UDC 316.42+330.34+519.711.2

PRINCIPAL COMPONENT ANALYSIS FOR STUDYING THE WORLD SECURITY PROBLEM

T. POMERANTSEVA, A. BOLDAK

This research is a continuation of the work [1], in which the list of ten most essential global threats to the future of mankind have been presented. The initial data on each threat are taken from the respectable international organizations data bases. Then, we defined the summarized impact of the examined ten global threats totality on different countries based on cluster analysis method with the purpose of selecting groups of the countries with "close" performances of summarized threats. By using the Minkovsky type metric the foresight of the future global conflicting has been executed. To facilitate the analysis and make it easier we use the method of Principal Component Analysis (PCA) which allows reduce variables with many properties to several hidden factors. The analysis shows that currently the most considerable threats for most countries are the reduction of energy security, worsening of balance between bio capacity and human demands and the incomes inequality between people and countries.

INTRODUCTION

In the work [1] the impact of system world conflicts on sustainable development is studied in the global context. On the basis of data analysis pertaining to the global conflicts taking place from 705 B.C. till now the regularity of their flow is determined. It is shown that the sequence of life cycles of system world conflicts is subordinate to the law of Fibonacci series, and the intensity of these conflicts, depending on a level of technological evolution of a society, builds up under the hyperbolic law. By using the revealed regularities we attempt to foresee the upcoming world conflict, called "the conflict of the XXI century" and analyze its nature and principal performances: - durations, main phases of the flow and intensity.

The totality of main global threats generating the conflict of the XXI century is given. These global threats are: ES — Energy Security; FB — Footprint and Biocapacity Balance; GINI — Incomes Inequality; GD — Global Diseases; CM — Child Mortality; CP — Corruption Perception; WA — Water Access; GW — Global Warming; SF — State Fragility; ND — Natural Disasters. By the cluster analysis method we define the impact of the above threats on different countries and on twelve large groups of countries (civilizations according to Huntington) combined by common culture features. Assumptions are made as to possible scenarios in the course of the conflict of the XXI century and after its termination.

Since it is difficult to analyze the security of this or that country simultaneously in the space of ten global threats, to make the research more convenient and demonstrative we use the Principal Component Analysis (PCA). This method makes it possible to reduce analysis of many properties to some hidden factors determining these properties. In this case the security of a country may be presented in a simplified form not by all ten global threats, but some most significant factors.

APPLICATION OF THE PRINCIPAL COMPONENT METHOD FOR THE ANALYSIS OF THE IMPACT OF GLOBAL THREATS TOTALITY ON SUSTAINABLE DEVELOPMENT

The example of sustainable development global simulation [2] presents global threats and degree of their impact on different countries. Let us format table 1 in the form of the initial data matrix, X_N^m, $N=106$, $m=10$, in such a way that its lines X_i, $i=\overline{1,N}$ correspond to the analyzed countries, and the columns X^j, $j=\overline{1,m}$ contain the values of threats (indicators) PX_k, $k=\overline{1,m}$, $m=10$. Then, for each country there will be the corresponding vector $X_i=\langle x_i^1, x_i^2, \ldots, x_i^m \rangle$ of threats values (the upper index corresponds to the threat's ordinal number).

The purpose of the given study conducted with application of the principal component method is finding out and interpreting latent common factors with simultaneous goal to minimize both their number and the degree of dependence PX_i on their specific residual random components. Suppose that each threat PX_i is a result of impact m' of hypothetical and one characteristic factor [3]: $PX_i=\sum_{j=1}^{m'} q_j^i F_j + e_i$, $i=\overline{1,m}$, where q_j^i — factor loadings; F_j — factors to be defined; e_i — characteristic factor for the *i*-th initial feature representing independent random value with zero mathematical expectation and finite variance.

The expression for PX_i may be presented in matrix form:

$$X_N^m = VQ^T + E, \text{ where} \qquad (1)$$

1. V — matrix of factor scores; Q — matrix of factor loadings; E — matrix of residuals.

Searching of principal components is reduced to finding the matrix decomposition X_N^m in the form (Lindsay I. Smith, 2002): $X_N^m = TP^T + E$, where T — matrix of scores with dimension $N \times m'$ $(m' \le m)$. Each line of this matrix is a projection of data vector X_i^m on m' of principal components. Number of lines — N corresponds to the number of vectors of the initial data. Number of columns or number of principal components vectors selected for projection is equal m'. P — loadings matrix of dimension $m' \times m$, where m' — number of lines (data space dimension); m — number of columns (number of vectors of principal components selected for projection); E — matrix of residuals.

Matrix of scores assigns a set of vectors $T_i = \langle t_i^j \rangle$, $i=\overline{1,N}$, $j=\overline{1,m'}$, determining projectors of vectors X_i^j, $i=\overline{1,N}$, $j=\overline{1,m}$ in the principal components space (number of components is equal $m' \le m$). Matrix of loadings assigns the mapping of the initial space basis in principal components space. The principal component method allows find such mapping $R^m \xrightarrow{F} R^{m'}$, that $m' \le m$ and $\sum_i \sum_j e_{ij}^2 \to \min$ for all possible T and P [3].

Defining principal components is connected with calculation of eigenvectors of the covariance matrix [3, 4], defined as:

$$C = (c_{ij} : c_{ij} = \operatorname{cov}(PX_i, PX_j)), \quad i=\overline{1,m}, \quad j=\overline{1,m}, \qquad (2)$$

where $\mathrm{cov}(PX_i, PX_j) = \dfrac{\sum_{k=1}^{N}(x_k^i - \overline{X^i})(x_k^j - \overline{X^j})}{N-1}$ — covariance of parameters PX_i and PX_j.

For selection of sufficient number $m' \le m$ of principal components a cumulative variance is often used [5]:

$$D_i = \frac{\sum_{j=1}^{i}\lambda_j}{m},\ i = \overline{1,m}, \tag{3}$$

where λ_j, $j = \overline{1,m}$ — eigenvalues of covariance matrix C are used.

Preliminary analysis of principal components is given in Table 1.

Table 1. Analysis of principal components

Value	Eigenvalues	Total variance, %	Comulative Eigenvalues	Comulative, %
1	5,065629	50,65629	5,065629	50,65629
2	1,331475	13,31475	6,397103	63,97103
3	1,065071	10,65071	7,462175	74,62175

We shall define the sufficient number of principal components by using the "slide rocks" criterion suggested by [6]. "Slide rocks" is a geological term to define rock debris accumulated in the lower part of a rocky slope. Using this analogy it is possible to show graphically (Fig. 1) the eigenvalues presented in table 1. It is necessary to find such a place in the plot where a decrease of eigenvalues left to right is maximally slow. It is supposed that to the right from this point only "factorial slide rocks" are located. In accordance with this criterion only 2 or 3 factors may be left.

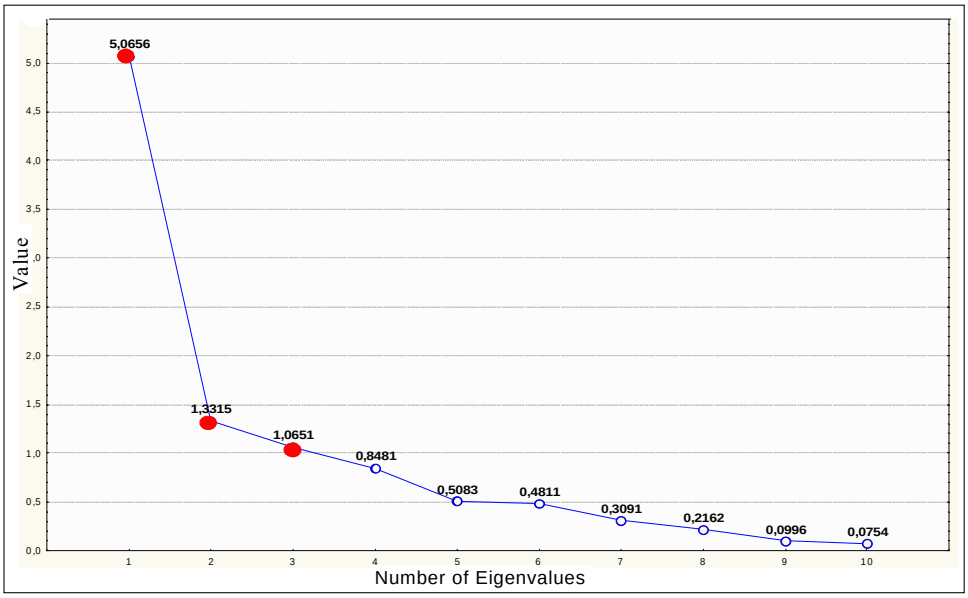

Fig. 1. Defining principal components by using "slide rocks" criterion

As seen from the above presented data it is sufficient to use three first principal components (the eigenvalues corresponding to them are indicated in red) to represent the data variability higher than 74 %.

Definition of factor loadings. Now let us analyze principal components and consider solving a problem with three factors. For this we consider correlations between threats and factors (or "new" variables) which are calculated by the formula [7]:

$$r_{k,l} = \frac{\sum_{i=1}^{N}(x_i^k - \overline{X^1})(x_i^l - \overline{X^l})}{\sqrt{\sum_{i=1}^{N}(x_i^k - \overline{X^k})^2}\sqrt{\sum_{i=1}^{N}(x_i^l - \overline{X^l})^2}}, \qquad (4)$$

where $r_{k,l}$ — correlation coefficient of parameters X^l and X^k; $\overline{X^1}$, $\overline{X^k}$ — average values of parameters X^l and X^k; $\overline{X^l} = \frac{\sum_{i=1}^{N} x_i^l}{N}$; $\overline{X^k} = \frac{\sum_{i=1}^{N} x_i^k}{N}$.

The correlation coefficient itself does not have informal interpretation. However, its square called the coefficient of determination shows to what extent variations of dependent characteristics may be explained by variations of an independent one. It is thought that correlation coefficients which by their module are more that 0.7 indicate a strong connection (in this case coefficients of determination > 50%, i.e. one characterististics determines the other more than by half. Correlation coefficients which by their module are less that 0.7, but more than 0.5 indicate that connection is average (in this case the coefficients of determination are less than 50%, but more than 25%). At last, correlation coefficients which by their module are less than 0.5 indicate a weak connection (here the coefficients of determination are less than 25 %). Table 2 shows the values of correlation coefficients between principal factors and initial threats. The coefficients corresponding to strong connections are indicated in red.

From Table 2 it is seen that the first factor to greater extent correlates with threats than the second and third factors. It should be expected, since, as it has been mentioned above, factors are defined sequentially and contain less and less total variance.

Table 2. Correlation coefficients between principal factors and initial threats

Variable	Factor 1	Factor 2	Factor 3
ES	0,208964	0,817502	0,342974
FB	– 0,855800	0,412124	0,053021
GINI	– 0,355499	0,105301	– 0,716591
CP	– 0,856876	0,248258	– 0,003646
NA	– 0,809616	– 0,315140	0,210144
GW	0,723432	– 0,392527	– 0,006533
CM	– 0,844045	– 0,267343	– 0,024123
ND	– 0,326707	– 0,285766	0,615743
SF	– 0,899250	– 0,086816	– 0,005283
GD	– 0,788874	– 0,080839	– 0,084617
Expl. Var	5,065629	1,331475	1,065071
Prp. Totl	0,506563	0,133147	0,106507

Interpretation of factor structure. It is convenient to carry out interpretation of factors (principal components) by using a diagram where threats are shown as vectors the coordinates of which correspond to factor loadings (Fig. 2).

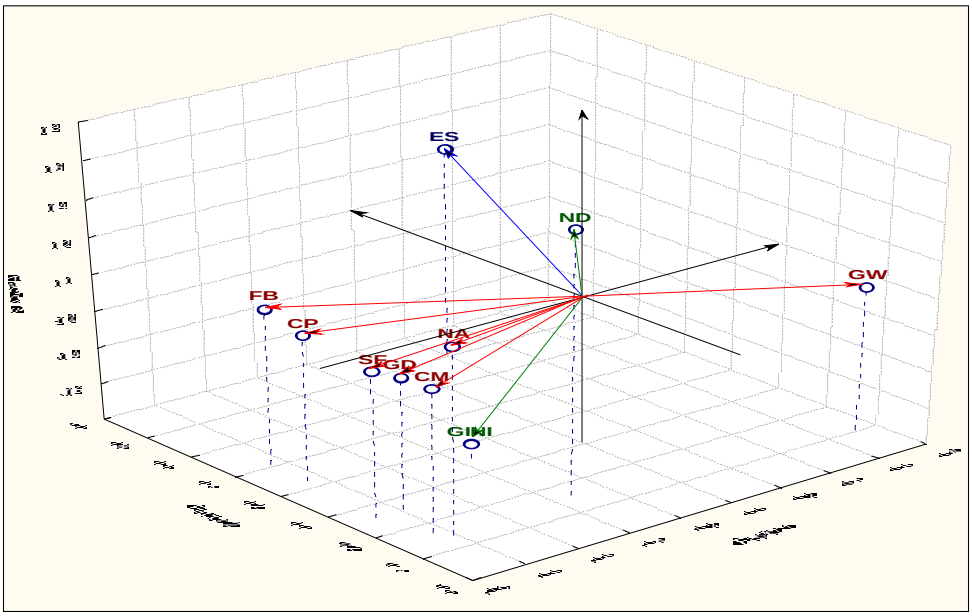

Fig. 2. Interpretation of threats in coordinates of principal components

In accordance with maximum factor loadings threats may be divided into three categories (red, blue and green coulours). The first group of threats includes: FB, CP, SF, GD, NA, CM, GW. As seen in fig. 2 these threats are in the plane of the first and second factors. It means that for more detail analysis it is advisable to show them in the projection on this plane (Fig. 3).

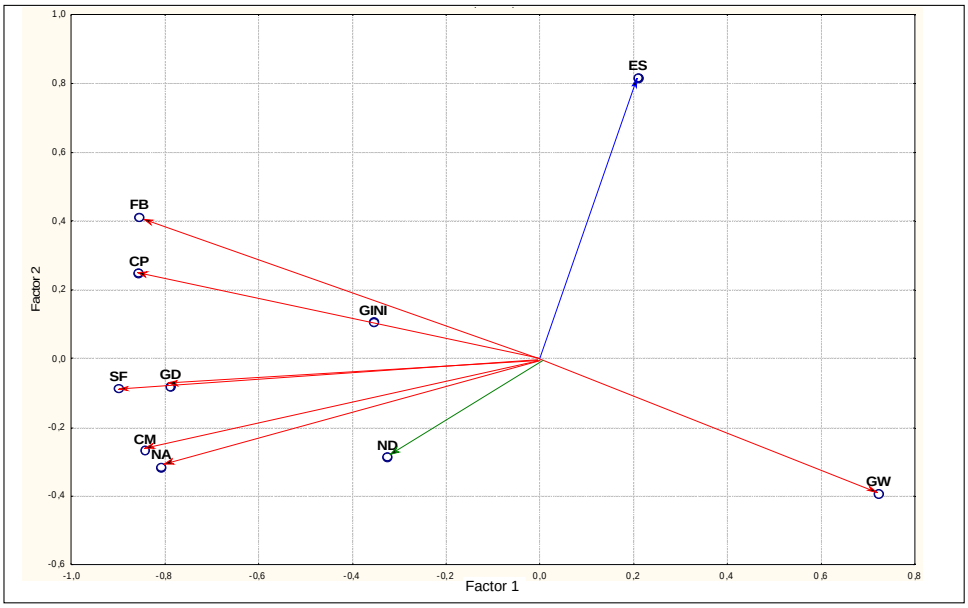

Fig. 3. Projection of threats on the plane of the first and second factors

As seen from Fig. 3 the pairs of vectors SF-GD, FB-GW are practically colinear, which indicates their high degree of dependence. It is interesting that we study only two factors, then the pair of vectors CP-GINI may be considered as colinear. It should be also noted that the vector ES is orthogonial to FB (GW).

It means that:

- between level of energy security (ES), balance of biological capacity of the Earth and people's needs (FB) and CO2 emissions(GW) the dependence is inconsiderable;
- balance between biological capacity of the Earth and people's needs(FB) and CO2 emissions (GW) has negative correlation;
- level of state fragility (SF)) is closely connected with level of global diseases vulnerability(GD);
- corruption perception index (CP) is closely connected with level inequality between people and countries (GINI) in the context determined by the first and second factors.

The most significant global threats are defined by using factor loadings of the initial list of threats. For this it is necessary to select such factors which have maximum loading by absolute value on the first, second and third factors. This choice ensured the definition of maximum impact of initial threats under condition of their maximum independence on the aggregated indicator (Minkovsky norm) of these threats (Fig. 4).

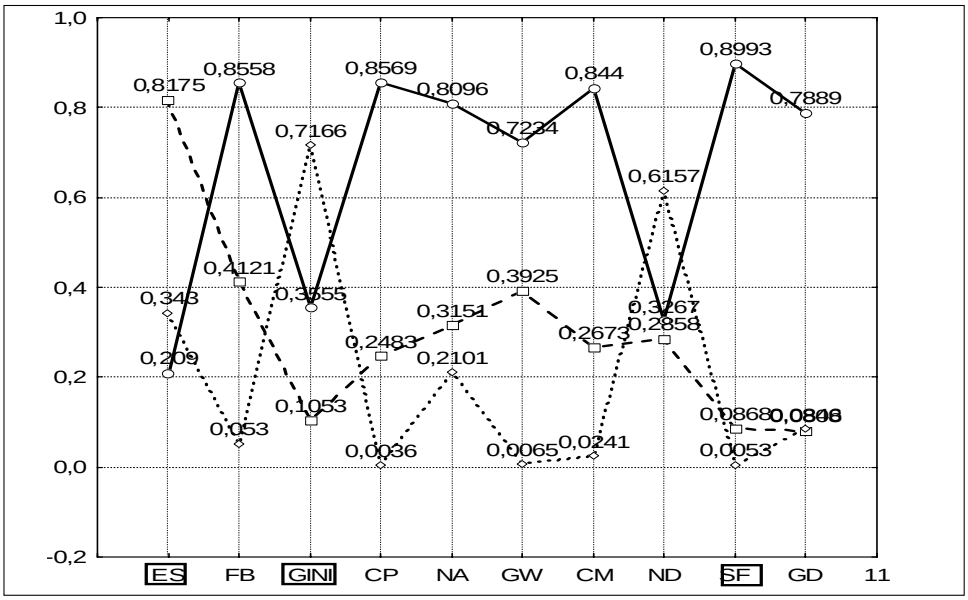

Fig. 4. Definition of most significant global threats

In accordance with the indicated approach such threats are SF, ES, GINI, (Fig. 4) i.e. the most significant threats in descending order are **state fragility, global decrease of energy security and growing inequality between people and countries**.

Clustering of countries by the level of global threats and the corresponding graphic interpretation is done in the plane of the first and second factors. For this purpose we cluster countries by the degree of their remoteness from threats (Minkovsky norm) using the clustering method of *K*-averages.

As seen from Fig. 5 the isolines which assign the Minlovsky norm approximation are practically orthogonal to the first factor axis. It gives the ground to state that the first factor values mostly determine the countries' remoteness from global threats (Fig. 5).

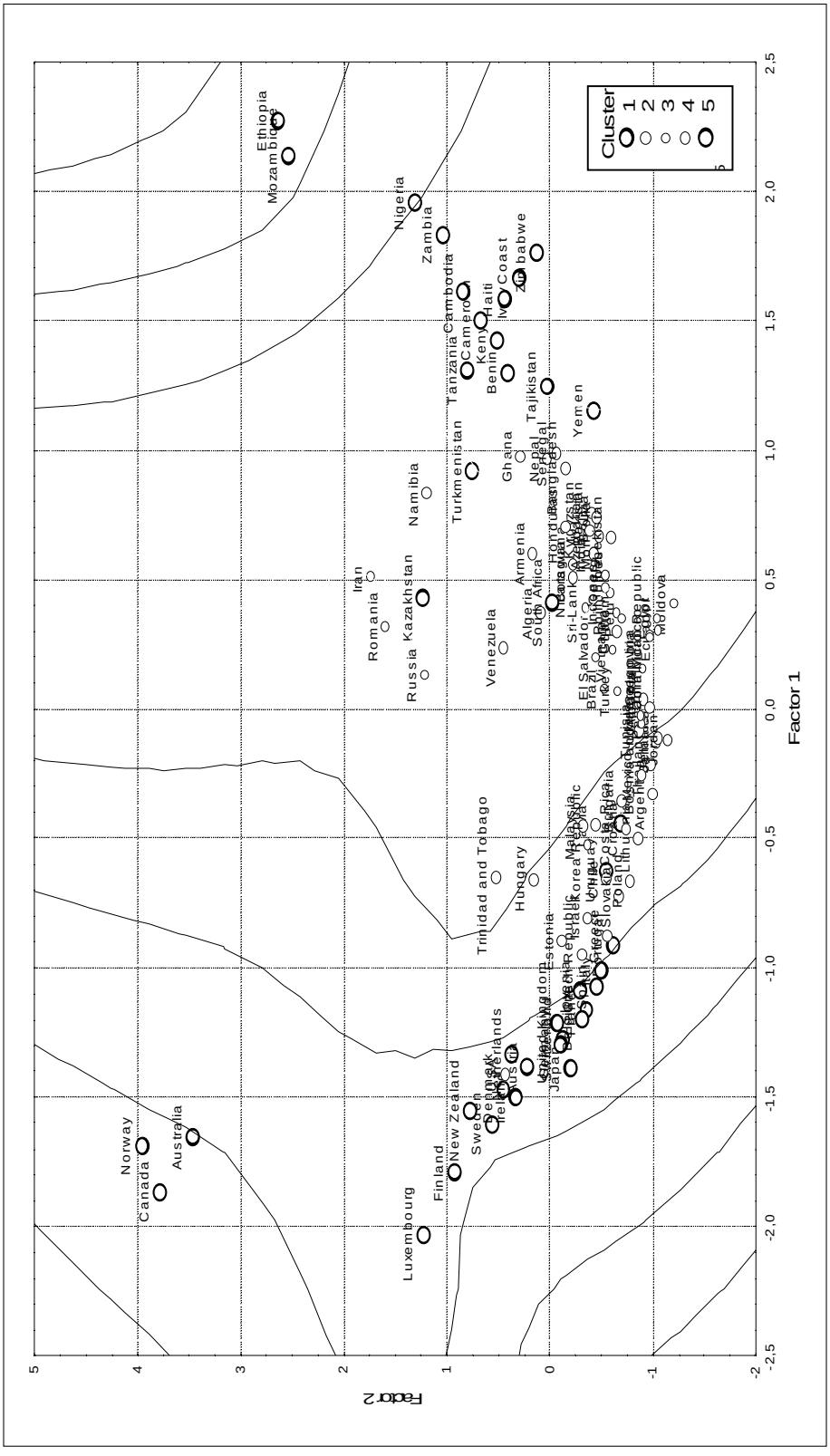

Fig. 5. Interpretation of global threats in the plane of the first and second factors

RESEARCHING THE DEPENDENCE OF COUNTRIES' NATIONAL SECURITY ON PARTICULAR THREATS BY USING MODIFIED METHOD OF WEIGHTED LOCAL CORRELATION

Let us consider that the quantitative value of Minkovsky norm for this or that country is an estimate of its national security level. We define the level of Minkovsky norm dependence on initial threats by calculating the corresponding correlation coefficients (Table 3).

Table 3. Correlation coefficients between Minkovsky norm and global threats

Varuable	ES	FB	GINI	CP	NA	GW	CM	ND	SF	GD
Minkovsky norm	– 0,16	0,80	0,31	0,82	0,83	– 0,54	0,83	0,40	0,89	0,78

The calculated correlation coefficients show a high degree of dependence of Minkovsky norm on initial threats, but at the same time do not answer the question what risks the countries are running from the point of view of their approaching various threats. The reason is the averaging of correlation coefficients on the entire data sample.

For detailed analysis of global threats the countries may face, it is necessary to localize the sample on which correlation is estimated. It is natural to assume that this sample should include "alike" countries the degree of similarity of which may be estimated as, for example, a Euclidean distance in the space of threats. The second assumption is connected with the idea that the closer is a country to the point in which the correlation is analyzed; the higher is the degree of the country's indicators impact on the correlation coefficient.

In accordance with the above assumptions we define the weighted mean [8] as:

$$m(X,W) = \frac{\sum_i w_i x_i}{\sum_i w_i}, \tag{5}$$

where X — data sample; W — weighted function.

If we define W, as function depending on distance, for example,

$$W(x,t) = e^{-\lambda d(x,t)}, \tag{6}$$

in which: $d(x,t)$ — distance between points x, $t \in R^n$, and λ — distribution parameter and substitute in (5), then we get the expression for calculating the weighted localized mean in point t for sample X:

$$m(X,t) = \frac{\sum_i e^{-\lambda d(t,x_i)} x_i}{\sum_i e^{-\lambda d(t,x_i)}}, \; x_i \in X. \tag{7}$$

Similarly, we can define the weighted localized covariation:

$$\operatorname{cov}(X,Y,t)=\frac{\sum_i e^{-\lambda d(t,x_i)}(x_i-m(X,t))(y_i-m(Y,t))}{\sum_i e^{-\lambda d(t,x_i)}}. \quad (8)$$

And we define the weighted localized correlation (WLC):

$$\operatorname{corr}(X,Y,t)=\frac{\operatorname{cov}(X,Y,t)}{\sqrt{\operatorname{cov}(X,X,t)\operatorname{cov}(Y,Y,t)}}. \quad (9)$$

The distribution parameter of weights λ may be chosen in such a way that it is possible to restrict the impact area of point's located at large distances. For example, we assume that points located at mean distance from the point where WLC is calculated have the weight equal S (distribution scale). I.e.

$$e^{-\lambda(t)m(d_t)}=s \text{, then } \lambda(t)=\frac{\ln(s)}{m(d_t)}, \quad (10)$$

where $m(d_t)$ — mean distance from the sample points to point t. Examples of weights distribution for different values of mean distance and distribution scale are given in Figs. 6, 7.

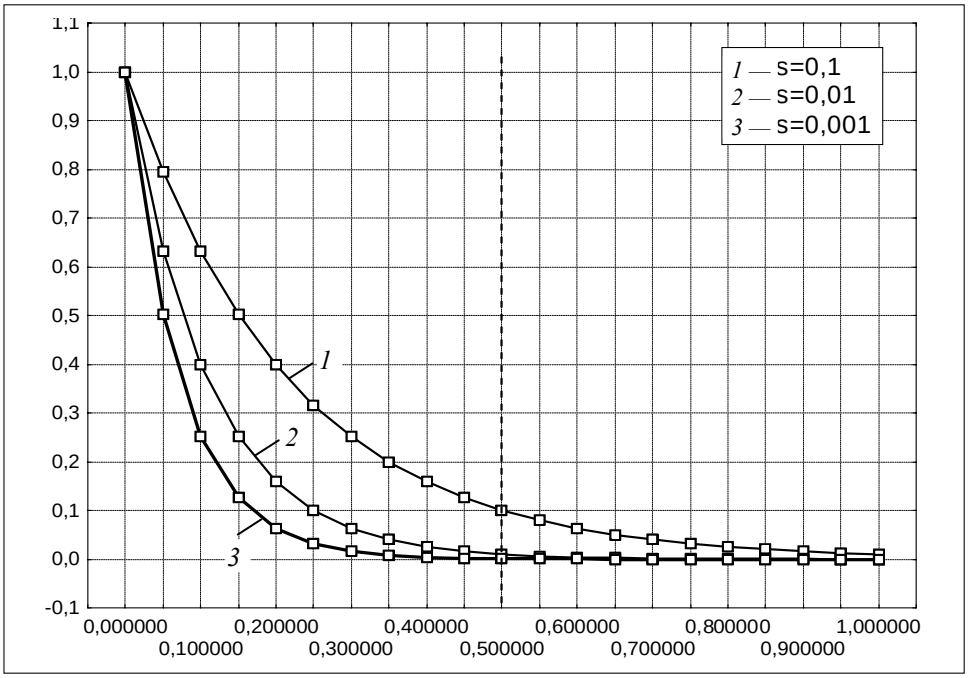

Fig. 6. Weights distribution for mean distance equal 0,5

With distribution scale equal 1, WLC coincides with Pearson product-moment correlation coefficient. As seen from (10), the weights distribution parameter is calculated for each point t, which is a sample point. And for each new point the mean distance value is calculated $m(d_t)$ anew. Hence, the suggested method of estimating threats local dependence is adaptive. The interpretation of WLC values is presented in Table 4.

Table 4. Interpretation of values of weighted localized correlation (WLC)

Value of WLC	Behavior of global threats under study	Interpretation
[– 1.0, – 0.5)	High degree of negative correlation (more than 25 %). The growth of one threat is connected with reduction of the other	With a decrease of a particular threat the general remoteness from the totality of global threats considerably decreases. The studied threat has low (as compared to others) contribution to the general remoteness from global threats
[– 0.5, – 0.3)	Mean degree of negative correlation (9–25%). The growth of one threat is connected with reduction of the other	With a decrease of a particular threat the general remoteness from the totality of global threats considerably decreases at the mean degree
[– 0.3, 0.3]	Low degree of correlation (less than 9%)	It is possible to speak about an inconsiderable dependence of the degree of remoteness from the totality of global threats on the studied threat
(0.3, 0.5]	Mean degree of positive correlation (9 – 25 %). The growth of one threat is connected with the growth of other	With a decrease of the particular threat the general remoteness from global threats increases at the mean degree
(0.5, 1.0]	High degree of positive correlation. The growth of threat is connected with the growth of other (more than by 25%)	With a decrease of the particular threat the general remoteness from the totality of global threats considerably increases. The studied threat considerably influences the general remoteness from the totality of global threats

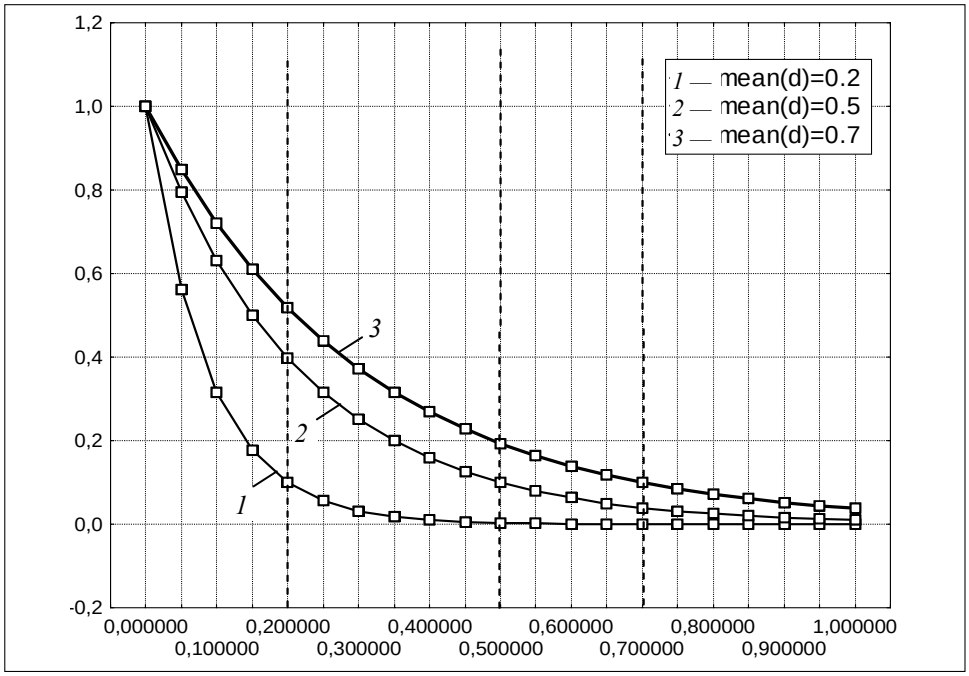

Fig. 7. Weights characteristics for scale distribution equal 0,1

Figs. 8–10 present the plotted values of weighted localized correlation (WLC) between Minkovsky norm and most significant threats, respectively: SF, ES и GINI.

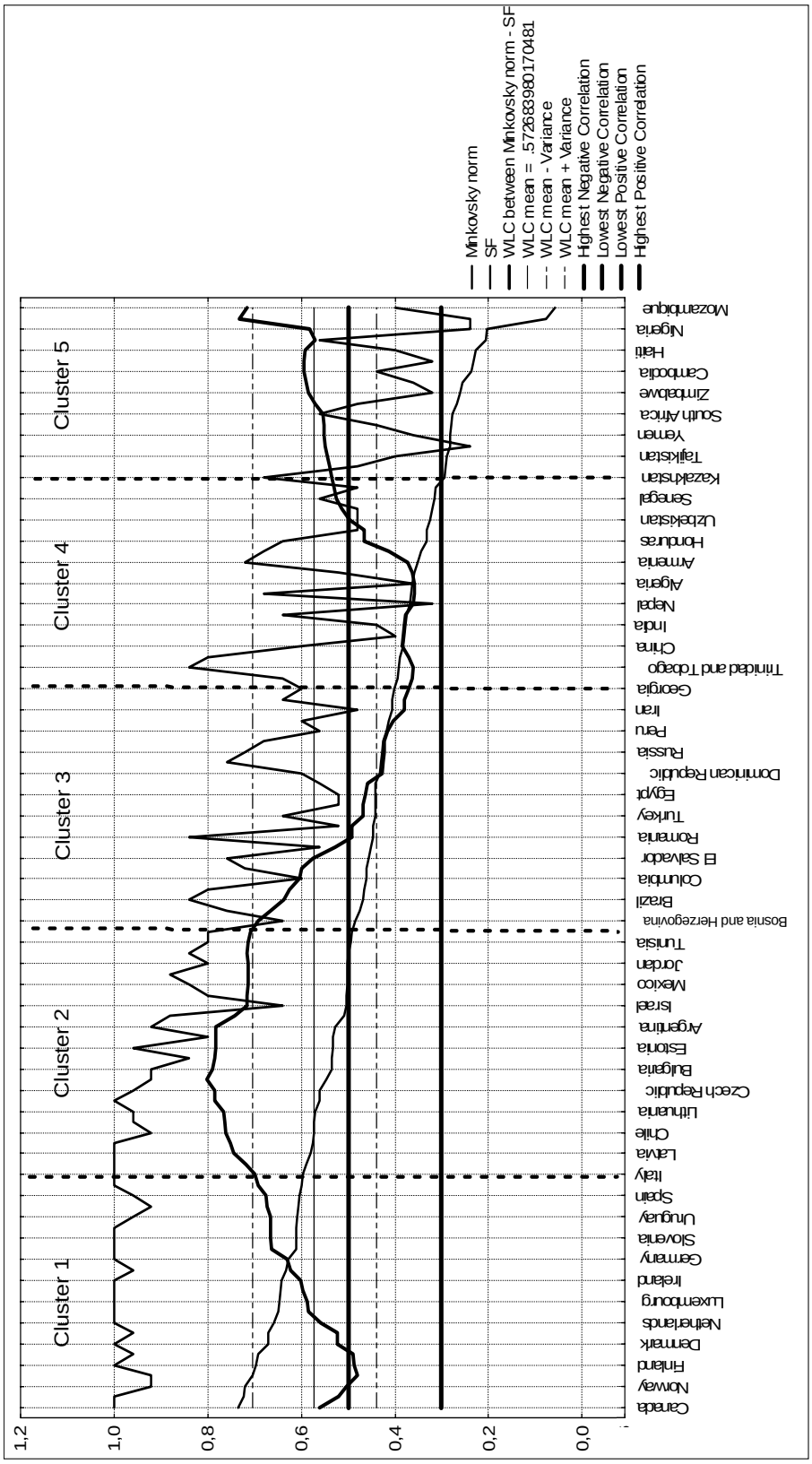

Fig. 8. Values of WLC between Minkovsky norm and state fragility (SF)

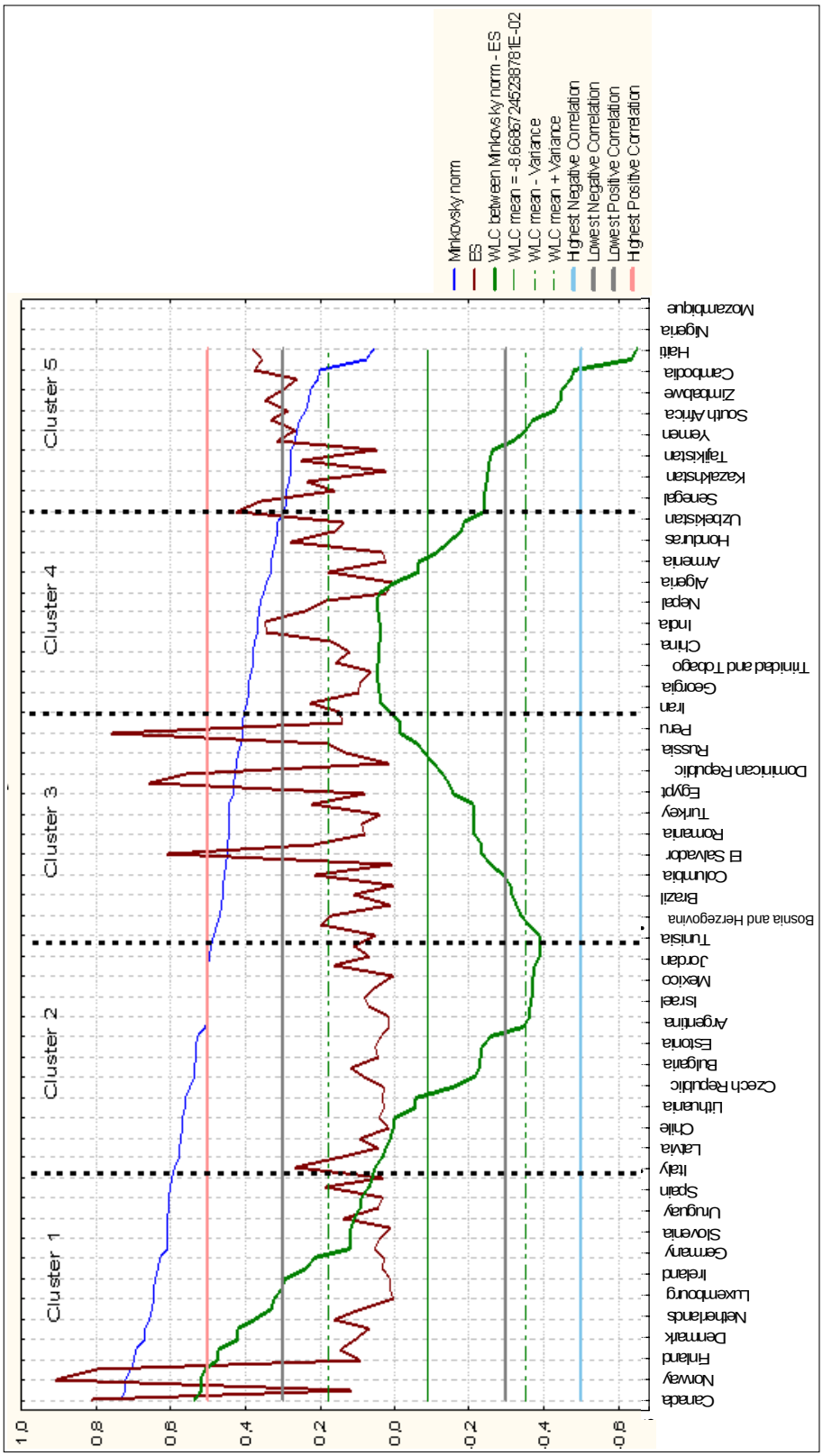

Fig. 9. Values of WLC between Minkovsky norm and energy security (ES)

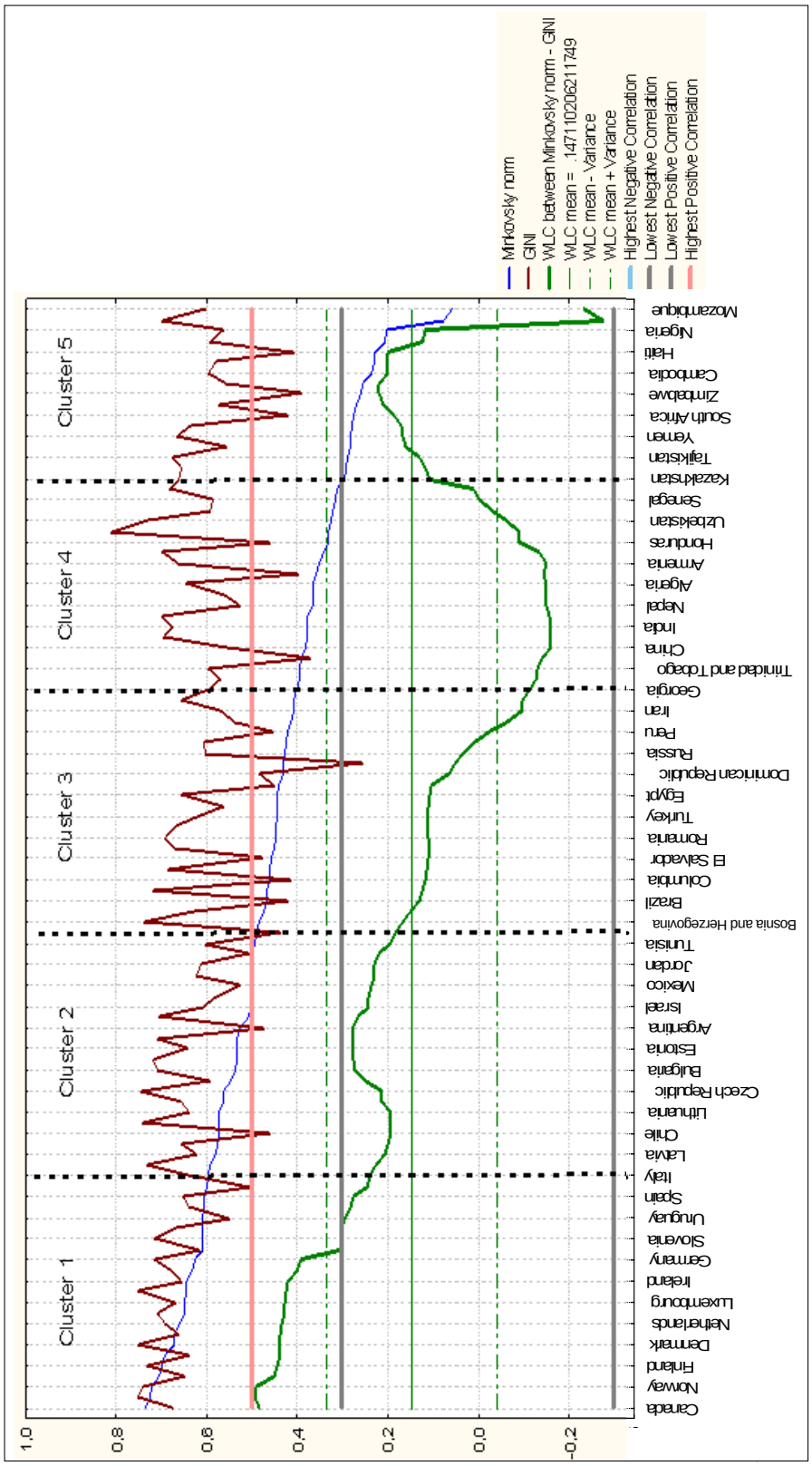

Fig. 10. Values of WLC between Minkovsky norm and population inequality (Gini)

As seen from Fig. 8 the level of state fragility (SF) for most countries has considerable impact on their national security.

As to the impact of energy security on the level of national security (Fig. 9), the following groups of countries may be identified [9]:

- A group of countries with high level of ES and high values of Minkovsky norm (Canada, Sweden, Norway, Australia) for which energy security considerably influences their national security.
- An adjacent group (Finland, New Zealand, Denmark, Switzerland, Netherlands, Austria, Luxembourg, Japan), for which a mean level of dependence between energy security and Minkovsky norm is observed.
- A group of countries for which this dependence is weak.
- A group of countries with mean level of national security (Belarus, Israel, Thailand, Mexico, Jamaica, Jordan, Malaysia, Tunisia, Panama, Bosnia and Herzegovina, Vietnam, Brazil, Ukraine, Columbia, Korea Republic), for which there exist threats more serious than energy security.
- A group of countries with low level of national security (Kenya, Zimbabwe, Cameroon, Cambodia, Zambia, Haiti, Turkmenistan, Nigeria), for which both energy security and other threats are equally important.
- A group of most problem countries (Ethiopia, Mozambique), where the level of energy security at least extent determines the level of national security.

As to the impact of population inequality on national security (fig.10) it is possible to identify a group of countries (Canada, Sweden, Norway, Australia, Finland, New Zealand, Denmark, Switzerland, Netherlands, Austria, Luxembourg, Japan, Ireland, France, Germany, Portugal, Slovenia, Belgium), for which a mean positive correlation between this threat and Minkovsky norm is observed. For the rest of countries this correlation is insignificant.

CONCLUSIONS

1. Since it is very complicated to analyze security of this or that country simultaneously in the space of ten global threats the principal component analysis (PCA) was used. This method allowed reducing ten global threats influencing the general level of national security (in the sense of Minkovsky norm) to three hidden factors determining this characteristic. The application of this approach allowed considerably facilitate research of national security, reducing it to the analysis in the space of three determining factors.

2. By using this method a comprehensive study of national security of different countries was carried out in the space of three determining factors. Factor loadings were defined by calculating coefficients of correlation between principal factors and initial threats. Clustering of countries was made according to the level of global threats, and three most significant threats were defined influencing national security of most countries: state fragility (SF), energy security (ES) and people's inequality (Gini). Graphic interpretation of global threats was done in the space of three principal components. The factor structure of threats was studied, and the degrees of dependence between main groups were defined.

3. The method of weighted localized correlation was modified, which allowed carry out research of the dependence of national security level (Minkovsky norm) on particular global threats. By using this method the dependence between Minkovsky norm and most significant threats were analyzed in detail, in particular, state fragility (SF), energy security (ES) and people's inequality (Gini). Recommendations were made for different countries regarding strengthening their national security.

REFERENCES

1. *Zgurovsky M.Z., Gvishiani A.G.* Sustainable development global simulation. Report 2008. (http://wdc.org.ua/en/node/357) Kiev.:Polytechnika, 2008. — 363 p.
2. *System* Analysis and Decisions, The example of sustainable development global simulation, 2009, from Word Wide Web: http://systemdecisions.com/index.php?option=com_content&view=article&id=30&Itemid=27.
3. *Lindsay I Smith.* A tutorial on Principal Components Analysis, 2002, URL: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components. pdf.
4. *Strang Gilbert.* Linear algebra and its applications, Thomson, Brooks/Cole, Belmont, CA, ISBN 0-030-10567-6, 2006.
5. *Jambu M.* Exploratory and multivariate data analysis. Academic Press., 1991.
6. *Cattel R.B.* The screen test for the number of factors. Multivariate Behavioral Research, 1, 245–276, 1966.
7. *Harman H.H. Modern factor analysis.* Chicago: University of Chicago Press, 1966.
8. *A MATLAB* Toolbox for computing Weighted Correlation Coefficients, 2008, http://www.mathworks.com/matlabcentral/fileexchange/20846.
9. *Pomerantseva T., Boldak A.* Human Security Analysis Under Impact of the Totality of Global Threats on Sustainability. 5-th International EURO-Mini Conference "Knowledge-Based Technologies and OR Methodologies for Strategic Decisions of Sustainable Development" (KORSD-2009), Vilnius, Lithuania, September 30 – October 3, 2009. — P. 164–169.

Received 09.07.2009

From the Editorial Board: the article corresponds completely to submitted manuscript.