Investigation of data quality in the problem of calculating the composite index of a system from a series of observations

T V Zhgun
Yaroslav-the-Wise Novgorod State University, 41, ul. B.St. Petersburgskaya, Veliky Novgorod, Russian Federation

E-mail: Tatyana.Zhgun@novsu.ru

Abstract. The article discusses the use of the finite difference method for assessing the quality of data in the problem of calculating a composite indicator of system quality based on a number of observations. For this technique to be applicable, the data must be approximated with polynomials of lower degrees than the number of observations minus one. The assumption is tested empirically on a specific data set. 37 variables characterizing the quality of life of the population of Russia for 2010-2017 are considered. The dependences of the quality of data approximation on the degree of polynomial regression are analyzed. The results of the numerical experiment make it possible to draw a conclusion about the legitimacy of evaluating data errors using the finite difference method. The use of the finite difference apparatus for analyzing of the fetch shows the presence of fatal errors from 0.59% to 28.92%. Therefore, obtaining the composite characteristics of objects on the basis of such data must necessarily take into account the presence of a fatal error. In particular, the number of parameters characterizing the system should be large enough to compensate for random errors with averaging.

1. Introduction
The problem of determining a complex indicator of complex systems arises both when studying physical phenomena, technical systems, and when solving problems of managing socio-economic systems. When describing stochastic dynamical systems in problems of hydrodynamics, magneto hydrodynamics, astrophysics, plasma physics, radio physics, integral quantities characterizing such systems are their main characteristics. For example, all conservation laws of continuum in mechanics and electrodynamics are written for integral quantities. Integral characteristics describe the dynamics in general, allowing to abstract from the side effects associated with the randomness of indicators, distorted with noise and are the key to understanding structure formation in stochastic dynamic systems [1].

The construction of an integral indicator introduces order relations on a multidimensional set of objects and allows to compare the quality of objects. The common goal of most composite indicators is the ranking of objects (countries) and their comparative analysis [2–6]. A huge number of methods used to assess the quality of poorly structured systems [2] indicates dissatisfaction with the results and the need for further research in this area [4, 7].

The construction of composite indices without the use of a priori information to obtain the weight coefficients makes it possible to obtain objective indicators of the system under study. However, the use of formal methods for determining the weight coefficients in the event of errors (distortions) will lead to a change in the calculated coefficients. Therefore, one of the reasons for the insufficient quality of composite indices may be the unsatisfactory quality of the data used.
The fundamental difference in the calculation of composite indices for weakly structured systems is the uncertainty of the quality of the data used, in contrast to the calculation of characteristics, for example, technical systems for which the measurement error is known in advance. The authors note the presence of a large number of errors in statistical data when calculating composite indices [7–9]. Nevertheless, it is the statistical data containing fatal errors that currently represent the best estimates of the available real values in social systems [7].

The article discusses the use of the finite difference method for assessing the quality of data in the problem of calculating a composite indicator of system quality based on a number of observations. This technique requires the data to be approximated with polynomials of lower degrees than the number of observations minus one. The assumption is tested empirically on a specific data set. The dependences of the quality of data approximation on the degree of polynomial regression are analyzed. The results of the numerical experiment make it possible to conclude that it is legitimate to estimate the data errors using the finite difference method.

2. Statement of the problem

Let’s consider the construction of an integral estimate of a system of \( m \) objects, for which tables of object descriptions for a number of observations are known - matrices of dimension \( m \times n \)

\[
A^T = \{a_{ij}^T\}_{i,j=1}^{n,m}, \quad t = 1, \ldots, T.
\]

The matrix element \( a_{ij}^T \) is the value of the \( j \)-th indicator of the \( i \)-th object, the vector \( \mathbf{a}_i^T = (a_{i1}^T, \ldots, a_{in}^T) \) is the description of the \( i \)-th object at the moment \( t \). For each moment \( t \), the vector of integral indicators has the form

\[
\mathbf{q}_t = A^T \cdot \mathbf{w}_t,
\]

or, for the \( i \)-th object at the moment \( t \)

\[
\mathbf{q}_t^i = \sum_{j=1}^n w_{ij} \cdot \mathbf{a}_i^j
\]

where \( \mathbf{q}_t = (q_{1t}, q_{2t}, \ldots, q_{mt})^T \) is the vector of integral indicators at the moment \( t \), \( \mathbf{w}_t = (w_{11}, w_{12}, \ldots, w_{m1})^T \) is the vector of weights of indicators for the moment \( t \), \( A^T \) is the matrix of preprocessed data for the moment \( t \). The numerical characteristics of the system were preliminarily subjected to unification - bringing the values of variables to the segment \([0, 1]\) according to the principle “the more, the better”.

To construct an integral indicator of the quality of the system, it is required to find the weights of the indicators \( \mathbf{w}_t \) for each time point that adequately reflect the properties of the system under consideration. That is the determinable weighting factors should reflect the structure of the system being assessed. This interpretation of weighting factors eliminates one of the main uncertainties in the design of an integral indicator. If the composite system quality indices are determined for a series of consecutive observations, then we are dealing with a change in data. This change in data over time is caused with both a change in the situation and random errors in data logging [10].

One of the simplest methods for analyzing the structure of the system under study is the principal component method (PCA). Principal component space is optimal for modeling internal data structures. The multivariate analysis technique, which works well for evaluating technical systems, often gives an unreliable result when constructing composite indices of semi-structured systems. In particular, computed composite indices are extremely unstable [11–14]. This may be due to data errors. The presence of a number of recorded measurements \( A^T = \{a_{ij}^T\}, t = 1, \ldots, T \) allows us to estimate this error.

3. Estimating fatal data errors

Data quality is a generalized concept that reflects the degree of their suitability for solving a specific problem [15, 16]. In accordance with the ISO 9000: 2015 standard, the main quality criteria are
completeness, reliability, accuracy, consistency, availability and timeliness [17]. Abnormal values and noises are cited as the main problems causing data quality degradation. These shortcomings do not disrupt the work of data processing algorithms, but generate incorrect analysis results.

In world statistical practice, there is no generally accepted definition of data quality as a result of statistical activity. However, the generally accepted components of the modern concept of statistical quality are the concepts of data accuracy and reliability. In practical terms, there is no single and complete measure of the reliability and accuracy of the results of statistical observations, therefore, several forms of its expression are used. Based on practical needs, the degree of accuracy of a value is usually characterized with its variance, standard error, and coefficient of variation. But these measures of accuracy poorly characterize the reliability and the presence of possible registration errors. Such errors can be estimated using the finite difference apparatus.

Let \( y_i \) is the exact (unknown) value of the measured quantity determined for a number of observations \( i = 1, \ldots, k \); \( y_i^* \) – measured value containing an error; \( e_i = y_i^* - y_i \) – measurement error. The error in statistical data is random. The value \( e_i \) is unknown and cannot be calculated from the recorded observations. However, the maximum of the errors can be estimated.

Let’s denote the maximum error for all observations \( \varepsilon = \max_i |e_i| \). Then the measured value \( y_i^* \) lies in the range \( y_i - \varepsilon \leq y_i^* \leq y_i + \varepsilon \). Let’s consider the first finite differences of the approximate quantities

\[
\Delta y_i = \Delta \left( y_i^* \right) = y_i^* - y_i
\]

\[
\Delta y_i = (y_{i+1} + e_{i+1}) - (y_i + e_i) = (y_{i+1} - y_i) - (e_{i+1} - e_i)
\]

Considering that \( |e_{i+1} - e_i| \leq |e_{i+1}| + |e_i| \leq 2 \cdot \varepsilon \) is the modulus of the approximate finite difference, where \( \Delta_i = y_{i+1} - y_i \) are the first finite differences of unknown exact quantities. Further, the second finite difference of approximate values \( \Delta y_i^2 = \Delta y_i - \Delta y_i \) is estimated similarly \( |\Delta y_i^2| \leq |\Delta y_i| + 4 \cdot \varepsilon \). The last calculated \( k \)-th approximate finite difference satisfies the estimate

\[
|\Delta y_i^k| \leq |\Delta y_i^1| + 2^k \cdot \varepsilon
\]  

It is known that for smooth functions the value of the finite difference tends to zero as the order of the difference increases. In particular, for a polynomial of degree \( k \) \( P_k(x) = a_k \cdot x^k + a_{k-1} \cdot x^{k-1} + \ldots + a_0 \) the following relations are held:

\[
\Delta^{k+1}(P_k(x)) = 0 \text{ and } \Delta^k(P_k(x)) \approx a_k \cdot h^k,
\]

where \( h \) is the table step. If the values of the measured function from measurement to measurement do not change too quickly (the function is continuous and the derivatives of higher orders are limited), the function can be approximated with a polynomial of a low degree and the values of the exact finite differences \( \Delta^k \) tend to zero with increasing order. This means that the calculated values of the approximate finite differences provide an estimate of the initial error:

\[
|\Delta y_i^k| \leq 2^k \cdot \varepsilon
\]  

We denote

\[
\varepsilon^* = |\Delta y_i^k| / 2^k
\]

Taking into account that the value \( \varepsilon \) estimated the errors of the values of the function from above, \( |e_i| \leq \varepsilon \) and according to inequality (4) \( \varepsilon \geq |\Delta y_i^k| / 2^k \), then there are two options for estimating \( \varepsilon \):
The real relationship between the values $|e_i| \leq e \leq e^*$ can be obtained from a numerical experiment. To do this, let's consider a model example.

Table 1 shows the result of calculating high-order finite differences for the exact specification of the function $f(x) = \ln(x + 4)$ and in the case of introducing errors in its values. In the absence of errors, already the third finite differences of a precisely given function turn to zero (highlighted in Table 1). If an error of the order of $\varepsilon = 0.01$ is introduced into each value of the function, then the final difference of the seventh order for the approximate values of the function is $\Delta^7 = 0.25$ (highlighted in the table) and the error estimate is $\varepsilon^* = 0.002$ with a real error $\varepsilon = 0.1$. So, the calculated estimate $\varepsilon^*$ turned out to be less than the real error $\varepsilon$, i.e. the following inequality is held to estimate the error

$$\varepsilon^* = \frac{\left| \Delta^k \right|}{2^k} \leq \varepsilon \quad \text{(6)}$$

If the error $\varepsilon = 1$ is introduced into one value of the function and the exact value of the function $f(0.3) = 1.459$ is replaced with the approximate value $f^*(0, 3) = 2.459$ (highlighted in Table 1), then the seventh finite difference for the approximate values of the function, in this case, it will be $\Delta^7 = 35$, and the error estimate by formula (3) will be $\varepsilon^* = 35/2^7 = 0.276$. The computed estimate $\varepsilon^*$ turned out to be less than the real error $\varepsilon$.

**Table 1.** Examples of calculating high-order finite differences.

| $x$  | $f(x)$ | $\Delta^1$ | $\Delta^2$ | $\Delta^3$ | $\Delta^4$ | $\Delta^5$ | $\Delta^6$ | $\Delta^7$ |
|------|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 0    | 1.386  | 0.025       | -0.001      | 0           | 0           | 0           | 0           | **0**       |
| 0.1  | 1.411  | 0.024       | -0.001      | 0           | 0           | 0           | 0           |
| 0.2  | 1.435  | 0.024       | -0.001      | 0           | 0           | 0           |
| 0.3  | 1.459  | 0.023       | -0.001      | 0           | 0           |
| 0.4  | 1.482  | 0.022       | 0           | 0           |
| 0.5  | 1.504  | 0.022       | 0           |
| 0.6  | 1.526  | 0.022       |
| 0.7  | 1.548  |             |             |             |             |

Values of the function are rounded to the nearest hundredth

| $x$  | $f^*(x)$ | $\Delta^1$ | $\Delta^2$ | $\Delta^3$ | $\Delta^4$ | $\Delta^5$ | $\Delta^6$ | $\Delta^7$ |
|------|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 0    | 1.380    | 0.03        | -0.01       | 0.01        | 0           | -0.03       | 0.1         | **-0.25**   |
| 0.1  | 1.410    | 0.02        | 0           | 0.01        | -0.03       | 0.07        | -0.15       |
| 0.2  | 1.430    | 0.02        | 0.01        | -0.02       | 0.04        | -0.08       |
| 0.3  | 1.450    | 0.03        | -0.01       | 0.02        | -0.04       |
| 0.4  | 1.480    | 0.02        | 0.01        | -0.02       |
| 0.5  | 1.500    | 0.03        | -0.01       |
| 0.6  | 1.530    | 0.02        |
| 0.7  | 1.550    |             |             |             |             |             |

**Single emission**

| $x$  | $f^*(x)$ | $\Delta^1$ | $\Delta^2$ | $\Delta^3$ | $\Delta^4$ | $\Delta^5$ | $\Delta^6$ | $\Delta^7$ |
|------|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 0    | 1.386    | 0.025       | -0.001      | 1           | -4          | 10          | -20         | **35**      |
| 0.1  | 1.411    | 0.024       | 0.999       | -3          | 6           | -10         | 15          |
| 0.2  | 1.435    | 1.024       | -2.001      | 3           | -4          | 5           |
| 0.3  | **2.459**| -0.977      | 0.999       | -1          | 1           |
| 0.4  | 1.482    | 0.022       | 0           |
| 0.5  | 1.504    | 0.022       |
| 0.6  | 1.526    | 0.022       |
| 0.7  | 1.548    |             |             |             |             |             |             |
This means that the calculated value of $\varepsilon^*$ is a lower estimate of the possible error and may be a characteristic of the sample under study. We can say that this value is, in a sense, a measure of the randomness of the data in the sample. If the values of the measured quantities are preliminarily given for the interval $[0, 100]$, then the value $\varepsilon^*$ calculated by formula (3) will characterize the relative sampling error. The actual error may exceed this value.

**4. Influence of the smoothness of the function on the error estimate**

Let the variable $x_{ij}$ be represented with observations $x_{ij}(1), x_{ij}(2), \ldots, x_{ij}(T)$ that implement the unknown dependence of the functioning of the system under consideration with some errors $x_{ij}(t) = x_{ij}(t) + \varepsilon_{ij}(t)$. The function $x_{ij}(t)$ on the interval $t \in [1, T]$ can be approximated with a polynomial of degree $n$ $x_{ij}(t) \approx P_n(t) = a_n \cdot t^n + a_{n-1} \cdot t^{n-1} + \ldots + a_0$. Then the measured value is represented with its approximate value, which contains an error $x_{ij}(t) \approx P_n(t) + \varepsilon_{ij}(t)$. From the values $x_{ij}(1), x_{ij}(2), \ldots, x_{ij}(T)$, approximate finite differences can be calculated up to the order of $T-1$ inclusively: $\Delta_{ij}^{-1} = \Delta_{ij}^{-1}(x_{ij}(1)) \equiv \Delta_{ij}^{-1}(P_n(1) + \varepsilon_{ij}(t))$.

If the degree of the approximating function $n$ is less than $T-1$ $n < T-1$, then the last exact finite difference becomes zero $\Delta_{ij}^{-1}(P_n(1)) = 0$. Then the calculated approximate finite difference is $(\Delta_{ij}^{-1} = \Delta_{ij}^{-1}(P_n(1)) + \varepsilon_{ij}(t))$ and the estimate is fair according to (4, 5)

$$\left| (\Delta_{ij}^{-1})^{T-1} \right| \leq 2^{T-1} \cdot \varepsilon_{ij} = 2^{T-1} \cdot \varepsilon_{ij}, \quad \varepsilon_{ij} = \max_{t} \left| \varepsilon_{ij}(t) \right|. $$

So, the calculated estimate of the data error for the parameter $x_{ij}$ over the observation interval $t \in [1, T]$ is determined with the relation

$$\varepsilon_{ij}^* = \left| (\Delta_{ij}^{*})^{T-1} \right| / 2^{T-1}, \quad \text{at that } \varepsilon_{ij}^* \leq \varepsilon_{ij} \quad \text{where } \varepsilon_{ij} = \max_{t} \left| \varepsilon_{ij}(t) \right|.$$ (7)

The calculated value $\varepsilon_{ij}^*$ is an estimate from below of the possible error in registering the $j$-th parameter for the $i$-th object and can be a characteristic of the quality of the fetch under study. If the values of the studied quantities $x_{ij}(1), x_{ij}(2), \ldots, x_{ij}(T)$ are preliminarily given for the interval $[0, 100]$, then $\varepsilon_{ij}^*$ it will characterize the relative fetching error.

So, if the function allows approximation with polynomials of lower degrees than the number of observations minus one, and the value of the last approximate finite difference is nonzero $(\Delta_{ij}^{*})^{T-1} \neq 0$, then this value of the approximate finite difference is determined with the distortions of the values of the variable introduced during registration of observations, which can be estimated by (7).

Obviously, the functional dependence of the measured data can be unambiguously restored from the available $T$ values with an interpolation polynomial of degree $T-1$: $P_{T-1}(x) = a_{T-1} \cdot x^{T-1} + a_{T-2} \cdot x^{T-2} + \ldots + a_0$. In this case, the value of the last exact finite difference is determined with the leading coefficient of the interpolation polynomial

$$\Delta_{ij}^{-1} \approx \Delta_{ij}^{-1}(P_{T-1}(x)) = a_{T-1}$$

If the function being modified allows approximation with polynomials of lower degrees, then the value of the last exact finite difference should be zero $\Delta_{ij}^{-1}(P_{T-1}(x)) = 0$. If the function allows approximation with polynomials of lower degrees, and the value of the last approximate finite difference does not vanish $(\Delta_{ij}^{*})^{T-1} \neq 0$, then this value is determined with the distortions of the values of the variable.
introduced during the registration of observations. The assumption that the data can be approximated with polynomials of lower degrees than the number of observations minus one can be tested empirically on a specific dataset.

5. Application of regression analysis to describe experimental data
The selection of mathematical models that best describe the experimental data is the task of regression analysis. In classical regression analysis [17–22], the model is represented in the following form

\[ y = G(x, \beta) + \varepsilon \]  

(8)

where \( x = (x_1, x_2, \ldots, x_m) \) is the vector of input (independent) variables; \( y \) is the random output (dependent) variable; \( \beta = (\beta_1, \beta_2, \ldots, \beta_m) \) is the unknown vector of model coefficients; \( \varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m) \) is the random variable (random disturbance, error, noise), taking into account the influence of random factors. It is assumed that random perturbations are normally distributed with parameters \( M(\varepsilon) = 0, \quad \sigma^2 = \sigma^2 = \text{const} \).

Obviously, due to the action of random factors, model (8) cannot accurately predict the value of the output variable for the given values of the input variables. Therefore, there is no reason to speak of a “true” model in the full sense of the word. Usually, the “true” value of the output variable is understood as its conditional mathematical expectation at the given values of the input variables:

\[ M[y|x] = M[(G(x, \beta) + \varepsilon)|x] = M[G(x, \beta)] + M[\varepsilon] = G(x, \beta) \]

(9)

Relation (9) is a theoretical regression model - the regression equation for \( y \) relative to \( x \). The main task of regression analysis is to identify and approximate the mathematical description of the causal relationship \( G(x, \beta) \) between the output and input variables. For this purpose, based on statistical data (sample \( \{x_i, y_i\}, \quad i = 1 \ldots T \) ), the least squares method is used to construct a statistical analogue of relationship (9) - an empirical regression model

\[ y' = G'(x, b) \]

(10)

the adequacy of which, i.e., the degree of its correspondence to statistical data, is assessed using the sample coefficient of determination

\[ R^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \]

(11)

The coefficient \( R^2 \) shows how much of the variation \( y \) is due to the regression model. It follows from relation (11) that the maximum possible value \( R^2 \) is limited with the measurement error, the presence of unaccounted factors and does not depend on the composition of the input variables and the structure of the model. For example, if 10% of the variation in the output variable is due to the measurement error of the variables, then the maximum possible value \( R^2 \) cannot exceed 0.9 for any model structure. In the presence of statistical noise, a large value \( R^2 \) is provided due to the excessive complication of the model, which, instead of the existing regularity, describes random errors.

6. Using of polynomial regression to describe experimental data
To make sure that the estimates of the distortions of the values of the variable introduced during the registration of observations are correct, you need to make sure that the recorded values can be approximated with polynomials of lower degrees than the number of observations minus one. Then the calculated value of the last approximate finite difference \( (\Delta_{ij}^*)^{T-1} \neq 0 \), will determine the estimate of the distortions of the variables introduced during the registration of observations according to formula (7).

Let us consider in table 2 a set of variables characterizing the quality of life of the population [18] for 2010-2017. These variables were used to assess the quality of life of the Russian population in many studies, for example, in [20]. For correct comparison, the values of the variables are given on the segment
[0, 100]. The resulting estimate in this case is the relative error value in percent. We consider eight observations in total. It is necessary to check whether on this interval the observational data are approximated with a polynomial of degree at most sixth.

It is necessary to calculate for each indicator the values of the approximate seventh finite differences for all objects according to (5) \( \varepsilon^*_j = \left| \frac{1}{A_{jj}} \right|^{2/7} \). Further, for each indicator, we define the maximum observed value of the distortion of the variable as the maximum value of the obtained values for all subjects: \( \varepsilon^*_j = \max_i \left| \varepsilon^*_{ij} \right| \). The calculated value is an estimate of the observed distortions of the variable values. The obtained values together with the list of variables are shown in Table 2. The minimum observed error of the observed variables is 0.59%, and the maximum is 28.92%.

**Table 2.** Variables for calculating the integral index of the quality of life and the observed estimate of the error in the data for the 2010-2017 sample.

| Variable number | Variable Name | Eps |
|-----------------|---------------|-----|
| **Block 1: Population’s welfare** | | |
| 1 | Pre capita GDP–living wage ratio, units | 1.76 |
| 2 | Per capita income purchasing power relative to living wage, % | 3.70 |
| 3 | Share of people with incomes below living wage, % | 2.76 |
| 4 | The ratio of average income of the richest 20% to the poorest 20% (R/P 20) | 2.45 |
| 5 | Number of cars per 1,000 people | 2.50 |
| 6 | Share of families on waiting lists for housing, % | 14.84 |
| 7 | Total area of housing resources per resident (m2/10 people) | 3.71 |
| 8 | Share of dilapidated housing, % | 8.04 |
| 9 | Public road density (km/10,000 km2) | 1.40 |
| **Block 2: Population quality** | | |
| 10 | Life expectancy at birth, years | 4.43 |
| 11 | Mortality rate, infant (per 1,000 live births) | 5.78 |
| 12 | Population growth rate, per 1,000 people | 1.84 |
| 13 | Deaths caused by communicable, parasitic diseases and TB per 100,000 people | 0.78 |
| 14 | Deaths caused by neoplasms per 100,000 people | 0.59 |
| 15 | Deaths caused by cardiovascular diseases per 100,000 people | 0.68 |
| 16 | Deaths caused by respiratory diseases per 100,000 people | 1.72 |
| 17 | Deaths caused by digestive system diseases per 100,000 people | 1.33 |
| 18 | Incidence of injuries, intoxication and other external causes per 100,000 people | 6.38 |
| 19 | Number of disabled people per 1,000 people | 0.93 |
| 20 | Incidence of congenital anomalies per 1,000 people | 8.50 |
| 21 | Specialists with higher education employed in economy, % | 2.40 |
| 22 | Labor force productivity (GRP per average annual number of employed in economy, thousand rubles/person) | 1.69 |
| 23 | Graduates from higher and vocational educational institutions per 1,000 people | 1.67 |
| **Block 3: Social quality** | | |
| 24 | Unemployment, % | 3.64 |
| 25 | Employers engaged in harmful and hazardous working conditions in the average annual number of employed in economy, % | 21.86 |
| 26 | Number of employees injured at work resulting in death or loss of earning capacity for 1 or more days per 1,000 employees | 6.18 |
| 27 | Net migration per 10,000 people | 11.13 |
| 28 | Intentional homicides per 100,000 people | 7.75 |
Incidence of intentional infliction of grievous bodily harm per 100,000 people 4.45
Incidence of rape per 100,000 people 28.92
Incidence of robbery and theft per 100,000 people 3.55
Incidence of larceny or embezzlement per 100,000 people 4.20
Number of registered with drug and substance abuse per 100,000 people 4.14
Number of registered with alcohol abuse per 100,000 people 1.93
Number of infected with TB per 100,000 people 1.79
Mortality from external causes per 100,000 people 10.93
Number of people with mental disorders per 100,000 people 1.63

The largest values of the observed errors are among the most socially significant indicators. The expected is a low quality of data characterizing migration growth. But the statistics of registration of families to be on filed with housing contain even more errors. There are many mistakes in registering old and dilapidated housing, congenital anomalies, the number of people employed in hazardous working conditions, the number of rapes and deaths from external causes.

To make sure that the assumption about the possibility of approximating the considered data with a polynomial of degree no more than sixth is valid, it is sufficient to numerically evaluate the quality of the polynomial regression for subjects providing the maximum value of the approximate seventh finite difference for each variable. Let us demonstrate this for variable 6. For this variable, check whether the observational data for it can be approximated with a polynomial of degree at most six. The quality of the approximation increases with the growth of the degree of polynomial regression (table 3). With an increase in the degree of the approximating polynomial, the number of objects with a relatively low coefficient of determination decreases. For polynomial regression of the sixth degree for 83 objects out of 85, the quality of the approximation can be considered good, for the remaining two - satisfactory. The analyzed dependences for the remaining variables confirm the conclusion about the validity of the observed estimates of the data errors presented in table 2.

Table 3. Dependence of the quality of approximation on the degree of polynomial regression for variable 6.

| Number of subjects | Polynomial Regression Degree |
|--------------------|----------------------------|
| R² >= 0.7          | 1  | 2  | 3  | 4  | 5  | 6  |
| 44                 | 61 | 72 | 79 | 82 | 83 |

7. Conclusion
Solving the problem of constructing the integral characteristic of the system requires a detailed understanding of the influence of the errors of the data used on the calculated characteristics. Even a small perturbation of the original data can cause a significant change in the weighting factors when using multivariate analysis methods. The reason for this may be the presence of fatal errors in the data used.

The finite difference method allows to estimate the presence of errors for a number of observations. For this technique to be valid, the data must be approximated with polynomials of lower degrees than the number of observations minus one. The validity of the assumption was tested empirically on a specific dataset of 37 variables that characterize the quality of life of the population of Russia for 2010-2017. Analysis of the sample shows the presence of observed fatal errors from 0.59% to 28.92%. Consequently, obtaining various composite characteristics of objects based on such data must necessarily take into account the presence of a fatal data error. In particular, the number of parameters that characterize the system should be large enough to compensate for random errors by averaging. The 37 parameters proposed to characterize the quality of life are quite sufficient to compensate for measurement errors. The characteristic of individual blocks, in which there are 9, 14 and 14 variables,
will be sensitive to fatal errors. The used method for calculating the composite indicator should also take into account the presence of fatal data errors.

References

[1] Klyatskin V I 2011 Integral characteristics: a key to understanding structure formation in stochastic dynamic systems Phys. Usp. 54 441–464 DOI: 10.3367/UFN.e.0181.201105a.0457a [In Russian]
[2] Bandura R Composite Indicators and Rankings: Inventory 2011 (Working Paper) p 257
[3] Saltelli A, Munda G and Nardo M 2006 From Complexity to Multidimensionality: the Role of Composite Indicators for Advocacy of EU Reform L1 Rev. of Business and Economic Literature 0(3) 221–235
[4] 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 URL: http://repub.eur.nl/pub/50510/ISD-WP-2012-4.pdf
[5] 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 URL: http://feb.kuleuven.be/rebel/jaargangen/2001-2010/2006/TEM%202006-3/TEM_2006-3_03_Saltelli.pdf
[6] 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
[7] Nardo M, Saisana M, Saltelli A and Tarantola S 2005 Tools for composite indicators building (Joint Research Centre, Ispra, Italy) 1–134
[8] Krishnan V 2015 Development of a Multidimensional Living Conditions Index (LCI) Social Indicators Research 120 2 455–481
[9] 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
[10] Zhgun T V 2019 Complex index of a system's quality for a set of observations Journal of Physics: Conference Series 1352 (1), article id. 012064 Pub Date: October 2019 DOI: 10.1088/1742-6596/1352/1/012064
[11] Becker W, Saisana M, Paruolo P and Vandecasteele I 2017 Weights and importance in composite indicators: Closing the gap. Ecological Indicators, 80 12–22. URL: https://doi.org/10.1016/J.ECOLIND.2017.03.056
[12] Paruolo P, Saisana M and Saltelli A 2013 Ratings and rankings: voodoo or science? Journal of the Royal Statistical Society: Series A (Statistics in Society) 176 (3) 609–634 https://doi.org/10.1111/j.1467-985X.2012.01059.x
[13] 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) 103–109
[14] Zhgun T V 2019 Method for evaluating the robustness of rankings generated by composite indicators Journal of Physics: Conference Series 1352 (1) article id. 012064 Pub Date: October 2019 DOI: 10.1088/1742-6596/1352/1/012065
[15] Batini C and Scannapieca M 2006 Data quality (Berlin: Springer-Verlag, Germany) 19–31
[16] Herzog T N, Scheuren F J and Winkler W E 2007 What is Data Quality and Why Should We Care? Data Quality and Record Linkage Techniques (Springer New York) 7–15 URL: https://doi.org/10.1007/0-387-69505-2_2
[17] Wang R Y, Kon H B and Madnick S E 1993 Data quality requirements analysis and modeling Data Engineering, 1993. Proceedings. Ninth International Conference on IEEE 670–677
[18] Weisberg S 2014 Applied Linear Regression 4th ed Hoboken (NJ: Wiley) p 368
[19] Seber G A F and Lee A J 2003 Linear Regression Analysis 2 nd ed Hoboken (NJ: Wiley) p 582
[20] Hoffmann Jh P and Shafer K 2015 Linear Regression Analysis. Assumptions and Applications (Washington: NASW Press) p 240
[21] Bingham N H and Fry J M 2010 *Regression* (London: Springer) p 284
[22] Montgomery D C, Peck E A and Vining G G 2013 *Introduction to Linear Regression Analysis.* Fifth ed (New York: Wiley) p 836
[23] Isakin M A 2006 Modification of the K-means method with an unknown number of classes *Applied econometrics* 4 62–70 [In Russian]
[24] 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 [In Russian]