УДК 519.25
DOI: 10.25559/SITITO.019.202302.393-402

Оригинальная статья

Метрики оценки качества числовых параметров динамических систем

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

Аннотация

Проблема неопределенности качества входных данных, описывающих систему, является одной из наиболее существенных проблем при построении систем управления сложными объектами. Еще более остро такая проблема стоит при управлении слабо формализованными (мягкими) системами. Критически важным компонентом управления качеством данных является разработка метрик, информирующих потребителей о характеристиках качества, которые наиболее важны для оценки степени пригодности данных к использованию. В статье предлагаются такие параметры для измерения качества данных, как точность данных, которая определяется как совпадение характеристики набора данных с неискаженными характеристиками реального объекта, и достоверность данных, которая определяется как несовпадение характеристики набора данных с характеристиками объекта, все регистрируемые параметры абсолютно случайны. Приводятся формулы для определения мер этих параметров качества, использующие аппарат конечных разностей. Предлагаемая методика предоставляет достаточно формализованный и вычислительно несложный алгоритм оценки качества совокупности входных параметров слабо формализованной динамической системы. Предлагаемые оценки являются эффективными метриками качества, анализ которых позволяет инициировать алгоритм управления, выделяющий полезный сигнал из зашумленного потока данных. Предлагаемая методика применена для анализа совокупности статистических данных, характеризующих качество жизни населения субъектов Российской Федерации за 2009–2019 годы. Анализ показывает, что значительное число рассматриваемых параметров имеет значительную ошибку регистрации и недостаточную степень достоверности. Следовательно, использование таких данных, как основы для принятия решений, без учета имеющихся искажений привносит ошибки в оценки и прогнозы и, как следствие, приводит к значительному снижению качества принимаемых управленческих решений. В частности, вычисление композитных индексов качества системы по однократному наблюдению по данным статистических измерений с помощью математических методов не предполагает устранения имеющейся шумовой компоненты данных, вследствие чего полученный результат может быть неправдоподобным.

Ключевые слова: количественный математико-статистический анализ, качество данных, ошибки данных, метод конечных разностей, композитные индексы

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Для цитирования: Жгун Т. В. Метрики оценки качества числовых параметров динамических систем // Современные информационные технологии и ИТ-образование. 2023. Т. 19, № 2. С. 393-402. doi: https://doi.org/10.25559/SITITO.019.202302.393-402

Original article

Metrics for Assessing the Quality of Numerical Parameters of Dynamic Systems

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

Abstract

The problem of uncertainty concerning the quality of input data describing the system is one of the most significant problems in the construction of control systems for complex objects. This problem is even more acute when managing poorly formalized systems. A critical component of data quality management is the development of metrics that inform consumers about the quality characteristics that are most important for assessing the degree of suitability of the data for use. The article suggests such parameters for measuring data quality as data accuracy, which is defined as the similarity of the characteristics of a data set with non-distorted characteristics of a real object, and data reliability, which is defined **as** the discrepancy between the characteristics of a data set with the characteristics of an object, for which all recorded parameters are absolutely random**.** Formulas for determining the measures of these quality parameters using the finite difference apparatus are given. The proposed methodology provides a fairly formalized and computationally simple algorithm for evaluating the quality of a set of input parameters of a complex dynamic system. The proposed estimates are effective quality metrics, the analysis of which allows you to initiate a control algorithm that extracts a useful signal from a noisy data stream. The analysis shows that a significant number of the parameters under consideration have a significant registration error and an insufficient degree of reliability. Consequently, the use of such data as a basis for decision-making, without taking into account the existing distortions, introduces errors in estimates and forecasts and, as a result, leads to a significant decrease in the quality of management decisions. In particular, the calculation of composite system quality indices based on a single observation of statistical measurements using mathematical methods does not imply the elimination of the existing noise component of the data, as a result of which the result obtained may be implausible.

Keywords: ice accretion, surface computational mesh, remeshing

Conflict of intersts: The author declares no conflict of interest.

For citation: Zhgun T. V. Metrics for Assessing the Quality of Numerical Parameters of Dynamic Systems. *Modern Information Technologies and IT Education.* 2023;19(2):393-402. doi: https://doi.org/10.25559/SITITO.019.202302.393-402

Введение

При исследовании сложных систем приходится ставить и решать как хорошо формализованные в математических терминах задачи (жесткие системы), так и слабо структурированные задачи, выражаемые на естественном языке и решаемые эвристическими методами (слабо формализованные системы). «Классические» подходы к управлению системами строятся на том предположении, что можно получить пусть сложную, но точную аналитическую форму функциональной зависимости входных и выходных параметров системы управления с последующим уточнением значений, входящих в функциональную зависимость коэффициентов[1] [1, 2].

Одной из наиболее существенных проблем при построении систем управления сложными системами разных типов (жесткими и слабо формализованными) является неопределенность характеристик объекта управления. Примерами жестких систем с четкой формализацией являются технические системы и технологические процессы. Неопределенность характеристик объекта управления обычно не учитывается при управлении такими системами. Например, при механической обработке деталей в машиностроении неполнота априорной и текущей информации об объекте управления обуславливается изменением режимов работы оборудования, нестабильностью материалов заготовок деталей и режущего инструмента, нестабильностью характеристик оборудования и т. д. Для компенсации возможных нежелательных изменений входных параметров требуется существенное снижение интенсивности рабочих режимов оборудования, что недопустимо при его высокой стоимости. Существующие методы построения процессов управления процессами механообработки при выработке управления чаще всего совсем не используют оценку входных воздействий на систему либо используют оценку входных воздействий на систему лишь для некоторых входных параметров системы[2].

В управлении слабо формализованными (мягкими) системами входными параметрами обычно являются статистические данные. Проблема неопределенности качества входных данных в этом случае стоит еще более остро, так как для таких систем не только неизвестны характеристики качества входных параметров системы, но и номенклатура входных параметров является отдельной проблемой. Наличие этой проблемы подтверждает, в частности, изменение год от года списка регистрируемых показателей в статистических справочниках. Например, в статистическом сборнике Росстата «Здравоохранение», выходящем с периодичностью один раз в два года, номенклатура показателей сборника за 2019 год только в 87 % случаев совпадает с номенклатурой сборника за 2009 год.

Традиционный подход к оптимальному управлению динамическими системами сводится к синтезу оптимальных алгоритмов (законов) управления на стадии проектирования управления объектом [3]. Однако неопределенность входных параметров управляемой системы диктует необходимость адаптивной оптимизации алгоритмов управления в реальном времени в зависимости от реальных характеристик входных параметров управляемой системы. Очевидно, что практическая реализация алгоритмов, реализующих традиционный подход, предполагает достаточную формализуемость и вычислительную простоту процедур синтеза оптимальных управлений.

Процедура синтеза адаптивной системы управления обычно разбивается на два этапа[3]: синтез основного контура и синтез контура адаптации. Контур адаптации необходимым элементом содержит блок анализа входных параметров. Оценка качества совокупности всех входных параметров должна быть естественной частью блока анализа входных параметров [4]. Решение этой задачи при синтезе адаптивной системы управления также предполагает достаточную формализуемость и вычислительную простоту алгоритма оценки качества входных параметров.

1. Постановка задачи

Определение качества данных затруднено из-за множества контекстов, в которых используются данные, а также из-за различных точек зрения на эту проблему среди производителей данных, органов регистрации данных и потребителей данных. Разногласие мнений относительно того, какие именно параметры определяют качество данных, определяется сложной и неоднородной природой данных и областью их применения [5]. В 2021 году рабочая группа Data Quality of DAMA Netherlands исследовала определения параметров качества данных из разных источников. Результатом является список из 60 параметров качества данных[4]. Такое количество параметров говорит скорее о том, что эти определения качества данных не могут служить для формирования контура адаптации в системе управления.

Data Quality (качество данных) — характеристика, показывающая степень пригодности данных к использованию. Соответствующими международному стандарту качества данных ISO 8000 считаются «переносимые данные, удовлетворяющие предъявляемым требованиям». Обычно данные считают высококачественными, если они пригодны для предполагаемого использования в операциях, принятии решений и планировании. Согласно другому подходу, данные считаются высококачественными, если они правильно представляют события или объекты реального мира, к которым эти данные относятся [6–12]. Многие ведущие исследователи в области качества

[1] Цыпкин Я. З. Адаптация и обучение в автоматических системах. М. : Наука, 1968. 399 с.; Загоруйко Н. Г. Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999. 270 с.

[2] Зориктуев В. Ц., Лютов А. Г. Управление процессами механообработки деталей авиационных двигателей в условиях неопределенности. М. : Изд-во МАИ, 2003. 120 с.

[3] Фрадков А. Л. Адаптивное управление в сложных системах: беспоисковые методы. М. : Наука, 1990. 296 с.

[4] A Management System for Data Quality: Towards a Solid, Coherent and Standardised Approach / P. van Nederpelt [et al.]. [Электронный ресурс]. URL: https://www.dama-nl.org/wp-content/uploads/2022/05/A-Management-System-for-Data-Quality-DAMA-NL-v2.2-EN-1.pdf (дата обращения: 14.02.2023); Black A., van Nederpelt P. Dimensions of Data Quality. DAMA NL Foundation, 2020. 113 p. [Электронный ресурс]. URL: https://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf дата обращения: 14.02.2023).

данных предлагают в своих публикациях весьма тщательно проработанные наборы характеристик для измерения и оценки качества данных[5] [13–19].

Приведенные определение характеристик качества данных крайне расплывчаты и не могут служить для формирования контура адаптации в системе управления. Нужны численные характеристики качества данных, позволяющие принимать решения.

Для управления слабо формализованными (мягкими) системами органы управления обычно используют данные, предоставляемые органами регистрации данных. В мировой статистической практике нет общепринятого определения качества данных как результата статистической деятельности. Традиционный подход определяет качество статистических данных как их соответствие требованиям полноты, достоверности и сопоставимости. Эти параметры плохо формализованы и не могут служить для формальной оценки качества статистической информации. В последней четверти XX века в мировой статистике был принят подход к определению качества статистической информации, когда качество определяется соответствием потребностям и ожиданиям пользователей.

Российские органы статистики, предоставляя обширные массивы числовой информации, характеризующей в динамике разные аспекты функционирования слабо формализованных социально-экономических систем страны, совсем не приводят объективных оценок качества публикуемых данных. Согласно «Положению о Росстате» российская статистическая служба обязана предоставлять официальную статистическую информацию без указания критериев качества.

Росстат ведет работы по оценке качества первичной информации. В 2003 году разработаны «Методологические рекомендации по расчету и анализу рейтинговых оценок качества результатов проведения обследований по формам федерального государственного статистического наблюдения». Оценка должна определить уровень качества проводимых статистических обследований. Однако конкретные результаты такого оценивания качества государственной статистики неизвестны, и вопрос об уровне качества публикуемых статистических данных остается открытым.

2. Эффективные метрики качества данных

Критически важным компонентом управления качеством данных является разработка метрик, информирующих потребителей о характеристиках качества, которые наиболее важны для оценки степени пригодности данных к использованию. Измеримых параметров всегда имеется в избытке, но далеко не все из них актуальны и стоят времени и труда, затрачиваемых на их измерение и учет [20]. При разработке метрик качества данных следует учитывать следующие характеристики:

- измеримость: параметры качества должны быть измеримыми, ожидаемые результаты должны поддаваться количественному определению в рамках дискретного диапазона значений;
- значимость для потребителя: прежде всего, результаты измерений должны интересовать потребителей данных, из множества доступных для измерения параметров системы далеко не все могут быть переведены в полезные для контура управления метрики;
- контролируемость: при выходе значения измеряемого параметра за пределы установленного допуска контур адаптации должен выделить в потоке зашумленных данных неискаженный сигнал (например, инициировать процедуру улучшения данных или параметров работы алгоритма обработки входных данных). Если же введенная метрика не обеспечивает функционирования контура управления, то она, возможно, является излишней.

В справочнике[6] приведен набор общепринятых измерений качества данных с определениями и описаниями подходов к их измерению. Называются такие характеристики: актуальность, допустимость, полнота, разумность, согласованность, соответствие, уникальность, целостность. Все перечисленные характеристики качества являются абстрактными понятиями с никак не проверяемыми критериями соответствия требованиям, так как отсутствуют четкие определения степени актуальности информации, допустимости информации и т. д. Более очевидной характеристикой качества данных в приведенном списке является «полнота» данных, но и она нуждается в определении объективной меры.

3. Мера оценки точности сигнала

Во всех случаях оценка «качества данных» представляет собой сравнение фактического состояния конкретного набора данных с желаемым состоянием данных (данных без дефектов). Для хорошо формализованных жестких систем эталонные данные без дефектов обычно определяются экспертами по стандартизации, законами и нормативными актами [7]. Для мягких систем этот подход невозможен в принципе. Эталонные значения параметров такой системы определить невозможно. Но для таких систем можно говорить о том, насколько точно имеющиеся данные описывают реальное положение вещей. Качественны те данные, которые достаточно точно представляют конкретную систему [7–12] и, следовательно, не имеют ошибок регистрации или имеют допустимо малую ошибку регистрации. Такое измерение качества обладает свойствами эффективной метрики качества: точность данных измерима, значима и контролируема.

В сложившейся практике степень точности величины обычно характеризуется ее дисперсией, стандартной ошибкой, коэффициентом вариации. Но эти меры не позволяют судить о наличии возможных ошибок в системе регистрации данных. Однократное наблюдение проблему наличия ошибок регистрации оставляет открытой. Но наблюдения за динамической системой позволяют решить эту проблему, применяя аппарат конечных разностей. Предполагается, что регистрируемые данные не случайны, а являются числовыми характеристиками некоторого неслучайного процесса, который на интервале

[5] Lehmann C., Roy K., Winter B. The State of Enterprise Data Quality: 2016. Perception, Reality an the Future of DQM. 451 Research, commissioned by Blazent, 2016. 21 p. [Электронный ресурс]. URL: https://siliconangle.com/files/2016/01/Blazent_State_of_Data_Quality_Management_2016.pdf (дата обращения: 14.02.2023).

[6] DAMA-DMBOK : Свод знаний по управлению данными. Второе издание / Dama International [пер. с англ. Г. Агафонова]. М. : Олимп-Бизнес, 2020. 828 с.

наблюдения может быть аппроксимирован достаточно гладкой функцией, например полиномом. Справедливость этого предположения, т. е. качество аппроксимации, может быть проверено экспериментально для любого набора данных [21-23].

Определение

Определим **точность данных** как **совпадение характеристики набора данных с характеристикой желаемого состоянием данных**, т. е. с неискаженными характеристиками реального объекта (явления).

Мерой точности данных назовем **максимальную из оценок ошибок наблюдения** измеряемого в ряде наблюдений параметра.

Напомним, что при наличии ряда наблюдений $y_0, y_1, ..., y_k$ конечной разностью первого порядка называют разность двух последовательных значений измеряемой величины: $\Delta^1_i = y_{i+1} - y_i$. Аналогично конечной разностью k-го порядка называют разность $\Delta^k_i = \Delta^{k-1}_{i+1} - \Delta^{k-1}_i$. Конечная разность k-го порядка выражается через значения функции и биномиальные коэффициенты следующим образом

$$\Delta^k_i = y_{i+k} - C^1_k y_{i+k-1} + C^2_k y_{i+k-2} - C^3_k y_{i+k-3} + C^4_k y_{i+k-4} - \ldots + (-1)^k y_i$$

В реальных условиях вместо точных значений параметра $y_0, y_1, ..., y_k$ доступны его приближенные значения $y^*_0, y^*_1, ..., y^*_k$, и, соответственно, вместо точных значений конечных разностей Δ^k_i исследователю доступны значения приближенных конечных разностей Δ^{*k}_i.

Ошибка измерений в i-ом наблюдении $\varepsilon_i = y^*_i - y_i$ имеет случайный характер, ее величина неизвестна и не может быть непосредственно вычислена по фиксируемым наблюдениям. Обозначим максимальную величину ошибки для ряда наблюдений $y^*_0, y^*_1, ..., y^*_k$

$$\varepsilon = \max_i \left| \varepsilon_i \right| .$$

Рассмотрим первые конечные разности приближенных величин

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

Модуль приближенной конечной разности $\left| \Delta^*_i \right| \le \left| \Delta_i \right| + 2 \cdot \varepsilon$, а для последней вычисленной по имеющимся значениям k-ой приближенной конечной разности справедлива оценка

$$\left| \Delta^{*k}_i \right| \le \left| \Delta^k_i \right| + 2^k \cdot \varepsilon \qquad (1)$$

Если значения измеряемого параметра от измерения к измерению меняются не слишком быстро и функцию, описывающую измеряемый параметр, можно аппроксимировать полиномом, степень которого менее k, то значения точных конечных разностей Δ^k_i с увеличением порядка разности стремятся к нулю. Справедливость предположения о возможности аппроксимации для измеряемых входных параметров проверяется экспериментально. При выполнении этого условия наблюдаемые значения приближенных конечных разностей обеспечивают оценку исходной погрешности:

$$\left| \Delta^{*k}_i \right| \le 2^k \cdot \varepsilon \qquad (2)$$

Очевидно, что максимальная из ошибок регистрации $\varepsilon \ge \left| \Delta^{*k}_i \right| / 2^k$.

Обозначим

$$\varepsilon^* = \left| \Delta^{*k}_i \right| / 2^k \qquad (3)$$

Очевидно, что определенное (3) значение $\varepsilon^* \le \varepsilon$ является оценкой снизу ошибки регистрации данных.

Рассмотрим теперь многомерный процесс. Значения величины $x_{ij} = x_{ij}(t)$ — значения j-го признака для i-го объекта в момент t, t = 0,..., k. Пусть переменная x_{ij} на рассматриваемом интервале представлена наблюдениями с некоторыми погрешностями $x^*_{ij}(0), x^*_{ij}(1), ..., x^*_{ij}(k)$, которые реализуют неизвестную сложную зависимость функционирования рассматриваемой системы с некоторыми погрешностями: $x^*_{ij}(t) = x_{ij}(t) + \varepsilon_{ij}(t)$, $\varepsilon_{ij} = \max_t \left| \varepsilon_{ij}(t) \right|$. Если регистрируется изменение параметра $j = 1...n$ во времени $t = 0$ k для системы объектов $i = 1...m$, то вычисленная оценка представления данных для параметра j объекта i на промежутке наблюдения определяется соотношением:

$$\varepsilon^*_{ij} = \left| \Delta^{*k}_j \right| / 2^k .$$

Вычисленное значение ε^*_{ij} является оценкой снизу возможной ошибки регистрации j-го параметра для i-го объекта. Характеристикой параметра j будет максимальная из наблюдаемых ошибок

$$\varepsilon^*_j = \max_i \left| \varepsilon^*_{ij} \right| \qquad (4)$$

Математическое моделирование показало, что оценка погрешности регистрации данных, полученная в серии испытаний, составляет около 70 % от величины вносимой погрешности.

4. Мера оценки достоверности сигнала

Определение

Определим **достоверность данных** как **несовпадение характеристики набора данных с характеристиками объекта с полным отсутствием определенности**, т. е. для объекта, все регистрируемые параметры абсолютно случайны.

Для определения меры достоверности данных рассмотрим случай, когда регистрируемые наблюдения никак не зависят друг от друга и их значения абсолютно произвольны в некотором диапазоне [0, a]. В этом случае два любых наблюдения представляют реализацию двух независимых случайных величин, равномерно распределенных на отрезке [0, a]. Оценим, какие значения может принимать математическое ожидание абсолютного значение конечной разности, вычисленной по таким наблюдениям.

Пусть (ξ, η) — двумерная случайная величина с плотностью распределения вероятности $f(x,y)$, $\zeta = \varphi(\xi, \eta)$ — функция двух случайных аргументов. Математическое ожидание функции двух случайных аргументов

$$M(\zeta) = \iint_{D(z)} \varphi(x, y) \cdot f(x, y) dx\, dy ,$$

где область $D(z)$ определяется соотношением: $\varphi(x, y) < z$.

Рассмотрим в качестве функции абсолютную величину разности двух случайных величин $\zeta = \varphi(\xi, \eta) = \left| \xi - \eta \right|$. Оценим математическое ожидание модуля разности двух случай-

ных величин:

$$M(\zeta) = \iint\limits_{D(z)} \varphi(x, y) \cdot f(x, y) dx\, dy = \iint\limits_{D(z)} |x - y| \cdot f(x, y) dx\, dy$$

Так как случайные величины независимы и равномерно распределены на интервале [0, *a*], то справедливо

$$M(\zeta) = \iint\limits_{D(z)} |x - y| \cdot f(x) \cdot f(y) dx\, dy = \frac{1}{a^2} \iint\limits_{D(z)} |x - y| \cdot dx\, dy =$$

$$\frac{1}{a^2}\left[\int\limits_{-a}^{0}\int\limits_{0}^{a+x}(y - x) dx\, dy + \int\limits_{0}^{a}\int\limits_{0}^{a-x}(x - y) dx\, dy\right] = \frac{a}{3} \quad (5)$$

Пусть теперь имеется *k* реализаций случайного процесса y_i, *i* = 0, ..., *k*. Будем считать, что *k* — нечетно, а общее число известных реализаций четно. *k+1* значение измеренной величины определяют *k-ю случайную разность*

$$\Delta^k = C_k^0 y_k - C_k^1 y_{k-1} + C_k^2 y_{k-2} - \cdots - C_k^2 y_2 + C_k^{k-1} y_1 - C_k^k y_0$$

Учитывая симметричность биномиальных коэффициентов

$$\Delta^k = C_k^0 (y_k - y_0) - C_k^1 (y_{k-1} - y_1) + C_k^2 (y_{k-2} - y_2) + \cdots\cdots$$

$$|\Delta^k| \le C_k^0 |y_k - y_0| + C_k^1 |y_{k-1} - y_1| + C_k^2 |y_{k-2} - y_2| + \ldots$$

Значит, для математического ожидания справедливо

$$M(|\Delta^k|) \le M\left(C_k^0 |y_k - y_0| + C_k^1 |y_{k-1} - y_1| + C_k^2 |y_{k-2} - y_2| + \ldots\right) =$$

$$= M(C_k^0 |y_k - y_0|) + M(C_k^1 |y_{k-1} - y_1|) + M(C_k^2 |y_{k-2} - y_2|) + \cdots =$$

$$= C_k^0 \cdot M|y_k - y_0| + C_k^1 \cdot M|y_{k-1} - y_1| + C_k^2 \cdot M(|y_{k-2} - y_2|) + \cdots$$

Все величины y_i, y_j — независимы и одинаково распределены и $M(|y_i - y_j|) = \frac{a}{3}$. Тогда

$$M(|\Delta^k|) \le M(|y_i - y_j|) \cdot \left(C_k^0 + C_k^1 + C_k^2 \cdot + \cdots C_k^{[k/2]}\right) = \frac{a}{3} \cdot 2^{k-1} = \frac{a}{6} \cdot 2^k$$

Значит, если число известных реализаций *k+1* четно, то

$$M(|\Delta^k|) \le \frac{a}{6} \cdot 2^k \quad (6)$$

Полученная оценка (6) дает эталон меры для оценки достоверности параметров системы. Это величина, характеризующая абсолютно случайный процесс на входе системы. Сравнение характеристик исследуемого процесса с этой мерой позволит сделать заключение о наличии или отсутствии отклонений входных параметров от параметров случайного процесса, т.е. о вкладе случайного компонента в регистрируемые данные. Для имеющейся реализации наблюдаемого процесса y_i, *i* = 0, ..., *k*. отношение

$$\mu = \frac{6 \cdot |\Delta^{*k}|}{a \cdot 2^k} \cdot 100\% \quad (7)$$

характеризует вклад случайного компонента в регистрируемые данные и является мерой достоверности набора данных. Превышение пятипроцентного порога позволяет сделать вывод о недостаточной достоверности данных.

Рассмотрим теперь многомерный процесс. Значения величины $x_{ij} = x_{ij}(t)$ — значения *j*-го признака для *i*-го объекта в момент *t*, *t* = 1, ..., k. Пусть переменная x_{ij} на рассматриваемом интервале представлена наблюдениями с некоторыми погрешностями $x^*_{ij}(0), x_{ij}{}^*(1) \ldots, x^*_{ij}(k)$, $x^*_{ij}(t) = x_{ij}(t) + \varepsilon_{ij}(t)$. Приближенная *k*-я разность параметра *j* для объекта *i* определяется по ряду наблюдений аналогично через регистрируемые с ошибкой данные:

$$\Delta^{*k}_{ij} = x^*_{ij}(k) - C^1_k x^*_{ij}(k\text{-}1) + C^2_k x^*_{ij}(k\text{-}2) - C^3_k x^*_{ij}(k\text{-}3) + \ldots + (-1)^k x^*_{ij}(0)$$

Определение

Величина отношения математического ожидания модуля последней приближенной разности к аналогичной характеристике случайного процесса

$$\mu_j = \frac{6 \cdot M(|\Delta^{*k}_j|)}{a \cdot 2^k} \cdot 100\% \quad (8)$$

является **мерой достоверности** *j*-го параметра входных данных.

Отношение (8) характеризует относительный вклад случайного компонента в исследуемый процесс для переменной *j*. Значение этой величины более 5 % в наборе данных будет свидетельствовать о значительном уровне случайной компоненты в сигнале и о необходимости применять методы устранения случайных искажений — методы шумоподавления — для анализа сигнала. Введенное измерение качества обладает свойствами эффективной метрики качества: достоверность данных измерима, значима и контролируема.

Введенные метрики характеризуют разные стороны наблюдаемого процесса.

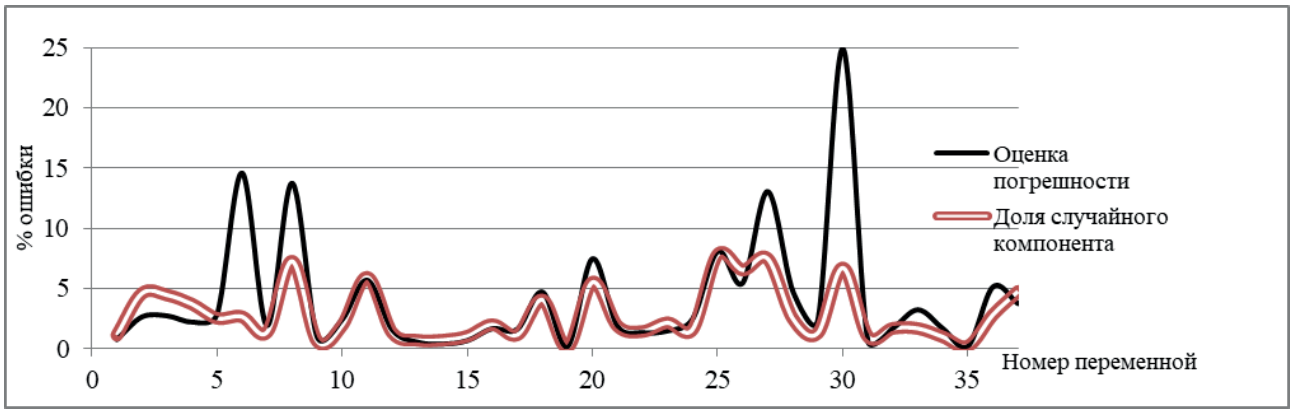
5. Результаты

В таблице 1 приведены оценки рассматриваемого набора данных, характеризующих качество жизни населения регионов России за 2009–2019 годы по введенным измерениям качества. Данные взяты из сборников Росстата. Все данные рассматриваются на стандартном интервале и значение *a* = 100. При предварительной обработке из набора данных удалены явные выбросы и при необходимости применена логарифмическая трансформация данных [24, 25]. Значения метрик, превышающие 5 %, выделены цветом в таблице.

У девяти параметров из 37 погрешность регистрации данных превышает допустимые 5 %. Наибольшую ошибку регистрации имеют переменные «численность смертей при несчастных случаях на производстве» — ε_{25} = 8,08, «коэффициент миграции» — ε_{27} =13,05, «доля ветхого и аварийного жилья» — ε_8 =13.74, «доля семей, состоящих на учете на получение жилья» — ε_6 =14,58, «число зарегистрированных изнасилований на 100 тысяч человек» ε_{30} =24,90. При этом ошибку регистрации менее 1 % имеют 8 показателей (минимальная ошибка 0,18 % у показателя «число инвалидов на 1000 человек»), ошибку регистрации менее 2 % имеют 17 показателей. Вполне ожидаемо, что уровень безработицы фиксируется менее точно, чем число умерших от новообразований, но точнее, чем число умышленных убийств.

Семь параметров из 37 показали превышение пятипроцентной доли случайности в сигнале. Содержательный анализ

полученных значений может указать на недостоверные переменные. Например, показатель № 18 «заболеваемость от травм» (достоверность μ_8 = 4,07) может быть случайным показателем, а доля ветхого жилья (μ_{18} =7.27) в периоде наблюдения не может быть случайной величиной, поэтому μ_8.значение должно превышать μ_{18}. Наблюдаемые характеристики связаны обратным соотношением, что характеризует переменную 18 как недостаточно достоверную. Следует отметить, что все переменные, для которых превышен порог достоверности, также имеют превышение и порога точности, а обратное неверно (рис. 1). Одновременное нарушение условий $\varepsilon_j < 5$ и $\mu_j < 5$ указывает на переменные с наименьшим качеством. В этом наборе это переменные 25, 27, 8, 30, 26, 11, 20.



Р и с. 1. Гистограммы распределения для переменных с разными характеристиками асимметрии

F i g. 1. Distribution histograms for variables with different skewness characteristics

Источник: составлено автором.

Source: Compiled by the author.

Т а б л и ц а 1. **Оценки качества набора данных**

T a b l e 1. **Dataset quality ratings**

№ п/п	Переменные	Оценка погрешности	Доля случайного компонента
1	ВРП на душу населения с учетом инфляции, тысяч рублей	0,77	1,13
2	Отношение среднедушевых денежных доходов к прожиточному минимуму	2,64	4,60
3	Доля населения с доходами ниже прожиточного минимума	2,74	4,46
4	Отношение доходов 20 % самых богатых и 20 % самых бедных	2,22	3,71
5	Обеспеченность собственными легковыми автомобилями на 1000 человек	2,83	2,53
6	Доля семей, состоящих на учете на получение жилья	**14,58**	2,67
7	Общая площадь жилищного фонда на одного жителя	1,93	1,38
8	Доля ветхого и аварийного жилья	**13,74**	**7,27**
9	Плотность автомобильных дорог общего пользования	0,99	0,65
10	Ожидаемая продолжительность жизни при рождении	2,43	1,94
11	Число умерших детей в возрасте до 1 года на 1000 родившихся	**5,70**	**5,93**
12	Коэффициент естественного прироста на 1000 человек	1,58	1,33
13	Умерших от инфекционных болезней и туберкулеза на 100 тысяч человек	0,55	0,71
14	Число умерших от новообразований на 100 тысяч человек	0,39	0,72
15	Умерших от болезней системы кровообращения на 100 тысяч человек.	0,68	1,05
16	Число умерших от болезней органов дыхания на 100 тысяч человек	1,68	2,01
17	Число умерших от болезней органов пищеварения на 100 тысяч человек	1,63	1,16
18	Заболеваемость от травм и других внешних причин на 100 тысяч человек	4,71	4,07
19	Число инвалидов на 1000 человек	0,18	0,13
20	Зарегистрировано врожденных аномалий на 1000 человек	**7,48**	**5,53**
21	Доля специалистов с высшим образованием к занятым в экономике	1,85	1,90
22	Отношение ВРП к численности занятых в экономике, тыс. руб/чел	1,34	1,44

№ п/п	Переменные	Оценка погрешности	Доля случайного компонента
23	Численность студентов высших и средних учебных заведений на 1000 человек	1,49	2,17
24	Уровень безработицы, %	2,48	1,61
25	Численность смертей при несчастных случаях на производстве на 1000 работающих	**8,08**	**7,82**
26	Численность пострадавших при несчастных случаях на производстве на 1000 работающих	**5,46**	**6,54**
27	Коэффициент миграционного прироста на 10 тысяч человек	**13,05**	**7,47**
28	Число зарегистрированных умышленных убийств на 100 тысяч человек	4,81	2,51
29	Число фактов умышленного причинения тяжкого вреда здоровью на 100 тысяч человек	2,36	1,30
30	Число зарегистрированных изнасилований на 100 тысяч человек	**24,90**	**6,72**
31	Число разбоев, грабежей, краж на 100 тысяч человек	0,66	1,03
32	Зарегистрированных присвоений или растрат на 100 тысяч человек	1,71	1,70
33	Состоящих на учете: наркомания и токсикомания на 100 тысяч человек	3,24	1,69
34	Состоящих на учете: алкоголизм в расчете на 100 тысяч человек	1,67	0,99
35	Больных туберкулезом в расчете на 100 тысяч человек	0,22	0,22
36	Число больных с диагнозом сифилиса на 100 тысяч человек	**5,14**	2,83
37	Число больных психическими расстройствами на 100 тысяч человек	3,75	4,76

Источник: составлено автором.

Source: Compiled by the author.

Заключение

Проблема неопределенности качества входных данных, описывающих систему, является одной из наиболее существенных проблем при построении систем управления сложными объектами. Еще более остро такая проблема стоит при управлении слабо формализованными (мягкими) системами. Неопределенность входных параметров управляемой системы диктует необходимость адаптивной оптимизации алгоритмов управления в реальном времени в зависимости от характеристик неопределенности входных параметров управляемой системы. Возможность практической реализации подобных алгоритмов предполагает достаточную формализуемость и вычислительную простоту процедур синтеза оптимальных управлений, которые в качестве необходимого элемента содержат блок анализа входных параметров.

Критически важным компонентом управления качеством данных является разработка метрик, информирующих потребителей о характеристиках качества, которые наиболее важны для оценки степени пригодности данных к использованию. Измеримых параметров всегда имеется в избытке, но далеко не все из них актуальны и стоят времени и труда, затрачиваемых на их измерение и учет. Эффективные метрики качества данных должны поддаваться количественному определению (измеримость), быть полезными для контура управления (значимость) и при выходе значения измеряемого параметра за пределы установленного допуска инициировать алгоритм управления, выделяющий полезный сигнал из зашумленного потока данных (контролируемость).

Предлагаемая методика предоставляет достаточно формализованный и вычислительно несложный алгоритм оценки качества совокупности входных параметров сложной динамической системы. Предлагаемые оценки являются эффективными метриками качества, анализ которых позволяет инициировать алгоритм управления, выделяющий полезный сигнал из зашумленного потока данных.

Предлагаемая методика применена для анализа совокупности статистических данных, характеризующих качество жизни населения субъектов Российской Федерации. Анализ показывает, что значительное число рассматриваемых параметров имеет значительную ошибку регистрации и недостаточную степень достоверности. Следовательно, использование таких данных как основы для принятия решений, без учета имеющихся искажений, привносит ошибки в оценки и прогнозы и, как следствие, приводит к значительному снижению качества принимаемых управленческих решений и способно свести к нулю их возможный позитивный эффект. В частности, вычисление композитных индексов качества системы по однократному наблюдению по объективным данным статистических измерений с помощью математических методов не предполагает устранения имеющейся шумовой компоненты данных, вследствие чего полученный результат может быть неправдоподобным.

References

[1] Skałkowski K., Zieliński K. Automatic Adaptation of SOA Systems Supported by Machine Learning. In: Camarinha-Matos L.M., Tomic S., Graça P. (eds.) Technological Innovation for the Internet of Things. DoCEIS 2013. *IFIP Advances in Information and Communication Technology*. Vol. 394. Berlin, Heidelberg: Springer; 2013. p. 61-68. https://doi.org/10.1007/978-3-642-37291-9_7

[2] Stanisz T., Drożdż S., Kwapień J. Complex systems approach to natural language. *Physics Reports*. 2024;1053:1-84. https://doi.org/10.1016/j.physrep.2023.12.002

[3] Hangos K.M., Tuza Zs. Optimal control structure selection for process systems. *Computers & Chemical Engineering.* 2001;25(11-12):1521-1536. https://doi.org/10.1016/S0098-1354(01)00716-5

[4] Xue L., Liu Z.-G. Adaptive Control for Complex Systems with Dynamics and Time-Varying Powers. *Complexity.* 2023;2023:2127312. https://doi.org/10.1155/2023/2127312

[5] Bian J., Lyu T., Loiacono A., Viramontes T.M., Lipori G.Y., Guo Y., Wu Y. Prosperi M., George T.J., Harle C.A., Shenkman E.A. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *Journal of the American Medical Informatics Association*. 2020;27(12):1999-2010. https://doi.org/10.1093/jamia/ocaa245

[6] Azeroual O., Abuosba M. Improving the Data Quality in the Research Information Systems. *International Journal of Computer Science and Information Security.* 2017;15(11):82-86. Available at: https://dspacecris.eurocris.org/retrieve/2415/Azeroual_IJCSIS_201711.pdf (accessed 14.02.2023).

[7] Fürber C. Data Quality. In: Data Quality Management with Semantic Technologies. Wiesbaden: Springer Gabler; 2016. p. 20-55. https://doi.org/10.1007/978-3-658-12225-6_3

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

[9] 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. New York, NY: Springer; 2007. p. 7-15. https://doi.org/10.1007/0-387-69505-2_2

[10] Wang R.Y., Kon H.B., Madnick S.E. Data quality requirements analysis and modeling. In: Proceedings of the 9th International Conference of Data Engineering. Vienna, Austria; 993. p. 670-677. Available at: https://web.mit.edu/tdqm/www/tdqmpub/IEEEDEApr93.pdf (accessed 14.02.2023).

[11] Redman T.C. Data Driven: Profiting from Your Most Important Business Asset. Harvard Business Press; 2008. 272 p.

[12] Fadahunsi K.P., Akinlua J.T., O'Connor S., Wark P.A., Gallagher J., Carroll C., Majeed A., O'Donoghue J. Protocol for a systematic review and qualitative synthesis of information quality frameworks in eHealth. *BMJ Open.* 2019;9(3):e024722. https://doi.org/10.1136/bmjopen-2018-024722

[13] Redman T. Data Quality: The Field Guide. Digital Press; 2001. 260 p.

[14] English L.P. Improving Data Warehouse and Business Information Quality: Methods For Reducing Costs And Increasing Profits. John Wiley and Sons; 1999. 544 p.

[15] Jugulum R. Competing with High Quality Data. Wiley; 2014. 307 p.

[16] Caballero I., Gualo F., Rodríguez M., Piattini M. BR4DQ: A methodology for grouping business rules for data quality evaluation. *Information Systems*. 2022;109:102058. https://doi.org/10.1016/j.is.2022.102058

[17] Batini C., Scannapieco M. Data Quality: Concepts, Methodologies and Techniques. *Data-Centric Systems and Applications*. Berlin: Springer; 2006. 262 p. https://doi.org/10.1007/3-540-33173-5

[18] Myers D. The Value of Using the Dimensions of Data Quality. *Information Management*. 2013. p. 1-5. Available at: https://dqmatters.com/_download/2013-04-01_the-value-of-using-the-dimensions-of-data-quality(DanMyers).pdf?src=imdm2013 (accessed 14.02.2023).

[19] Sebastian-Coleman L. Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Morgan Kaufmann; 2013. 376 p. https://doi.org/10.1016/C2011-0-07321-0

[20] Wang J., Liu Y., Li P., Lin Z., Sindakis S., Aggarwal S. Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. *Journal of the Knowledge Economy*. 2023. https://doi.org/10.1007/s13132-022-01096-6

[21] Zhgun T.V. Evaluation of Statistical Data Quality in the Problem of Calculating the Inte-gral Characteristic of a System for a Number of Observations. *Modern Information Technologies and IT-Education*. 2020;16(2):295-303. (In Russ., abstract in Eng.) https://doi.org/10.25559/SITITO.16.202002.295-303

[22] 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 Seriesthis*. 2020;1658(1):012082. https://doi.org/10.1088/1742-6596/1658/1/012082

[23] Zhgun T.V. Data transformations when constructing a composite system quality index. *Journal of Physics: Conference Seriesthis*. 2021;2052:012058. https://doi.org/10.1088/1742-6596/2052/1/012058

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

[25] Zhgun T.V. The Application of Data Transformations in the Calculation of a Composite Index of a System's Quality. *Modern Information Technologies and IT-Education*. 2021;17(3):550-563. (In Russ., abstract in Eng.) https://doi.org/10.25559/SITITO.17.202103.550-563

Поступила 14.02.2023; одобрена после рецензирования 20.04.2023; принята к публикации 24.05.2023.
Submitted 14.02.2023; approved after reviewing 20.04.2023; accepted for publication 24.05.2023.

Об авторе:

Жгун Татьяна Валентиновна, доцент кафедры прикладной математики и информатики института 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), Cand. Sci. (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.