УДК 519.25
DOI: 10.25559/SITITO.17.202103.550-563

Научная статья

Применение преобразований данных при вычислении композитного индекса качества системы

Т. В. Жгун
ФГБОУ ВО «Новгородский государственный университет имени Ярослава Мудрого», г. Великий Новгород, Российская Федерация
173003, Российская Федерация, Новгородская область, г. Великий Новгород, ул. Большая Санкт-Петербургская, д. 41
Tatyana.Zhgun@novsu.ru

Аннотация

В статье анализируются особенности данных, используемых при вычислении композитных индексов сложных систем. Метод главных компонент дает объективное резюме набора данных, однако метод критичен к качеству исходных данных. Одним из основных моментов в критике использования методов многомерного анализа для получения весов композитных индексов является, в частности, отсутствие социально-экономической интерпретации получаемых весовых коэффициентов в случае, если метод главных компонент определяет отрицательные знаки весовых коэффициентов. Показано, что известные статистические характеристики, такие как коэффициент асимметрии, коэффициент вариации, характеристики наличия/отсутствия нормального распределения данных не позволяют выявить аномальные переменные, влияющие на результат применения метода главных компонент для вычисления весовых показателей. Для аномальных переменных выбросы на верхнем уровне нивелируют все остальные значения показателя. Логарифмическое преобразование в таком случае позволяет избежать потерь значений. В современных исследованиях логарифмическое преобразование хотя и упоминается как возможное при вычислении композитных индексов, однако конкретных рекомендаций по его применению не дается, и оно практически не применяется на практике. В работе обосновано применение логарифмического преобразования данных, аномальный характер которых выявляется с помощью тепловых карт и подтверждается применением аналитического критерия отношения сигнал / шум, оценивающего переменные в стандартном диапазоне, не содержащем нуля. Предложенный вид логарифмического преобразования позволяет уйти от проблемы больших отрицательных значений, характерных для применения логарифмического преобразования на стандартном единичном интервале. Эффективность использования логарифмического преобразования для оценки качества слабо формализованных систем демонстрируется на примере исследования качества жизни населения РФ с помощью авторской модификации метода главных компонент. Показано, что применение логарифмической коррекции для аномальных переменных решает проблему с отрицательностью весовых коэффициентов и перераспределяет значения получаемых весовых коэффициентов в сторону более корректной социально-экономической интерпретации.

Ключевые слова: композитные индексы, метод главных компонент, тепловые карты, отношение сигнал/шум, коэффициент асимметрии, логарифмическое преобразование данных, весовые коэффициенты

Автор заявляет об отсутствии конфликта интересов.

Для цитирования: Жгун, Т. В. Применение преобразований данных при вычислении композитного индекса качества системы / Т. В. Жгун. – DOI 10.25559/SITITO.17.202103.550-563 // Современные информационные технологии и ИТ-образование. – 2021. – Т. 17, № 3. – С. 550-563.

Original article

The Application of Data Transformations in the Calculation of a Composite Index of a System's Quality

T. V. Zhgun
Yaroslav-the-Wise Novgorod State University, Veliky Novgorod, Russian Federation
41 Bolshaya St. Petersburgskaya Str., Veliky Novgorod 173003, Russian Federation Tatyana.Zhgun@novsu.ru

Abstract
The paper examines features of data used when calculating composite indexes of complex systems. Principal component analysis gives an objective summary of the dataset, but is sensitive to the quality of the data. One of the main critiques of using multidimensional analysis when calculating the weights of composite indexes is ambiguity in the socio-economic interpretation of negative weight coefficients. The paper shows that known statistical characteristics of the data, such as the coefficient of asymmetry, the coefficient of variation and the presence/absence of a normal distribution of data do not allow the identification of anomalous variables. Anomalous data is considered such if the upper range outliers neutralize all other values for an indicator. Such data affects the calculated the weights and can be identified by using heatmaps. The logarithmic transformation of anomalous variables eliminates the peculiarities of the distribution of such data. The paper proposes an analytical criterion for determining the anomalous data. The criterion evaluates the signal-to-noise ratios of variables in a fixed range that does not contain zero. The justification of using the logarithmic transformation when assessing the quality of weakly formalized systems is demonstrated in the example of using the author's modification of the PCA when studying the quality of life of the population of Russia's regions. The paper shows that the use of logarithmic correction for anomalous variables eliminates the negativity of the weight coefficients and results in a redistribution of weights with a more correct socio-economic interpretation.

Keywords: composite index, principal component analysis, signal-to-noise ratio, skewness factor, heat maps, logarithmic transformation, weighting coefficients

The author declares no conflict of interest.

For citation: Zhgun T.V. The Application of Data Transformations in the Calculation of a Composite Index of a System's Quality. *Sovremennye informacionnye tehnologii i IT-obrazovanie = Modern Information Technologies and IT-Education.* 2021; 17(3):550-563. DOI: https://doi.org/10.25559/SITITO.17.202103.550-563

Введение

Задача определения композитных индексов сложных систем возникают при исследовании физических явлений, технических систем, при решении проблем управления социальными и социально-экономическими системами. Количество композитных индексов, используемых на сегодняшний день для измерения разнообразных латентных характеристик слабо формализованных социальных и социально-экономических систем измеряется сотнями. Согласно данным ООН, к 2011 году существовало 290 композитных индексов, применяющихся для и комплексной оценки государств[1]. Большое число используемых на практике композитных индексов для оценки качества разнообразных систем является явным признаком их практической значимости в практике управления сложными системами [1].

Распространение сводных показателей в различных областях общественной жизни влечет возникновение вопросов относительно их точности и надежности. Чувствительность композитных индексов к ошибкам исходных данных может привести к искажению информации о результатах функционирования оцениваемой системы и к ошибочным управленческим и политическим решениям. Следовательно, совершенствование методов построения композитных индексов является важной проблемой исследования как с теоретической, так и с оперативной точек зрения.

Принципиальным отличием вычисления композитных индексов для социально-экономических систем является неопределенность качества используемых данных, в отличие от вычисления характеристик технических и физических систем, для которых заранее известна погрешность измерения и законы распределения данных. Особенности используемого пула статистических данных при автоматическом применении методов многомерного анализа часто могут приводить к недостоверным результатам. Получение разнообразных характеристик объектов должно обязательно учитывать особенности используемых данных при проведении необходимых предварительных преобразованиях данных.

Преобразование данных является одним из важнейших процессов при вычислении показателя качества систем. От того, насколько грамотно она будет проведена, зависит достоверность и точность результатов. Для некоторых данных характерно распределение, когда большая часть значений сосредоточена вблизи среднего значения, но имеется небольшое количество аномально больших значений или аномально малых значений. При вычислении интегрального показателя качества системы выбросы на верхнем уровне значений показателя нивелируют остальные значения, а малые значения игнорируются. Логарифмическая функция преобразования данных в таком случае позволяет избежать потерь значений, уменьшая (сжимая) большие значения в наборе данных и растягивая меньшие значения. Однако в современных исследованиях логарифмическое преобразование хотя и упоминается как возможное, но конкретных рекомендаций по его применению не дается, и исследователи это преобразование на практике практически не применяют.

Метод главных компонент (*Principle Components Analysis, PCA*) является объективным способом определения весовых коэффициентов композитного индекса, не использующим дополнительно субъективную априорную информацию для определения весовых коэффициентов. Хотя *PCA* является лучшим выбором для получения эмпирического резюме набора данных, однако всеобщего признания исследователей эти методики не завоевали. Причиной недостаточной востребованности методики многомерного анализа является, кроме относительной сложности применяемого математического аппарата, нерешенность некоторых методологических проблем. В частности, одним из основных моментов в критике использования методов многомерного анализа для получения весов композитных индексов является отсутствие социально-экономической интерпретации получаемых весовых коэффициентов в случае, если *PCA* определяет знаки весовых коэффициентов, противоположные ожидаемым.

Визуализация данных с помощью тепловых карт служит для наглядного сравнения множества показателей при оценке изменения композитных индексов качества систем. Также визуализация данных выявляет особенности данных, требующих нестандартных методов преобразований данных. В качестве аналитического критерия аномальности исходных данных предлагается отношение сигнал/шум, которое оценивает переменные в стандартном диапазоне. Применение логарифмической трансформации желательно для переменных, у которых характеристика сигнала менее характеристик шума, т.е. вычисленное значение отношения сигнал/шум менее единицы. Применение логарифмической коррекции для переменных такого типа при использовании метода главных компонент для определения весовых коэффициентов решает проблему с отрицательностью весовых коэффициентов и, кроме того, перераспределяет значения получаемых весовых коэффициентов в сторону более корректной социально-экономической интерпретации..

Целью исследования является выявления особенностей данных, требующих нестандартных методов предварительной обработки при вычислении композитных индексов сложных систем, аналитическое обоснование использования логарифмического преобразования переменных при вычисления композитных индексов слабо формализованных систем и демонстрация эффективности использования логарифмического преобразования для анализа и оценки качества слабо формализованных систем на примере исследования качества жизни населения РФ с помощью многомерного анализа.

Статья имеет следующую структуру. В разделе 1 представлен обзор использования методов многомерного анализа для вычисления композитных индексов. В разделе 2 оценивается специфика набора статистических данных на примере сета данных, традиционно характеризующих качество жизни населения Российской Федерации во многих исследованиях. Распределение переменных иллюстрируется с помощью тепловых карт. Рассматриваются характеристики данных, позволяющие судить об наличии/отсутствия нормального распределения, рассматривается коэффициент вариации как харак-

[1] Bandura R. Composite Indicators and Rankings: Inventory 2011. Technical Report. United Nations Development Programme – Office of Development Studies, 2011. 257 p. URL: http://nebula.wsimg.com/beb0f5b0e9b6f0c298ae72e7519b22a4?AccessKeyId=EA00C1BFA208EDA3747F&disposition=0 (дата обращения: 14.07.2021).

теристика набора данных и предлагается в качестве характеристики набора данных использовать отношение сигнал/шум, имеющее прозрачную интерпретацию. Предлагается критерий оценки качества набора данных на основании отношения сигнал/шум. В разделе 3 излагается обсуждение применения логарифмического преобразования данных при построении композитных индексов. В разделе 4 продемонстрированы результаты применения логарифмической коррекции для вычисления интегрально показателя качества жизни населения РФ за период 2010 – 2017 гг. Общие выводы по результатам работы приводятся в заключительном разделе.

Использование многомерного анализа для вычисления композитных индексов

Рассмотрим систему из *m* объектов, где каждый объект характеризуется *n* описаниями, т.е. система характеризуется матрицей описаний размерностью *m*×*n* $A = \{ a_{ij} \}$. Численные характеристики системы приведены на отрезок [0, 1] по принципу: «чем больше, тем лучше». Для решения задачи управления требуется дать мотивированную интегральную оценку каждого наблюдаемого объекта.

Хотя в литературе разработаны различные функциональные формы правил агрегирования композитных индексов [2], [3], в стандартной практике составной показатель *CI* объекта *i* обычно рассматривается как взвешенная линейная свертка набора переменных[2]:

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

Или в матричной форме

$$CI = A \cdot w \quad (2)$$

где $CI = \langle CI_1, CI_2, \ldots, CI_m \rangle^T$ – вектор интегральных индикаторов для объектов системы, $w = \langle w_1, w_2, \ldots, w_n \rangle^T$ – вектор весов показателей. Обычно требуется, чтобы $\sum_{j=1:n} w_j \cdot = 1$ и $0 \le w_j \le 1$, $j = 1 \ldots n$. Для построения искомого композитного индекса качества системы требуется найти веса показателей w_j

Многомерный анализ впервые был использован для объединения показателей в единый индекс при разработке индекса здоровья в работе [4] и широко использовался исследователями с конца 1990 годов при разработке других композитных индексов. Метод главных компонент (*Principle Components Analysis, PCA*) и факторный анализ (*Factor Analysis, FA)* являются наиболее часто используемыми многомерными статистическими методами для формирования весов составных индексов[3]. Многие авторы для получения весов с использованием PCA или FA используются факторные нагрузки переменных на первую компоненту, поскольку «вариация, объясненная первой главной компонентой, обычно достаточна для адекватного представления исходных переменных»[4].

Проекция на первую главную компоненту определяла вычисляемый индекс при конструировании Индексов социально-экономический статуса: *(Socio-Economic Status Indices, SES*) [5], [6], Индекса экологической устойчивости [7], индекса качества жизни в Европе [8] и России[5] и других композитных индексов[6] [5], [9]-[11]. Однако если объясняющая способность первой компоненты составляет менее 55%, то для получения весов следует включать последующие компоненты, чтобы предотвратить потерю полезной не избыточной информации [12], [13].

Исследователи Организации экономического сотрудничества и развития[7]: (OECD, 2003) для формирования составного индикатора используют факторный анализ, где количество извлеченных факторов объясняло бы более 50% общей дисперсии. Значение композитного индекса в этом случае определяют только значимые нагрузки выбранных главных факторов после вращения, а незначимые нагрузки обнуляются. Эта методика становится широко применяется при вычислении композитных индексов[8] [4], [8], [11], [14].

Часто *PCA* и *FA* используются взаимозаменяемо, но если исследователь нуждается в эмпирическом обобщении набора данных, которое объясняет максимальную дисперсию с помощью уникального математического решения, то предпочтительнее использовать *PCA* [14], Руководство по построению составных индексов[9] рекомендует использовать именно *PCA* при разработке составных индексов, поскольку «достоинством метода является простота и он позволяет получить весовые коэффициенты, отражающие информационное содержание отдельных показателей». Т.е. *PCA* является полезным методом выбора переменных, которые должны быть включены в синтетический индекс [8]. Кроме того, использование PCA позволяет избежать дублирования информации, поскольку компоненты ортогональны [8]. *PCA* является объективным способом определения весовых коэффициентов композитного индекса, не используя дополни-

[2] Handbook on Constructing Composite Indicators. Methodology and user guide. OECD, European Commission, Joint Research Centre, 2008. 158 p. URL: https://knowledge4policy.ec.europa.eu/sites/default/files/jrc47008_handbook_final.pdf (дата обращения: 14.07.2021).

[3] Greyling T. A composite index of quality of life for the Gauteng city-region: a principal component analysis approach. Department of Economics and Econometrics, University of Johannesburg, 2013. 47 p. URL: https://cdn.gcro.ac.za/media/documents/occasional_paper_7.pdf (дата обращения: 14.07.2021).

[4] Айвазян С. А. Анализ качества и образа жизни населения: эконометрический подход. М.: Наука, 2012. 432 с.

[5] Там же.

[6] Lindman C., Sellin J. Measuring Human Development: The Use of Principal Component Analysis in Creating an Environmental Index: Bachelor Thesis. Uppsala: Department of Statistics, Uppsala University, 2011. URL: http://uu.diva-portal.org/smash/get/diva2:464378/FULLTEXT03.pdf (дата обращения: 14.07.2021).

[7] Freudenberg M. Composite Indicators of Country Performance: A Critical Assessment. STI Working Paper 2003/16. Industry Issues. DSTI/DOC(2003)16. OECD, 2003. doi: https://doi.org/10.1787/405566708255

[8] Nicoletti G., Scarpetta S., Boylaud O. Summary Indicators of Product Market Regulation with an Extension to Employment Protection Legislation (December 1999). OECD, ECO Working Paper No. 226. ECO/WKP(99)18. OECD, 1999. doi: http://dx.doi.org/10.2139/ssrn.201668; Nardo M., Saisana M., Saltelli A., Tarantola S. Tools for Composite Indicators Buildings. Report EUR 21682 EN. OECD, European Commission – Joint Research Centre, Ispra, 2005. 132 p. URL: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC31473/EUR%2021682%20EN.pdf (дата обращения: 14.07.2021).

[9] Handbook on Constructing Composite Indicators. Methodology and user guide. OECD, European Commission, Joint Research Centre, 2008. 158 p. URL: https://knowledge4policy.ec.europa.eu/sites/default/files/jrc47008_handbook_final.pdf (дата обращения: 14.07.2021).

тельно субъективную априорную информацию, необходимую для определения весовых коэффициентов. Это основной метод определения весов, который способствует выбору ограниченного числа индикаторных переменных, представляющих данные и обеспечивающих измерение качества описываемой системы. Хотя *PCA* является лучшим выбором для получения эмпирического резюме набора данных [14], однако всеобщего признания исследователей эти методики не завоевали. Причиной этого является, в частности, нерешенность некоторых методологических проблем

Авторы[10] [22] обращают внимание, что различные методы извлечение основных компонент и различные способы вращения подразумевают различные значимые переменные, а значит, разные веса переменных при вычислении композита, следовательно, и различные значения вычисляемого интегрального индикатора. Веса переменных изменяются по мере изменения данных с течением времени. Следовательно, методика определения весовых коэффициентов с помощью факторного анализа не может применяться для сравнения характеристик описываемых объектов в динамике, так как даже при фиксированных методах извлечения факторов и способа вращения, факторный анализ для разных наблюдений системы структуру главных факторов (значимые нагрузки) определяет по-разному, что делает бессмысленным межвременные сравнения.

Еще одним основным моментом в критике использования методов многомерного анализа для получения весов композитных индексов является отсутствие социально-экономической интерпретации получаемых весовых коэффициентов[11] [8]. В частности, применение метода главных компонент *PCA* может давать знаки веса переменных, противоположные ожидаемым. При построении индекса экологической устойчивости отрицательные веса некоторых показателей, полученных с помощью *PCA*, явились причиной отказа от этой методики и были использованы равные веса вместо весов, полученных при помощи *PCA*[12].

В независимых исследованиях по оценке социально-экономического положения с помощью *PCA* [15] были получены противоположные ожиданию веса: отрицательные для переменных «Размер жилища», «Размер доходов» и положительные для уровня безработицы и возраста жилища. При этом вес переменной «Образование» – положительный.

При наличии отрицательных факторных нагрузок неопределенным является выбор направления полученных главных факторов или главных компонент. Очевидно, что при изменении направления компонент с факторными нагрузками разных знаков итоговые веса в композитном индексе будут определены по-разному. Одним из способов решения этой проблемы является уход от отрицательных весовых коэффициентов. Для этого в работе[13] выбранные значимые нагрузки возводят в квадрат.

В исследовании[14] авторы считают, что при наличии в главных компонентах показателей разных знаков не следует придавать значения знаку факторных нагрузок. В работе [11] для формирования композитного индикатора были выбраны пять факторов, которые объяснили 55,7 % от общей вариации. Но только в двух факторах нагрузки имеют одинаковые знаки. Выбор направления факторов автором не обоснован. При этом стоимость жилья, средний доход, размер жилья, наличие жилья в собственности в социально-экономическом индексе имеют отрицательные нагрузки и их увеличение уменьшает величину вычисляемого интегрального показателя, а низкий уровень дохода имеет положительный весовой коэффициент.

В работах[15] [11], [15], [16] при вычислении композитного индекса авторы также суммируют нормализованные факторы с факторными нагрузками разных знаков. Выбор направления факторов никак не объясняется.

Методы многомерного анализа, дающие потрясающие надежные результаты для комплексной оценки физических, технических, естественных систем, не показывают столь же впечатляющих результатов при оценке слабо формализованных социально-экономических систем. Одной из причин столь значительной разницы результата применения одинакового подхода к системам разного типа может быть в специфике используемых данных.

Статистические данные представляют собой результат измерения и значит, неизбежно содержат неустранимые ошибки измерения, связанные с точностью измерительного инструмента (белый гауссовский шум). роме того, любому типу данным присущи другие виды искажений – выбросы данных, вызванные внешними причинами: сбоями техники, ошибками при регистрации, сознательные искажения (импульсный шум). Для систем разного типа для борьбы с шумами успешно используются одинаковые методы шумоподавления, которые могут работать при наличии **нескольких временных реализаций** описываемой системы. Для однократного наблюдения шумы устранить невозможно. Следовательно, можно использовать те же методы шумоподавления для систем, описываемых статистическими данными.

Рассмотрим набор переменных, характеризующий качество

[10] Nardo M., Saisana M., Saltelli A., Tarantola S. Tools for Composite Indicators Buildings. Report EUR 21682 EN. OECD, European Commission – Joint Research Centre, Ispra, 2005. 132 p. URL: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC31473/EUR%2021682%20EN.pdf (дата обращения: 14.07.2021).

[11] Mishra S. K. A Comparative Study of Various Inclusive Indices and the Index Constructed by the Principal Components Analysis. MPRA Paper, 2007. No. 3377. URL: https://mpra.ub.uni-muenchen.de/3377 (дата обращения: 14.07.2021).

[12] Jacobs R., Smith P., Goddard M. Measuring performance: An examination of composite performance indicators. The University of York: Centre for Health Economics, Technical Paper Series, 2004. vol. 29. 112 p. URL: https://www.york.ac.uk/che/pdf/tp29.pdf (дата обращения: 14.07.2021).

[13] Greyling T. A composite index of quality of life for the Gauteng city-region: a principal component analysis approach. Department of Economics and Econometrics, University of Johannesburg, 2013. 47 p. URL: https://cdn.gcro.ac.za/media/documents/occasional_paper_7.pdf (дата обращения: 14.07.2021); Nicoletti G., Scarpetta S., Boylaud O. Summary Indicators of Product Market Regulation with an Extension to Employment Protection Legislation (December 1999). OECD, ECO Working Paper No. 226. ECO/WKP(99)18. OECD, 1999. doi: http://dx.doi.org/10.2139/ssrn.201668

[14] Foa R., Tanner J.C. Methodology of the Indices of Social Developmen (No. 2012-04). ISD Working Paper Series, 2012. 79 p. URL: http://hdl.handle.net/1765/50510 (дата обращения: 14.07.2021).

[15] Исакин М. А. Модификация метода k-средних с неизвестным числом классов // Прикладная эконометрика. 2006. № 4(4). С. 62-73. URL: https://www.elibrary.ru/item.asp?id=9482376 (дата обращения: 14.07.2021).

жизни населения[16]. Эти переменные использовались для оценки качества жизни населения России во многих исследованиях, например[17] [17]. Список переменных приведен в табл.1.

Особенности используемого пула статистических данных при автоматическом применении методов многомерного анализа часто могут приводить к недостоверным результатам. Получение разнообразных характеристик объектов на основании статистических данных должно обязательно учитывать наличие неустранимой ошибки данных. В работах автора [18]-[20], [23] приводится авторская модификация метода главных компонент для определения изменения качества системы по ряду наблюдений, учитывающий наличие шума в исходных данных. Метод дает устойчивые и достоверные результаты. Обсуждение метода не входит в проблематику данной статьи.

В исследование далее будет использоваться авторская методика построения композитного индекса качества системы по ряду наблюдений. По набору переменных, предложенному в работе[18], строилась комплексная оценка качества жизни населения регионов Российской Федерации за 2010-2017 годы. Рассмотрение именно этой композитной характеристики качества сложный систем позволяет говорить о наличии или отсутствии социально-экономической интерпретации результатов, а именно о социально-экономической интерпретации вычисляемых весовых коэффициентов линейной свертки показателей (1), так как именно этот момент является основным в критике применения многомерного анализа при вычислении композитных индексов систем.

Таблица 1. **Переменные для вычисления композитного индекса качества жизни**

Table 1. **Variables for calculating the composite index of quality of life**

№ п/п	Переменные
	Блок 1: Уровень благосостояния населения
1	ВРП на душу населения с учетом инфляции, тыс. руб.
2	Отношение среднедушевых денежных доходов к прожиточному минимуму
3	Доля населения с доходами ниже прожиточного минимума
4	Отношение доходов 20% самых богатых и 20% самых бедных
5	Обеспеченность собственными легковыми автомобилями на 1000 человек
6	Доля семей, состоящих на учете на получение жилья
7	Общая площадь жилищного фонда на одного жителя
8	Доля ветхого и аварийного жилья
9	Плотность автомобильных дорог общего пользования
	Блок 2: Качество населения
10	Ожидаемая продолжительность жизни при рождении
11	Число умерших детей в возрасте до 1 года на 1000 родившихся.
12	Коэффициент естественного прироста на 1000 человек
13	Умерших от инфекционных болезней и туберкулеза на 100 тысяч человек
14	Число умерших от новообразований на 100 тысяч человек
15	Умерших от болезней системы кровообращения на 100 тысяч человек.
16	Число умерших от болезней органов дыхания на 100 тысяч человек
17	Число умерших от болезней органов пищеварения на 100 тысяч человек
18	Заболеваемость от травм и других внешних причин на 100 тысяч человек
19	Число инвалидов на 1000 человек населения
20	Зарегистрировано врожденных аномалий на 1 000 человек
21	Доля специалистов с высшим образованием к занятым в экономике
22	Отношение ВРП к численности занятых в экономике, тыс. руб./чел.
23	Численность студентов высших и средних учебных заведений на 1000 населения
	Блок 3: Качество социальной сферы
24	Уровень безработицы (%)
25	Численность смертей при несчастных случаях на производстве на 1000 работающих
26	Численность пострадавших при несчастных случаях на производстве на 1000 человек работающих
27	Коэффициент миграционного прироста на 10 тысяч человек
28	Число зарегистрированных умышленных убийств на 100 тысяч человек
29	Число фактов умышленного. причинения тяжкого вреда здоровью на 100 тысяч человек
30	Число зарегистрированных изнасилований на 100 тысяч человек.
31	Разбоев, грабежей, краж из квартир граждан в расчете на 100 тысяч человек
32	Зарегистрированных присвоений или растрат в расчете на 100 тысяч человек
33	Состоящих на учете наркомания и токсикомания на 100 тысяч человек.
34	Состоящих на учете с диагнозом алкоголизм в расчете на 100 тысяч человек
35	Больных, инфицированных туберкулезом в расчете на 100 тысяч человек.
36	Число больных с диагнозом сифилиса на 100000 человек
37	Число больных психическими расстройствами на 100 тысяч человек.

[16] Айвазян С. А., Степанов В. С., Козлова М. И. Измерение синтетических категорий качества жизни населения региона и выявление ключевых направлений совершенствования социально-экономической политики (на примере Самарской области и ее муниципальных образований) // Прикладная эконометрика. 2006. № 2(2). С. 18-84. URL: https://www.elibrary.ru/item.asp?id=9482361 (дата обращения: 14.07.2021).

[17] Жгун Т. В. Алгоритм построения интегрального индикатора качества сложной системы для ряда последовательных наблюдений // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. 2017. Т. 6, № 1. С. 5-25. doi: https://doi.org/10.14529/cmse170101

[18] Айвазян С. А., Степанов В. С., Козлова М. И. Измерение синтетических категорий качества жизни населения региона и выявление ключевых направлений совершенствования социально-экономической политики (на примере Самарской области и ее муниципальных образований) // Прикладная эконометрика. 2006. № 2(2). С. 18-84. URL: https://www.elibrary.ru/item.asp?id=9482361 (дата обращения: 14.07.2021).

Оценка особенностей данных

Исследование вида распределения данных

Отметим особенность статистических данных, участвующих в определении качества социально-экономических систем. Обычно в качестве разных типов данных рассматриваются данные с нормальным распределением и данные, не подчиняющиеся нормальному закону.

Для определения близости эмпирического распределения к нормальному закону используют показатели эксцесса и асимметрии. Вычисление асимметрии позволяет установить симметричность распределения случайной величины относительно математического ожидания. Для этого находят третий центральный момент, характеризующий асимметрию закона распределения случайной величины. Если он равен нулю, то случайная величина симметрично распределена относительно математического ожидания. Поскольку третий центральный момент имеет размерность случайной величины в кубе, то вводят безразмерную величину – коэффициент асимметрии:

$$A_s = m_3 / \sigma_2^3 \qquad (3)$$

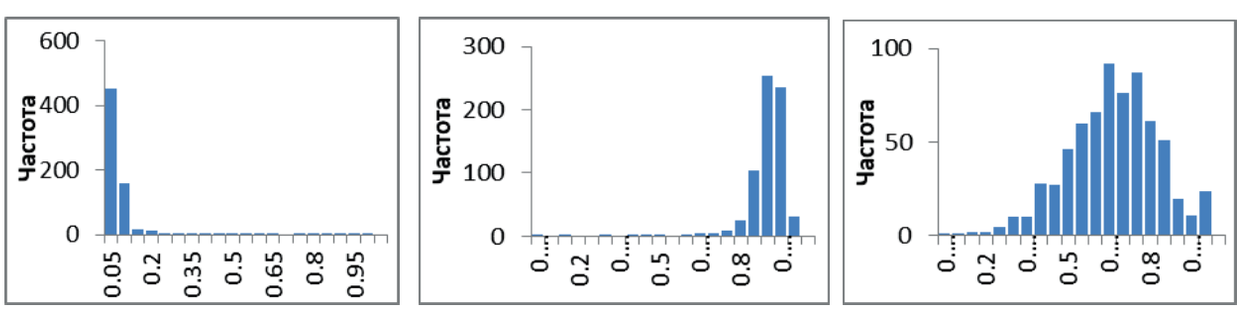
где σ_2 – стандартное выборочное отклонение, m_3 – центральный эмпирический момент третьего порядка: $m_3 = (\sum_{i=1}^{n}(x-\bar{x}_2)^3)/n$, $\bar{x}_2$ – выборочная средняя. При этом принята следующая условная градация: если коэффициент асимметрии $|A_s| < 0.25$, то асимметрия незначительна, если вычисленное значение коэффициента асимметрии $0.25 < |A_s| < 0.5$, то асимметрия умеренная, и если коэффициент асимметрии $|A_s| > 0.5$, то асимметрия существенная. При значительных отклонениях показателей асимметрии от нуля нельзя признать совокупность однородной, а распределение близким к нормальному.

Коэффициент асимметрии, определенный для рассматриваемых переменных в табл. 2, показывает, что распределение только 5 переменных из рассматриваемых 37 можно характеризовать как умеренную асимметрию, где $0.25 < |A_s| < 0.5$. Для всех остальных переменных коэффициент асимметрии говорит о большом перекосе и полном отсутствии нормального распределение переменных.

Таблица 2. **Коэффициенты асимметрии переменных**

Table 2. **Coefficients of asymmetry of variables**

Номер блока	Номер переменной в блоке													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	5.3	0.5	1.5	0.3	0.3	2.7	-0.8	3.4	4					
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	-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



а) б) в)

Рис. 1. Гистограммы распределения для переменных с разными характеристиками асимметрии:
а) для переменной 1 (As = 5.3), б) для переменной 30 (As = –5.0), в) для переменной 31 (As = 0.018)

Fig. 1. Distribution histograms for variables with different asymmetry characteristics:
а) for variable 1 (As = 5.3), б) for variable 30 (As = -5.0), в) for variable 31 (As = 0.018)

Наиболее значительную асимметрию имеют переменные 1, 8, 9, 22, 23, 27, 28, 29. Обратим внимание, что выборку, составленную из коэффициентов асимметрии можно охарактеризовать как умеренно асимметричную: среднее значение коэффициента асимметрии $\bar{A}_s = -0.27$, а коэффициент асимметрии для этой выборки $A_s(A_s) = 0.37$.

Факт, что в рассматриваемом случае все данные не подчиняются закону нормально распределения, подтверждается тестом Шапиро-Уилка. Для всех переменных вычисленное значение p_{value} лежит в диапазоне 0.0000 – 0.0167 и значительно менее $p = 0.05$. Это значит, что различие рассматриваемых распределений с нормальным сильно статистически значимо, и гипотезу о нормальном распределении переменных необходимо отвергнуть. Проверяя условие нормальности распределения данных, необходимо хорошо представлять себе, в каких случаях его выполнение является критическим для применения конкретного статистического метода. Так, например, метод главных компонент не требует, чтобы данные были распределены нормально [13].

На рис.1 приведены гистограммы частот переменных, имею-

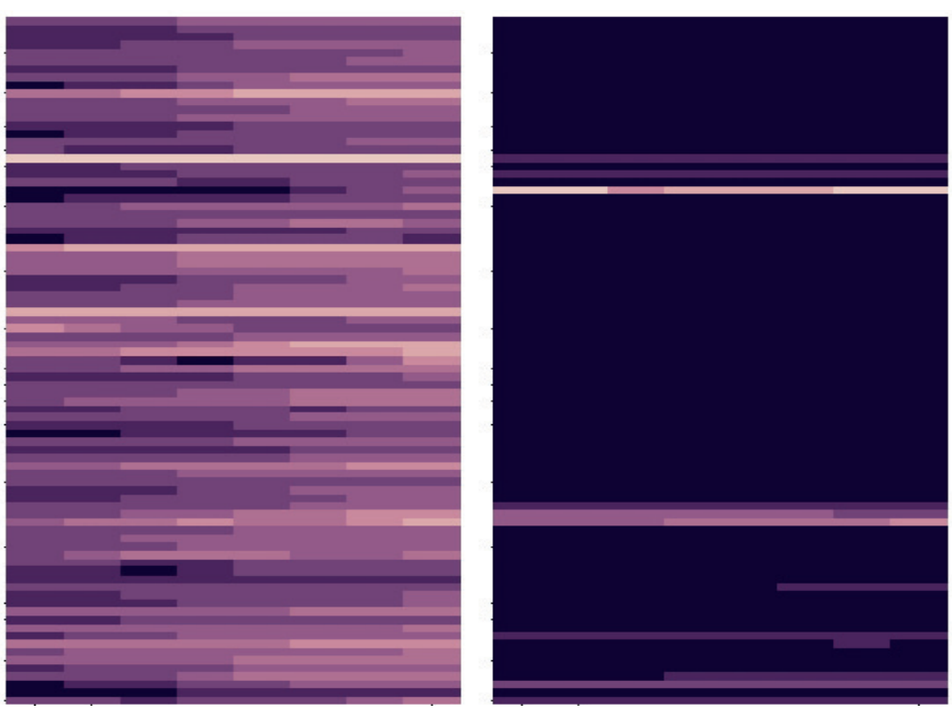
щих разные типы распределения: 1 «ВРП на душу населения», 30 «Число зарегистрированных изнасилований на 100 тысяч человек» и 31 «Число разбоев, грабежей, краж из квартир граждан в расчете на 100 тысяч человек». Это пример переменных с самыми выдающимися коэффициентами перекоса: самое значительное смещение влево (переменная 1) – положительная (правосторонняя) асимметрия; самое значительное смещение вправо (переменная 30) – отрицательная (левосторонняя) асимметрия и максимальная близость к нормальному распределению (переменная 31).

Однако переменные с одинаково сильной асимметрией ведут себя по-разному. Для переменной 1 видим существенное преобладание маленьких значений: 620 значений из рассматриваемых 680 сгруппированы в диапазоне 0 – 0.1 единичного отрезка, а переменная 30 такой особенностью не обладает. Именно значительное преобладание значений в узком диапазоне является той особенностью данных, которое требует дополнительных преобразований. Коэффициенты асимметрии, характеристики наличия/отсутствия нормального распределения данных эту особенность не смогли выявить. Для таких аномальных статистических данных принципиально не наличие большого разброса значений переменных, которое устраняется масштабированием, а явное наличие кластеров разновеликих значений. Данные такого типа в статистике встречаются довольно часто. Это характеристики денежных величин, заболеваемости болезнями разного типа, характеристика плотности дорожного покрытия и др.

Применение тепловых карт для исследования особенностей данных

Такую особенность данных хорошо видно на тепловых картах (*heatmap*) переменных. Визуализация с помощью тепловых карт служит для наглядного сравнения множества показателей и выявления особенностей данных, требующих нестандартных методов дальнейшей работы с данными и необходимых преобразований данных. Одним из инструментов представления многомерных данных являются тепловые карты. Исследователи в разных областях используют тепловые карты, чтобы представить многомерные данные [21]. Визуализация с помощью тепловых карт служит для наглядного сравнения множества показателей и выявления особенностей данных, требующих нестандартных методов дальнейшей работы с данными и необходимых преобразований данных.

При представлении матрицы данных с помощью тепловых карт числа в ячейках таблицы заменяются на заливку этих ячеек цветом по определенному принципу. Численные значения величин предварительно приведены в единую шкалу, при этом большее значение параметра соответствует более высокому качеству системы и изображается более светлым оттенком. Данные с большим разбросом хорошо видны на тепловых картах. Они выражены либо аномально темным (рис. 2 б) либо аномально светлыми тепловыми картами, которые резко контрастируют с представлением данных другого типа (рис. 2 а). Анализ тепловых карт выявил аномальность переменных 1, 8, 9, 22. Строки тепловой карты соответствуют субъектам РФ, столбцы – это годы. На рис. 2 представлен диапазон наблюдений 2010-2017 годы. Посветление карты на рис.2 а свидетельствуют об улучшение ситуации со временем.



а) б)

Р и с. 2. Тепловая карта изменения типичной переменной 21 (а) и аномальной переменной 22 (б)

F i g. 2. Heat map of a typical change in a variable 21 (a) and an anomalous variable 22 (б)

Отношение сигнал/шум для характеристики данных

Коэффициент вариации позволяет судить об однородности совокупности, высокая степень которой обеспечивает объективность и достоверность показателей, об устойчивости значений признака, типичности средней и др. Однако для характеристики данных как сигнала более употребительно отношение сигнал/шум, обратное коэффициенту вариации – *SNR,* характеризующее отношение среднего значения амплитуды сигнала к амплитуде шума. Среднее значение является мерой сигнала, а стандартное отклонение является мерой шума.

$$SNR = m / \sigma \quad (4)$$

SNR является наиболее важным параметром, характеризующим уровень помех в системе. Различить сигнал и шум возможно в случае, если параметры сигнала превышают параметры шума, т.е. если $SNR > 1$.

Для сравнения переменные следует привести к единому масштабу. Обычно переменные приводят на единичный отрезок. В этом случае для аномальных переменных вычисленное значение математического ожидания может быть близко к нулю, что даст и близкое к нулю значение *SNR*. Кроме этого, результат логарифмирования на единичном диапазоне сильно искажает распределение значений переменных и неопределен для минимального (нулевого) значения показателя. Чтобы не получить нулевых характеристик математического ожидания и потом сделать возможным логарифмирования, анализируемые данные стоит рассматривать в сигнальной форме. В этом случае минимальное значение переменной один, максимальное – 100. Соответствующее преобразование переменной имеет следующий вид:

$$x'_i = \frac{99 \cdot (x_i - m)}{(M - m)} + 1, \text{ где } M = max\, x_i,\ m = max\, x_i \quad (5)$$

Максимум и минимум переменной берутся по выборке для всех наблюдений. При возможном применении операции логарифмирования преобразование (5) позволит избежать давления малых значений переменных, которые после непосредственного логарифмирования переменных на стандартном отрезке [0,1] имели бы большие отрицательные значения.

Преобразование (5) данных из статистической формы в сигнальную изменяет характеристики данных, в частности отношение сигнал / шум и делает возможным корректное сопоставление разных данных. Для статистических данных значение $SNR < 1$ имели 7 показателей (выделены в табл. 3), а те же показатели в сигнальной форме выделяют 4 таких показателя: 1, 8, 9, 22. Именно для этих переменных была выявлена аномальность распределения с помощью тепловых карт. Именно эти показатели имеют аномальное распределение значений переменных, которое может вносить искажения в работу метода главных компонент. Все эти показатели имеют значительный коэффициент асимметрии, но не самый большой. Наибольший перекос в рассматриваемой выборке имели переменные 1, 9, 22, 30 (табл. 2). Для выявленных аномальных показателей необходимо преобразование, устраняющее имеющуюся особенность распределения.

Т а б л и ц а 3. **Изменение отношения сигнал/шум при преобразовании данных из статистической формы в сигнальную**

T a b l e 3. **Change in the signal-to-noise ratio when converting data from a statistical form to a signal form**

Номер блока	Тип данных	Номер переменной в блоке													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	статистические данные	***0.7***	4.6	2.8	6.2	3.9	1.6	6.8	***0.8***	***0.7***					
	сигнальная форма	**0.7**	2.3	1.8	2.5	3	1.5	3.5	***0.9***	***0.8***					
2	статистические данные	25.5	2.7	***0***	2	1.6	4.4	2.7	3.3	2.9	1.2	1.5	5.5	***1***	3.5
	сигнальная форма	4.7	1.9	***2***	1.5	3.7	2.7	2.6	3.1	2.6	3.5	1.4	2.4	***0.8***	2.8
3	статистические данные	1.4	1.6	2.4	***–0.1***	1.6	1.3	***1***	2.3	1.8	1.9	2.1	2.1	1.3	3.4
	сигнальная форма	1.4	1.5	2.3	***3.6***	1.5	1.4	***1.1***	2.2	1.6	1.9	2.2	1.5	1.3	2

Логарифмическая коррекция

Преобразование данных является одним из важнейших процессов при вычислении показателя качества систем. От того, насколько грамотно она будет проведена, зависит достоверность и точность результатов. Для некоторых данных характерно распределение, когда большая часть значений сосредоточена вблизи среднего значения, но имеется небольшое количество аномально больших значений или аномально малых значений. При вычислении интегрального показателя качества системы выбросы на верхнем уровне значений показателя нивелируют остальные значения, а малые значения игнорируются. Логарифмическая функция преобразования данных в таком случае позволяет избежать потерь значений, уменьшая (сжимая) большие значения в наборе данных и растягивая меньшие значения. Визуально логарифмическое преобразованных данных позволяет человеческому глазу более четко визуализировать слабые сигналы (значения серого), поэтому логарифмическое преобразование стало стандартной процедурой обработки сигналов оптической когерентной томографии[19].

В работах[20] [22] логарифмическое преобразование упоминает-

[19] Freudenberg M. Composite Indicators of Country Performance: A Critical Assessment. STI Working Paper 2003/16. Industry Issues. DSTI/DOC(2003)16. OECD, 2003. doi: https://doi.org/10.1787/405566708255

[20] Handbook on Constructing Composite Indicators. Methodology and user guide. OECD, European Commission, Joint Research Centre, 2008. 158 p. URL: https://knowledge4policy.ec.europa.eu/sites/default/files/jrc47008_handbook_final.pdf (дата обращения: 14.07.2021); Foa R., Tanner J.C. Methodology of the Indices of Social Developmen (No. 2012-04). ISD Working Paper Series, 2012. 79 p. URL: http://hdl.handle.net/1765/50510 (дата обращения: 14.07.2021).

ся в ряду других нелинейных преобразований, часто используемым для уменьшения асимметрии положительных данных. Логарифмическое преобразование рекомендуется применять, в случае, если диапазон значений показателя имеет большой разброс или он положительно смещен. Преобразование авторы рекомендуют осуществлять по формуле $y = log(x)$ для положительных переменных $x > 0$. Авторы отмечают, что логарифмическое преобразование сжимает правую часть распределения. А в случае близости к нулю исходных значений переменных после трансформации (*log transformation*) они становятся в основном отрицательными. Это преобразование приводит к присвоению более высоких весов переменным с малыми значениями и уменьшает вес переменных с высокими показателями. Решение использовать или не использовать логарифмическое преобразование принимается пользователем.

Работа[21] предназначена служить «методологическим компаньоном» для построения сводных показателей. В статье рассматриваются методы построения составных показателей и присущие им недостатки. Автор отмечает, что при вычислении композитных индексов переменные должны быть стандартизированы (приведены на единый интервал). Кроме того, автор советует применять преобразование переменных, чтобы избежать доминирования экстремальных значений и частично исправить проблемы с качеством данных. Величины, далекие от среднего значения или от нормального диапазона, с большей вероятностью отражают плохие исходные данные. Такие сильно искаженные распределения могут быть выравнены с помощью логарифмических преобразований. Однако конкретных рекомендаций для применения логарифмической трансформации не дается.

Работа [24] посвящена вопросу построения композитных индикаторов. Авторы объявляют свой целью выработку рекомендаций практикующим специалистам в этой области. Основное внимание исследования уделяется нелинейным преобразованиям исходных данных. Логарифмическое преобразование авторы относят к классу силовых преобразований (*Power Transformations, PTs*), описываемых семейством функций:

$$T_p(x_i) = \begin{cases} a \cdot x^p + b, & \text{если } p \neq 0 \\ a \cdot \log(x) + b, & \text{если } p = 0 \end{cases}, \text{ где } b \neq 0 \text{ и } a \neq 0.$$

Отмечая, что *PT* широко используются в статистических моделях для анализа экспериментальных данных с целью стабилизации дисперсии, восстановления нормальности и устранения отсутствия аддитивности, авторы не дают рекомендаций о выборе параметров силового преобразования, ограничиваясь замечанием, что выбор подходящего значения p обусловлен целью исследования и характером исходных данных. В общем случае правильное значение параметра p получается графическими методами, которые могут быть использованы для грубой оценки исходных преобразований. Для построения композитных индикаторов исходные данные часто преобразуются с помощью логарифмического преобразования ($p = 0$), главным образом потому, что оно линеаризует данные и уменьшает асимметрию. Один из недостатков логарифмического преобразования заключается в том, что исходные данные изначально нормализованы и концентрируются в узком диапазоне между нулем и единицей, и тогда преобразованные данные становятся большими отрицательными числами [25]. Такая ситуация искажает общую картину.

Подводя итоги, можно сказать, что слова о возможности/желательности применения логарифмической трансформации данных остаются общими словами, поскольку ни в одной из вышеупомянутых работ при вычислении конкретных композитных индексов это преобразование не применялось, хотя все используемые переменные имеют явную асимметрию. Рассмотренный выше критерий аномальности исходных данных на основании отношения сигнал/шум, оценивающий переменный в стандартном диапазоне сигнальной формы, дает четкие указания о желательности применения логарифмической трансформации переменных, для которых вычисленное значение $SNR < 1$.

Логарифмическая функция имеет важную особенность – позволяет сжимать диапазон значений переменных, имеющих большие вариации в значениях (следовательно, низкое отношение сигнала к шуму). Использование логарифма позволяет узкий диапазон малых значений преобразовать в более широкий диапазон, одновременно уменьшая диапазон больших значений. Недостатком метода является то, что логарифм малых чисел отрицателен по знаку и имеет большие абсолютные значения. Это приводит к очень высокой чувствительности методов обработки сигнала к погрешностям в области малых значений исходных данных. Для устранения этого недостатка значения функции преобразования должны иметь только положительные значения. Для этого значения переменных сначала приводятся на интервал 1 – 100 (5), а потом логарифмируются. Такая трансформация данных меняет их *SNR*- характеристики (табл. 4). Они стали уверенно более единицы.

В рассматриваемом случае в первом блоке логарифмическая коррекция целесообразна для переменных 1, 8, 9 – табл. 4. Для этих переменных среднее значение оказывается менее среднеквадратичного отклонения. Во втором блоке аномальной оказывается переменная 22. Итого в наборе их 37 переменных логарифмической коррекции потребовали 4 переменные.

Таблица 4. ***SNR*- характеристики переменных первого блока для разных представлений данных**
Table 4. **SNR-characteristics of the variables of the first block for different data representations**

Форма данных	Номер переменной								
	1	2	3	4	5	6	7	8	9
сигнальная форма	**0.66**	2.33	1.79	2.46	3.04	1.5	3.46	***0.88***	***0.78***
логарифмическая коррекция	2.26							2.69	3.08

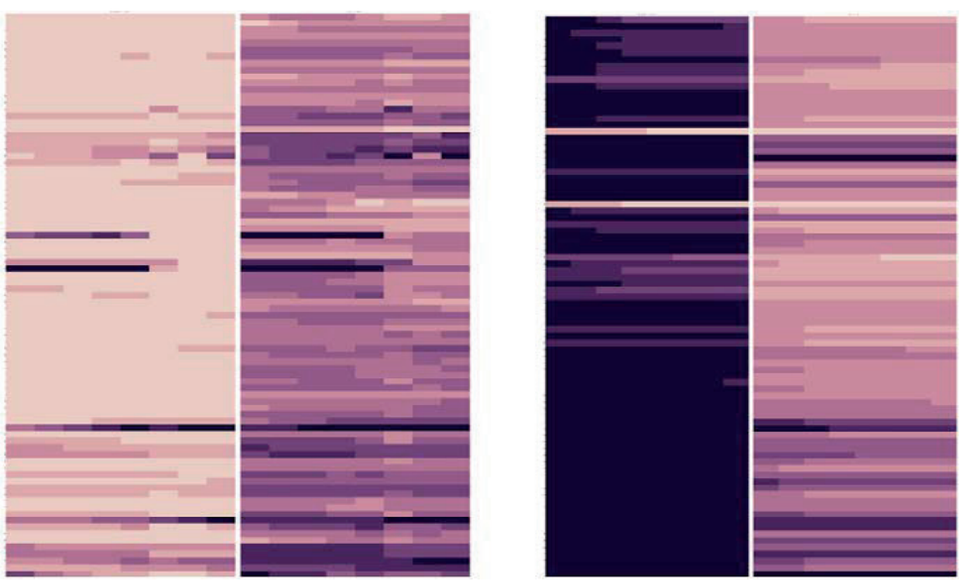
[21] Aiello F., Attanasio M. Some Issues in Constructing Composite Indicators [Электронный ресурс] // VIII International Meeting on Quantitative Methods for Applied Sciences. University of Siena, Certosa di Pontignano, 2006. URL: http://www3.unisi.it/eventi/dmq2006/paper/Aiello_Attanasio.pdf (дата обращения: 14.07.2021).

Результат применения логарифмической коррекции

Для рассматриваемого набора данных (табл. 1) с помощью авторской модификации *PCA* построения композитного индекса изменения качества системы по ряду наблюдений, учитывающей наличие шума исходных данных, были вычислены композитные показатели качества жизни населения России за 2010 – 2017 годы двумя способами: без логарифмический коррекции и с применением логарифмической коррекции к выявленным аномальным переменным:

- 1 «ВРП на душу населения»;
- 8 «Доля ветхого и аварийного жилья»;
- 9 «Плотность автомобильных дорог общего пользования»;
- 22 «Отношение ВРП к численности занятых в экономике».

Переменные преобразовывались по формуле (5). Предварительно явные выбросы из данных были удалены. Приведение переменных в единый диапазон учитывало характер их связи с показателем качества системы. Например, переменная 19 «Число инвалидов на 1 000 человек» имеет отрицательный характер связи с показателем качества. Следовательно, положительный весовой коэффициент этой переменной характеризует увеличение качества системы при уменьшении показателя. На рис. 4 приведены тепловые карты изменения переменных 8 и 9 после логарифмической коррекции. После коррекции распределение этих переменных нельзя назвать атипичным, данные имеют достаточную дифференциацию.



а) б)

Р и с. 4. Изменение переменных 8 (а) и 9 (б) за 2010-2017 годы до и после логарифмической коррекции

F i g. 4. Cyange of variables 8 (a) and 9 (б) during 2010-2017 before and after logarithmic correction

В табл. 5 и 6 приведены значения весовых коэффициентов до и после логарифмической коррекции. Отрицательные знаки весовых коэффициентов без логарифмической коррекции имеют три переменные: «ВРП на душу населения», «Отношение доходов 20% самых богатых и 20% самых бедных», «Число инвалидов на 1 тысячу человек.

Т а б л и ц а 5. **Весовые коэффициенты переменных Блока 1 «Уровень благосостояния населения» до и после логарифмической коррекции**

T a b l e 5. **Weight coefficients of the variables of Block 1 "The level of well-being of the population" before and after the logarithmic correction**

№ п/п	Переменные	Вес	
		без коррекции	после коррекции
1	Отношение ВРП на душу населения к прожиточному минимуму	**-0.41**	**0.73**
2	Отношение среднедушевых денежных доходов к прожиточному минимуму	0.89	0.53
3	Доля населения с доходами ниже прожиточного минимума	0.86	0.9
4	Отношение доходов 20% самых богатых и 20% самых бедных	**-0.3**	**0.19**
5	Обеспеченность собственными легковыми автомобилями на 1 тысячу человек .	0.61	**1.59**
6	Доля семей, состоящих на учете на получение жилья	0.74	**1.51**
7	Общая площадь жилищного фонда на одного жителя	1.1	0.79
8	Доля ветхого и аварийного жилья	1.09	0.88
9	Плотность автомобильных дорог общего пользования	**1.4**	0.43

Отрицательный весовой коэффициент последней переменной, который определяют данные без логарифмической коррекции, соответствует увеличению качества системы при увеличении числа инвалидов. Что противоречит интуитивному представлению о качестве жизни населения. Применение логарифмической коррекции для аномальных переменных решает проблему с отрицательностью весовых коэффициентов. Все веса переменных после проведения логарифмической коррекции неотрицательны (табл. 5 и 6).

Кроме того, отметим, что применение логарифмической коррекции иначе назначило весовые коэффициенты в блоке. Например, в первом блоке наибольший весовой коэффициент имела переменная «Плотность автомобильных дорог общего пользования», а после логарифмической коррекции самой весомой стала переменная «Обеспеченность собственными легковыми автомобилями на 1 тысячу человек», во втором блоке соответственно наиболее весома была переменная «Число умерших от болезней органов дыхания на 100 тысяч человек», а стала «Ожидаемая продолжительность жизни при рождении». Распределение весов после логарифмической коррекции представляется более корректным.

То же можно сказать о весах переменных второго блока. Если до логарифмической коррекции наиболее значимыми были переменные «Число умерших от болезней органов пищеварения на 100 тысяч человек» и «Доля специалистов с высшим образованием к занятым в экономике», а «Ожидаемая продолжительность жизни при рождении» была лишь четвертой по значимости, то после логарифмической коррекции всего лишь одной переменной интерпретация переменных значительно изменяется. Наиболее важной (и с большим отрывом) оказывается переменная «Ожидаемая продолжительность жизни при рождении», второй по значимости «Заболеваемость от травм и других внешних причин», и далее «Число умерших детей в возрасте до 1 года». А «Доля специалистов с высшим образованием к занятым в экономике» перестала играть существенную роль. При этом качество системы после логарифмической коррекции возрастает при уменьшении числа инвалидов, как и ожидалось априорно.

Таблица 6. **Весовые коэффициенты переменных Блока 2: «Качество населения» до и после логарифмической коррекции**
Table 6. **Weighting coefficients of the variables of Block 2: "Quality of the population" before and after the logarithmic correction**

№ п/п	Переменные	Вес	
		без коррекции	После коррекции
10	Ожидаемая продолжительность жизни при рождении	1.16	**2.11**
11	Число умерших детей в возрасте до 1 года на 1 тысячу человек.	0.71	1.03
12	Коэффициент естественного прироста 1 тысячу человек	0.34	0.44
13	Число умерших от инфекционных болезней и туберкулеза на 100 тысяч человек	0.33	0.97
14	Число умерших от новообразований на 100 тысяч человек	0.23	0.63
15	Число умерших от болезней системы кровообращения на 100 тысяч человек	0.27	0.45
16	Число умерших от болезней органов дыхания на 100 тысяч человек	1.63	0.37
17	Число умерших от болезней органов пищеварения на 100 тысяч человек	**1.69**	0.99
18	Заболеваемость от травм и других внешних причин на 100 тысяч человек	1.2	1.65
19	Число инвалидов на 1 тысячу человек	**-0.33**	**0.65**
20	Зарегистрировано врожденных аномалий на 1 тысячу человек.	0.33	0.33
21	Доля специалистов с высшим образованием к занятым в экономике	1.35	0.51
22	Отношение ВРП к численности занятых в экономике, тыс. руб./чел)	0.42	0.13
23	Количество выпускников высших и средних учебных заведений на 1 тысячу человек	0.15	0.74

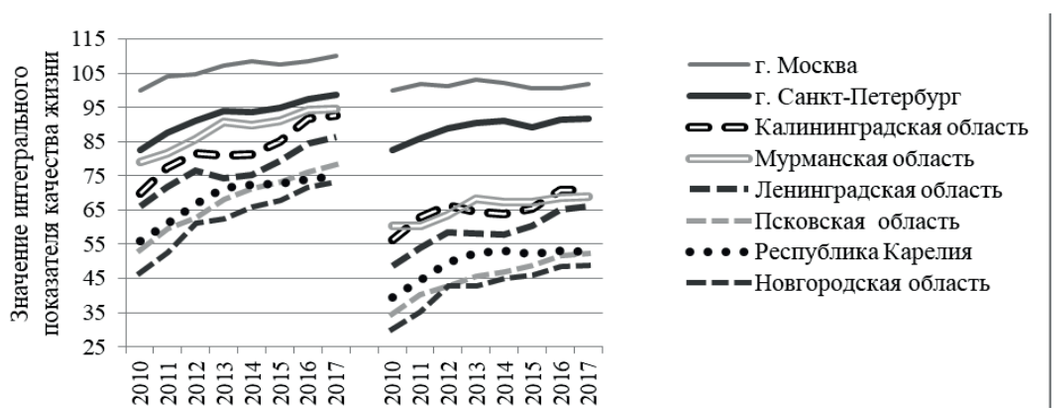


Рис. 5. Интегральные индикаторы качества жизни населения некоторых субъектов РФ до (слева) и после (справа) применения логарифмической коррекции
Fig. 5. Integral indicators of the quality of life of the population of some federal subjects of the Russian Federation before (left) and after (right) the application of a logarithmic correction

На рис. 5 представлены интегральные показатели качества жизни населения России за 2010 – 2017 годы для некоторых субъектов России. Значение интегрального показателя на первом году наблюдения приведены на отрезок [1, 100]. Минимальное значение показателя – один – в 2010 году у Тывы, максимальное значение 100 – у Москвы. Логарифмическая коррекция более наглядным сделала разделение кластеров – столицы и приморские регионы стали в большем отдалении от внутренних регионов. Также уменьшилось номинальные значения показателя, что тоже согласуется с интуитивным представлением о ситуации.

Заключение

Принципиальным отличием вычисления композитных индексов для социально-экономических систем является неопределенность качества используемых данных, в отличие от вычисления характеристик технических и физических систем, для которых заранее известна погрешность измерения и законы распределения данных. Особенности используемого пула статистических данных при автоматическом применении методов многомерного анализа часто могут приводить к недостоверным результатам. Получение разнообразных характеристик объектов на основании статистических данных должно обязательно учитывать особенности данных.

Метод главных компонент является объективным способом определения весовых коэффициентов композитного индекса, не использующим дополнительно субъективную априорную информацию для определения весовых коэффициентов. Однако метод критичен к качеству исходных данных. Одним из основных моментов в критике использования методов многомерного анализа для получения весов композитных индексов является, в частности, отсутствие социально-экономической интерпретации получаемых весовых коэффициентов в случае, если *PCA* определяет знаки весовых коэффициентов, противоположные ожидаемым.

Преобразование данных является одним из важнейших процессов при вычислении показателя качества систем. Для некоторой части статистических данных характерно распределение со значительной асимметрией. При вычислении интегрального показателя качества системы выбросы на верхнем уровне значений показателя нивелируют остальные значения, а малые значения игнорируются. Логарифмическая функция преобразования данных в таком случае позволяет избежать потерь значений, уменьшая (сжимая) большие значения в наборе данных и растягивая меньшие значения.

Визуализация данных с помощью тепловых карт служит для наглядного сравнения множества показателей при оценке изменения качества жизни и выявления особенностей данных, требующих нестандартных методов дальнейшей работы с данными и необходимых преобразований данных.

В работе обосновано применение логарифмического преобразования данных, аномальный характер которых выявляется с помощью тепловых карт и подтверждается применением аналитического критерия отношения сигнал / шум, оценивающего переменные в стандартном диапазоне. Предложенный вид логарифмического преобразования позволяет уйти от проблемы больших отрицательных значений, характерных для применения логарифмического преобразования на стандартном интервале [0, 1]. Применение логарифмической коррекции для аномальных переменных решает проблему с отрицательностью весовых коэффициентов и перераспределяет значения получаемых весовых коэффициентов в сторону более корректной социально-экономической интерпретации.

References

[1] Greco S., Ishizaka A., Tasiou M., Torrisi G. On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. *Social Indicators Research*. 2019; 141(1):61-94. (In Eng.) DOI: https://doi.org/10.1007/s11205-017-1832-9

[2] Diewert W.E. Exact and superlative index numbers. *Journal of Econometrics*. 1976; 4(2):115-145. (In Eng.) DOI: https://doi.org/10.1016/0304-4076(76)90009-9

[3] Silver M., Webb B. The measurement of inflation: Aggregation at the basic level. *Journal of Economic and Social Measurement.* 2002; 28(1-2):21-35. (In Eng.) DOI: https://doi.org/10.3233/JEM-2003-0185

[4] Hightower W.L. Development of an Index of Health Utilizing Factor Analysis. *Medical Care.* 1978; 16(3):245-255. (In Eng.) DOI: https://doi.org/10.1097/00005650-197803000-00006

[5] McKenzie D.J. Measuring Inequality with Asset Indicators. *Journal of Population Economics*. 2005; 18(2):229-260. (In Eng.) DOI: https://doi.org/10.1007/s00148-005-0224-7

[6] Vyas S., Kumaranayake L. Constructing Socio-Economic Status Indices: How to Use Principal Components Analysis. *Health Policy and Planning*. 2006; 21(6):459-468. (In Eng.) DOI: https://doi.org/10.1093/heapol/czl029

[7] Manly B.F.J., Navarro J.A.A. Multivariate Statistical Methods: A Primer. 4th Ed. Chapman and Hall/CRC; 2016. 269 p. (In Eng.)

[8] Somarriba N., Pena B. Synthetic Indicators of Quality of Life in Europe. *Social Indicators Research*. 2009; 94(1):115-133. (In Eng.) DOI: https://doi.org/10.1007/s11205-008-9356-y

[9] Filmer D., Pritchett L.H. Estimating Wealth Effects Without Expenditure Data – Or Tears: An Application to Educational Enrollments in States of India. *Demography*. 2001; 38(1):115-132. (In Eng.) DOI: https://doi.org/10.1353/dem.2001.0003

[10] Houweling T.A.J., Kunst A.E., Mackenbach J.P. Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter? *International Journal for Equity in Health.* 2003; 2(1):1-12. (In Eng.) DOI: https://doi.org/10.1186/1475-9276-2-8

[11] Krishnan V. Constructing a Multidimensional Socioeconomic Index and the Validation of It With Early Child Developmental Outcomes. In: Ed. by Management Association, Information Resources. *Early Childhood Development: Concepts, Methodologies, Tools, and Applications*. IGI Global, Hershey, PA; 2019. p. 130-165. (In Eng.) DOI: https://doi.org/10.4018/978-1-5225-7507-8.ch008

[12] Rencher A.C. Methods of Multivariate Analysis. 2nd Ed. A John Wiley & Sons, Inc. Publication; 2002. 738 p. (In Eng.)

[13] Jolliffe I.T. Principal Component Analysis. *Springer Series in*

Statistics. Second Edition. N.-Y.: Springer; 2002. 488 p. (In Eng.) DOI: https://doi.org/10.1007/b98835

[14] Tabachnick B.G., Fidell L.S. Using Multivariate Statistics. 5 Ed. Boston: Pearson Education; 2007. 980 p. (In Eng.)

[15] Fukuda, Y., Nakamura, K., & Takano, T Higher mortality in areas of lower socioeconomic position measured by a single index of deprivation in Japan. *Public Health*. 2007; 121(3):163-173. (In Eng.) DOI: https://doi.org/10.1016/j.puhe.2006.10.015

[16] Maity S., Kachari S. Socioeconomic status and the factors influencing the Socio-economic status of Bodo tribes:A case study of Udalguri District, Assam. *Socioeconomica*. 2015; 4(8):371-394. (In Eng.) DOI: https://dx.doi.org/10.12803/SJSECO.48132

[17] Molchanova E.V., Kruchek M.M., Kibisova Z.S. Building of the rating assessments of the Russian Federation subjects by the blocks of socio-economic indicators. *Economic and Social Changes: Facts, Trends, Forecast*. 2014; (3):196-208. (In Eng.) DOI: https://doi.org/10.15838/esc/2014.3.33.15

[18] Zhgun T.V. 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*. 2017; 10(2):214-235. (In Eng.) DOI: https://doi.org/10.15838/esc.2017.2.50.12

[19] Zhgun T.V. Complex index of a system's quality for a set of observations. *Journal of Physics: Conference Series*. 2019; 1352(1):012064. (In Eng.) DOI: https://doi.org/10.1088/1742-6596/1352/1/012064

[20] Wilkinson L., Friendly M. The History of the Cluster Heat Map. *The American Statistician*. 2009; 63(2):179-184. (In Eng.) DOI: https://doi.org/10.1198/tas.2009.0033

[21] Li X., Liang S., Zhang J. Acceleration of OCT Signal Processing with Lookup Table Method for Logarithmic Transformation. *Applied Sciences*. 2019; 9(7):1278-1286. (In Eng.) DOI: https://doi.org/10.3390/app9071278

[22] Mundo G., Nardo M. Noncompensatory/nonlinear composite indicators for ranking countries: a defensible setting. *Applied Economics*. 2009; 41(12):1513-1523. (In Eng.) DOI: https://doi.org/10.1080/00036840601019364

[23] Zhgun T.V. Investigation of data quality in the problem of calculating the composite index of a system from a series of observations. *Journal of Physics: Conference Series*. 2020; 1658:012082. (In Eng.) DOI: https://doi.org/10.1088/1742-6596/1658/1/012082

[24] Zhao H., Lu L., He Z., Chen B. Adaptive recursive algorithm with logarithmic transformation for nonlinear system identification in α-stable noise. *Digital Signal Processing*. 2015; 46:120-132. (In Eng.) DOI: https://doi.org/10.1016/j.dsp.2015.08.004

[25] Klyatskin V.I. Integral characteristics: a key to understanding structure formation in stochastic dynamic systems. *Physics-Uspekhi*. 2011; 54(5):441-464. (In Eng.) DOI: https://doi.org/10.3367/UFNe.0181.201105a.0457

Поступила 14.07.2021; одобрена после рецензирования 21.08.2021; принята к публикации 10.09.2021.
Submitted 14.07.2021; approved after reviewing 21.08.2021; accepted for publication 10.09.2021.

Об авторе:

Жгун Татьяна Валентиновна, доцент кафедры прикладной математики и информатики, Институт электронных и информационных систем, ФГБОУ ВО «Новгородский государственный университет имени Ярослава Мудрого» (173003, Российская Федерация, Новгородская область, г. Великий Новгород, ул. Большая Санкт-Петербургская, д. 41), кандидат физико-математических наук, доцент, **ORCID: https://orcid.org/0000-0002-7518-6925,** Tatyana.Zhgun@novsu.ru

Автор прочитал и одобрил окончательный вариант рукописи.

About the author:

Tatyana V. Zhgun, Associate Professor of the Department of Applied Mathematics and Computer Science, Institute of Electronic and Information Systems, Yaroslav-the-Wise Novgorod State University (41 Bolshaya St. Petersburgskaya Str., Veliky Novgorod 173003, Russian Federation), Ph.D. (Phys.-Math.), Associate Professor, **ORCID: https://orcid.org/0000-0002-7518-6925,** Tatyana.Zhgun@novsu.ru

The author has read and approved the final manuscript.