Methodological Approaches to Data Pre-Processing
Formalization for Statistical Analysis

D A Pisareva¹, E A Pakhomova² and O V Rozhkova³

¹PhD student, Dubna State University, Russia
²Doctor of Economic Sciences, Candidate of Technical Sciences, Professor, Dubna State University, Russia
³Director of the Centre For Educational Services Support, Candidate of Economic Sciences Degree Applicant, Dubna State University, Russia

E-mail: uni-dubna@mail.ru

Abstract. The topic of data processing is not accidental for us. Initially, we were engaged in the development of mathematical tools for assessing the impact of the University of Science city on the effectiveness of Regional development (USR). The theme turned out to be similar to the Triple Helix (TH) model of H. Etzkowitz. The main difference of the TH from the USR: the TH gives deep processes interpretations, the USR – the economic-mathematical modeling tools. In our opinion both models supplement each other: the USR interpretation results become brighter by means of the TH formulations. Then we began to develop a toolkit for the TH model with its adaptation to the conditions of Russia. Our articles are available on the Official Website of the Triple Helix Association Russian Chapter. It was important to adapt the apparatus of continuous functions to discrete functions, since official statistics are discrete in nature. Working with statistics revealed methodological problems. The paper offers methodological approaches to formalizing data pre-processing (the method of time series linking) in order to unify them and to further process them by modeling and forecasting. It seems that this formalization eliminates the appropriate methodological flaw in the literature.

1. Introduction
The quality of economic forecasts based on statistical data depends on data pre-processing. The main task of the latter is to ensure compatibility of time series levels. It is the comparability of the series levels, that makes it possible to form a sample of the size, suitable for econometric models building.

Analyzing [1][2][3][4][5][6][7], we can present the classification of time series in the form of the flowchart ‘figure 1’.
The variety of time series forming options explains the difference in levels. The uniformity of the time series is affected by changes in [7][8][9][10][11]:

- Borders of economic entities (for example, the statistical data for the Moscow region for 2018 will be different, compared to previous years, since over the past two years a number of urban districts were removed from the Moscow region and assigned to the city of Moscow).
- Methods for calculating indicators. This problem is clearly highlighted in the work of V.A. Bessonov: "Thus, data on GDP production in the structure of the all-Union classifier of branches of the national economy (OKONH) are now available only since 1995, whereas previously there were official estimates of dynamics in relation to 1991. The loss of this data is usually attributed to the transition to a more advanced method. However, this transition was not accompanied by re-calculation of retrospective estimates, using the new methodology, nor by a concatenation of time series of previous and new estimates. The data was simply discarded. There is no doubt, that the quality of early assessments was poor. However, it is better to have bad grades than not to have any" [11].
- Measurement units (in some ways, this reason can be attributed to the previous one, but it is highlighted separately for updating attention).
- Prices for value indicators (for example, inflation).
- Completeness of covering the phenomenon under study (that is, the possible loss of process components, for example, will distort statistical metrics).
- the moment of data registration (usually manifested in overturning processes).
- The duration of time intervals (that is, the violation of the homogeneity of the process stages, which refers to analog time series, is resolved by periodization of dynamics [9]).

Of the above points, the most common problem of comparability of time series levels is the use of various methods for calculating indicators and prices, which we encountered in the study of the cyclical development of the socio-economic system [12].

2. Methodology
In order to account for these changes to ensure comparability of the series levels in [5][9][10][11], the following method of time series linking is proposed: several time series of the same indicator with disparate levels are combined into one, longer series in two ways, which are presented only as concrete examples, without mathematical formalization of the problem, which, in our opinion, is a methodological disadvantage in the field of data processing.

In this article, the noted methodological disadvantage is eliminated by formalizing the method of time series linking; the result of this operation is viewed in the example.
2.1. The absolute values series linking

Let us have two time series of absolute values of the same indicator: \( \{a_i\}, \{b_j\}, i = 1, n, j = 1, m, n, m \in \mathbb{N}_0 \) and \( a_n = b_n \), which means that the data is not comparable at level \( n \). To ensure comparability, it is necessary that \( a_n = b_n \); there may be cases:

A) \( a_n \) takes the value \( b_n \);
B) \( \text{takes the value } a_n \).

Write the series as:

\[
\{a_i\} = \{a_1, a_2, ..., a_n\};
\]

\[
\{b_j\} = \{b_n, b_{n+1}, ..., b_m\}.
\]

In order for \( a_n \) to take the value \( b_n \), it is necessary and sufficient to divide \( a_n \) by itself and to multiply by \( b_n \):

\[
\text{However, this is not sufficient for the entire row (1), since the last element of this row will not be comparable to the previous elements. To eliminate this situation, let's divide all the elements of the series (1) by } a_n \text{ and multiply by } b_n. \text{ These mathematical operations can be performed with a number of discrete data, forming an array–vector, as operations with vectors. We will receive:}
\]

\[
\frac{a_1}{a_n}, \frac{a_2}{a_n}, ..., \frac{a_{n-1}}{a_n}, b_n, \quad (3)
\]

Now we can write a linked series by adding (3) elements (2):

\[
\left\{ \frac{a_1}{a_n}*b_n, \frac{a_2}{a_n}*b_n, ..., \frac{a_{n-1}}{a_n}*b_n, b_{n+1}, ..., b_m \right\}
\]

or

\[
\left\{ \frac{a_1}{a_n}, \frac{b_n}{a_n}, \frac{a_2}{a_n}, ..., \frac{b_{n-1}}{a_n}, b_{n+1}, ..., b_m \right\}
\]

In the row (4) we denote the linking co-efficient \( \frac{b_n}{a_n} = q \) and get:

\[
\text{For case (B), the elements of series (2) must be divided by } b_n, \text{ multiplied by } a_n \text{ (here } q = \frac{a_n}{b_n} \text{); and the resulting series must be supplemented with elements of series (1):}
\]

\[
\{a_1, a_2, ..., a_n, b_{n+1}*q, ..., b_m+q\}
\]

Rows (5) and (6) are linked in absolute units. The choice of method is determined by the purpose of studying the economic process.

2.2. The relative values series linking

To avoid errors in econometric model building related to the dimension of indicators, it is desirable to express statistical data in relative units (indices). This procedure can be performed both at the first stage before the series linking and after the absolute units series linking.

Indeed, the procedure of "de – dimensionalization", or relativization, is the division of all levels of the series by a level that is taken as the basic (1 or 100%) and, in fact, it does not matter at what stage the division operation will be performed due to the combinative property of operations on vectors. The choice of the basic level may be due to the selection of some feature, for example, additional factual content, for example, the economic crisis. In this case, as a rule, the crisis or the year following the crisis are chosen for the base level, when there is a dramatic change in the indicator dynamics.

Let's show this in a formalized manner. Let the series (1), (2) be given in absolute units. Initially, we will "de-size" these rows, and then we will link them. Let's assume that the base level is \( k \in [1, n] \).

Then, the elements of series (1), (2) will take the form:

\[
\left\{ \frac{a_1}{a_k}, \frac{a_2}{a_k}, ..., \frac{a_n}{a_k} \right\}, \left\{ \frac{b_n}{a_k}, \frac{b_{n+1}}{a_k}, ..., \frac{b_m}{a_k} \right\}
\]

Linking the series in (7), for example, in the way (A), we will get a linked series in relative values:
If we apply the "de-dimensionalization" procedure to the linked series (5), we will get a time series:
\[
\left\{ \frac{a_1 b_n}{a_k}, \frac{a_2 b_n}{a_k}, ..., \frac{a_k b_n}{a_k}, \frac{a_n b_n}{a_k}, \frac{b_{n+1}}{a_k}, ..., \frac{b_m}{a_k} \right\} = \left\{ \frac{a_1 q}{a_k}, \frac{a_2 q}{a_k}, ..., \frac{a_n q}{a_k}, a_k, ..., \frac{b_{n+1}}{a_k}, ..., \frac{b_m}{a_k} \right\}.
\] (8)

This is the series (8).

For the case (B), we will get a linked series in relative units:
\[
\left\{ \frac{a_1 q}{a_k}, \frac{a_2 q}{a_k}, ..., \frac{a_n q}{a_k}, \frac{b_{n+1}}{a_k}, ..., \frac{b_m}{a_k} \right\}, \text{ где } q = \frac{b_n}{a_n}.
\] (9)

If the basic level is \( k \in [n, m] \), then the elements of the series (1), (2) must be divided by \( b_k \) and then linked using the formulas (5) or (6).

2.3. Considering the inflation impact in values

It should be noted that in order to consider price changes in time series for values, for example, inflation, in [7][8][9][10][11] it is proposed to calculate the levels of the series in conditionally constant (comparable) prices. For this purpose, the values of the original time series at step \( k \) are divided by the base index-deflator of the same step [13]. This operation assumes the presence of statistical data of deflator indices of the studied indicator. However, statistics do not contain information about deflator indices for all indicators - in this case, price index data can be used to consider inflation impact. As a rule, statistics show chain deflator indices, and in order to find the basic deflator indices, you need to use the formula [14]:
\[
GJ(k) = GJ(k-1)J(k).
\] (10)

where \( GJ(k) \) is the base index-deflator of period \( k \), \( GJ(k-1) \) – the base deflator of \((k-1)\) period, \( J(k) \) is the chain deflator of \( m \)-period.

That is, for the value time series (1), (2), the real (deflated) time series in absolute units will take the form:
\[
\left\{ \frac{a_1}{GJ(1)}, \frac{a_2}{GJ(2)}, ..., \frac{a_n}{GJ(n)} \right\} = \left\{ \frac{a_1 b_k}{b_k}, \frac{a_2 b_k}{b_k}, ..., \frac{a_n b_k}{b_k} \right\}.
\] (11)

\[
\left\{ \frac{b_{n+1}}{GJ(n+1)}, ..., \frac{b_m}{GJ(m)} \right\} = \left\{ \frac{b_{n+1}}{b_k}, ..., \frac{b_m}{b_k} \right\}.
\] (12)

Series (11), (12) can also be linked in absolute units using formulas (5) or (6), or in relative units.

As a rule, the deflation operation is rarely used, using time series in relative units (in prices of the same period) to study phenomena, the essence of which is to compare the price level with the selected base level. In this case, the constructed model will reflect the situation typical for the base period. For forecasting this means that the resulting forecast is valid for the reference period. If the inflation impact is taken into account, the forecast values are obtained as net values, and having inflation forecast, it is possible to calculate the inflated forecast values of the indicator.

2.4. Series linking and deflating on the example of statistical data of macroeconomic parameters

Let’s consider the operations of series linking and deflating on the example of statistical data of macroeconomic parameters. Let us have statistical data on the indicators "Gross Domestic Product" (GDP) and "Average Annual Number of Employed Persons" (Employment) for the period 1999-2019 (Table 1).

| Period (year) | GDP (billion rubles) | GDP (billion rubles) | Employment (thousand people) | Employment (thousand people) | Employment (thousand people) | Employment (thousand people) |
|---------------|----------------------|----------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| 1999-2011     | 2011-2016rr.         | 1999-2000            | 2000-2005                   | 2005-2015                   | 2015-2019                   |                             |

*Table 1. Statistical Data*
The analysis of table 1 shows that for econometric modeling, the time series of GDP and Employment for the period 1999-2019 in the presented form is quite difficult to use for the following reasons.

First, for GDP, the data differ from the results for previous years by the magnitude of changes related to the implementation of the international methodology for assessing housing services produced and consumed by homeowners; for assessing fixed capital consumption based on its current market value; and the alignment of data on exports and imports with the balance of payments data developed according to methodology 6 of the IMF’s balance of payments and international investment position guide (BPM6); updating data on the results of the development of the basic “Input-output” tables for 2011, including the assessment of domestic workers’ services, as well as data starting from 2011 contain changes related to the implementation of the provisions of the System of National Accounts (SNA) 2008 regarding the accounting of research and development results and weapons systems [15].

Second, Employment data are taken from various sources: Russian Statistical Yearbook (RSU) 2005 contains data for 1995-2004, where the data is generated prior to 2003 excluding the national population census (VPN) 2002; RFE 2006 contains data for 2000-2005, where the data prior to 2003 calculated with regard to the outcomes of the VPN 2002, Rosstat data on the average annual number of formed 2005-2011 given VPN 2010, while data for the years 2015-2016 calculated in accordance with the updated methodology of calculation of balance of labour resources and labour measurement. The increase in the average annual number of employees is due to changes in the estimate of the number of employees that are not reflected in the statistical reports of organizations and individual entrepreneurs [15].

So, the statistical data in Table 1 should be presented in a comparable form.

### 2.4.1. Time series forming on the example of GDP macroeconomic indicator

According to the classification, the GDP data series is an isolated complete digital time series of absolute units in the value of the aggregate indicator. In addition to the fact that the series must be linked and brought to relative units, due to its value expression, it is necessary to correct the levels of

| Year | GDP (in absolute units) | GDP (in relative units) |
|------|------------------------|------------------------|
| 1999 | 4823                   | 63963                  |
| 2000 | 7306                   | 64327                  |
| 2001 | 8944                   | 64517                  |
| 2002 | 10831                  | 64980                  |
| 2003 | 13208                  | 65574                  |
| 2004 | 17027                  | 65979                  |
| 2005 | 21610                  | 66407                  |
| 2006 | 26917                  | 66683                  |
| 2007 | 33248                  | 67047                  |
| 2008 | 41277                  | 67922                  |
| 2009 | 38807                  | 68397                  |
| 2010 | 46309                  | 67418                  |
| 2011 | 55967                  | 67493                  |
| 2012 | 60283                  | 67644                  |
| 2013 | 68164                  | 67922                  |
| 2014 | 73134                  | 68397                  |
| 2015 | 79200                  | 67901                  |
| 2016 | 83387                  | 67813                  |
| 2017 | 85918                  | 68389                  |
| 2018 | 91843                  | 72425                  |
| 2019 | 104630                 | 70949                  |

*Russian Statistical Yearbook https://rosstat.gov.ru/folder/210/document/12994
the series, removing the inflation impact. Therefore, we first deflate the series of GDP, then link (where the linking coefficient $k = 1.077$) and give the index form (Table 2).

**Table 2. Time Series of GDP in Comparable Form, Taking into Account Deflation**

| Period (year) | GDP deflator chain index | Basic GDP deflator index | Deflated data 1999-2011 | Deflated data 2011-2019 | Linked deflated time series of GDP (billion rubles) | Linked deflated time series of GDP (%) |
|---------------|--------------------------|--------------------------|--------------------------|--------------------------|---------------------------------------------------|---------------------------------------|
| 1999          | 172.5                    | 21.4                     | 22582                    | 24323                    | 54.71                                             |
| 2000          | 137.6                    | 29.4                     | 24858                    | 26775                    | 60.22                                             |
| 2001          | 116.5                    | 34.2                     | 26121                    | 28135                    | 63.28                                             |
| 2002          | 115.6                    | 39.6                     | 27364                    | 29473                    | 66.29                                             |
| 2003          | 113.8                    | 45.0                     | 29324                    | 31585                    | 71.04                                             |
| 2004          | 120.3                    | 54.2                     | 31424                    | 33847                    | 76.13                                             |
| 2005          | 119.3                    | 64.6                     | 33429                    | 36007                    | 80.99                                             |
| 2006          | 115.2                    | 74.5                     | 36145                    | 38932                    | 87.57                                             |
| 2007          | 113.8                    | 84.7                     | 39232                    | 42257                    | 95.05                                             |
| 2008          | 118.0                    | 100.0                    | 41277                    | 44459                    | 100.00                                            |
| 2009          | 102.0                    | 102.0                    | 38046                    | 40980                    | 92.17                                             |
| 2010          | 114.2                    | 116.5                    | 39755                    | 42821                    | 96.31                                             |
| 2011          | 115.9                    | 135.0                    | 41456                    | 44652                    | 100.43                                            |
| 2012          | 108.9                    | 147.0                    | 46364                    | 46364                    | 104.28                                            |
| 2013          | 105.3                    | 154.8                    | 47240                    | 47240                    | 106.25                                            |
| 2014          | 107.5                    | 166.4                    | 47589                    | 47589                    | 107.04                                            |
| 2015          | 107.2                    | 178.4                    | 46740                    | 46740                    | 105.13                                            |
| 2016          | 102.8                    | 183.4                    | 46847                    | 46847                    | 105.37                                            |
| 2017          | 105.3                    | 193.1                    | 47557                    | 47557                    | 106.97                                            |
| 2018          | 111.1                    | 214.6                    | 48765                    | 48765                    | 109.68                                            |
| 2019          | 103.8                    | 222.7                    | 49412                    | 49412                    | 111.14                                            |

Let’s plot the graphs for the source data and the GDP time series in a comparable form in order to visualize the analysis ‘figure 2’.
As an experiment, we will plot a graph not for deflated source data, but for simply linked series, presented in index form, and will compare it with the graph of the original statistical data ‘figure 3’.

Comparison of Fig. 2 with Fig. 3 suggests that the deflation procedure is not necessary, since it violates the picture of the rate of change of the indicator, and a linked series of non-deflated relative
data practically describes the actual situation. Therefore, the choice of the methodology for series linking with or without considering inflation impact depends directly on a problem formulated by a researcher. However, it is worth paying attention to the fact that to build a qualitative forecast for value economic indicators, it is necessary to consider inflation impact in the form of building an inflation forecast and its concatenation with the object under study.

2.4.2. Time series forming on the example of the macroeconomic indicator "Average Annual Number Of Employed Persons"

According to the classification (Figure 1), the "Employment" data series is an isolated complete digital time series of absolute units for a particular indicator. Since the time series "Employment" will be formed from four statistical data series (Table 1), linking will be performed three times with coefficients: \( k_1 = 1.003 \) for the period linking 1999-2005, \( k_2 = 0.998 \) for the period linking 1999-2015, \( k_3 = 1.059 \) for the final period linking 1999-2019. The result of linking and "de-dimensioning" is shown in Table 3.

| Period (year) | Linked time series Employment (thousand people) | Linked time series Employment (%) |
|---------------|-----------------------------------------------|----------------------------------|
| 1999          | 67827                                         | 93.64                            |
| 2000          | 68213                                         | 94.17                            |
| 2001          | 68703                                         | 94.85                            |
| 2002          | 69331                                         | 95.72                            |
| 2003          | 69759                                         | 96.31                            |
| 2004          | 70211                                         | 96.93                            |
| 2005          | 70618                                         | 97.49                            |
| 2006          | 71004                                         | 98.03                            |
| 2007          | 71930                                         | 99.31                            |
| 2008          | 72433                                         | 100.00                           |
| 2009          | 71397                                         | 98.57                            |
| 2010          | 71476                                         | 98.68                            |
| 2011          | 71636                                         | 98.90                            |
| 2012          | 71979                                         | 99.37                            |
| 2013          | 71908                                         | 99.27                            |
| 2014          | 71815                                         | 99.15                            |
| 2015          | 72425                                         | 99.99                            |
| 2016          | 72065                                         | 99.49                            |
| 2017          | 71843                                         | 99.18                            |
| 2018          | 71562                                         | 98.80                            |
| 2019          | 70949                                         | 97.95                            |

That the rate of data change for linked time series in relative units is greater than the rate of data change for linked time series in absolute units, which, in turn, is more than rate of date changes for the source statistics, we can see on the basis of the indicator change dynamics (figure 4). This leads to the conclusion that, using comparable series formed from more than two statistical data series for econometric modeling, it is necessary to analyze the reasons for the indicator dynamics and take it into account while forecasting. Of course, this fact must be carefully checked on the data of other parameters, not only experimentally, but also from formal mathematics point of view.
3. Results
In this work, the following methods of preliminary processing of statistical data in a formalized form were studied: linking the series in absolute units, linking the series in relative units, bringing the series to comparable prices, deflating. The study revealed that deflating slows the rate of indicator dynamics, but multiple linking, on the contrary, accelerates indicator dynamics rate.

4. Conclusion/Recommendations
The conceptual conclusion of the presented authors’ tools, which lie in the field of big data cognition, both structured and unstructured, concerns preliminary data analysis by unifying statistics for the purpose of further data processing by modeling and forecasting.

5. References
[1] Ivanov Yu N and other 2002 Economic statistics (Moscow: INFRA-M) p 480.
[2] Granger K and Hatanaka M 1972 Spectral analysis of time series in economics (Moscow: "Statistics"-M) p 312.
[3] Anderson T 1976 Statistical analysis of time series (Moscow: MIR PUBLISHING) p 755.
[4] Carried R and Enoxon L 1982 Applied Time Series Analysis. Basic methods (Moscow: PUBLISHING "MIR") p 428.
[5] Yudina A V Statistics, Available at: https://abc.vvsu.ru/books/statistika_up/page0015.asp
[6] Loskutov A Yu Time Series Analysis: Course of Lectures, Available at: http://chaos.phys.msu.ru/loskutov/PDF/Lectures_time_series_analysis.pdf.
[7] Arora, Amandeep Singh, Linesh Raja, and Barkha Bahl. "Data centric security approach: A way to achieve security & privacy in cloud computing." Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT) 2018.
[8] Minashkin V G, Shamoilova R A, Sadovnikova N A, Moiseikina L G and Rybakova E S 2008 Theory of statistics: Educational-methodical complex (Mpscow: Ed. Center EAIOI) p 296.
[9] Mkhitaryan V S, Dubrova T A, Minashkin V G and other 2004 Statistics: Textbook for students institutions of environments (Moscow: Publishing Center "Academy") p 272.
[10] Eliseeva I I and Yuzbashev M M 2004 General theory of statistics (Moscow: Finance and statistics) p 656.
[11] Bessonov V A 2014 What will modern Russian statistics preserve for history? Nat. issled. University Higher School of Economics. Electron. text. Dan. (660 KB) (Moscow: Ed. House of the Higher School of Economics) (Series WP2 “Quantitative Analysis in Economics”) p 39 (Preprint WP2 /2014/02/) Available at: https://www.hse.ru/data/2014/04/30/1322594622/WP2_2014_02.pdf.
[12] Vazhenina V S, Pakhomova E A and Pisareva D A 2016 Application of harmonic and spectral analysis to identify the main development cycles of the socio-economic system (for example, agriculture) National interests: priorities and security No. 11 (344) pp 4–14.
[13] Livshits V N 2013 Systematic analysis of the market reform of the non-stationary economy of Russia: 1992–2013 (Moscow: LENAND) 640 p.
[14] Livshits V N, Lychagina T A and Pakhomova E A 2011 Financial management. Fundamentals of evaluating the effectiveness of investment projects. (Dubna State University) 183 p.
[15] Federal State Statistics Service, Available at: http://www.gks.ru/.