УДК 519.25, 330.46
DOI: 10.25559/SITITO.16.202002.295-303

Оценка качества статистических данных в задаче вычисления интегральной характеристики системы по ряду наблюдений

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

Аннотация

Построение интегральной характеристики системы по ряду наблюдений, фиксирующих состояние системы с некоторыми ошибками, можно рассматривать как задачу выделения полезного сигнала на фоне шумов. Сигналом в этом случае являются весовые коэффициенты линейной свертки показателей, которые должны отражать постоянную структуру оцениваемой системы в рассматриваемом периоде. Для ряда наблюдений методы многомерного анализа по-разному определяют структуру главных компонент (главных факторов). Даже небольшое возмущение исходных данных может вызвать значительное изменение весовых коэффициентов при использовании методов многомерного анализа. Причиной этого может являться наличие неустранимых погрешностей используемых данных. Поэтому существенным является вопрос об оценке качества используемых данных в задаче вычисления композитного индикатора качества системы по ряду наблюдений, а именно, оценка наличия в регистрируемых измерениях значительных неустранимых ошибок. В статье рассматривается применение метода конечных разностей для оценки неустранимых ошибок регистрации данных по ряду наблюдений. Для применения этой методики необходимо, чтобы данные допускали приближение полиномами более низких степеней, чем число наблюдений минус один. Справедливость предположения проверена эмпирически на конкретном наборе данных. Рассматриваются 37 переменных, характеризующих качество жизни населения России за 2010-2017 годы. Проанализированы зависимости качества аппроксимации данных от степени полиномиальной регрессии. Результаты численного эксперимента позволяют сделать вывод о правомерности оценки неустранимых погрешностей данных с помощью метода конечных разностей. Применение аппарата конечных разностей для анализа выборки показывает наличие неустранимых ошибок в исследуемой выборке от 0,59% до 28,92 %. Следовательно, получение композитных характеристик объектов на основании таких данных должно обязательно учитывать наличие неустранимой ошибки. В частности, число параметров и число наблюдений, характеризующих систему, должно быть достаточно большим, чтобы компенсировать случайные ошибки.

Ключевые слова: интегральные индикаторы, качество данных, ошибки данных, метод главных компонент, метод конечных разностей.

Для цитирования: Жгун, Т. В. Оценка качества статистических данных в задаче вычисления интегральной характеристики системы по ряду наблюдений / Т. В. Жгун. – DOI 10.25559/SITITO.16.202002.295-303 // Современные информационные технологии и ИТ-образование. – 2020. – Т. 16, № 2. – С. 295-303.

Evaluation of Statistical Data Quality in the Problem of Calculating the Integral Characteristic of a System for a Number of Observations

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

Abstract

The construction of the composite index of a system can be considered as a problem of separating signal from noise. The signal in this case is the weight coefficients of the linear convolution of indicators. The weights to be determined should reflect the structure of the system being evaluated. However, principal component analysis and factor analysis determine the structure of principal components and principal factors differently for different observations. The reason for this may be the presence of inevitable errors in the used data. A solution of the problem requires a detailed understanding of input data errors' influence on the calculated model's parameters. The article discusses the use of the finite difference method for evaluating statistical data quality in the problem of calculating the integral characteristic of a system for a number of observations. For this technique to be applicable, the data must be approximated with polynomials of lower degrees than the number of observations minus one. The assumption is tested empirically on a specific data set. 37 variables characterizing the quality of life of the population of Russia for 2010-2017 are considered. The dependencies of the quality of data approximation on the degree of polynomial regression are analyzed. The results of the numerical experiment make it possible to draw a conclusion about the legitimacy of evaluating data errors using the finite difference method. The use of the finite difference apparatus for analyzing the data shows the presence of fatal errors from 0.59% to 28.92%. Therefore, obtaining the composite characteristics of objects on the basis of such data must necessarily take into account the presence of a fatal error. In particular, the number of parameters characterizing the system should be large enough to compensate for random errors with averaging.

Keywords: composite index; data quality; data errors; principal component analysis, method of finite differences.

For citation: Zhgun T.V. Evaluation of Statistical Data Quality in the Problem of Calculating the Integral Characteristic of a System for a Number of Observations. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2020; 16(2):295-303. DOI: https://doi.org/10.25559/SITITO.16.202002.295-303

Введение

Задача определение комплексного показателя сложных систем возникают как при исследовании физических явлений, технических систем, так при решении проблем управления социально-экономическими системами. При описании стохастических динамических систем в задачах гидродинамики, магнитной гидродинамики, астрофизике, физике плазмы, в радиофизике интегральные величины являются основными характеристиками таких систем. Например, все законы сохранения в механике и электродинамике сплошных сред записываются для интегральных величин. Интегральные характеристики описывают динамику системы в целом, позволяя отвлечься от побочных эффектов, связанных со случайностью показателей, искаженных помехами и являются ключом к пониманию структурообразования в стохастических динамических системах [1].

Построение интегрального индикатора вводит отношения порядка на многомерном множестве объектов и позволяет сравнивать качество объектов. Общей целью большинства композитных показателей является ранжирование объектов (стран) и их сравнительный анализ [2-6]. Огромное количество методов, используемых для оценки качества слабо структурированных систем [2] свидетельствует о неудовлетворенности результатами и необходимости дальнейших исследований в этой области[1] [4, 7].

Построение композитных индексов без использования априорной информации для получения весовых коэффициентов позволяет получить объективные показатели исследуемой системы. Однако применение формальных методик определения весовых коэффициентов в случае наличие ошибок (искажений) используемых данных будет приводить к искажению и показателей, и весовых коэффициентов свертки показателей, определяющих интегральную характеристику. Следовательно, одной из причин недостаточного качества композитных индексов может являться неудовлетворительное качество используемых данных.

Принципиальным отличием вычисления композитных индексов для слабо структурированных систем является неопределенность качества используемых данных, в отличие от вычисления характеристик, например, технических систем, для которых заранее известна погрешность измерения. Наличие большого количества ошибок в статистических данных при вычислении композитных индексов отмечают авторы [7-9]. Тем не менее, именно статистические данные, содержащие неустранимые погрешности, в настоящее время представляют собой наилучшие оценки имеющихся реальных величин в социальных системах [7].

В статье рассматривается применение метода конечных разностей для оценки качества данных в задаче вычисления композитного индикатора качества системы по ряду наблюдений. Для применения этой методики необходимо, чтобы данные допускали приближение полиномами более низких степеней, чем число наблюдений минус один. Справедливость предположения проверена эмпирически на конкретном наборе данных. Проанализированы зависимости качества аппроксимации данных от степени полиномиальной регрессии. Результаты численного эксперимента позволяют сделать вывод о правомерности оценки погрешностей данных с помощью метода конечных разностей.

1 Постановка задачи

Рассмотрим построение интегральной оценки системы из m объектов, для которой известны таблицы описаний объектов для ряда наблюдений – матрицы размерностью $m \times n$ $A^t = \left\{ a_{ij}{}^t \right\}_{i,j=1}^{n,m}$, $t = 1,..., T$. Элемент матрицы $a_{ij}{}^t$ — значение j-го показателя i-го объекта в момент t. Для каждого момента t вектор интегральных показателей имеет вид

$$q^t = A^t \cdot w^t, \quad (1)$$

Или же, для i–го объекта в момент t

$$q^t{}_i = \sum_{j=1}^{n} w^t{}_j \cdot a^t{}_{ij} \quad (2)$$

где $q^t = \left\langle q^t{}_1, q^t{}_2, \ldots, q^t{}_m \right\rangle^T$ – вектор интегральных индикаторов момента t, $w^t = \left\langle w^t{}_1, w^t{}_2, \ldots, w^t{}_m \right\rangle^T$ – вектор весов показателей для момента t, A^t – матрица предварительно обработанных данных для момента t. Численные характеристики системы предварительно подвергнуты унификации – приведению значений переменных на отрезок [0, 1] по принципу «чем больше, тем лучше».

Для построения интегрального индикатора качества системы требуется найти веса показателей w^t для каждого момента времени, адекватно отражающие свойства рассматриваемой системы. Т.е. определяемые весовые коэффициенты должны отражать **структуру** оцениваемой системы. Такая трактовка весовых показателей устраняет одну из главных неопределенностей при конструировании интегрального индикатора. Если определяется композитные индексы качества системы для ряда последовательных наблюдений, то имеем дело с изменением данных. Это изменение данных во времени вызывается как изменением ситуации, так и случайными ошибками регистрации данных [10].

Одним из наиболее простых методов анализа структуры исследуемой системы является метод главных компонент (МГК). Пространство главных компонент оптимально для моделирования внутренней структуры данных. Методика многомерного анализа, прекрасно работающая для оценки технических систем, часто дает недостоверный результат при построении композитных индексов слабоструктурированных систем. В частности, вычисляемые композитные индексы чрезвычайно неустойчивы [11-14]. Причиной этого может быть наличие ошибок данных. Наличие ряда зафиксированных измерений $A^t = \left\{ a_{ij}{}^t \right\}$, $t = 1,..., T$ позволяет оценить эту ошибку.

Оценка неустранимых ошибок данных

Качество данных — обобщенное понятие, отражающее степень их пригодности к решению определенной задачи [15, 16]. В соответствии со стандартом ISO 9000:2015 основными критериями качества являются полнота, достоверность, точность, согласованность, доступность и своевременность [17]. Аномальные значения и шумы называют в качестве основных проблем, вызывающих снижение качества данных. Эти недо-

[1] Sharpe, A. Literature Review of Frameworks for Macro-indicators. Ottawa, Canada: Centre for the Study of Living Standards; 2004. [Электронный ресурс]. URL: https://core.ac.uk/download/pdf/7033316.pdf (дата обращения: 21.06.2020).

статки не нарушают работу алгоритмов обработки данных, но порождают некорректные результаты анализа.

В мировой статистической практике нет общепринятого определения качества данных как результата статистической деятельности. Однако общепринятыми компонентами современного понятия качества статистических являются понятия точности и достоверности данных. В практических терминах не существует единой и полной меры достоверности и точности результатов статистических наблюдений, поэтому используются несколько форм ее выражения. Исходя из практической потребности, степень точности величины обычно характеризуется ее дисперсией, стандартной ошибкой, коэффициентом вариации. Но эти меры точности плохо характеризуют достоверность и наличие возможных ошибок регистрации. Такие ошибки можно оценить с применением аппарата конечных разностей.

Пусть y_i — точное (неизвестное) значение измеряемой величины, определяемой для ряда наблюдений $i = 1,..., k$; y_i^* – измеренное значение, содержащее ошибку, $\varepsilon_i = y_i^* - y_i$ — ошибка измерений. Ошибка статистических данных имеет случайный характер. Величина — неизвестна и не может быть вычислена по фиксируемым наблюдениям. Однако максимальную из ошибок можно оценить.

Обозначим максимальную величину ошибки для всех наблюдений $\varepsilon = \max_i |\varepsilon_i|$. Тогда измеренная величина y_i^* лежит в диапазоне $y_i - \varepsilon \le y_i^* \le y_i + \varepsilon$. Рассмотрим первые конечные разности приближенных величин $\Delta_i^* = \Delta(y_i^*)$

$$\Delta_i^* = y_{i+1}^* - y_i^* = (y_{i+1} + \varepsilon_{i+1}) - (y_i + \varepsilon_i) = (y_{i+1} - y_i) - (\varepsilon_{i+1} - \varepsilon_i) = \Delta_i + (\varepsilon_{i+1} - \varepsilon_i)$$

Учитывая, что $|\varepsilon_{i+1} - \varepsilon_i| \le |\varepsilon_{i+1}| + |\varepsilon_i| \le 2 \cdot \varepsilon$, модуль приближенной конечной разности $|\Delta_i^*| \le |\Delta_i| + 2 \cdot \varepsilon$, где $\Delta_i = y_{i+1} - y_i$ – первые конечные разности неизвестных точных величин. Далее, вторая конечная разность приближенный значений $\Delta_i^{*2} = \Delta_{i+1}^* - \Delta_i^*$ оценивается аналогично $|\Delta_i^{*2}| \le |\Delta_i^2| + 4 \cdot \varepsilon$. *Для* последней вычисленной *k*-ой приближенной конечной разности справедлива оценка

$$|\Delta_i^{*k}| \le |\Delta_i^k| + 2^k \cdot \varepsilon \quad (3)$$

Известно, что для гладких функций величина конечной разности стремится к нулю при увеличении порядка разности. В частности, для полинома степени *k* $P_k(x) = a_k \cdot x^k + a_{k-1} \cdot x^{k-1} + ... + a_0$ выполняются следующие соотношения:

$$\Delta^{k+1}(P_k(x)) = 0 \text{ и } \Delta^k(P_k(x)) \cong a_k \cdot h^k,$$

где *h* — шаг таблицы. Если значения измеряемой функции от измерения к измерению меняются не слишком быстро (функция непрерывна и производные старших порядков ограничены), функцию можно аппроксимировать полиномом невысокой степени и значения точных конечных разностей $\varnothing_i^k$ с увеличением порядка стремятся к нулю. Значит, вычисленные значения приближенных конечных разностей обеспечивают оценку исходной погрешности:

$$|\Delta_i^{*k}| \le 2^k \cdot \varepsilon \quad (4)$$

Обозначим $\varepsilon^* = |\Delta_i^{*k}| / 2^k$ (5)

Таблица 1. **Примеры вычисления приближенных конечных разностей**
Table 1. **Examples of calculating approximate finite differences**

Точное значение функции								
x	*f (x)*	Δ^1	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6	Δ^7
0	1,386	0,025	–0,001	0	0	0	0	***0***
0,1	1,411	0,024	–0,001	0	0	0	0	
0,2	1,435	0,024	–0,001	0	0	0		
0,3	1,459	0,023	–0,001	0	0			
0,4	1,482	0,022	0	0				
0,5	1,504	0,022	0					
0,6	1,526	0,022						
0,7	1,548							
Значения функции округлены до сотых								
x	*f*(x)*	Δ^{*1}	Δ^{*2}	Δ^{*3}	Δ^{*4}	Δ^{*5}	Δ^{*6}	Δ^{*7}
0	1,380	0,03	–0,01	0,01	0	–0,03	0,1	***–0,25***
0,1	1,410	0,02	0	0,01	–0,03	0,07	–0,15	
0,2	1,430	0,02	0,01	–0,02	0,04	–0,08		
0,3	1,450	0,03	–0,01	0,02	–0,04			
0,4	1,480	0,02	0,01	–0,02				
0,5	1,500	0,03	–0,01					
0,6	1,530	0,02						
0,7	1,550							
Единичный выброс								
x	*f*(x)*	Δ^{*1}	Δ^{*2}	Δ^{*3}	Δ^{*4}	Δ^{*5}	Δ^{*6}	Δ^{*7}
0	1,386	0,025	–0,001	1	–4	10	–20	***35***
0,1	1,411	0,024	0,999	–3	6	–10	15	
0,2	1,435	1,024	–2,001	3	–4	5		
0,3	***2,459***	–0,977	0,999	–1	1			
0,4	1,482	0,022	0	0				
0,5	1,504	0,022	0					
0,6	1,526	0,022						
0,7	1,548							

Учитывая, что величина ε оценивала погрешности значений функции сверху, $|\varepsilon_i| \le \varepsilon$, а согласно неравенству (4), $\varepsilon \ge |\Delta_i^{*k}| / 2^k$, то возможны два варианта для оценки ε: $|\varepsilon_i| \le \varepsilon^* \le \varepsilon$ или $|\varepsilon_i| \le \varepsilon \le \varepsilon^*$. Реальное соотношение между значениями $|\varepsilon_i|$, ε, ε^* можно получить из численного эксперимента. Для этого рассмотрим модельный пример.

В таблице 1 приведен результат вычисления конечных разностей высокого порядка для точного задания функции *f (x) = ln(x+4)* и в случае внесения погрешностей в ее значения. При отсутствии погрешностей уже третьи конечные разности точно заданной функции обращаются в ноль (выделено в таблице 1). Если в каждое значение функции внесена погрешность порядка ε = 0.01, то конечная разность седьмого порядка для приближенных значений функции Δ^{*7} = 0.25 (выделено в таблице) и оценка погрешности составляет ε^* = 0.002 при реальной погрешности ε = 0.1. Итак, вычисленная оценка ε^* оказалась меньше реальной ошибки ε, т.е. выполняется следующее неравенство для оценки погрешности

$$\varepsilon^* = |\Delta_i^{*k}| / 2^k \le \varepsilon \qquad (6)$$

Если в одно значение функции внесена погрешность ε = 1 и точное значение функции *f(0,3)* = 1.459 заменено на приближенное *f*(0,3) = 2.459 (выделено в таблице 1), то седьмая конечная разность для приближенных значений функции в этом случае составит Δ^{*7} = 35, а оценка погрешности по формуле (3) составит ε^* = 35/ 2^7 = 0.276. Снова вычисленная оценка ε^* оказалась меньше реальной ошибки ε. Значит, вычисленное значение ε^* является оценкой снизу возможной ошибки и может являться характеристикой исследуемой выборки. Можно сказать, что эта величина является в некотором смысле мерой случайности данных в выборке. Если значения измеряемых величин предварительно приведены на отрезок [0, 100], то вычисляемая по формуле (5) величина ε^* будет характеризовать относительную неустранимую погрешность выборки. Реальная ошибка может превышать это значение.

Влияние гладкости функции на оценку погрешности

Пусть переменная x_{ij} представлена наблюдениями $x^*_{ij}(1), x^*_{ij}(2), ..., x^*_{ij}(T)$, которые реализуют неизвестную зависимость функционирования рассматриваемой системы с некоторыми погрешностями: $x^*_{ij}(t) = x_{ij}(t) + \varepsilon_{ij}(t)$. Функция $x_{ij}(t)$ на промежутке $t \in [1, T]$ может быть аппроксимирована полиномом степени n $x_{ij}(t) \approx P_n(t) = a_n \cdot t^n + a_{n-1} \cdot t^{n-1} + ... + a_0$. Тогда измеряемая величина представлена своим приближенным значением, содержащим ошибку $x^*_{ij}(t) \approx P_n(t) + \varepsilon_{ij}(t)$. По значениям $x^*_{ij}(1), x^*_{ij}(2), ..., x^*_{ij}(T)$ могут быть вычислены приближенные конечные разности до порядка T-1 включительно $(\Delta^*_{ij})^{T-1} = \Delta^{T-1}(x^*_{ij}(1)) \approx \Delta^{T-1}(P_n(1) + \varepsilon_{ij}(t))$.

Если степень аппроксимирующей функции *n* менее *T-1 n < T-1,* то последняя точная конечная разность обращается в ноль $\Delta^{T-1}_{ij}(P_n(1)) = 0$. Тогда вычисленная приближенная конечная разность $(\Delta^*_{ij})^{T-1} = \Delta^{T-1}_{ij}(P_n(1)) + \varepsilon_{ij}(t))^{\varepsilon}$ и справедлива оценка согласно (4, 5)

$$|(\Delta^*_{ij})^{T-1}| \le \Delta^{T-1}_{ij}(P_n(1)) + 2^{T-1} \cdot \varepsilon_{ij} = 2^{T-1} \cdot \varepsilon_{ij},$$
$$\varepsilon_{ij} = \max_t |\varepsilon_{ij}(t)| \cdot$$

Итак, вычисленная оценка погрешности данных для параметра x_{ij} на промежутке наблюдения $t \in [1, T]$ определяется соотношением

$$\varepsilon^*_{ij} = |(\Delta^*_{ij})^{T-1}| / 2^{T-1}, \text{ причем } \varepsilon^*_{ij} \le \varepsilon_{ij} \text{ где}$$
$$\varepsilon_{ij} = \max_t |\varepsilon_{ij}(t)| \qquad (7)$$

Вычисленное значение ε^*_{ij} является оценкой снизу возможной ошибки регистрации *j*-го параметра для *i*-го объекта и может являться характеристикой качества исследуемой выборки. Если значения исследуемых величин $x^*_{ij}(1), x^*_{ij}(2), ..., x^*_{ij}(T)$ предварительно приведены на отрезок [0, 100], то ε^*_{ij} будет характеризовать относительную погрешность выборки.

Итак, если функция допускает приближение полиномами более низких степеней, чем число наблюдений минус один, а значение последней приближенной конечной разности отлично от нуля $(\Delta^*_{ij})^{T-1} \ne 0$, то это значение приближенной конечной разности определяется искажениями значений переменной, вносимых при регистрации наблюдений, которые могут быть оценены по (7).

Очевидно, что измеряемых данных функциональная зависимость может быть однозначно восстановлена по имеющимся *T* значениям интерполяционным полиномом степени *T*-1: $P_{T-1}(x) = a_{T-1} \cdot x^{T-1} + a_{T-2} \cdot x^{T-2} + ... + a_0$. При этом значение последней точной конечной разности определяется старшим коэффициентом интерполяционного полинома $\Delta^{T-1} \approx \Delta^{T-1}(P_{T-1}(x)) = a_{T-1}$.

Если изменяемая функция допускает приближение полиномами более низких степеней, тогда значение последней точной конечной разности должно быть нулевым $\Delta^{T-1}(P_{T-1}(x)) = 0$. Если функция допускает приближение полиномами более низких степеней, а значение последней приближенной конечной разности не обращается в ноль: $(\Delta^*)^{T-1} \ne 0$, то это значение определяется искажениями значений переменной, вносимых при регистрации наблюдений. Предположение, что данные допускают приближение полиномами более низких степеней, чем число наблюдений минус один, может быть проверено эмпирически на конкретном наборе данных.

Применение регрессионного анализа для описания экспериментальных данных

Подбор математических моделей, наилучшим образом описывающих экспериментальные данные является задачей регрессионного анализа. В классическом регрессионном анализе [17-22] модель представляется в следующем виде

$$y = G(x, \beta) + \varepsilon \qquad (8)$$

где $x = (x_1, x_2, ..., x_m)$ – вектор входных (независимых) переменных; y – случайная выходная (зависимая) переменная; $\beta = (\beta_1, \beta_2, ..., \beta_m)$ – неизвестный вектор коэффициентов модели; $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_m)$ – случайная величина (случайное возмущение, ошибка, шум), учитывающая влияние случайных факторов. Предполагается что случайные возмущения распределены нормально с параметрами: $M(\varepsilon) = 0$, $\sigma_i^2 = \sigma^2 = const$. Очевидно, что в силу действия случайных факторов модель (8) не может точно прогнозировать значение выходной переменной при заданных значениях входных переменных. Следовательно, нет оснований говорить об «истинной» модели в полном смысле этого слова. Обычно под «истинным» значением выходной переменной понимают его условное математиче-

ское ожидание при заданных значениях входных переменных:

$$M\left[\, y \,|\, x \,\right] = M\left[\, (G(x,\beta)+\varepsilon) \,|\, x \,\right] = M\left[\, G(x,\beta) \,\right] + M\left[\varepsilon\right] = G(x,\beta) \quad (9)$$

Соотношение (9) представляет собой теоретическую регрессионную модель — уравнение регрессии *y* относительно *x*. Основной задачей регрессионного анализа является выявление и приближенное математическое описание причинно-следственной зависимости $G(x,\beta)$ между выходной и входными переменными. С этой целью на основании статистических данных (выборки $\{x_i, y_i\}$, $i = 1,...T$) методом наименьших квадратов строится статистический аналог соотношения (9) – эмпирическая регрессионная модель

$$y' = G'(x,b) \quad (10)$$

качество которой, т. е. степень ее соответствия статистическим данным оценивается с помощью выборочного коэффициента детерминации

$$R^2 = \sum_{i=1}\left(y_i' - \bar{y}\right)^2 / \sum_{i=1}\left(y_i - \bar{y}\right)^2,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (11)$$

Коэффициент R^2 показывает, какая часть вариации *y* обусловлена регрессионной моделью. Из соотношения (11) следует, что максимально возможное значение R^2 ограничено погрешностью измерений, наличием неучтенных факторов и не зависит от состава входных переменных и структуры модели. Например, если 10% вариации выходной переменной обусловлено погрешностью измерения переменных, то максимально возможное значение R^2 не может превышать 0,9 при любой структуре модели. При наличии статистического шума большое значение R^2 обеспечивается за счет излишнего усложнения модели, которая вместо существующей закономерности описывает случайные ошибки.

Применение полиномиальной регрессии для описания экспериментальных данных

Чтобы убедиться в справедливости оценки искажениями значений переменной, вносимых при регистрации наблюдений, нужно убедиться, что регистрируемые значения допускают приближение полиномами более низких степеней, чем число наблюдений минус один. Тогда вычисленное значение последней приближенной конечной разности $(\Delta_{ij}^*)^{T-1} \neq 0$, определит оценку искажений переменных, вносимых при регистрации наблюдений по формуле (7).

Рассмотрим в таблице 3 набор переменных, характеризующий качество жизни населения [18] за 2010-2017 годы. Эти переменные использовались для оценки качества жизни населения России во многих исследованиях, например в [20]. Значения переменных для корректности сравнения приведены на отрезок [0, 100]. Полученная оценка в этом случае есть относительное значение ошибки в процентах. Всего рассматриваем восемь наблюдений. Нужно проверить, аппроксимируются ли на этом интервале данные наблюдений полиномом степени не более шестой.

Чтобы убедиться в справедливости предположения о возможности аппроксимации рассматриваемых данных полиномом степени не более шестой, достаточно численно оценить качество полиномиальной регрессии для субъектов, обеспечивающих максимальное значение приближенной седьмой конечной разности для каждого блока. Продемонстрируем это для первого блока, где максимальную оценку имеет переменная 6. Для этой переменной проверим, аппроксимируются ли данные наблюдений полиномом степени не более шестой.

Вычислительный эксперимент показывает, что качество аппроксимации увеличивается с ростом степени полиномиальной регрессии (Таблица 2): с ростом степени аппроксимирующего полинома количество объектов с относительно невысоким коэффициентом детерминации уменьшается. Для полиномиальной регрессии шестой степени для 83 объектов из 85 качество приближения можно признать хорошим, для двух оставшихся – удовлетворительным. Проанализированные зависимости для остальных переменных подтверждают вывод о правомерности наблюдаемых оценок погрешностей данных, представленных в таблице 3.

Таблица 2. **Зависимость качества аппроксимации от степени полиномиальной регрессии для переменной 6**

Table 2. **Dependence of the quality of approximation on the degree of polynomial regression for variable 6**

Степень полиномиальной регрессии	1	2	3	4	5	6
Количество субъектов, для которых $R^2 \geq 0{,}7$	44	61	72	79	82	83

Вычислим для каждого показателя значения приближенных седьмых конечных разностей для всех объектов согласно (5) . Далее для каждого показателя определим максимальное наблюдаемое значение искажения переменной как максимальное значение полученных величин по всем субъектам: $\varepsilon_j^* = \max_i \left| \varepsilon_{ij}^* \right|$. Вычисленная величина является оценкой наблюдаемых искажений значений *j* –ой переменной. Полученные значения для каждого показателя вместе со списком переменных приведены в таблице 2. Минимальная наблюдаемая ошибка наблюдаемых переменных составляет 0,59%, а максимальная 28, 92 %.

Таблица 3. **Наблюдаемая оценка погрешности данных по выборке 2010-2017 гг.**

Table 3. **Observed estimate of data error for the 2010-2017 sample**

№ п/п	Переменные	*Eps*
	Блок 1: Уровень благосостояния населения	
1	ВРП на душу населения с учетом инфляции, тысяч рублей	1,76
2	Отношение среднедушевых денежных доходов к прожиточному минимуму	3,7
3	Доля населения с доходами ниже прожиточного минимума	2,76
4	Отношение доходов 20% самых богатых и 20% самых бедных	2,45
5	Обеспеченность собственными легковыми автомобилями на 1000 человек	2,5

№ п/п	Переменные	*Eps*
6	Доля семей, состоящих на учете на получение жилья	***14,84***
7	Общая площадь жилищного фонда на одного жителя	3,71
8	Доля ветхого и аварийного жилья	8,04
9	Плотность автомобильных дорог общего пользования	1,4
	Блок 2: Качество населения	
10	Ожидаемая продолжительность жизни при рождении	4,43
11	Число умерших детей в возрасте до 1 года на 1000 родившихся.	5,78
12	Коэффициент естественного прироста на 1000 человек	1,84
13	Умерших от инфекционных болезней и туберкулеза на 100 тысяч человек	0,78
14	Число умерших от новообразований на 100 тысяч человек	0,59
15	Умерших от болезней системы кровообращения на 100 тысяч человек.	0,68
16	Число умерших от болезней органов дыхания на 100 тысяч человек	1,72
17	Число умерших от болезней органов пищеварения на 100 тысяч человек	1,33
18	Заболеваемость от травм и других внешних причин на 100 тысяч человек	6,38
19	Число инвалидов на 1000 человек	0,93
20	Зарегистрировано врожденных аномалий на 1 000 человек	***8,5***
21	Доля специалистов с высшим образованием к занятым в экономике	2,4
22	Отношение ВРП к численности занятых в экономике, тыс. руб./чел.	1,69
23	Численность студентов высших и средних учебных заведений на 1000 человек	1,67
	Блок 3: Качество социальной сферы	0
24	Уровень безработицы, %	3,64
25	Численность смертей при несчастных случаях на производстве на 1000 работающих	21,86
26	Численность пострадавших при несчастных случаях на производстве на 1000 человек работающих	6,18
27	Коэффициент миграционного прироста на 10 тысяч человек	11,13
28	Число зарегистрированных умышленных убийств на 100 тысяч человек	7,75
29	Число фактов умышленного причинения тяжкого вреда здоровью на 100 тысяч человек	4,45
30	Число зарегистрированных изнасилований на 100 тысяч человек.	***28,92***
31	Число разбоев, грабежей, краж на 100 тысяч человек	3,55
32	Зарегистрированных присвоений или растрат на 100 тысяч человек	4,2
33	Состоящих на учете: наркомания и токсикомания на 100 тысяч человек.	4,14
34	Состоящих на учете: алкоголизм в расчете на 100 тысяч человек	1,93
35	Больных туберкулезом в расчете на 100 тысяч человек.	1,79
36	Число больных с диагнозом сифилиса на 100 тысяч человек	10,93
37	Число больных психическими расстройствами на 100 тысяч человек.	1,63

Ожидаемым является низкое качество данных, характеризующих миграционный прирост (11,13%). Но статистика регистрации семей, состоящих на учете для получения жилья, содержит еще больше ошибок (14,84%). Много ошибок при регистрации ветхого и аварийного жилья (8,04%), врожденных аномалий (8,50%), смертей при несчастных случаях на производстве (21,86%), числа изнасилований (28,92%) и заболеваемости сифилисом (10,93%. Можно сказать, что максимальные ошибки регистрации данных демонстрируют наиболее социально весомые индикаторы. Более того, можно предположить, что такими наиболее информативными показателями являются как раз показатели, для которых оценка неустранимых ошибок превышает 5%. Неустранимую ошибку рассматриваемых данных будет в целом характеризовать наибольшая из ошибок – 28,92 %. Еще раз подчеркнем, что вычисленная оценка является оценкой снизу. Определение композитных индексов качества таких систем должно учитывать наличие неустранимых ошибок данных.

Заключение

Построение интегральной характеристики системы требует детального понимания влияния погрешностей используемых данных на вычисляемые характеристики. Даже небольшое возмущение исходных данных может вызвать значительное изменение весовых коэффициентов при использовании методов многомерного анализа. Причиной этого может являться наличие неустранимых погрешностей используемых данных. Оценить наличие погрешностей для ряда наблюдений позволяет метод конечных разностей. Для правомерности применения этой методики необходимо, чтобы данные допускали приближение полиномами более низких степеней, чем число наблюдений минус один. Справедливость предположения проверена эмпирически на конкретном наборе данных из 37 переменных, характеризующих качество жизни населения России за 2010-2017 годы. Анализ выборки показывает наличие неустранимых ошибок от 0,59% до 28,92 %. Следователь-

но, получение разнообразных композитных характеристик объектов на основании таких данных должно обязательно учитывать наличие неустранимой ошибки данных. В частности, число параметров, характеризующих систему, должно быть достаточно большим, чтобы компенсировать случайные ошибки усреднением. 37 параметров, предлагаемых для характеристики качества жизни, вполне достаточны для компенсации ошибок измерения. Характеристика отдельных блоков, в которых 9, 14 и 14 переменных, будет чувствительна к неустранимым ошибкам. Используемая методика вычисления композитного индикатора тоже должна учитывать наличие неустранимых ошибок данных.

References

[1] 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

[2] 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

[3] Bandura R. Composite indicators and rankings: inventory 2011. Tech. rep., United Nations Development Programme – Office of Development Studies; 2011. (In Eng.)

[4] Saltelli A., Mundo G., Nardo M. From Complexity to Multidimensionality: The Role of Composite Indicators for Advocacy of EU Reform. *Review of Business and Economic Literature*. 2006; LI(3):221-235. Available at: https://ideas.repec.org/a/ete/revbec/20060303.html (accessed 21.06.2020). (In Eng.)

[5] Foa R., Tanner J.C. Methodology of the Indices of Social Development. *ISD Working Paper Series*. 2012; 04. International Institute of Social Studies of Erasmus University Rotterdam (ISS), The Hague. Available at: http://repub.eur.nl/pub/50510/ISD-WP-2012-4.pdf (accessed 21.06.2020). (In Eng.)

[6] 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

[7] Auerbach A.J., Gorodnichenko Yu., Murphy D. Macroeconomic Frameworks. *NBER Working Paper*. 2019; 26365. (In Eng.)

[8] Nardo M., Saisana M., Saltelli A., Tarantola S. Tools for Composite Indicators Building. Joint Research Centre, Ispra, Italy; 2005. Available at: https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/tools-composite-indicators-building (accessed 21.06.2020). (In Eng.)

[9] Krishnan V. Development of a Multidimensional Living Conditions Index (LCI). *Social Indicators Research*. 2015; 120(2):455-481. (In Eng.) DOI: https://doi.org/10.1007/s11205-014-0591-0

[10] Jacobs R., Goddard M., Smith P. Measuring Performance: An Examination of Composite Performance Indicators: A report for the Department of Health. York, UK: Centre for Health Economics, University of York; 2004. Available at: https://www.york.ac.uk/che/pdf/tp29.pdf (accessed 21.06.2020). (In Eng.)

[11] 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

[12] Becker W., Saisana M., Paruolo P., Vandecasteele I. Weights and importance in composite indicators: Closing the gap. *Ecological Indicators*. 2017; 80:12-22. (In Eng.) DOI: https://doi.org/10.1016/j.ecolind.2017.03.056

[13] Paruolo P., Saisana M., Saltelli A. Ratings and rankings: voodoo or science? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2013; 176(3):609-634. (In Eng.) DOI: https://doi.org/10.1111/j.1467-985X.2012.01059.x

[14] Mazziotta M., Pareto A. On The Construction Of Composite Indices By Principal Components Analysis. *RIEDS - Rivista Italiana di Economia, Demografia e Statistica - Italian Review of Economics, Demography and Statistics*. 2016; 70(1):103-109. Available at: http://www.sieds.it/listing/RePEc/journl/2016LXX_N1_RIEDS_103-109_Mazziotta_Pareto.pdf (accessed 21.06.2020). (In Eng.)

[15] Zhgun T.V. Method for evaluating the robustness of rankings generated by composite indices. *Journal of Physics: Conference Series*. 2019; 1352(1):012064. (In Eng.) DOI: https://doi.org/10.1088/1742-6596/1352/1/012065

[16] Batini C., Scannapieca M. Data Quality Dimensions. In: *Data Quality. Data-Centric Systems and Applications*. Springer, Berlin, Heidelberg; 2006. p. 19-49. (In Eng.) DOI: https://doi.org/10.1007/3-540-33173-5_2

[17] Herzog T.N., Scheuren F.J., Winkler W.E. What is Data Quality and Why Should We Care? In: *Data Quality and Record Linkage Techniques*. Springer, New York, NY; 2007. p. 7-15. (In Eng.) DOI: https://doi.org/10.1007/0-387-69505-2_2

[18] Wang R.Y., Kon H.B., Madnick S.E. Data quality requirements analysis and modeling. In: *Proceedings of IEEE 9th International Conference on Data Engineering*. Vienna, Austria; 1993. p. 670-677. (In Eng.) DOI: https://doi.org/10.1109/ICDE.1993.344012

[19] Weisberg S. Applied Linear Regression. 4th ed. Hoboken, NJ: Wiley; 2014. (In Eng.)

[20] Seber G.A.F., Lee A.J. Linear Regression Analysis. 2 nd ed. John Wiley & Sons, Inc.; 2003. (In Eng.) DOI: https://doi.org/10.1002/9780471722199

[21] Hoffmann J.P, Shafer K. Linear Regression Analysis. Assumptions and Applications. Washington: NASW Press; 2015. (In Eng.)

[22] Bingham N.H., Fry J.M. Regression: Linear Models in Statistics. London: Springer; 2010. (In Eng.) DOI: https://doi.org/10.1007/978-1-84882-969-5

[23] Montgomery D.C., Peck E.A., Vining G.G. Introduction to Linear Regression Analysis. Fifth ed. New York: Wiley; 2012. (In Eng.)

[24] Isakin M.A. Modification of the *K*-means method with an unknown number of classes. *Applied Econometrics*. 2006; (4):62-73. Available at: https://www.elibrary.ru/item.asp?id=9482376 (accessed 21.06.2020). (In Russ., abstract in Eng.)

[25] Aivazian S., Stepanov V., Kozlova M. Measuring the synthetic categories of quality of life in a region and identification of main trends to improve the social and economic policy

(Samara region and its constituent territories). *Applied Econometrics*. 2006; (2):18-84. Available at: https://www.elibrary.ru/item.asp?id=9482361 (accessed 21.06.2020). (In Russ., abstract in Eng.)

Поступила 21.06.2020; принята к публикации 27.08.2020; опубликована онлайн 30.09.2020.
Submitted 21.06.2020; revised 27.08.2020; published online 30.09.2020.

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, Россия, 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, Russia), 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.