Data transformations when constructing a composite system quality index

T V Zhgun
Yaroslav-the-Wise Novgorod State University, Veliky Novgorod, Russian Federation
E-mail: Tatyana.Zhgun@novsu.ru

Abstract. The features of the data distribution can significantly affect the composite characteristics of objects, so composite indexes of objects must necessarily take into account the features of the data. Some types of data are characterized by distributions with a significant anomaly, when the vast majority of observations are concentrated near the boundary values. This type of data cannot always be characterized by an asymmetry coefficient. In addition, if the values of a variable are approximately symmetric with respect to zero or are concentrated near zero, the sample cannot also be characterized by the coefficient of variation. The paper proposes a transformation that allows us to identify the anomalous nature of variables using the signal-to-noise ratio. Variables are evaluated in the standard range, which is shifted to the right relative to zero. If it is necessary to logarithm, such a transformation will avoid the pressure of small values of variables that, after direct logarithm, would have large negative values. The application of logarithmic correction for the detected anomalous variables redistributes the values of the obtained weighting coefficients in the direction of a more correct interpretation and, in particular, solves the problem with the negativity of the weighting coefficients.

1. Introduction
The development of methods for constructing composite indexes that characterize the integrative properties of weakly formalized systems is a feature of the modern stage of modeling complex systems. In general, these properties – productivity, stability, well-being – are called the quality of the system. In public practice, more than three hundred composite indexes are now used to assess the quality of poorly formalized systems [1]. The main purpose of using composite indexes is to rank the studied objects according to the calculated indicator of the system quality [1-8]. A huge number of methods used to assess the quality of poorly structured systems [1] indicate dissatisfaction with the results and the need for further research in this area [3, 6].

The construction of composite indexes is associated with a number of methodological difficulties. Among these difficulties, the following can be distinguished:

– selection of the initial indicators forming the composite index;
– transformation of the initial indicators for the purpose of their comparability with each other and for the purpose of further application of formal methods;
– the choice of the method for determining the weight coefficients [9].

The necessary transformations of the initial data remain the least studied [10]. Data transformation is one of the most important processes when calculating the quality indicator of systems. The features of the data distribution can significantly affect the composite characteristics of objects, so obtaining the characteristics of objects must necessarily take into account the features of the data. The reliability and accuracy of the results depends on how well the data transformations are carried out. The obvious
transformation is for the purpose of comparability of indicators among themselves – data scaling. As a rule, a linear transformation to a single scale is used. The most commonly used scale is from zero to one (Normalization, unification). The use of multivariate analysis methods requires "standardization" (Standard scores (or Z-scores)) of variables: subtraction of the mean and division by the standard deviation. Other types of transformations, especially nonlinear transformations, do not have clear recommendations for applying a specific data transformation [6, 11-14].

The purpose of the study is to design a data scaling transformation that allows us to evaluate the specifics of the data used to build a composite index. Namely, it allows you to identify anomalies in the distribution of variables. The analytical criterion for the anomaly of the initial data is based on the signal-to-noise ratio, which is the inverse of the covariance coefficient. The criterion evaluates variables in the standard range, confirms the assumption that it is desirable to use a logarithmic transformation of variables for which the calculated signal/noise value is less than one. The use of logarithmic correction for anomalous variables solves the problem with the negativity of the weighting coefficients and redistributes the values of the resulting weighting coefficients towards a more correct socio-economic interpretation.

2. Problem statement
Let the system be characterized by a matrix of descriptions of dimension $m \times n$ $A = \{a_{ij}\}$. Although various functional forms of composite index aggregation rules have been developed in the literature [15-17], in standard practice, the composite indicator $CI$ of an object $i$ is usually considered as a weighted linear convolution of a set of variables [18]:

$$CI_i = \sum_{j=1}^{n} w_j \cdot a_{ij}$$  (1)

Or in matrix form

$$CI = A \cdot w$$

where $CI = \{CI_1, CI_2, ..., CI_m\}^T$ — is the vector of integral indicators for the objects of the system, $w = \{w_1, w_2, ..., w_n\}^T$ — is the vector of weights of indicators. To build the desired composite index of the system quality, it is necessary to find the weights of the $w_j$ indicators.

The numerical characteristics of the system are preliminarily reduced to a single scale. Usually, the values of variables are given for the segment $[0, 1]$ according to the following rule. If the initial indicator of the $j$-th indicator of the $i$-th object $x_{ij}$ is associated with the quality property of the system by a monotone dependence, then the variables $x_{ij}$ are transformed according to the rule:

$$a_{ij} = s_j + (1 - s_j) \cdot \frac{x_{ij} - m_j}{M_j - m_j}$$  (2)

where $s_j = 0$, if the optimal value of the $j$-th indicator is the maximum and $s_j = 1$, if the optimal value of the $j$-th indicator is the minimum, $m_j$ is the smallest value of the $j$-th indicator, $M_j$ is the largest value of the $j$-th indicator.

If the initial indicator is associated with the analyzed integral quality property by a non-monotonic dependence (i.e., there is a value $x_{ij}^{opt}$ within the range of variation of the $j$-th indicator, at which the highest quality is achieved), then the value of the corresponding unified indicator is calculated by the formula:
\[ a_{ij} = 1 - \frac{x_{ij} - x_{j}^{\text{opt}}}{\max ((M_{j} - x_{j}^{\text{opt}}), (x_{j}^{\text{opt}} - m_{j}))} \].

(3)

An example of such an indicator is, for example, the unemployment rate when determining the composite index of quality of life. The zero value of the unemployment rate is not optimal. Indicators of zero unemployment are considered harmful for the development of the economy. The absence of unemployment means, in particular, the lack of competition and, as a result, a decrease in the quality of labor performance. So, although at first glance it seems that the unemployment rate of 0% is an ideal value, in fact, a small unemployment is desirable. According to experts, the ideal unemployment rate is 5%. This level allows the average income of the population to grow, and maintain a good social and economic situation. Similarly, an infinitely large indicator of the average age of the population cannot be optimal, since this means that there are no young people in the population structure.

3. Processing of the type of variables distributions

We note the peculiarity of the statistical data involved in determining the quality of poorly formalized systems. The main question that researchers ask is the question of whether the data under study belongs to a normal distribution.

To determine the proximity of the empirical distribution to the normal law, the indicators of kurtosis and asymmetry are used. The calculation of the asymmetry allows us to establish the symmetry of the distribution of a random variable relative to the mathematical expectation. To do this, we find the third central moment that characterizes the asymmetry of the distribution law of a random variable. If it is equal to zero, then the random variable is symmetrically distributed relative to the mathematical expectation. Since the third central moment has the dimension of a random variable in a cube, a dimensionless value is introduced — the asymmetry coefficient:

\[ A_{s} = \frac{m_{3}}{\sigma_{s}^{3}} \]

(4)

where \( m_{3} \) — is the central empirical moment of the third order: \( m_{3} = \frac{1}{n} \sum_{i=1}^{n} (x_{i} - \overline{X})^{3} \),

\( \overline{X} \) — sample average, \( \overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_{i} \), \( \sigma_{s} \) — standard sample deviation, \( \sigma_{s} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_{i} - \overline{X})^{2}} \).

In this case, the following conditional gradation is adopted: if \( |A_{s}| < 0.25 \), then the asymmetry is insignificant, if the calculated value of the asymmetry coefficient is \( 0.25 < |A_{s}| < 0.5 \), then the asymmetry is moderate, and if the asymmetry coefficient \( |A_{s}| > 0.5 \), then the asymmetry is significant. With significant deviations of the asymmetry indicators from zero, it is impossible to recognize the aggregate as homogeneous, and the distribution is close to normal.

Let's illustrate the properties of the data set used on a specific sample. To illustrate, let's consider a set of 37 variables that characterize the quality of life of the population of the subjects of the Russian Federation for 2010-2017. These variables have been repeatedly used by researchers to assess the quality of life of the Russian population in many studies, for example [19]. The list of variables is given in [20]. Table 1 shows the coefficients of the asymmetry of the variables. Many variables have significant asymmetry.

The asymmetry coefficient defined for the variables under consideration in Table 1 shows that the distribution of only 5 variables out of 37 can be characterized as moderate asymmetry, where \( 0.25 < |A_{s}| < 0.5 \). For all other variables, the asymmetry coefficient indicates a large skew and a complete absence of a normal distribution of variables.
Table 1. Variables skewness ratio.

| Block number | The number of the variable in the block |
|--------------|----------------------------------------|
| 1            | 1 | 5.3 | 0.5 | 1.5 | 0.3 | 0.3 | 2.7 | -0.8 | 3.4 | 4 |
| 2            | 2 | -0.5 | -2.4 | 1.8 | -1.7 | 1.1 | 0.3 | -0.3 | 0.2 | -0.4 | -0.2 | -2 | 1.5 | 5.2 | 0.3 |
| 3            | 3 | -4.9 | -2.8 | -0.3 | -1.3 | -3.4 | -3.7 | -5 | 0 | -0.6 | -1.3 | -1.8 | -1.8 | -2.7 | -0.5 |

The variables have a particularly significant asymmetry 1, 8, 9, 22, 23, 27, 28, 29. Note that the sample made up of the asymmetry coefficients of these variables can be characterized as moderately asymmetric: the average value of the asymmetry coefficient of such a sample \( A_s = -0.27 \), and the asymmetry coefficient for this sample \( A_s = 0.37 \).

The fact that in this case all the data do not obey the law of normal distribution is confirmed by the Shapiro-Wilk test. For all variables, the calculated \( p \)-value lies in the range 0.0000 – 0.0167 significantly less than \( p = 0.05 \). This means that the difference between the considered distributions and the normal one is strongly statistically significant, and the hypothesis of a normal distribution of variables must be rejected. When checking the condition of normality of the data distribution, it is necessary to have a good idea in which cases its fulfillment is critical for the application of a specific statistical method. For example, the principal component method does not require that the data be distributed normally [21]. The calculated coefficients of asymmetry show that the data under consideration have features, but do not allow us to make recommendations about the necessary data transformations.

Figure 1 shows the histograms of the frequencies of three variables: variable 1 “GRP per capita”, variable 30 “The number of registered rapes per 100 thousand people” and variable 31 "The number of robberies, robberies, and thefts from citizens’ apartments per 100 thousand people". This is an example of variables with the most prominent skew coefficients: the most significant shift to the left (Figure 1a) is positive (right — sided) asymmetry; the most significant shift to the right (Figure 1b) is negative (left — sided) asymmetry and the maximum proximity to the normal distribution (Figure 1c).

In Figure 1a, for variable 1, we see a significant predominance of small values: 620 values out of the 680 considered are grouped in the range 0-0.1. In some sense, such data can be called anomalous. The asymmetry could not characterize this feature – variables 1 (Figure 1a) and 30 (Figure 1b) have almost the same value of the absolute value of the asymmetry coefficient, but completely different distributions. For anomalous data, it is not the presence of a large spread of variable values that is eliminated by scaling, but the obvious presence of clusters of different values. This type of data is quite common in statistics. These are the characteristics of monetary values, the incidence of diseases of various types, the characteristics of the density of the road surface, etc. The presence of variables of this type can give an implausible result when calculating the composite index. It is necessary to find a
numerical characteristic of the sample, indicating the abnormality of the sample and the need to use non-standard data transformations.

4. Data characteristics: $CV$ and $SNR$

The Coefficient of Variation is used for analytical comparison of data sets. The Coefficient of Variation (Coefficient of Variation, $CV$) is a measure of the relative spread of a random variable. It shows what proportion is the average spread of a random variable from the average value of this value. The coefficient of variation is defined as the dimensionless ratio of the standard deviation $\sigma$ to the mean $m$:

$$CV = \frac{\sigma}{m}$$  \hspace{1cm} (5)

When only a sample of data is available, the $CV$ can be estimated using the ratio of the sample standard deviation $\sigma_s$ to the sample mean $\bar{X}$

$$CV = \frac{\sigma_s}{\bar{X}}$$  \hspace{1cm} (6)

The more the value of the coefficient of variation, the relatively more the spread and less alignment of the studied values. In the sample under consideration, only four indicators have a coefficient of variation less than 10%: 10, 21, 23, 27, and for 26 variables, the coefficient of variation exceeds 33%. This means that almost the entire sample is heterogeneous and far from a normal distribution.

The disadvantage of this indicator is its sensitivity to changes in the average, if it is close to zero. This will happen if the values of the variable are approximately symmetric with respect to zero or are concentrated near zero (as in the example discussed above, Figure 1a). In this case, a small change in the mathematical expectation can provide an infinitely large increase in the coefficient of variation.

Note also that the linear transformation of variables changes the coefficient of variation, its values change when scaling. Therefore, for the correct analysis of the data, the values should be considered in a single range, which gives the correct characteristics of the sample.

The coefficient of variation allows us to judge the homogeneity of the population, a high degree of which ensures the objectivity and reliability of the indicators, the stability of the values of the attribute, the typicality of the average, etc. However, to characterize the data as a signal, the signal-to-noise ratio is more commonly used the inverse of the coefficient of variation – $SNR$, which characterizes the ratio of the average value of the signal amplitude to the noise amplitude. The average value is a measure of the signal, and the standard deviation is a measure of the noise.

$$SNR = \frac{m}{\sigma}$$  \hspace{1cm} (7)

$SNR$ is the most important parameter that characterizes the level of interference in the system. It is possible to distinguish a signal among noisy data if the signal parameters exceed the noise parameters, i.e. if $SNR > 1$. The same ratio should be considered for the characteristics of any measured data, including data recorded by state statistics bodies.

In order not to get zero characteristics of the mathematical expectation and then make it possible to logarithm, the analyzed data should be considered in a special form in which the range of variables is shifted to the right relative to zero and the possible logarithm is correct. In this case, it is more convenient to take not the range $[0, 1]$ (or $[0, 100]$), but an interval that does not contain zero – $[1, 100]$. This form of data representation will be called the signal form. In this case, the minimum value of the variable will be one, and the maximum value will be 100. The corresponding transformation of the variable of the original variable $x_{ij}$ to the interval $[1, N]$ has the following form.

If the initial indicator of the $j$-th indicator of the $i$-th $x_{ij}$ object is associated with the quality property of the system by a monotonous increasing dependence, then the initial variables $x_{ij}$ are transformed according to the rule:
$$a_{ij} = 1 + \frac{N-1}{M_j - m_j} \left( x_{ij} - m_j \right),$$  \hspace{1cm} (8)$$

$m_j$ is the smallest value of the $j$-th indicator, $M_j$ is the largest value of the $j$-th indicator.

If the initial indicator of the $j$-th indicator of the $i$-th object is associated with the quality property of the system by a monotonous decreasing dependence, then

$$a_{ij} = N - \frac{N-1}{M_j - m_j} \left( x_{ij} - m_j \right),$$  \hspace{1cm} (9)$$

If the initial indicator is associated with the analyzed integral quality property by a non-monotonic dependence (i.e., there is a value $x_{j opt}$ within the range of variation of the $j$-th indicator, at which the highest quality is achieved), then the value of the corresponding indicator is calculated by the formula:

$$a_{ij} = \left( N - \frac{N-1}{\max \left( \left( M_j - x_{j opt} \right), \left( x_{j opt} - m_j \right) \right)} \left| x_{ij} - x_{j opt} \right| \right).$$  \hspace{1cm} (10)$$

With possible logarithmization, such a transformation will avoid the pressure of small values of variables that, after direct logarithmization, would have large negative values.

Converting data from a statistical form to a signal form (8-10) changes the signal-to-noise ratio. For statistical data, the value of $\text{SNR} < 1$ had 7 indicators: indicators 1, 8, 9, 12, 22, 27, 30. All these indicators have a significant coefficient of asymmetry (highlighted in Table 1). Tables 2.3 demonstrate how the SNR value changes with different data transformations.

**Table 2.** Changing the signal-to-noise ratio during the variables transformation of the third block.

| Type of data | The number of the variable |
|--------------|----------------------------|
|              | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
| Initial     | 0.69 | 4.60 | 2.85 | 6.24 | 3.86 | 1.51 | 6.81 | **0.82** | **0.71** |
| Unified     | 0.58 | 2.27 | 5.38 | 5.86 | 2.98 | 5.55 | 3.40 | 5.56 | 0.71 |
| Signal form | **0.66** | 2.33 | 1.79 | 2.46 | 3.04 | 1.50 | 3.46 | **0.88** | **0.78** |

For the statistical data of the first block, the value of $\text{SNR} < 1$ had 3 indicators: 1, 8, and 9 (highlighted in Table 2). When converting to the segment $[0, 1]$ (unifying), the variable 8 changed the $\text{SNR}$ value from 0.82 to 5.56, but the conversion to the signal form again gives the value of this indicator less than one. In the third block, changing the data form changed the properties of variable 27 (Table 3). The value of $\text{SNR}=0.09$ for the initial data indicates that the noise level in this data exceeds the signal level and does not allow making correct conclusions from this data. When converting to the signal form, the $\text{SNR}$ value increased to 3.64.

**Table 3.** Changing the signal-to-noise ratio when converting variables of the third block.

| Type of data | The number of the variable |
|--------------|----------------------------|
|              | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| Initial     | 1.61 | 2.37 | **0.09** | 1.64 | 1.33 | **1.00** | 2.31 | 1.83 | 1.94 |
| Unified     | 6.40 | 3.48 | 8.56 | 6.97 | 8.52 | 8.22 | 3.37 | 4.00 | 4.86 |
| Signal form | 1.47 | 2.30 | **3.64** | 1.48 | 1.39 | **1.10** | 2.25 | 1.60 | 1.95 |
As a result, the value $SNR < 1$ of the data in the signal form has 4 indicators: 1, 8, 9, and 22. All these indicators have a significant coefficient of asymmetry. However, the asymmetry of these indicators is not the most significant in the initial sample (Table 4). The variable 30 with the second largest coefficient of asymmetry does not require additional transformations. The selected indicators have an abnormal distribution of variable values, which can distort the calculation of composite indices. To eliminate this particular abnormality of variables, and not to eliminate the asymmetry, a logarithmic transformation should be used.

Table 4. Comparison of variable parameters.

| The number of the variable | 1   | 8   | 9   | 12  | 22  | 27  | 30  |
|----------------------------|-----|-----|-----|-----|-----|-----|-----|
| SNR, initial data          | 0.69| 0.82| 0.71| 0.18| 0.91| 0.09| 1.00|
| SNR, signal form           | **0.66** | **0.88** | **0.78** | 2.02 | **0.85** | 3.64 | 1.10 |
| $A_s$, initial data        | 5.3 | 3.4 | 4   | 1.8 | 5.2 | -1.3 | -5  |

In [11-14], the logarithmic transformation is mentioned among a number of other nonlinear transformations often used to reduce the asymmetry of positive data. However, this transformation was not applied in any of the above-mentioned works when calculating specific composite indexes, although all the variables used have an obvious asymmetry.

5. Results

For the dataset with the author's modification of PCA (principal component analysis) [22] construction of the composite index of quality change system on a number of observations, taking into account the noise of the original data, were calculated composite indicators of quality of life of the population of Russia in 2010 and 2017 in two ways: no logarithmic correction and with the use of a logarithmic correction to the identified anomalous variables:

– 1 “GRP per capita”;
– 8 “decrepit housing”;
– 9 “the density of public roads”;
– 22 “The ratio of GRP to the number of people employed in the economy”.

Table 5. Weight coefficients of the variables of the first block before and after logarithmic correction.

| №  | Indicator                                                  | Weight of indicator without correction | Weight of indicator after correction |
|----|------------------------------------------------------------|----------------------------------------|-------------------------------------|
| 1  | Pre capita GDP–living wage ratio, units                     | -0.411                                 | 0.732                               |
| 2  | Per capita income purchasing power relative to living wage, %| 0.893                                  | 0.529                               |
| 3  | Share of people with incomes below living wage, %           | 0.858                                  | 0.901                               |
| 4  | The ratio of average income of the richest 20% to the poorest 20% (R/P 20) | **-0.321**                              | **0.188**                           |
| 5  | Number of cars per 1 000 people                             | 0.606                                  | **1.588**                           |
| 6  | Share of families on waiting lists for housing, %           | 0.736                                  | 1.506                               |
| 7  | Total area of housing resources per resident (m²/10 people) | 1.104                                  | 0.791                               |
| 8  | Share of dilapidated housing, %                             | 1.092                                  | 0.883                               |
| 9  | Public road density (km/10,000 km²)                        | **1.401**                              | 0.432                               |

The conversion of indicators to the interval [1, 100] took into account the nature of their relationship with the system quality indicator. Negative signs of weighting coefficients without logarithmic
correction have three variables: “The ratio of GRP per capita to the subsistence minimum”, “The ratio of income of the richest 20% and the poorest 20%”, “The number of disabled people per 1 thousand people”. The negative weighting coefficient of the last variable, which is determined by the data without logarithmic correction, corresponds to an increase in the quality of the system with an increase in the number of disabled people. This contradicts the intuitive idea of the quality of life of the population. The application of logarithmic correction for anomalous variables solves the problem with the negativity of the weighting coefficients. All the weights of the variables of the first block after the logarithmic correction of the anomalous variables are non-negative (Table 5).

6. Conclusion

Data transformation is one of the most important processes when calculating the quality indicator of systems. The features of the data distribution can significantly affect the composite characteristics of objects, so obtaining the characteristics of objects must necessarily take into account the features of the data.

Some types of data are characterized by distributions with a significant anomaly, when the vast majority of observations are concentrated near the boundary values. This type of data can not always be characterized by an asymmetry coefficient. In addition, if the values of a variable are approximately symmetric with respect to zero or are concentrated near zero, the sample cannot also be characterized by the coefficient of variation. The signal-to-noise ratio, which is the inverse of the coefficient of variation, is the most important parameter that characterizes the level of interference in the system.

It is possible to distinguish a signal among noisy data if the signal parameters exceed the noise parameters, i.e. if \( SNR > 1 \). The same ratio should be considered for the characteristics of any measured data, including statistical data. The paper proposes a transformation that allows us to identify the anomalous nature of variables using the signal - to-noise ratio. Variables are evaluated in the standard range, which is shifted to the right relative to zero. If it is necessary to logarithm, such a transformation will avoid the pressure of small values of variables that, after direct logarithm, would have large negative values. The use of logarithmic correction for the detected anomalous variables redistributes the values of the obtained weighting coefficients in the direction of a more correct interpretation and, in particular, can solve the problem with the negativity of the weighting coefficients.

References

[1] Bandura R 2011 Composite Indicators and Rankings: Inventory (Working Paper) p 257
[2] Saltelli A, Munda G and Nardo M 2006 From complexity to multidimensionality: the role of composite indicators for advocacy of EU reform Rev. of Business and Economic Literature LI (3) 221–35
[3] Foa R and Tanner J C 2012 Methodology of the indices of social development ISD Working Paper Series from International Institute of Social Studies of Erasmus University Rotterdam (ISS), The Hague. 4 66 Available at: http:// repub.eur.nl/pub/50510/ISD-WP-2012-4.pdf
[4] Saltelli A, Munda G and Nardo M 2006 From complexity to multidimensionality: the role of composite indicators for advocacy of EU reform Tijdschrift voor Economie en Management LI 3 Available at: http://feb.kuleuven.be/rebel/jaargangen/2001-2010/2006/TEM%202006-3/TEM_2006-3_03_Saltelli.pdf
[5] Sharpe A 2004 Literature Review of Frameworks for Macro-indicators (Ottawa, Canada: Centre for the Study of Living Standards) URL: https://ideas.repec.org/p/sls/resrep/0403.html
[6] Nardo M, Saisana M, Saltelli A and Tarantola S 2005 Tools for composite indicators building (Joint Research Centre, Ispra, Italy) pp 1–134
[7] Verma A, Angelini O and Di Matteo T 2020 A new set of cluster driven composite development indicators EPJ Data Sci. 9 8 DOI: https://10.1140/epjds/s13688-020-00225-y
[8] Mazziotta M and Pareto A 2016 On the construction of composite indices by principal components analysis *Rivista Italiana di Economia Demografia e Statistica, Gennaio-Marzo* LXX (1) 103–09

[9] Becker W, Saisana M, Paruolo P and Vandecasteele I 2017 Weights and importance in composite indicators: Closing the gap. *Ecological Indicators* 80 12–22 DOI: 10.1016/J.ECOLIND.2017.03.056

[10] Talukder B W, Hipel K W and vanLoon G 2017 Developing composite indicators for agricultural sustainability assessment: effect of normalization and aggregation techniques. *Resources* 6 (4) 66 DOI: 10.3390/resources6040066

[11] OECD, European Commission, Joint Research Centre (2008) *Handbook on constructing composite indicators. Methodology and user guide*. (Paris: OECD Publisher) Available at: http://composite-indicators.jrc.ec.europa.eu

[12] Jacobs R, Smith P and Goddard M 2004 *Measuring performance: An examination of composite performance indicators*. (York, UK: Centre for Health Economics, University of York) p 124

[13] Freudenberg M 2003 Composite indicators of country performance: a critical assessment *OECD science, technology and industry working papers* 2003/16

[14] Aiello F and Attanasio M 2006 Some issues in constructing composite indicators retrieved from: Available at: http://www3.unisi.it/events/dmq2006/paper/Aiello_Attanasio.pdf

[15] Diewert W E 1976 Exact and superlative index numbers *Journal of Econometrics* 4 (2) 115–45

[16] Munda G 2012 Choosing aggregation rules for composite indicators *Social Indicators Research* 109 (3) 337–54 DOI: 10.1007/s11205-011-9911-9

[17] Special issue on index number aggregation 2002 *Journal of Economic and Social Measurement* 28 (1–2)

[18] Freudenberg M 2003 *OECD Composite indicators of country performance: a critical assessment*. DST/IND(2003)5 (Paris)

[19] Ajvazjan S A, Stepanov V S and Kozlova M I 2009 Applied Econometrics Measurement of synthetic categories of quality of life of the population of the region and identify key areas of improvement in socio-economic policy (in the Samara region and its municipalities an example). *Applied econometrics* 3 (19) 18–84

[20] Zhgun T V 2017 Building an integral measure of the quality of life of constituent entities of the Russian Federation using the principal component analysis *Economic and Social Changes: Facts, Trends, Forecast* 10 (2) 214–35 DOI: 10.15838/esc/2017.2.50.12

[21] Jolliffe I T 2010 *Principal Component Analysis* Second Edition (NY: Springer)

[22] Zhgun T V 2019 Complex index of a system’s quality for a set of observations MMPAM’2019 *IOP Conf. Series: Journal of Physics: Conf. Series* 1352 012064 DOI:10.1088/1742-6596/1352/1/012064