 52-67, doi: https://doi.org/10.1016/j.trd.2014.10.007

The Links between Unemployment and Self-Employment: Evidence from the EU Countries

Rita Remeikiene[1], Ligita Gaspareniene[2], Romualdas Ginevicius[3], Milan Robin Patak[4]

[1,2] Mykolas Romeris University, Ateities str. 20, 08303 Vilnius, Lithuania, E-mail: rita.remeikiene@mruni.eu; ligitagaspareniene@mruni.eu

[3] Vilnius Gediminas Technical University, Sauletekio av. 11, 10223 Vilnius, Lithuania, E-mail: romualdas.ginevicius@vgtu.lt

[4] University College of Business, Spalena 76/14, 110 00 Praha 1, E-mail: patak@vso-praha.eu

Abstract: the purpose of this work is to research the links between the unemployment and self-employment rates in the EU member states and provide recommendations on how to manage these links. The results of the empirical study have established that measures to reduce the global unemployment rate, suitable for all EU Member States, cannot be placed on the labor market. Calculations have shown that in the period of 2007-2017, which covered the financial crisis and the upturn, EU countries need to be grouped into specific groups and for each group of countries select appropriate measures to reduce unemployment. The first group of EU countries experienced a "pull" factor effect, the second was the effect of the "push" factor, on unemployment, thus making in the first group of countries more effective self-employment measures to reduce unemployment while the second group of countries - supporting self-employment measures was ineffective in combating unemployment problems.

Keywords: self-employment; unemployment; "push" and "pull" factors; EU

1 Introduction

The current situation in the European labor market has been significantly affected by a combination of the financial crisis and the crash of the global economy in 2008. From a volatile economic context, observed over the last ten years, the EU labor market has deeply deteriorated: with reference to the data of [1], the unemployment rate in the EU-28 was 0.6 points higher in 2017 than it was in 2007.

The problem of the high unemployment rate calls for the development of measures that would promote the involvement of the European population into the labor market. One of the main aims defined in the Europe 2020 strategy is to have 75 percent of the active population (aged from 20 to 64) working [2]. Apart from reduction of labor taxation or the support to newly established enterprises (subsidization, exemption from taxes, etc.), promotion of self-employment and atypical forms of employment is also considered to be one of the measures to help to fight unemployment [3].

Although the policy of turning unemployment into self-employment remains an ambiguous issue in economics (a large number of the self-employed is considered as one of the features of developing rather than advanced economies [4], it is still believed that the policy of this kind may significantly increase the probability of a person being employed, raise personal income level and reduce the probability of a person being unemployed [5]. To make the policy of turning unemployment into self-employment successful, it is extremely important to clearly understand the links between unemployment and self-employment.

Previous studies that focused on the dynamic relationship between unemployment and self-employment in different countries provide rather contradictory results: depending on the recession push or entrepreneurial pull approach followed, some authors [6-9] found a positive and statistically significant relationship between the two phenomena under research, while others reported about a negative [10-12] insignificant [13] relationship or the existence of the relationship was not confirmed, especially as far as it concerns particular countries [14].

Unambiguity and inconclusiveness of the results of previous studies, calls for a more comprehensive analysis in this area. ***The primary purpose*** of this article is to research the links between the unemployment and self-employment rates in the EU member states and provide the recommendations on how to manage these links. For fulfilment of the defined purpose, the following ***objectives*** were raised:

1) Review the literature on the links between unemployment and self-employment.
2) Select and substantiate the methodology of the research; 3) to introduce the results of the empirical research on the links between unemployment and self-employment in the EU member states and provide the recommendations on how to manage these links.

The methods of the research include systematic and comparative literature review, Spearman’s correlation coefficient and multiple regression.

2 The Links between Unemployment and Self-Employment: Literature Review

According to [4], the links between unemployment and self-employment are determined by the structure of employment in a certain country/region and frictions in the labor market. Since it is usually difficult for job seekers to deal with strong labor market frictions, they start treating job search as a relatively less attractive alternative in comparison to self-employment. Minding the fact that much higher unemployment rates and more severe labor market frictions are inherent to developing rather than advanced economies, many authors [15] [16] [17] [4] and others, believe that developing economies have systematically higher self-employment rates than advanced economies. Nevertheless, affected by economic upheavals or industrial declines, advanced economies can also have the challenging periods of the frictions in their labor markets, as it, for instance, could be observed in the EU just after the stroke of the global economic crisis in 2008. Hence, the models of frictional labor markets have to be adjusted to advanced economies.

Previous studies provide rather inconclusive results on the direction and significance of the links between unemployment and self-employment. The review of previous findings has been presented in Table 1.

Table 1

The review of previous findings on the links between unemployment and self-employment

Author(-s), year	Research purpose	Research methods	Findings
[7]	To investigate the dynamic relationship between self-employment and unemployment rates	A two-equation vector autoregression model	The research confirmed the existence of the interdependence between unemployment and self-employment
[5]	To evaluate the effectiveness of two self-employed activity start-up programs for the unemployed	Administrative data analysis with a time lag, survey	Self-employed activity programs increase the probability of being employed, reduce the probability of being unemployed and raise personal income
[14]	To provide further time series evidence on the links between unemployment and self-employment	Autoregressive Distributed Lag (ARDL) approach	The empirical results confirmed existence of the links between unemployment and self-employment in 7 OECD countries out of 28
[4]	To determine the links between wage	Extended standard DMP search and	Variation in labor market frictions can

	employment, unemployment and self-employment	matching model	explain almost the entire variation in both unemployment and self-employment, i.e. unemployment and self-employment show the joint variation under the conditions of the frictions in the labor market
[9]	To research the long-term links between unemployment and self-employment	Panel cointegration methods	There exists a positive and statistically significant relation between unemployment and self-employment
[6]	To research the dynamic relationship between unemployment and self-employment rates	A two-equation vector autoregression model	The research confirmed the existence of the interdependence between unemployment and self-employment
[11]	To research the dynamic relationship between unemployment and self-employment in Turkey	Cointegration test, vector, error correction model	There exists a long-term relationship between unemployment and self-employment
[12]	To research the effect of economic conditions on self-employment in Canada	Asynchronous variation in economic conditions across time and provinces	The relationship between the provincial unemployment and self-employment is negative
[13]	To research the links between unemployment and self-employment by considering the role of entrepreneurship training	Survey, regression models	The results provided limited support to the hypothesis that entrepreneurship training can be effective in combating unemployment
[8]	To research the links between the entrepreneurial cycles and the national economic cycles	Aggregation, a panel framework, Granger causality test	The entrepreneurial cycle is positively affected by the national unemployment cycle

Source: compiled by the authors

First of all, it should be noted, that the links between unemployment and self-employment are analyzed by following “push” and “pull” approaches. The “push” approach, also known as the refugee effect, desperation effect, recession push or unemployment push, affirms that when unemployment rate is rising, an increasingly higher number of people start having difficulties in finding paid jobs (or wage jobs), and these difficulties, in their turn, lead to-self-employment as an alternative [11]. Under the conditions of high unemployment rate, the unemployed have lower opportunity costs, which may push them to undertake the risk associated with a business start-up [13]. In this case, unemployment and self-employment show a positive and statistically significant relationship [6] [7] [8] [9] [4]. The "pull" approach, also known as the prosperity pull or entrepreneurial effect, suggests that since self-employment promotes business activities, it also stimulates a population's inclination to and readiness for entrepreneurial activities, which, in its turn, causes a decline in unemployment in subsequent periods [11] and contributes to the rise of the minimum wage [18-19], i.e. the prosperity pull (or entrepreneurial) effect manifests as the reduction in unemployment rates caused by increasing self-employment. In the latter case, unemployment is negatively related to self-employment [6] [10] [7] [12].

On the basis of one, another or both above-introduced approaches, the majority of literature sources confirm the interdependence between unemployment and self-employment (as it can be seen in Table 1), although the results of the studies carried out by [6] and [7] disclosed that the "push" effect is considerably stronger than the "pull" effect. These results were confirmed by [11] who found the evidence for the existence of the causality running from self-employment to unemployment rate (i.e. the existence of an entrepreneurial effect was confirmed), but at the same time noted that it was not possible to accurately assess the strength of the entrepreneurial effect due to the involvement of unpaid workers in self-employment which might plausibly disguise the exact entrepreneurial effect.

It should also be noted that the interdependence between unemployment and self-employment may vary during the periods of economic (labor market) stability and upheaval. According to [4], the channel of labor market frictions is important since the "variation in labor market frictions can account for a large fraction of the univariate and joint variation in self-employment and unemployment rates across countries observed in the data" (p. 5). In other words, as different countries can be undergoing different periods of an economic (labor market) cycle, they may show different results of the interdependence between unemployment and self-employment. For instance, by employing panel cointegration methods, [9] investigated the COMPENDIA dataset, developed for a wide range of European OECD countries over the period 1990-2011. The results of their study disclosed a positive and statistically significant long-term relation between unemployment and self-employment which was observed in more than half of the countries under consideration. Nevertheless, the relation between unemployment and self-employment was found to be negative or statistically insignificant for the rest of the countries in the sample, which shows that the relation between the phenomena under research can be bidirectional and depend on the economic (labor market) cycle in a particular country. The differences in the stage of an economic (labor market) cycle in the countries under research may also explain the inconclusiveness of the findings of [14] study which confirmed existence of the links between unemployment and self-employment only in 7 out of 28 OECD countries.

In addition, some of the studies revealed that the strength and direction of the interdependence between unemployment and self-employment may depend on some other factors besides the stage of an economic (labor market) cycle. [4] notes the impact of economic policies which determine the flow value of unemployment, the changes in tax-variation caused profitability of own-account work, the enforcement of business size regulations and the size of unemployment benefits (transfers to job seekers reduce the attractiveness of entrepreneurial activities). [20] highlights the impact of demographic factors by stating that the links between unemployment and self-employment can be determined by gender: the authors found that unlike men, women's self-employment decisions were very sensitive to the sources of household income rather than to the general economic conditions.

To summarize, the theoretical recession-push hypothesis suggests a positive, while the prosperity-pull hypothesis proposes a negative relationship between unemployment and self-employment. Due to the involvement of unpaid workers in self-employment which might plausibly disguise the exact entrepreneurial effect, the influence of "push" factors on the relationship between unemployment and self-employment is recognized to be considerably stronger than the influence of "pull" factors. The strength and direction of the links between the two phenomena under research are mainly affected by the stage of an economic (labor market) cycle, which plausibly determines the differences in cross-country findings. Apart from that, some literature sources also confirm the impact of economic policies (the flow value of unemployment, the changes in tax-variation caused profitability of own-account work, the enforcement of business size regulations and the size of unemployment benefits) and demographic factors (gender).

3 Research Methodology

To achieve the purpose of the article - a statistically significant relationship between unemployment and self-employment in the EU countries Spearman's Correlation Coefficient (rS) was selected for investigating the strength of the phenomena in terms of legality. The calculations include the unemployment rate, expressed in thousands of individuals (y) and self-employment (x), expressed in thousands of individuals in the EU-28 in the period of 2007-2017.

Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. In a sample it is denoted by and is by design constrained as follows and its interpretation is similar to that of Pearsons, e.g. the closer is to the stronger the monotonic relationship. Correlation is an effect size and so we can verbally describe the strength of the correlation using the following guide for the absolute value of: .00-.19 “very weak”; .20-.39 “weak”; .40-.59 “moderate”; .60-.79 “strong”; .80-1.0 “very strong”.

In order to investigate the weight of self-employed people with unemployment rates, is being used the linear multiply regression model. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable *x* is associated with a value of the dependent variable *y*. The population regression line for *p* explanatory variables $x_1, x_2, \ldots, x_p$ is defined to be $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$. This line describes how the mean response μ_y changes with the explanatory variables. The observed values for *y* vary about their means μ_y and are assumed to have the same standard deviation σ. The fitted values $b_0, b_1, \ldots, b_p$ estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ of the population regression line.

4 The Results of the Empirical Research

In the first stage of the empirical study, Spearman's correlation coefficient for the EU-28 countries where Yt is the unemployment rate for the period 2007-2017, Yt-1 is the unemployment rate in the last year, X - self-employed persons, thus, it is possible to classify EU countries in three groups:

1 group. Countries with a negative relationship between the unemployment rate and the level of self-employment, i.e. the unemployment rate reduces the number of self-employed people and vice versa. This group includes countries like: having a **very strong connection** Greece ($r_S = -0.939$, $p = 0.000$/su Y_{t-1}), Lithuania ($r_S = -0.927$, $p = 0.000$ / su Y_{t-1}), Ireland ($r_S = -0.903$, $p = 0.000$ / su Y_{t-1}), Latvia ($r_S = -0.855$, $p = 0.002$), Spain ($r_S = -0.806$, $p = 0.005$); Italy ($r_S = -0.879$, $p = 0.002$/ su Y_{t-1}); **strong connection** Estonia ($r_S = -0.733$, $p = 0.016$), Cyprus ($r_S = -0.790$, $p = 0.007$/su Y_{t-1}), Croatia ($r_S = -0.733$, $p = 0.016$/su Y_{t-1}).

Based on “push“ and “pull“ theories, the 1 group can be classified as pull-based theory, i.e. “pull” factors are those which make the choice of self-employment more attractive to paid employment. For employees who voluntarily leaves the job, there is a greater chance of becoming self-employed. In addition, longer-term unemployment tends to be linked to increased probability of self-employment. Thus, for the first group of countries, only those unemployment reduction measures which are focused on the positive impact that business start-up funding/partial funding has on implementation of ideas, dreams or competences gained at a hired work may serve as the measures of self-employment promotion. This means, the European Commission, in the context of reducing strategic unemployment levels, recommends that the assigned countries propose new business development programs in the business cycle recovery/rise phases, because only in this period their return on the efficiency of their use would be highest.

2 group. Countries with a positive relationship between the unemployment rate and the level of self-employment, i.e. the rise in unemployment rates contributes to the growth of self-employment and vice versa. This group includes countries like: having a **very strong connection** Luxembourg ($r_S = 0.911$, $p = 0.000$), France ($r_S = 0.863$, $p = 0.01$/ su Y_{t-1}), Netherlands ($r_S = 0.842$, $p = 0.002$/ su Y_{t-1}); **strong connection** Austria ($r_S = 0.721$, $p = 0.019$), Germany ($r_S = 0.758$, $p = 0.011$/ su Y_{t-1}), Sweden ($r_S = 0.614$, $p = 0.05$) and Belgium ($r_S = 0.685$, $p = 0.01$).

According to [20] "Push" factors are typically those associated with being pushed out of paid employment into a less preferred self-employed situation, and are thus positively associated with increases in the unemployment rate and unemployment durations. The most common "push" set of hypotheses suggests that workers are primarily pushed into self-employment by weak economic job prospects. The 2 group of countries revealed that the population chooses self-employment as a necessity and not as a voluntary choice, while a positive relationship means that as unemployment grows, self-employment also increases, and vice versa. This means, the European Commission in order to reduce the unemployment rate in these countries should not focus on self-employment promotion measures as a way to reduce unemployment.

3 group. Countries that have not recorded statistically significant relationships between unemployment and self-employment (Czech Republic, Denmark, Finland, Hungary, Malta, Poland, Portugal, Romania, Slovakia, Slovenia, UK, Bulgaria).

It can be presumed that the statistically insignificant links cause greater effects of some other labor market factors while solving the problems of unemployment. For instance, sufficient unemployment benefits or corporate taxes discourage from looking for other employment alternatives, such as self-employment [22].

The results of multiple regression evaluated in the second stage of the empirical study are presented in Table 2.

Table 2

Multiple regression results

Countries	Equation	Explanation
*1 group**		
Lithuania	y = 577,231-3,098*Self-employment	With a decrease of 1 thousand in self-employment, the unemployment rate is increasing by 3,098 thousand people.
Ireland	y = 1,160,443-3,163*Self-employment	With a decrease of 1 thousand in self-employment, the unemployment rate is increasing by 3,163 thousand people.

Spain	y = 24,740.714-6.831*Self-employment	With a decrease of 1 thousand in self-employment, the unemployment rate is increasing by 6,831 thousand people.
Italy	y = 18,210,727-3,178*Self-employment	With a decrease of 1 thousand in self-employment, the unemployment rate is increasing by 3,178 thousand people.
Estonia	y = 262,288-3,836*Self-employment	With a decrease in self-employment of 1 thousand, the unemployment rate increases by 3,836 thousand. individuals.
Cyprus	y = 146,885-1,941*Self-employment	With a decrease of 1 thousand in self-employment, the unemployment rate is increasing by 1,941 thousand people.
Croatia	y = 1,528,717-0,024*GDP-0,774* Self-employment	With a decrease of 1 thousand in self-employment, the unemployment rate is increasing by 774 people. The standardized beta coefficients showed that the impact of GDP (-0.764) and self-employment (-0.643) on the trends in the unemployment rate are close.
2 group.*		
France	y = -2638,897+1,963*Self-employment	With an increase in self-employed employment of 1 thousand, the unemployment rate increases by 1,963 thousand people.
Netherlands	y = -465,204+0,822*Self-employment	With an increase in self-employed employment of 1 thousand, the unemployment rate increases by 822 people.
Belgium	y = -302,751+1,160* Self-employment	With an increase in self-employed employment of 1 thousand, the unemployment rate increases by 1,160 thousand people.

* With regard to Greece, Latvia, Luxembourg, Austria, Germany, Sweden, the multi-regression equation self-employment was not statistically significant.

The calculations checked by Multiregresine analysis revealed that country classification or grouping according to certain criteria (in this case “push“ and “pull“) allows identifying the specifics of countries in combating negative phenomena such as unemployment. 2 group countries are classified as economies in developed countries, which joined the EU in period of 1958 to 1981. Their economy is robust, so the "emission" of the EU-wide unemployment reduction measures into the labor market through the self-employment prism will not be effective and will not reach the target group of unemployed. Meanwhile, Group 1 countries joined the EU in 1981 and later. The economy of these countries is not very stable, therefore the decrease in the autonomy employment significantly increases the number of the unemployed.

Conclusions

This work represents a preliminary research on the relationship between unemployment and self-employment (positive or negative) between the countries and how they become self-employed and illustrates methods to combat unemployment more effectively.

The empirical statistical significance of the relationship between unemployment and the level of self-employment, in all age groups, has revealed that universal measures for reducing unemployment levels that are suitable for all EU members cannot be placed on the labor market. The calculations show that during the period of 2007-2017, which included the financial crisis and the upturn, EU countries need to be grouped into specific groups and for each group of countries select appropriate measures to reduce unemployment.

The first group of countries that includes Greece, Lithuania, Latvia, Estonia, Ireland, Spain, Italy, Cyprus and Croatia, noted the following connections: 1) All joined the EU after 1981 (except Italy), the economy is not as stable as the countries in 2 group, and the statistically significant negative relationship between unemployment and the level of employment ("pull") has been obtained, suggesting that business support programs aimed at a person who has worked for a long period of time in hired work, a short-term unemployed, would more effectively help to reduce unemployment; 2) Countries in 2 group joined the EU before 1981, their economies are stable, as evidenced is their economic development level, and the statistically significant link between unemployment and self-employment levels is positive ("push"). This means, measures aimed at reducing unemployment towards self-employment would increase unemployment, which would result EU funding to reduce the unemployment rate, rather than attaining the strategic objective of reducing unemployment in the EU; 3) In 3 group countries, it would be useful to conduct more in-depth studies on reducing unemployment and to find meaningful links to other, non-self-employed, for example, the relationship between unemployment and the corporate tax rate.

References

[1] Eurostat, "The EU in the world – labor market" (2018) Retrieved from Internet: https://ec.europa.eu/eurostat/statistics-explained/index.php/The_EU_in_the_world_-_labour_market#Unemployment_rate-

[2] European Commission, "Apie Europos Sąjungos politiką. Užimtumas ir socialiniai reikalai. Investavimas į darbo vietų kūrimą, įtrauktį ir socialinę politiką" [About the policies of the EU. Employment and social affairs. Investing in job creation, involvement and social policies] (2014) Retrieved from Internet: https://europa.eu/european-union/file/538/download_lt?token=3DFANlGb

[3] J. Poor, S. Vinogradov, G. G. Tözsér, I. Antalik, Z., Horbulák, T. Juhász, I. É. Kovács, K. Némethy, R. Machová. "Atypical forms of employment on Hungarian-Slovakian border areas in light of empirical researchers". Acta Polytechnica Hungarica 14 (7), pp. 123-141, 2017

[4] M. Poschke, “Wage employment, unemployment, and self-employment across countries”. *International Growth Centre*, (2018) Retrieved from Internet: https://www.theigc.org/wp-content/uploads/2018/05/Poschke-2018-Working-Paper.pdf

[5] H. J. Baumgartner, M. Caliendo, “Turning unemployment into self-employment: effectiveness of two start-up programmes”, *Oxford Bulletin of Economics and Statistics*, Vol. 70, No. 3, pp. 347-373, 2008

[6] D. B. Audretsch, D. B., M. A. Carree, R. Thuric, A. van Steel, “Does self-employment reduce unemployment?”, *CEPR Discussion Paper* no. 5057, pp. 1-17, 2005

[7] A. R. Thurik, M. A. Carree, A. van Steel, D. B. Audretsch, “Does self-employment reduce unemployment?” *Journal of Business Venturing,* Vol. 23, No. 6, pp. 673-686, 2008, doi: https://doi.org/10.1016/j.jbusvent.2008.01.007

[8] P. D. Koellinger, A. R. Thurik, “Entrepreneurship and the business cycle”. *The Review of Economics and Statistics*, Vol. 94, No. 4, pp. 1143-1156, 2012

[9] G. Saridakis, M. A. Mendoza, R. I. M. Torres, J. Glover, “The relationship between self-employment and unemployment in the long-run: a panel cointegration approach allowing for breaks”. *Journal of Economic Studies*, Vol. 43, No. 3, pp. 358-379, 2016, doi: https://doi.org/10.1108/JES-11-2013-0169

[10] D. Glocker, V. Steiner, ”Self-employment: a way to end unemployment? Empirical evidence from German pseudo-panel data”. *IZA Discussion Paper* No. 2561, pp. 1-25, 2007, Retrieved from Internet: http://ftp.iza.org/dp2561.pdf

[11] Y. Ozerkek, F. Dogruel, “Self-employment and unemployment in Turkey”. *Topics in Middle Eastern and North African Economies*, Vol. 17, No. 1, pp. 133-152, 2015, Retrieved from Internet: https://pdfs.semanticscholar.org/129d/bed6e5ef21c54e148c5593fc8b3e48f6aff6.pdf?_ga=2.148791871.900472165.1539084415-1907629831.1539084415

[12] C. Pongpaiboon, “The effect of economic conditions on self-employment in Canada”. *A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Bachelor of Arts, Honours in the Department of Economics University of Victoria* (2017) Retrieved from Internet: https://www.uvic.ca/socialsciences/economics/assets/docs/honours/Cherry%20Pongpaiboon%20Thesis.pdf

[13] Michaelides, M.; Davis, S. (2016) From unemployment to self-employment: the role of entrepreneurship training. University of Cyprus, working paper No. 09-2016, Retrieved from Internet: http://papers.econ.ucy.ac.cy/RePEc/papers/09-16.pdf

[14] F. Halicioglu, S. Yolac, “Testing the impact of unemployment on self-employment: empirical evidence from OECD countries”. *MPRA paper* No. 65026, 2015, Retrieved from Internet: https://mpra.ub.uni-muenchen.de/65026/1/MPRA_paper_65026.pdf

[15] F. Caselli, “Accounting for cross-country income differences”. *In P. Aghion and S. N. Durlauf, eds., ‘Handbook of Economic Growth’*, North Holland, Amsterdam, 2005

[16] D. Gollin, “Nobody’s business but my own: self-employment and small enterprise in economic development”. *Journal of Monetary Economics,* Vol. 55, No. 2, pp. 219-233, 2007

[17] S. F. Hipple, “Self-employment in the United States”. *Monthly Labor Review*, Vol. 133, No. 9, pp. 17-32, 2010, https://www.bls.gov/opub/mlr/2010/09/art2full.pdf

[18] C. J. Blattman, S. Dercon, “Occupational choice in early industrializing societies: experimental evidence on the income and health effects of industrial and entrepreneurial work”. *IZA Discussion Paper* No. 10255, pp. 1-91, 2016, Retrieved from: https://www.econstor.eu/bitstream/10419/147941/1/dp10255.pdf

[19] V. Bassi, A. Nansamba, “Information frictions in the labor market: evidence from a field experiment in Uganda”, 2017, Retrieved from Internet: http://conference.iza.org/conference_files/GLMLICNetwork_2017/bassi_v10212.pdf

[20] A. M. Biehl, T. Gurley-Calvez, B. Hill, “Self-employment of older Americans: do recessions matter?” *Small Business Economics*, Vol. 42, No. 2, pp. 297-309, 2014

[21] P. S. J. Leonard, J. T. McDonald, J. C. H. Emery, “Push or Pull into Self Employment? Evidence from Longitudinal Canadian Tax Data”, 2017, Retrieved from Internet: https://www.unb.ca/fredericton/arts/nbirdt/_resources/pdfs/working-paper_push-or-pull-into-self-employment.pdf

[22] A. Zirgulis, T. Šarapovas, “Impact of corporate taxation on unemployment”. *Journal of Business and Management*, Vol. 18, No. 3, pp. 412-426, 2017