Dynamic classification: economic welfare growth in the EU during 1995–2004

I. Gertsbakh[1] & I. Yatskiv[2]

[1]*Department of Mathematics, Ben Gurion University, Israel*
[2]*Department of Computer Science,*
Transport and Telecommunication Institute, Latvia

Abstract

The purpose of dynamic classification in application to the economic development of EU countries is to work out an Economic Welfare Growth Index (*EWGI*), the use of which would make it possible to establish, for each year and each country in the study, the level of its economic welfare, to estimate the rate of advance of each country toward economic well-being, to compare different countries and to make forecasts. *EWGI* is defined as an "optimal" linear combination of leading socio-economic parameters, such as *GDP, CPI* (corruption transparency index), level of unemployment, inflation, average life expectancy. On the first stage we consider one "training" year and classify all EU countries by using standard clustering algorithms to single out two groups: P and R, having the "worst" and the "best" values of the parameters, respectively ("poor" and "rich"). On the second stage we use these groups for constructing the Fisher Discriminant Function (FDF). For sake of convenience the FDF is linearly transformed in such a way that the centers of P and R give the FDF values 0 and 10, respectively. This transformed FDF is taken as the *EWGI*. It has range [-2, 4] for P, [9, 14] for R and [4, 9] for countries in the "middle". The final output of our analysis is a collection of time series (graphs) representing the dynamics and the behaviour pattern of *EWGI* for each country in the period 1995-2004. Similar approach can be used also in other areas, such as transportation, education, health care, whenever the development in time is described by a multidimensional time series.
Keywords: economic welfare index, cluster analysis, Fisher discriminant function.

WIT Transactions on Information and Communication Technologies, Vol 37,
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060061

1 Introduction

1.1 The data and notation

The initial data are given as a three-dimensional array of entries having the following form:

$$\boldsymbol{y}(i; j = 1,..., p; t) = [x_{i1}(t),..., x_{ij}(t),..., x_{ip}(t)], \quad (1)$$

where i is the country number, $i = 1, ..., 25$; j is the number of an economic parameter, $j = 1, ..., 6$; t denotes year, $t = 1995, ..., 2004$. The total number of entries in (1) is $25 \cdot 6 \cdot 10 = 1500$. The "economic position" of a particular country in a particular year is described by a point (1) in 6-dimensional space.

Since the original data (1) consist of indices of various dimensions and of highly differing ranges, we will operate with *standardized* data: we replace $x_{ij}(t)$ by its standardized value $x^{\circ}_{ij}(t) = (x_{ij}(t) - m_j)/s_j$, where m_j and s_j are the averages and standard deviations of the j-th coordinate, respectively, obtained from so-called training sample, see below Section 2.1.

The goal of this paper is to develop a measure of socio-economic welfare for the EU countries. We call this measure the Economic Welfare Growth Index (*EWGI*). The *EWGI* is defined for country i and year t in the following general form:

$$EWGI(i,t) = A \cdot \sum_{j=1}^{6} w_j \cdot x^{\circ}_{ij}(t) + B. \quad (2)$$

($EWGI(i, t) - B$) in eqn (2) is, up to a constant factor, a scalar product of vectors $\boldsymbol{y}^{\circ}(i,t) = [x^{\circ}_{i1}(t), x^{\circ}_{i2}(t), ..., x^{\circ}_{ip}(t)]$ and $\boldsymbol{w} = [w_1,...,w_p]$. Geometrically, according to the definition of the scalar product, ($EWGI(i, t) - B$) equals, up to a multiplier, to the distance from the origin of the projection of point $\boldsymbol{y}^{\circ}(i,t)$ on the direction $\boldsymbol{w}$.

The coordinates of $\boldsymbol{y}$ in eqn (1) are the following socio-economic parameters: *GDP* - Gross domestic product per capita in US dollars calculated as the aggregate of production divided by the population size. *LABR* - Labor productivity per person employed (GDP in purchasing power per person employed relative to EU-25, where EU-25=100), in *%*. *LIFE* - Life expectancy at birth, total average, in years. *CPI* - Corruption Perception Index, based on perception of the degree of corruption as seen by business people, risk analysts and general public; it ranges between 10 (highly clean) and 0 (highly corrupt). *UNEMP* - Unemployment level in percents of the total active population. *INFL* - Inflation level (annual average rate of change in harmonized indices of customer prices), in *%*.

Five of these indices are available for EU countries for the period 1995-2004 on the United Nations Statistics Division site [1] for free download, and *CPI* – on the site of Transparency International (non-governmental organisation) [2].

These indices satisfy the following natural demands. They have a transparent social and economic meaning; they are widely accepted by the economist community and are available to general public; these indices are the main

ingredients of our understanding and perception of economic welfare and prosperity in modern society; these indices change in a monotone way from "bad" to "good", in the direction of increasing the well-being, i.e. greater values of *GDP, LABR, LIFE* and *CPI* always mean higher prosperity and well-being, as well as do smaller values of *UNEMP* and *INFL.*

1.2 The main idea

Any economic welfare index is, by definition, a scalar function of the corresponding economic parameters. Most natural way of constructing such a function is to use linear forms, i.e. to build the Index as a linear combination of these parameters with weights w_i, $i = 1,..., p$, as shown in eqn (2). Thus, finding these weights becomes the principal issue in constructing the desired Index.

The core of our approach is the following. We take one "typical" year in the period of study and apply cluster analysis (CA) to the data of this year. Before applying CA, the data are standardized. The CA singles out several groups of countries with "similar" values of their economic parameters. We concentrate on two extreme groups, which we will mark as "P" and "R". The P-group is a set of countries with the lowest values of *GDP, LABR, LIFE, CPI* and the highest values of *INFL* and *UNEMP*. Simply speaking, this is a group of less prosperous countries. Denote by $\boldsymbol{x}_{\circ P} = [x_1^{\circ P},...,x_6^{\circ P}]$ the center of this group. On the contrary, the "opposite" group marked "R" contains the set of countries with the highest values of *GDP, LABR, LIFE, CPI* and the lowest values of *INFL* and *UNEMP*. This is a group of "more prosperous" countries. Denote by $\boldsymbol{x}_{\circ R} = [x_1^{\circ R},...,x_6^{\circ R}]$ the center of this group. It will be assumed that

$$x_j^{\circ R} \geq x_j^{\circ P},\ j = 1,...,4; \qquad x_j^{\circ R} \leq x_j^{\circ P},\ j = 5,6. \tag{3}$$

Geometrically, the P and R groups are two sets of points in a 6-dimensional space. Selecting weights as in eqn (2) means *projecting* these points on the vector $[w_1, ..., w_6]$. We choose this vector to provide the *maximal separation* between the P and R groups in terms of their projections on this vector. The vector providing maximal separation is called Fisher vector and is defined as

$$\boldsymbol{f} = [f_1^*,...,f_6^*] = \boldsymbol{W}^{-1}(\boldsymbol{x}_{\circ R} - \boldsymbol{x}_{\circ P})', \tag{4}$$

where $\boldsymbol{W}$ is the pooled variance-covariance matrix of the groups P and R (see Gertsbakh [3]).

The quality of the separation is measured by so-called Mahalanobis distance D^2, see JoHnson and Wichern [4]: $D^2 = (\boldsymbol{x}_{\circ R} - \boldsymbol{x}_{\circ P})\boldsymbol{W}^{-1}(\boldsymbol{x}_{\circ R} - \boldsymbol{x}_{\circ P})'$. In practical terms, $D^2 > 20$ means that the separation is quite good.

So far we have a "static" mapping of the vector of economic parameters for one country into a scalar. The "dynamics" of the *EWGI* for the *i*-th country is obtained in a form of a time series by considering the sequence

$$[S_i] = A \cdot \sum_{j=1}^{6} f_i^* \cdot x_{ij}^{\circ}(t) + B, \text{ for } t = 1995, 1996, \ldots, 2004. \tag{5}$$

Remark 1. The success of our method rests on the assumption (3). Theoretically, one can imagine a situation in which this property does not hold true. For example, suppose that the training sample contains only two points with coordinates (0,1) and (1,0) in two-dimensional space. Each of the groups P and R contains only one point and eqn (3) is violated. In practice, however, this situation takes place in very rare cases.
Remark 2. We expect that the signs of the coordinates of the Fisher vector (4) are "correct", i.e. the weights of *GDP, LIFE, LABR* and *CPI* are positive, and the weights of *INFL* and *UNEMP* are negative. In practice this natural demand is not always satisfied. This defect, however, can be easily repaired by considering a suitable convex combination of the Fisher vector and the vector connecting the centers of groups P and R. The price for that is a relatively small decrease in the value of D^2, see Section 2.4.

1.3 Review of literature

Osberg and Sharpe in their recent work [5] suggested an index called *IEWB*-index of economic well-being. *IEWB* is a linear combination *with equal weights* of four components: consumption, wealth, equality and economic security. Each of these components has been elaborated as a result of a detailed analysis of numerous relevant factors and rescaled afterwards to the [0,1] interval. [5] presents data for seven countries: Australia, Canada, Germany, Norway, Sweden, UK and USA, for the period 1980-2001. It is remarkable that economic well-being incorporates also equality and social security components, which means including into the economic well-being also "social" factors.

The idea to replace multidimensional object description by a scalar value in an "optimal" way, using the Fisher discriminant function, was proposed in Gertsbakh [3] and applied to planning preventive maintenance actions of technical objects.

In Section 2 we describe the Algorithm of constructing the *EWGI*. Section 3 presents the application of the Algorithm to real data. Our main results will be an expression for the *EWGI*, its sensitivity analysis and time series of the *EWGI* for EU countries.

2 The algorithm of constructing the *EWGI*

2.1 Step 1: data preparation

Create a training sample as an "average" of years 1999 and 2000:

$$v_{ij} = (x_{ij}(1999) + x_{ij}(2000))/2, \quad i = 1,...25; j = 1,...,6. \tag{6}$$

Calculate means m_i and standard deviations s_i for this training sample and standardize the training sample and initial data according to the following equation:

$$v_{ij} = (v_{ij} - m_j)/s_j, x_{ij}^{\circ}(t) = (x_{ij}(t) - m_j)/s_j, \quad i = 1,...25; j = 1,...,6. \tag{7}$$

2.2 Step 2: cluster analysis

Apply the following principal clustering algorithms to the training sample: Ward's, Complete Linkage, Iterative k-means. Single out three clusters of countries. The first cluster (denoted as "P") contains countries with the smallest values of *GDP, LABR, CPI, LIFE* and the largest values of *INFL* and *UNEMP*. The second cluster (denoted as "R") contains the countries with the largest values of *GDP, LABR, CPI, LIFE* and the smallest values of *INFL* and *UNEMP*. The remaining countries constitute the cluster denoted as "M" (middle).

Usually, the three above clustering methods produce clusters whose contents may have some variations. Define the P and R clusters as the *intersection* of the corresponding clusters produced by different methods.

Compute the coordinates of the centers of P and R:

$$c_j^P = (\sum_{i\in P} v_{ij}^\circ)/|P|, \qquad c_j^R = (\sum_{i\in R} v_{ij}^\circ)/|R|, \qquad j = 1,...6. \tag{8}$$

Calculate the vector $\boldsymbol{c} = [c_1,...,c_6]$ connecting the centers of P and R as having coordinates $c_j = c_j^R - c_j^P, j = 1,...6.$

2.3 Step 3: finding the Fisher vector

Compute the Fisher vector $\boldsymbol{g}$, which provides the maximal separation of P and R, according to the following equation:

$$\boldsymbol{g} = [g_1^*,...,g_6^*] = \boldsymbol{W}^{-1}\boldsymbol{c}', \tag{9}$$

where $\boldsymbol{W}$ is the pooled variance-covariance matrix computed for P and R clusters. (We remind that CA is applied to *standardized* values of x_{ij}).

If the coordinates of vector $\boldsymbol{g}$ have "correct" signs (see Remark 2, Section 1) then GO To Step 5, ELSE go to Step 4.

2.4 Step 4: correcting the FDF

Consider the following family of vectors $\boldsymbol{f}_\alpha$: $\boldsymbol{f}_\alpha = \alpha\cdot\boldsymbol{c} + (1-\alpha)\cdot\boldsymbol{g}$ for $\alpha \in [0,1]$. Set $\alpha = 0.2(0.2)1$ and take the smallest value of α which provides that the vector $\boldsymbol{f}_\alpha$ has coordinates with "correct" signs. Denote this vector as $\boldsymbol{f}^* = [f_1^*,...,f_6^*]$.

The value of D^2 which corresponds to $\boldsymbol{f}^*$ equals $D^{2*} = (\boldsymbol{f}^*\cdot\boldsymbol{c}')^2 / \boldsymbol{c}\boldsymbol{W}\boldsymbol{c}'$.

2.5 Step 5: formula for the index

The *EWGI*(*i, t*) is defined according to eqn (2). We choose the constants *A* and *B* to provide that for the original (nonstandarized) center coordinates of groups P and R, $\bar{x}_j^P$ and $\bar{x}_j^R$, *j=1,...,6,* the Index would be equal 0 and 10, respectively. The corresponding equation is

$$EWGI(i,t) = 10 \cdot \frac{\sum_{j=1}^{6} f_j^*(x_{ij(t)} - \bar{x}_j^P)/s_j}{\sum_{j=1}^{6} f_j^*(\bar{x}_j^R - \bar{x}_j^P)/s_j}. \quad (10)$$

2.6 Step 6: time series and partial derivatives *∂EWGI(i,t)/∂xij(t)*

To obtain time series of the Index for country i, consider the values of $EWGI(i,t)$ for t = 1995, ..., 2004. The partial derivatives of $EWGI$ with respect to $x_{ij}(t)$ are

$$\partial EWGI(i,t)/\partial x_{ij}(t) = Const \cdot f_j^*/s_j. \quad (11)$$

Note that the derivatives in eqn (11) do not depend on i and t.

3 Application of DCA for constructing the *EWGI* for EU countries in the period 1995–2004

3.1 Step 1

We create the training sample as the average of 1999 and 2000 data. Table 1 presents the parameter averages and standard deviations.

Table 1: Means and standard deviations for training sample.

Variable	*GDP*	*LABR*	*LIFE*	*CPI*	*UNEMP*	*INFL*
m_j	16530.540	89.220	76.102	6.510	8.420	3.318
s_j	11112.330	31.726	2.776	1.990	4.208	2.865

3.2 Step 2

Three clustering algorithms were applied to the training sample. The results of the Complete Linkage method are presented on fig.1. The results of Ward algorithm are similar.

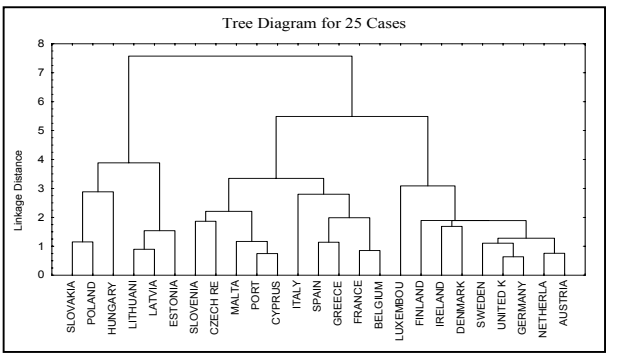

Figure 1: The results of classification by complete linkage method.

The Iterative k-means algorithm divides the countries in thee clusters, with 6 countries in P, 8 - in "Middle", and 11 – in R cluster. The descriptive statistics of these three clusters are presented in table 2. The cluster, called "R", consists of countries with highest values of indices *GDP*, *CPI, LABR*, *LIFE* and with the lowest values of indices *INFL*, *UNEMP*. The cluster called "P" consists of countries with the lowest values of indices *GDP*, *CPI, LABR*, *LIFE* and with the highest values of indices *INFL*, *UNEMP*. The remaining countries constitute the "Middle" cluster.

The classification results have some small deviations from method to method. We use a "robust" decision, i.e. the clusters are defined as follows. R and P consist of those countries, which were assigned to these clusters by all three methods; "*Middle*" cluster consists of those objects that do not belong to R and P. We also homogenize the data for R by taking out (for computational purposes only) *Luxembourg* that looks like an outlier. (The distance of this country to the center of R is twice as large as the distance of other countries). Finally, two "extreme" clusters are: P – Estonia, Hungary, Latvia, Lithuania, Poland and Slovakia; R – Austria, Belgium, Denmark, Germany, Finland, France, Ireland, Netherlands, Sweden, UK. The M - cluster contains Cyprus, Czechia, Greece, Italy, Malta, Portugal, Slovenia and Spain.

Table 2: Standardized descriptive statistics of the clusters.

	Cluster P (6)		Cluster M (8)		Cluster R (11)	
Variable	*Mean*	*Std.Dev.*	*Mean*	*Std.Dev.*	*Mean*	*Std.Dev.*
GDP	-1.150	0.061	-0.447	0.366	0.953	0.594
LABR	-1.334	0.314	-0.135	0.622	0.826	0.497
LIFE	-1.456	0.393	0.162	0.792	0.676	0.297
CPI	-1.094	0.461	-0.382	0.488	0.874	0.672
UNEMP	1.168	0.903	-0.018	0.683	-0.624	0.648
INFL	1.005	1. 513	-0.049	0.628	-0.512	0.289

3.3 Step 3

Using the Discriminant Analysis module of the package STATISTICA/WIN, we obtain the coefficients of the FDF. As we can see in table 3, five coefficients have correct signs but *UNEMP* has wrong sign, which contradicts the logic of its contribution to the Index. In this case we GO to step 4.

Table 3: Discriminant function coefficients.

	Coefficients for variables					
	GDP	*LABR*	*LIFE*	*CPI*	*UNEMP*	*INFL*
FDF	4.060	4.217	0.973	0.507	0.021	-1.138
f^*	3.669	3.805	1.205	0.799	-0.342	-1.213

3.4 Step 4

For repairing the "incorrect" sign, consider correction of the Fisher vector. The smallest value of α, which provides that the vector f_α has coordinates with "correct" signs, is equal to 0.2. The corrected vector f^* is presented in table 3. Using the within-group variance-covariance matrix for clusters P and R we obtain the value of D^2 =226.7. For the original FDF the Mahalanobis distance equals to 502.8. Our correction of the Fisher vector makes worse the separability of the clusters, but it still remains very good.

3.5 Step 5

Let us present the formula for the Index, see eqn (10). The coordinates of centroids for clusters P and R in the training sample (in nonstandardized form) are presented in first two rows of table 4. The standard deviations of variables in the training sample are given in the last row of table 1.

Table 4: The coordinates of centroids for clusters P and R.

Variable	*GDP*	*LABR*	*LIFE*	*CPI*	*UNEMP*	*INFL*
$\bar{x}_j^P$	3747.58	46.90	72.06	4.33	13.33	6.10
$\bar{x}_j^R$	25250.75	111.47	77.98	8.21	6.14	1.795

The constants for eqn (10) are: C=0.468 and B=16.942. Thus eqn (10) takes the following form:

$$\begin{aligned} &EWGI = GDP/6474.14 + LABR/17.82 + LIFE/4.92 + CPI/5.32 - \\ &- UNEMP/26.30 - INFL/5.05 - 16.94 = 0.000154 * GDP + 0.0561 * LABR + \\ &+ 0.203 * LIFE + 0.188 * CPI - 0.038 * UNEMP - 0.198 * INFL - 16.94 \end{aligned} \quad (12)$$

So, taking the values of variables from sample for Ireland (1999) - *GDP*=25054.5, *LABR*=122, *LIFE*=76.3, *CPI*=7.45, *UNEMP*=4.95, *INFL*=3.9, we obtain the value of *EWGI* equal to 9.71. Eqn (12) shows that the increase of *GDP* by 6474 dollars, or *LABR* by 17.8 units, or *CPI* by 5.32 units, etc., lead to the *same increase* of *EWGI* by 1 unit. It should be taken into account that in forecasting the *EWGI,* the increase of one of the variables, e.g. *GDP*, usually goes together with increase or decrease of other variables, e.g. *LABR* or *UNEMP*.

3.6 Step 6

To obtain time series of the Index values for country *i*, consider the values of *EWGI (i, t)* for *t* = 1995,1996, ..., 2004. The results in the form of three sets of time series are presented on figures 2 - 4.

On figure 2 we see the dynamics of *EWGI* for six countries of the P-group. These graphs allow comparing the rate of economic progress of the countries.

For instance, Slovakia had the best initial position at the beginning of the period, but did not keep it till the end of the period. Contrary, Lithuania had the lowest *EWGI* at the beginning of the period and close to the best in this group in 2004.

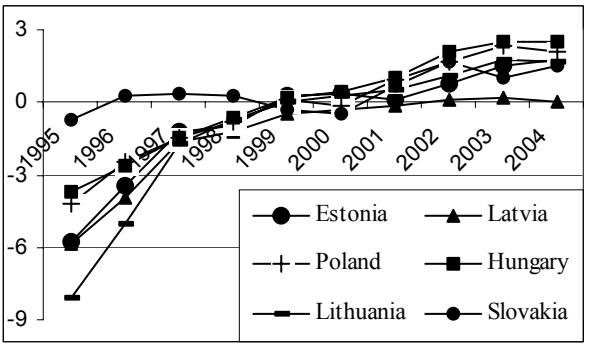

Figure 2: Time series of *EWGI* values for countries of cluster P.

On figure 3 we see the dynamics of *EWGI* for six countries of the R-group. Let us denote by *d* the overall increase of the Index during 10 years. The largest values of *d* have Ireland (*d*=5.7), Luxembourg - not shown - (*d*=4.5), UK (*d*=3.5) and Denmark (*d*=2.7). The smallest value has Germany (*d*=0.3).

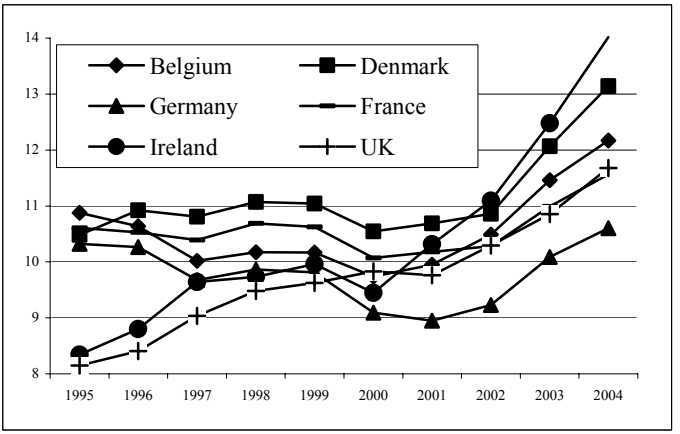

Figure 3: Time series of *EWGI* values for six countries of cluster R.

Figure 4 presents the comparison of *EWGI* of countries from different clusters. Special interest deserves Ireland, with largest *d* in groups R and M. Italy and Spain are on the top of the M-group, and the largest *d* has Slovenia (*d*=4.5), Greece (*d*=3.05), Spain (*d*=2.9) and Czechia (*d*=2.5).

The overall ranking of EU countries in 2004 is the following: Luxembourg (18.6), Ireland (14.0), Denmark (13.1), Sweden (12.5), Finland (12.2), Belgium

(12.2), UK (11.7), France (11.6), Netherlands (11.5), Austria (11.4), Germany (10.6), Spain (9.09), Italy (8.46), Greece (7.04), Cyprus (6.44), Malta (6.35), Slovenia (5.65), Portugal (5.43), Czechia (3.55), Hungary (2.56), Estonia (1.95), Poland (1.79), Lithuania (1.77), Slovakia (1.19), Latvia (0.086).

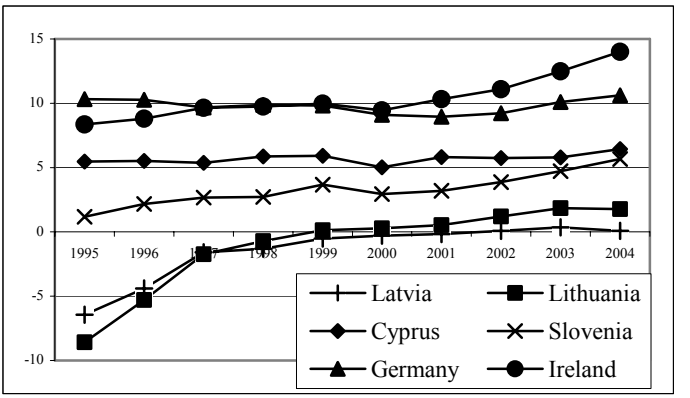

Figure 4: Examples of time series for six countries from different clusters.

4 Conclusions

We demonstrated how to obtain the Economic Welfare Growth Index (*EWGI*) using the suggested Dynamic Classification Algorithm (DCA). The DCA is a combination of cluster analysis and Fisher discrimination techniques applied to a training sample based on two year data average and extended to data samples for all years. The *EWGI* demonstrates the distinction between less and more prosperous countries, allows analysing and comparing their economic welfare progress, permits making forecasts about the changes in the *EWGI* as a result of the changes in the basic economic indices.

References

[1] Statistical Databases. http://unstats.un.org/unsd/snaama/dnllist.asp
[2] Corruption Surveys and Indices. http://ww1.transparency.org/surveys
[3] Gertsbakh, I.B., *Models of Preventive Maintenance*. North Holland: Amsterdam-New York - Oxford, pp.229-237, 1977.
[4] Johnson, R.A., & Wichern, D.W., *Applied Multivariate Statistical Analysis*, 4th ed, Prentice Hall: New York, pp.661-665, 1999.
[5] Osberg, L., & Sharpe, A., How Should We Measure the "Economic" Aspects of Well-Being? *Review of Income and Wealth*, Series 51, Number 2, June, pp.311-336, 2005.