https://doi.org/10.25559/SITITO.020.202402.455-466
УДК 519.25

Оригинальная статья

# Оценка качества системы регистрации данных

**Т. В. Жгун*, А. А. Жгун, Д. К. Проузи**
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

**Введение.** Проблема неопределённости качества системы регистрации данных, предоставляющих данные о системе, является одной из наиболее существенных проблем при построении систем управления сложными объектами. Наиболее остро такая проблема стоит при управлении слабо формализованными (мягкими) системами – социальными и социоэкономическими. Наличие такой неопределенности диктует необходимость адаптации алгоритмов управления в реальном времени в зависимости от характеристик неопределенности входных параметров управляемой системы. Неопределенность входных параметров управляемой системы определяется внутренними свойствами системы регистрации данных, представляемых обычно органами государственной статистики.

**Материалы и методы.** Предлагаемая в работе методика предоставляет формализованный и вычислительно несложный алгоритм оценки качества отдельных параметров и системы регистрации данных в целом по ряду наблюдений. Для выбранной характеристики качества определены меры (метрики), модели оценки и нормативные значения, в которых значения характеристики качества системы регистрации данных считаются нормой. Предлагаемая методика применена для анализа эффективности функционирования государственных статистических служб Российской Федерации, Великобритании, Швеции и Японии на основе анализа совокупности статистических данных, регистрируемых во всех рассматриваемых странах. Это данные, характеризующие смертность населения за 2013-2020 годы.

**Результаты исследования.** Анализ полученных результатов позволяет достоверно утверждать о разном качестве функционирования систем государственной статистики в рассматриваемых странах. Оценка системы регистрации данных в России близка к выбранному нормативному (пороговому) значению эталона. Оценка выделяет систему государственной статистики России как наилучшую из рассматриваемых и характеризует качество функционирования государственной статистики России как нормальное. Оценка качества системы регистрации данных Японии близка к этой оценке. Значит, качество функционирования этой системы тоже может быть охарактеризовано как удовлетворительное, но несколько худшее. Системы регистрации данных Великобритании и Швеции предоставляют данные менее качественные.

**Обсуждение и заключение.** Полученные оценки качества функционирования статистических служб рассматриваемых стран превышают нормативные значения, что свидетельствуют о недостаточном качестве их функционирования.

**Ключевые слова:** количественный математико-статистический анализ, качество данных, ошибки данных, метод конечных разностей, национальная государственная статистика

**Конфликт интересов:** авторы заявляют об отсутствии конфликта интересов.

**Для цитирования:** Жгун Т. В., Жгун А. А., Проузи Д. К. Оценка качества системы регистрации данных // Современные информационные технологии и ИТ-образование. 2024. Т. 20, № 2. С. 455-466. https://doi.org/10.25559/SITITO.020.202402.455-466

Original article

# Evaluating the Quality of Data Registration Systems

**T. V. Zhgun*, A. A. Zhgun, D. C. Prowse**
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

**Introduction.** The problem of uncertainty in the quality of data registration systems, which provide data about the system, is one of the most significant problems in building management systems for complex objects. This problem is most acute in the management of poorly formalized (soft) systems, such as social and socioeconomic systems. The presence of such uncertainty dictates the need to adapt control algorithms in real time depending on the characteristics of the uncertainty of the input parameters of the controlled system. The uncertainty in the input parameters of the controlled system is defined by the inherent properties of the data registration system, typically represented by government statistical bodies.

**Materials and Methods.** The methodology presented in this work offers a formalized and computationally simple algorithm for assessing the quality of individual parameters and the overall data registration system, based on a series of observations. For the chosen quality characteristic, specific measures (metrics), evaluation models, and standard values have been established, where the system's quality characteristic values are deemed standard. The proposed methodology is applied to analyze the effectiveness of the functioning of the state statistical services of the Russian Federation, the United Kingdom, Sweden, and Japan based on the analysis of a subset of statistical data registered in all the countries considered. These are data characterizing the mortality of the population for 2013-2020.

**Results.** The results of this analysis conclusively demonstrate varying quality levels in the state statistical systems of the countries in question. The assessment of the data registration system in Russia is close to the chosen normative (threshold) value of the standard. This evaluation ranks the Russian state statistical system as the best among those reviewed, characterizing its functioning as normal. Japan's data registration system's quality is close to Russia's, suggesting its performance is satisfactory but slightly inferior. In contrast, the data registration systems in the United Kingdom and Sweden deliver significantly lower quality data.

**Discussion and Conclusion.** The quality assessments of these countries' statistical services far exceed the normative values, indicating their functioning is subpar.

**Keywords:** composite quantitative statistical analysis, data quality, data error, method of finite differences, national statistical services

**Conflict of interests:** The authors declares no conflict of interest.

**For citation:** Zhgun T.V., Zhgun A.A., Prowse D.C. Evaluating the Quality of Data Registration Systems. *Modern Information Technologies and IT-Education.* 2024;20(2):455-466. https://doi.org/10.25559/SITITO.020.202402.455-466

## Введение

Государственные статистические службы занимают особое положение среди производителей статистической информации, т.е. являются источниками официальной статистической информации.

На протяжении длительного времени государственная статистика фиксирует множество параметров большого количества различных социальных и экономических систем. При синтезе адаптивной системы управления такими системами органам управления необходимо иметь объективную оценку качества отдельных регистрируемых параметров и всей системы регистрации данных в целом.

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. Качество данных является актуальной проблемой в области Интернета в целом [1-3], Интернета вещей [4-6], социальных сетей [7-9], искусственного интеллекта [10-12], промышленности [10], [13-14], информационных системах [15-17]. Наибольшее количество публикаций посвящено проблемам качества данных в области больших данных [18-20] и здравоохранения [21-24].

Data Quality (качество данных) – характеристика, показывающая степень пригодности данных к использованию. Соответствующими международному стандарту качества данных ISO 8000 считаются «переносимые данные, удовлетворяющие предъявляемым требованиям»[1]. 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[2] [25-29].

Разногласие мнений относительно того, какие именно параметры определяют качество данных, определяется сложной и неоднородной природой данных и областью их применения [30]. 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[3]. Такое количество параметров говорит скорее о том, что единого подхода для измерения качества не существует и измерение качества зависит от контекста использования этих данных. Следовательно, для формирования контура адаптации в системе управления подход к измерению качества данных должен формироваться контекстом задачи.

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 мировой статистической практике принята концепция качества данных, основанная на принципе максимального удовлетворения потребностей пользователей. Исходя из этого принципа и в соответствии с международными рекомендациями и стандартами[4] в качестве критериев качества статистической информации называет: востребованность; достоверность точность оценок показателей; своевременность; доступность; интерпретируемость; сопоставимость; согласованность.

Из обозначенных восьми позиций только одна имеет числовые характеристики – точность оценок. Точность отдельного параметра оценивается стандартной ошибкой среднего (*standard error*, *SE*), коэффициентом вариации и доверительным интервалом. Остальные характеристики в значительной степени являются субъективными и зависят от знаний экспертов. Для сравнения, стандарт ISO/IEC 25023:2016 качество программного обеспечения измеряет только количественными объективными, а не субъективными оценками экспертов: мерой размера ПО, мерой времени выполнения компонента, мерой усилий (производительность труда, трудоемкость и др.), мерой учета (количество ошибок, число отказов, ответов системы и др.)[5]. Введенные метрики позволяют объективно оценить совокупные качество программного продукта.

В мировой статистической практике отсутствует концепция оценки качества работы статистических служб. В современном мире во всех странах функционируют свои статистические службы. На 47-й сессии в Женеве 15 апреля 1992 г. Европейской экономической комиссией были утверждены основные принципы официальной статистики. Позже эти принципы (после внесения незначительных поправок) были одобрены Статистической комиссией ООН. В настоящее время эти десять принципов составляют общепризнанную основу для работы национальных статистических управлений, а также для статистической деятельности официальных международных организаций[6].

Наряду с принципами актуальности и объективности, профессионализма, конфиденциальности источников, экономической эффективности показателей, законности мер, в рамках

[1] ГОСТ Р ИСО 8000-100-2019 Качество данных. Часть 100. Основные данные. Обмен данными характеристик. Обзор: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 29 октября 2019 г. № 836-ст: введен впервые: дата введения 2020-05-01 / подготовлен ФГУП «СТАНДАРТИНФОРМ». М.: Стандартинформ, 2019. 20 с.

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

[3] DAMA-DMBOK: Data Management Body of Knowledge / ed. by. D. Henderson, S. Earley. 2nd Ed. New Jersey : Technics Publications, 2017. 624 p.

[4] Об утверждении Методологических положений по организации процессов производства официальной статистической информации : Приказ Росстата от 07.12.2018 № 732 [Электронный ресурс]. URL: https://www.consultant.ru/document/cons_doc_LAW_313411/ (дата обращения: 13.02.2024).

[5] ISO/IEC 25023:2016 Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of system and software product quality [Электронный ресурс] // ISO, 2016. URL: https://www.iso.org/standard/35747.html (дата обращения: 13.02.2024).

[6] Руководящие принципы в отношении комплексной экономической статистики. ST/ESA/STAT/SER.F/108. Нью-Йорк: Организация Объединенных Наций, 2013. 145 с. [Электронный ресурс]. URL: https://unstats.un.org/unsd/nationalaccount/docs/ies-guidelines-r.pdf (дата обращения: 13.02.2024).



которых функционируют статистические системы, сформулированные принципы требуют осуществления координации деятельности статистических ведомств на национальном и международном уровнях и осуществления двустороннего и многостороннего сотрудничества в области статистики. Несмотря на схожие цели и задачи, применение стандартизированных классификаторов и методов статистического обследования, методы проведения статистических обследований и системы показателей, применяемые в статистических ведомствах разных стран, весьма различаются. Это делает крайне затруднительной сравнение эффективности функционирования национальных статистических систем.

Эффективность функционирования национальных статистических систем может оцениваться экспертами. Эксперты журнала “The Economist” в 1993 г. составили рейтинг (табл. 1) национальных статистических служб 13 развитых стран, в основу которого легли такие факторы, как оперативность опубликования данных и точность показателя ВВП, измеряемого средним процентом корректировки показателя. В составлении рейтинга приняли участие руководители статистических служб, представители МВФ и других международных институтов, а также отдельные пользователи статистической информации. Логика авторов составления рейтинга непрозрачна и представляется субъективной. (Если Канада с минимальным изменением показателя ВВП и плохой оперативностью – 9 место из 13 – занимает лидирующую позицию, то почему рейтинг Австралии опережает Францию и Англию, которые имеют лучшие показатели по обоим измерениям?). Других оценок эффективности функционирования национальных статистических систем в литературе практически нет.

Т а б л и ц а 1. **Рейтинг систем государственной статистики по странам**
T a b l e 1. **Ranking of state statistical systems by country**

| № п/п | Страна | Рейтинг | % изменения ВВП при корректировке | Рейтинг оперативности |
|---|---|---|---|---|
| 1 | Канада | 1 | 0,2 | 9 |
| 2 | Австралия | 2 | 0,6 | 11 |
| 3 | Нидерланды | 3 | 1 | 5 |
| 4 | Франция | 4 | 0,3 | 5 |
| 5 | Швеция | 5 | нет данных | 7 |
| 6 | Англия | 6 | 0,3 | 2 |
| 7 | Германия | 6 | 0,8 | 2 |
| 8 | США | 6 | 3 | 1 |
| 9 | Япония | 9 | 0,8 | 10 |
| 10 | Швейцария | 10 | 0,4 | 4 |
| 11 | Италия | 11 | 0,3 | 8 |
| 12 | Испания | 12 | нет данных | 13 |
| 13 | Бельгия | 12 | нет данных | 12 |

*Источник:* The good statistics Quide // The Economist. 1993. September, 11. P. 68.

Итак, результаты оценивания качества данных государственной статистики неизвестны, и вопрос об уровне качества публикуемых статистических данных и вопрос о качестве функционирования системы регистрации данных остаются открытыми. Поэтому разработка показателей качества, позволяющих однозначно характеризовать рассматриваемую совокупность данных и систему регистрации данных в целом, является актуальной проблемой. Для набора характеристик качества должны быть определены соответствующие меры (метрики), модели их оценки и референсные (нормативные) значения для измерения отдельных атрибутов качества. Референсные значения представляют собой те пределы, в которых значения характеристики считаются нормой. Они могут выражаться либо конкретным диапазоном числовых параметров, в которые должен попасть результат, либо давать ответ «положительно» или «отрицательно» относительно рассматриваемого параметра качества.

## Меры качества данных

Дискуссия о количественных характеристиках порождает использования понятий меры и метрика, которые полезно различать. Известно, что мера является количественной характеристикой какого-либо свойства объекта. Метрика вычисляется по значениям опорных характеристик (мер) и позволяет оценить, в какой степени объект обладает заданными свойствами. Метрика позволяет проводить количественные сравнения. На основе вычисленной метрики можно сказать, что объект А ближе к объекту В, чем к объекту С, и принимать решения на основе этой информации. Что более важно в контексте рассматриваемой проблемы оценки качества данных: метрика позволяет характеризовать близость к некоторому эталонному объекту, целевому положению. В зависимости от соответствия целевому положению можно дать качественную характеристику оцениваемого объекта: ближе – лучше, дальше – хуже. Полезность метрики зависит от выбора метода

определения «расстояния». Различные метрики могут быть более или менее подходящими в зависимости от конкретного применения или конкретных свойств объектов, которые мы изучаем.

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. При разработке метрик качества данных следует учитывать следующие характеристики:

- измеримость: параметры качества должны быть измеримыми, ожидаемые результаты должны поддаваться количественному определению;
- значимость для потребителя: 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.

Метрики, обладающие перечисленными свойствами, назовем эффективными метриками.

В справочнике[7] 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 меры актуальности информации, меры допустимости информации и т.д. Более очевидной характеристикой качества данных в приведенном списке является «полнота» данных, но и она нуждается в определении объективной меры. Приведенные характеристики **не имеют меры для измерения**.

Единственным критерием качества статистических данных, имеющим меру для измерения, является точность оценок показателей, которая определяет ошибку, связанную с выборкой, и измеряется:

- стандартной ошибкой выборки;
- относительной стандартной ошибкой;
- доверительным интервалом;
- коэффициентом вариации оценки.

По наблюдаемой выборке определяется стандартная ошибка выборки, относительная стандартная ошибка и предельная ошибка выборки.

Стандартная ошибка среднего (*standard error, SE*) показывает, насколько отклоняется в среднем параметр выборочной совокупности от соответствующего параметра генеральной совокупности (истинного значения) и определяется формулой

$$SE = \frac{\sigma_в}{\sqrt{n}}. \quad (1)$$

где $\sigma_в$ – выборочная оценка стандартного отклонения.

Относительная стандартная ошибка *(relative standard error, RSE)* – это стандартная ошибка, деленная на среднее выборочное значение:

$$RSE = \frac{SE}{\bar{x}} = \frac{1}{\sqrt{n}} \cdot \frac{\sigma_в}{\bar{x}}. \quad (2)$$

Значение характеристики в процентах от среднего помогает показать, является ли важной ошибка измерения или нет. Для этой характеристики можно отметить наличие референсного значения: национальный центр статистики здравоохранения США NCHS не сообщает среднее значение, если относительная стандартная ошибка превышает 30% [31, 32]. Однако общепринятого нормативного значения для этого параметра нет.

Величина доверительного интервала $\left[\bar{x} - \Delta x, \bar{x} + \Delta x\right]$ для средней генеральной совокупности определяется предельной ошибкой *(limit error, LE)* выборки. Предельная ошибка выборки является максимально возможной при заданной доверительной вероятности ошибкой и рассчитывается по формуле:

$$LE = \Delta x = t \cdot SE = \frac{t}{\sqrt{n}} \cdot \sigma_в \quad (3)$$

где $t$ – коэффициент доверия, значения которого определяются доверительной вероятностью. Если объем выборки большой, можно применить знания о нормальном распределении при рассмотрении выборочного среднего. В этом случае обычно используется уровень доверительной вероятности 95% и $t = 1{,}96$.

Более информативно рассматривать относительную предельную ошибку *(relative limit error, RLE)* по отношению к выборочному среднему

$$RLE = \frac{\Delta x}{\bar{x}} = \frac{t \cdot SE}{\bar{x}} = \frac{t}{\sqrt{n}} \cdot \frac{\sigma_в}{\bar{x} \cdot}. \quad (4)$$

Для аналитического сравнения наборов данных с сильно отличающимися средними величинами используют коэффициент вариации. Коэффициент вариации *(Coefficient of Variation, CV)* – это мера относительного разброса случайной величины относительно среднего значения. Он показывает, какую долю составляет средний разброс случайной величины от среднего значения этой величины. Коэффициент вариации определяется как безразмерное отношение стандартного отклонения к выборочному среднему

$$CV = \frac{\sigma_в}{\bar{x}}. \quad (5)$$

Чем больше значение коэффициента вариации, тем относительно больший разброс и меньшая выравненность исследуемых значений. Если коэффициент вариации меньше 10%, то

[7] DAMA-DMBOK: Data Management Body of Knowledge / ed. by. D. Henderson, S. Earley. 2nd Ed. New Jersey : Technics Publications, 2017. 624 p.



изменчивость вариационного ряда принято считать незначительной, от 10% до 20% относится к средней, больше 20% и меньше 33% к значительной и если коэффициент вариации превышает 33%, то это говорит о неоднородности информации и необходимости исключения из выборки экстремальных значений.

Пример вычисления стандартных оценок точности для показателей, характеризующих некоторые стороны качества жизни населения 85 субъектов России и 48 префектур Японии за 2010-2019 годы приведен в таблице 2. Из 43 доступных показателей японской статистики только 10 показателей совпадают с номенклатурой показателей Росстата.

Для корректного сравнения переменные следует привести к единому масштабу [31]. Обычно переменные рассматривают на единичном отрезке [0, 1]. В этом случае вычисленное значение математического ожидания может быть близко к нулю, что даст и близкое к нулю значение вычисляемых характеристик. Кроме этого, результат логарифмирования на единичном диапазоне сильно искажает распределение значений переменных и не определен для минимального (нулевого) значения показателя. Чтобы избежать этих проблем, анализируемые данные стоит рассматривать на стандартном интервале, не содержащем ноль. Удобно рассматривать интервал [1,100]. Соответствующее преобразование переменной имеет следующий вид

$$x_i' = \frac{99 \cdot (x_i - m)}{(M - m)} + 1, \text{ где } M = max\, x_i, m = max\, x_i \qquad (6)$$

Максимум и минимум переменной берутся по выборке для всех наблюдений. При возможном применении операции логарифмирования преобразование (6) позволит избежать давления малых значений переменных, которые после непосредственного логарифмирования переменных на стандартном отрезке [0,1] имели бы большие отрицательные значения. Преобразование (6) делает возможным корректное сопоставление разных данных.

Таблица 2. **Стандартные характеристики для совпадающих показателей двух выборок России и Японии**

Table 2. **Standard characteristics for the matching indicators of two samples of Russia and Japan**

| Показатель | Оценка точности показателя | | | | | |
|---|---|---|---|---|---|---|
| | RSE, % | | RLE, % | | CV, % | |
| | Россия | Япония | Россия | Япония | Россия | Япония |
| Число зарегистрированных разбоев, грабежей, краж на 1 тыс. жителей | 0.5 | 8.6 | 0.9 | 16.9 | 4.5 | 59.7 |
| Число больных психическими расстройствами на 100 000 населения | 1.5 | 7.2 | 3 | 14.2 | 14.2 | 50.2 |
| Число умерших детей до 1 года на 1000 населения на 1000 родившихся живыми | 1.7 | 5.6 | 3.4 | 11 | 16 | 38.7 |
| Количество самоубийств (на 100 000 человек) | 1.8 | 7 | 3.6 | 13.8 | 17 | 48.6 |
| Численность пострадавших при несчастных случаях на производстве на 1000 работающих | 2.3 | 6.4 | 4.6 | 12.6 | 21.5 | 44.7 |
| ВРП на душу населения с учетом инфляции | 4.5 | 8 | 2 | 15.7 | 41.5 | 55.4 |
| Среднедушевые денежные доходы населения | 5 | 8 | 9.8 | 15.7 | 46.2 | 55.6 |
| Коэффициент естественного прироста | 5.5 | 4.4 | 10.7 | 8.7 | 50.4 | 30.7 |
| Плотность автомобильных дорог общего пользования (км дорог на 1000 км$^2$) | 5.8 | 11.7 | 11.3 | 22.9 | 53 | 80.9 |
| Число умерших от новообразований на 100 000 человек | 6.3 | 4.7 | 12.4 | 9.3 | 58.1 | 32.7 |

*Источник:* здесь и далее в статье все таблицы составлены авторами.

*Source:* Hereinafter in this article all tables were drawn up by the authors.

Все рассматриваемые характеристики, определяемые формулами (2)-(5), в сущности, являются производными коэффициента вариации, характеризующего внутреннюю природу данных, а не их качество. Например, значения коэффициента вариации для показателя «Плотность автомобильных дорог» по всей выборке за 10 лет для России составляет 53,0%, а для Японии 80,9%. Вычисленные значения коэффициента вариации не позволяют дать качественную характеристику оцениваемого параметра в сравнении лучше – хуже: нельзя заключить из полученных данных, что плотность автомобильных дорог в одной из рассматриваемых стран фиксируется лучше, чем в другой. Вычисленные значения коэффициента вариации для этого показателя говорят лишь о сильной неравномерности протяженности дорог в разных субъектах этих стран, определяемой географическими условиями. Аналогично, значения коэффициента вариации показателя «Среднедушевые денежные доходы населения» – для России 46,2%, для Японии 55,6% – говорит исключительно большом разрыве в доходах населения разных регионов, и ничего не сообщает о качестве самих данных.

Меры точности оценок показателей, измеряемые стандартной ошибкой выборки, относительной стандартной ошибкой, доверительным интервалом или предельной ошибкой выборки, коэффициентом вариации поддаются количественному опре-

делению, и, значит, измеримы. Однако результаты измерений не могут быть переведены в полезные для контура управления метрики, следовательно, вышеназванные характеристики не обеспечивают значимость и контролируемость, поэтому не могут быть названы эффективными метриками

## Метрики качества данных

Задачей исследования является определение качества набора данных, описывающих функционирование некоторой слабо формализованной системы и качество системы регистрации данных. Качество данных определяет система регистрации данных, в данном случае – это система государственной статистики. Очевидно, качество систем регистрации данных может быть определено только в сравнении «лучше – хуже», поэтому следует рассмотреть функционирование нескольких систем государственной статистики, предоставляющих данные по регионам на некотором временном интервале. В общем случае, оцениваться будут трехмерные кубы однотипных данных, предоставленных системами государственной статистики разных стран.

Для введения мер и метрик качества на основании имеющихся измерений будем сравнивать характеристики объекта с желательным идеальным объектом, описываемым без погрешностей и с нежелательным объектом, характеристики которого абсолютно случайны [32-34]. Качественны те данные, которые точно представляют конкретную систему [25-27] и, следовательно, не имеют ошибок регистрации и которые не похожи на характеристики случайного процесса.

**Определение.** Определим *точность данных* как меру совпадение характеристики набора данных с неискаженными характеристиками реального объекта (явления). Определим **меру достоверности данных** как меру несовпадения характеристики набора данных с характеристиками объекта (явления), все регистрируемые параметры которого абсолютно случайны.

Точность данных – ошибку регистрации – по ряду наблюдений можно оценить с применением аппарата конечных разностей. 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, $\Delta^k_i = \Delta^{k-1}_{i+1} - \Delta^{k-1}_i$ – конечная разность $k$-го порядка.

Обычно вместо точных значений параметра $y_n$ известны приближенные значения $y^*_n$, и, соответственно, вместо точных значений конечных разностей $\Delta^k_i$ – значения приближенных конечных разностей $\Delta^{*k}_i$. Ошибка измерений $\varepsilon_i = y^*_i - y_i$ имеет случайный характер, её величина неизвестна, но можно оценить по имеющимся наблюдениям максимальную из ошибок $\varepsilon$. 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 $\left|\Delta^{*k}_i\right| \le \left|\Delta^k_i\right| + 2^k\cdot\varepsilon$.

Если функцию, описывающую измеряемый параметр, можно аппроксимировать полиномом степени менее $k$, то значения точных конечных разностей $\Delta^k_i$ 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 и максимальная из ошибок регистрации $\varepsilon \ge \left|\Delta^{*k}_i\right| / 2^k$.

Рассмотрим теперь исследуемые данные [16]. Значения величины $x_{ij} = x_{ij}(t)$ – точные значения $j$–го признака, $j = 1...n$ для $i$–го объекта, $i = 1...m$ в момент $t$, $t = 0,...,$ k неизвестны и представлена наблюдениями с некоторыми погрешностями $x^*_{ij}(0), x^*_{ij}(1), ..., x^*_{ij}(k)$,: $x^*_{ij}(t) = x_{ij}(t) + \varepsilon_{ij}(t)$, $\varepsilon_{ij} = \max\left|\varepsilon_{ij}(t)\right|$. Вычисленная оценка представления данных для параметра $j$ объекта $i$ на промежутке наблюдения определяется соотношением: $\varepsilon^*_{ij} = |\Delta^{*k}_{ij}|/2^k$ Вычисленное значение $\varepsilon^*_{ij}$ является оценкой снизу возможной ошибки регистрации $j$-го параметра для $i$-го объекта. Характеристикой параметра $j$ будет максимальная из наблюдаемых ошибок

$$\varepsilon^*_j = \max_i\left|\varepsilon^*_{ij}\right| = \max_i\left|\Delta^{*k}_{ij}\right| / 2^k. \qquad (7)$$

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

Меру достоверности тоже оценим с помощью аппарата конечных разностей. Для этого рассмотрим поведение конечных разностей случайного процесса. Пусть имеется четное число реализаций случайного процесса $y_i$, $i = 0,..., k$. Можно показать, что если случайные величины независимы и равномерно распределены на интервале $[0, a]$, то математическое ожидание модуля $k$-ой конечной разности случайного процесса

$$M\left(\left|\Delta^k\right|\right) \le = \frac{a}{3}\cdot 2^{k-1} = \frac{a}{6}\cdot 2^k. \qquad (8)$$

Сравнение поведения абсолютных величин вычисленных приближенных конечных разностей с оценкой (8) даст оценку доли случайности в регистрируемых данных, т.е. оценит их достоверность.

**Определение.** *Мерой достоверности* наблюдаемого параметра $j$ является величина отношения математического ожидания модуля последней приближенной разности к аналогичной характеристике случайного процесса

$$\mu_j = \frac{6\cdot M\left(\left|\Delta^{*k}_j\right|\right)}{a\cdot 2^k}\cdot 100\%. \qquad (9)$$

Отношение (9) характеризует относительный вклад случайного компонента в исследуемый процесс для переменной $j$.

Если данные рассматриваются на едином интервале [1,100] [31], то меры точности, оценивающая абсолютную погрешность данных, будет совпадать с относительной оценкой погрешности. Полученные значения меры точности будут в том же диапазоне, что и переменные, что позволит трактовать их как проценты. Единый масштаб для оценок точности и достоверности позволяет указать референсные значения, превышение которых дадут основание отнести данные к недостаточно качественным. Следовательно, соотношения (7) и (9) определяют метрики качества данных. Значение введенных характеристик более 5% для набора данных будет свидетельствовать о значительном уровне искажений и случайной компоненты в сигнале и о необходимости применять методы устранения



случайных искажений – методы шумоподавления – для анализа сигнала.

Вычисленные оценки позволяют определить единую метрику качества каждой переменной

$$Q_j = (\varepsilon_j^2 + \mu_j^2)^{1/2} \quad (10)$$

Также вычисленные оценки позволят оценить характеристику системы регистрации данных, предоставившую исследуемые выборки. Среднее качество выборки из $n$ переменных, определяющее характеристику системы регистрации данных одной величиной

$$Q = \left( \sum_{j=1}^{n} (\varepsilon_j^2 + \mu_j^2)^{1/2} \right) / n . \quad (11)$$

## Результаты исследования

Показатели здравоохранения нации среди статистически показателей, характеризующих социальную систему, являются наиболее социально значимыми. Из всех регистрируемых параметров здравоохранения статистика смертности является краеугольным камнем в принятии решений на всех этапах управления системы и представляет наиболее значимый аспект здоровья населения. Правильное понимание и интерпретация статистики смертности позволяет формировать целевые политики, распределять ресурсы и оценивать общее здоровья населения. Однако целостность этих важных выводов зависит от качества исходных данных.

Один из подходов при оценке качества статистики смертности представлен в работе[8]. В работе выбираются «маркеры» – наиболее обще определенные параметры, например, «Симптомы, признаки и неточно обозначенные состояния или повреждения (без уточнений)» (по МКБ-9). Затем проводится сравнительный анализ повышения или понижения значений показателей, с использованием разбивки на различные категории и когорты. Выявляется «аномальное» по мнению экспертов в n-раз изменение определённого параметра, и этим числом характеризуется качество данных всей системы регистрации. Такой подход не может характеризовать качество системы в целом, как и не может характеризовать качество других, не рассмотренных параметров. По своей сути такой подход является субъективным.

Введенные метрики точности и достоверности данных являются объективными численными характеристиками и выгодно отличаются минимальным количеством предположений (а именно, предположением, что изменения параметра во времени описывается гладкой функцией). При проведении исследований, которые основаны на статистических данных, существует предположение об их широкой доступности. Однако, несмотря на наличие международных договоренностей, регламентирующих фиксацию данных, вопрос доступности стоит очень остро, особенно при работе с детализированными показателями. Так, в контексте данных о смертности, фиксируемых в максимально детальном разрешении (по причинам, полу, возрасту, региону), только британская и шведская статистика предоставляют информацию в удобно структурированной форме. В то же время, сервисы, способные предоставить подробную информацию в удобном для извлечения формате, зачастую либо «скрыты» в недоступных местах веб-сайтов, либо располагаются на веб-ресурсах других статистических органов, при этом взаимно не ссылаясь друг на друга.

С учетом вышеизложенного, был выбран временной интервал и показатели, обладающие приемлемой доступностью на этом интервале. Настоящее исследование включает анализ данных государственной статистики за 8-летний период (2013–2020) для административных территорий России, Японии, Великобритании и Швеции. Для Российской Федерации рассматриваются данные по 87 регионам (с учетом вариантов Тюменской и Архангельской областей как с автономными округами, так и без них), для Японии – по 47 префектурам, для Великобритании – по 174 унитарным административным единицам, для Швеции – рассматривается 21 лен.

В качестве источников данных выступили: для России – электронный выпуск ежегодного статистического бюллетеня «Естественное движение населения Российской Федерации»[9], для Великобритании – государственный агрегатор “nomis”[10], для Японии – государственный агрегатор “e-stat”[11] и данные, предоставленные Японским национальным институтом исследований народонаселения и социального обеспечения (IPSS)[12], для Швеции – Национальный совет по вопросам здравоохранения и социального обеспечения Швеции (Socialstyrelsen)[13].

Далее рассматриваются показатели смертности за 8-летний период (2013-2020) государственной статистики России, Великобритании, Японии и Швеции: В таблице 3 представлены рассчитанные значения точности данных для отдельных показателей государственной статистики смертности, полученные с использованием формулы (6). Так, показатель общей смертности обладает наименьшей точностью в Великобритании (11,2), в то время как в Японии он выявлен как наиболее точный (1,0). Из четырнадцати рассмотренных показателей Великобритания занимает последнее место по точности в 13 случаях, а Япония – в 8 случаев демонстрирует наибольшую

[8] Качество медико-статистических данных как проблема современного российского здравоохранения / В. Г. Семенова, Н. С. Гаврилова, Г. Н. Евдокушина, Л. А. Гаврилов // Общественное здоровье и профилактика заболеваний. 2004. № 2. С. 11-19.

[9] Бюллетень «Естественное движение населения Российской Федерации» [Электронный ресурс] // Федеральная служба государственной статистики, 2024. URL: https://rosstat.gov.ru/folder/11110/document/13269 (дата обращения: 13.02.2024).

[10] Nomis – Official Census and Labour Market Statistics : официальный сайт [Электронный ресурс] // Office for National Statistics, 2024. URL: https://www.nomisweb.co.uk (дата обращения: 13.02.2024).

[11] e-Stat: Portal Site of Official Statistics of Japan : официальный сайт [Электронный ресурс] // National Statistics Center, 2024. URL: https://www.e-stat.go.jp/en (дата обращения: 13.02.2024).

[12] National Institute of Population and Social Security Research (IPSS) : официальный сайт [Электронный ресурс]. URL: https://www.ipss.go.jp/index-e.asp (дата обращения: 13.02.2024).

[13] Socialstyrelsen : официальный сайт [Электронный ресурс] // The National Board of Health and Welfare, 2024. URL: https://www.socialstyrelsen.se/en (дата обращения: 13.02.2024).

точность показателей. Наименее точно во всей рассматриваемой выборке зафиксированы специфические показатели причин смертности в Великобритании: утопления на 100 тысяч населения (28,4), убийства (19,9) и транспортные происшествия на 100 тысяч населения (22,1). Можно сказать, что в Великобритании наименее точно фиксируются показатели смертности, относящиеся к классу внешних причин. Наиболее точно в рассматриваемой выборке зафиксированы специфические показатели причин смертности в Японии: Общая смертность (1,0), Новообразования (1,7), Болезни органов кровообращения (2,1).

Необходимо помнить, что метрика точности представляет собой нижнюю оценку ошибки регистрации и определяется самым слабым элементом системы регистрации.

Таблица 3. **Оценки точности данных**

Table 3. **Data Accuracy Assessments**

| Причины смертности | Метрика точности | | | |
|---|---|---|---|---|
| | Россия | Великобритания | Швеция | Япония |
| Общая смертность | 3.1 | 11.2 | 4.0 | 1.0 |
| Младенческая смертность | 6.9 | 20.6 | нет данных | 11.0 |
| Детская смертность | 3.9 | 24.1 | нет данных | 7.6 |
| Инфекционные, паразитарные болезни | 2.5 | 11.3 | 11.1 | 6.1 |
| Туберкулез | 3.0 | 27.1 | 3.5 | 9.7 |
| Новообразования | 6.2 | 15.5 | 8.4 | 1.7 |
| Болезни органов кровообращения | 4.2 | 4.4 | 4.4 | 2.1 |
| Болезни органов дыхания | 6.6 | 16.7 | 10.0 | 3.0 |
| Болезни органов пищеварения | 6.8 | 14.0 | 9.4 | 5.2 |
| Внешние причины | 4.0 | 15.8 | 11.2 | 3.9 |
| Транспортные случаи | 4.5 | 22.1 | нет данных | 9.9 |
| Утопления | 9.6 | 28.4 | 17.5 | 3.9 |
| Самоубийства | 13.7 | 11.3 | 16.1 | 8.0 |
| Убийства | 4.7 | 19.9 | 18.3 | 6.2 |

В таблице 4 представлены рассчитанные согласно формуле (8) оценки достоверности данных для тех же показателей государственной статистики смертности. Швеция отмечена как страна с наименьшей достоверностью данных в 10 из 11 случаях. С другой стороны, система регистрации данных России демонстрирует наибольшую достоверность для 10 из 14 рассматриваемых показателей. Самые низкие показатели достоверности наблюдаются для следующих категорий: в Швеции – смертность от убийств (40,6), смертность от утоплений (30,4), в Великобритании – показатель детской смертности (22,5), смертность от утоплений (22,1). Наиболее достоверными показателями во всей выборке является общая смертность в Японии (1,4).

В целом, минимальные различия (между странами) в достоверности данных наблюдаются для таких показателей, как смертность от болезни органов кровообращения и общая смертность. Наибольшие различия в достоверности данных замечены для показателей смертности от убийств, утоплений и смертности от болезни органов дыхания. Это свидетельствует о различиях в методологии и подходах к регистрации данных, применяемых в различных странах. Стоит учесть, что, в отличие от показателя точности, обозначающего нижнюю границу и отражающего слабейшее звено системы регистрации, показатель достоверности служит мерой среднего качества данных. Он дает более общую картину надежности данных, предоставляемых различными статистическими системами. С учетом этого, если позиция в рейтинге качества данного показателя меняется при переходе от оценки точности к оценке достоверности, это свидетельствует о наличии аномальных субъектов, искажающих общую статистику. Для российской статистики регионы с наименее качественными данными по причине самоубийств включают Республику Ингушетию, Ямало-Ненецкий и Чукотский автономные округа. В Великобритании регионы с наименее надежной статистикой по причине утопления включают Rotherham и Buckinghamshire, а по причине убийств – Ealing, Cambridgeshire, Leeds, Lincolnshire и Salford. Очевидным образом, причины ненадлежащего качества данных в разных станах объясняются разными причинами.

В таблице 5 представлена общая оценка качества систем регистрации данных. При выбранных допустимых пятипроцентных значениях точности и достоверности общее приемлемое качество системы регистрации оценивается величиной 7,1. Величины выше этого значения будут характеризовать систему регистрации данных как недостаточно эффективную. Оценка системы регистрации данных в России (7,9) близка к выбранному пороговому значению эталона и оценивает систему государственной статистики России как наилучшую из рассматриваемых и как функционирующую с должным качеством. Система регистрации данных Японии (9,6) также близка к этой оценке. Системы регистрации данных Великобритании и Швеции предоставляют данные значительно менее точные. Полученные средние оценки статистических служб этих стран свидетельствуют о недостаточном качестве их функционирования.

Таблица 4. **Оценки достоверности данных**
Table 4. **Data Reliability Assessments**

| Причины смертности | Метрика достоверности | | | |
|---|---|---|---|---|
| | Россия | Великобритания | Швеция | Япония |
| Общая смертность | 3.6 | 4.9 | 6.5 | 1.4 |
| Младенческая смертность | 8.1 | 17.9 | нет данных | 13.9 |
| Детская смертность | 6.4 | 22.5 | нет данных | 11.3 |
| Инфекционные, паразитарные болезни | 4.0 | 15.2 | 22.3 | 7.4 |
| Туберкулез | 2.7 | 14.4 | 4.5 | 14.5 |
| Новообразования | 4.7 | 6.6 | 12.0 | 2.1 |
| Болезни органов кровообращения | 4.5 | 5.4 | 6.9 | 3.3 |
| Болезни органов дыхания | 5.3 | 8.5 | 27.3 | 5.9 |
| Болезни органов пищеварения | 4.8 | 6.1 | 20.7 | 7.8 |
| Внешние причины | 3.9 | 13.4 | 18.7 | 5.6 |
| Транспортные случаи | 7.4 | 20.0 | нет данных | 9.5 |
| Утопления | 8.7 | 22.1 | 30.4 | 5.8 |
| Самоубийства | 6.1 | 16.8 | 23.7 | 9.1 |
| Убийства | 4.1 | 16.4 | 40.6 | 10.6 |

Таблица 5. **Суммарная оценка систем регистрации данных**
Table 5. **Summary assessment of data recording systems**

| Страна | Россия | Великобритания | Швеция | Япония |
|---|---|---|---|---|
| Показатель качества национальной системы регистрации данных | 7.9 | 22.3 | 22.1 | 9.6 |

## Обсуждение и заключение

Проблема неопределённости качества отдельных входных данных, описывающих систему и качества системы регистрации данных, предоставляющих данные о системе, является одной из наиболее существенных проблем при построении систем управления сложными объектами. Наиболее остро такая проблема стоит при управлении слабо формализованными (мягкими) системами – социальными и социоэкономическими. Наличие такой неопределенности диктует необходимость адаптации алгоритмов управления в реальном времени в зависимости от характеристик неопределенности входных параметров управляемой системы. Неопределенность входных параметров управляемой системы определяется внутренними свойствами системы регистрации данных, представляемых обычно органами государственной статистики.

Предлагаемая в работе методика предоставляет формализованный и вычислительно несложный алгоритм оценки качества отдельных параметров и системы регистрации данных в целом по ряду наблюдений. Для выбранной характеристики качества определены меры (метрики), модели оценки и нормативные значения, в которых значения характеристики качества системы регистрации данных считаются нормой. Предлагаемая методика применена для анализа эффективности функционирования государственных статистических служб Российской Федерации, Великобритании, Швеции и Японии на основе анализа совокупности статистических данных, регистрируемых во всех рассматриваемых странах. Это данные, характеризующие смертность населения за 2013-2020 годы.

Анализ полученных результатов показывает, что значительное число рассматриваемых параметров имеют 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. Кроме того, можно достоверно утверждать о разном качестве функционирования систем государственной статистики в рассматриваемых странах.

Оценка системы регистрации данных в России (7,9) близка к выбранному нормативному (пороговому) значению эталона и оценивает систему государственной статистики России как наилучшую из рассматриваемых и как функционирующую с должным качеством. Система регистрации данных Японии (9,6) также близка к этой оценке, и качество ее функционирования тоже может быть охарактеризовано как практически удовлетворительное, но несколько худшее. Системы регистрации данных Великобритании и Швеции предоставляют данные значительно менее качественные. Полученные оценки качества функционирования статистических служб этих стран значительно превышают нормативные значения, что свидетельствуют о недостаточном качестве их функционирования.

## References

[1] To A., Meymandpour R., Davis J.G., Jourjon G., Chan J. A Linked Data Quality Assessment Framework for Network Data. In: Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) (GRADES-NDA'19). New York, NY, USA: Association for Computing Machinery; 2019. Article number: 4. p. 1-8. https://doi.org/10.1145/3327964.3328493

[2] Mihindukulasooriya N., García-Castro R., Priyatna F., Ruckhaus E., Saturno N. A Linked Data Profiling Service for Quality Assessment. In: Blomqvist E., Hose K., Paulheim H., Ławrynowicz A., Ciravegna F., Hartig O. (eds.) The Semantic Web: ESWC 2017 Satellite Events. ESWC 2017. *Lecture Notes in Computer Science*. Vol. 10577. Cham: Springer; 2017. p. 335-340. https://doi.org/10.1007/978-3-319-70407-4_42

[3] Debattista J., Auer S., Lange C. Luzzu – A Methodology and Framework for Linked Data Quality Assessment. *Journal of Data and Information Quality*. 2016;8(1):4. https://doi.org/10.1145/2992786

[4] Zubair N., Niranjan A., Hebbar K., Simmhan Y. Characterizing IoT Data and its Quality for Use. *arXiv:*1906.10497. https://doi.org/10.48550/arXiv.1906.10497

[5] Luo T., Huang J., Kanhere S.S., Zhang J., Das S.K. Improving IoT Data Quality in Mobile Crowd Sensing: A Cross Validation Approach. *IEEE Internet of Things Journal*. 2019;6(3):5651-5664. https://doi.org/10.1109/JIOT.2019.2904704

[6] Karkouch A., et al. Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications.* 2016;73:57-81. https://doi.org/10.1016/j.jnca.2016.08.002

[7] Salvatore C., Biffignandi S., Bianchi A. Social Media and Twitter Data Quality for New Social Indicators. *Social Indicators Research*. 2021;156(2):601-630. https://doi.org/10.1007/s11205-020-02296-w

[8] Zengin O., Onder M.F. YouTube for information about side effects of biologic therapy: A social media analysis. *International Journal of Rheumatic Diseases*. 2020;23(12):1645-1650. https://doi.org/10.1111/1756-185X.14003

[9] Al Qundus J., Paschke A., Gupta S., Alzouby A.M., Yousef M. Exploring the impact of short-text complexity and structure on its quality in social media. *Journal of Enterprise Information Management*. 2020;33(6):1443-1466. https://doi.org/10.1108/JEIM-06-2019-0156

[10] Benedick P.-L., Robert J., Le Traon Y. A Systematic Approach for Evaluating Artificial Intelligence Models in Industrial Settings. *Sensors*. 2021;21(18):6195. https://doi.org/10.3390/s21186195

[11] Siegert V. Content-and Context-Related Trust in Open Multi-agent Systems Using Linked Data. In: Bakaev M., Frasincar F., Ko I.Y. (eds.) Web Engineering. ICWE 2019. *Lecture Notes in Computer Science*. Vol. 11496. Cham: Springer; 2019. p. 541-547. https://doi.org/10.1007/978-3-030-19274-7_42

[12] Bertossi L., Geerts F. Data Quality and Explainable AI. *Journal of Data and Information Quality*. 2020;12(2):11. https://doi.org/10.1145/3386687

[13] Xiao Q., Shan M., Xiao X., et al. Evaluation Model of Industrial Operation Quality Under Multi-source Heterogeneous Data Information. *International Journal of Fuzzy Systems*. 2020;22:522-547. https://doi.org/10.1007/s40815-019-00776-x

[14] Guo A., Liu X., Sun T. Research on Key Problems of Data Quality in Large Industrial Data Environment. In: Proceedings of the 3rd International Conference on Robotics, Control and Automation (ICRCA '18). New York, NY, USA: Association for Computing Machinery; 2018. p. 245-248. https://doi.org/10.1145/3265639.3265680

[15] 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/bitstream/11366/633/1/Azeroual_IJCSIS_201711.pdf (accessed 13.02.2024).

[16] Azeroual O., Saake G., Abuosba M. Data Quality Measures and Data Cleansing for Research Information Systems. *arXiv:*1901.06208. 2019. https://doi.org/10.48550/arXiv.1901.06208

[17] Timmerman Y., Bronselaer A. Measuring data quality in information systems research. *Decision Support Systems*. 2019;126:113138. https://doi.org/10.1016/j.dss.2019.113138

[18] Cappiello C., Samá W., Vitali M. Quality awareness for a Successful Big Data Exploitation. In: Proceedings of the 22nd International Database Engineering & Applications Symposium (IDEAS '18). New York, NY, USA: Association for Computing Machinery; 2018. p. 37-44. https://doi.org/10.1145/3216122.3216124

[19] Taleb I., Serhani M.A., Dssouli R. Big Data Quality: A Data Quality Profiling Model. In: Xia Y., Zhang L.J. (eds.) Services – SERVICES 2019. SERVICES 2019. *Lecture Notes in Computer Science*. Vol. 11517. Cham: Springer; 2019. p. 61-77. https://doi.org/10.1007/978-3-030-23381-5_5

[20] Ramasamy A., Chowdhury S. Big Data Quality Dimensions: A Systematic Literature Review. *Journal of Information Systems and Technology Management*. 2020;17:e202017003. https://doi.org/10.4301/S1807-1775202017003

[21] Pezoulas V.C., et al. Medical data quality assessment: On the development of an automated framework for medical data curation. *Computers in Biology and Medicine*. 2019;107:270-283. https://doi.org/10.1016/j.compbiomed.2019.03.001

[22] Terry A.L., et al. A basic model for assessing primary health care electronic medical record data quality. *BMC medical informatics and decision making.* 2019;19(1):30. https://doi.org/10.1186/s12911-019-0740-0

[23] Lee K., Weiskopf N., Pathak J. A framework for data quality assessment in clinical research datasets. *American Medical Informatics Association Annual Symposium Proceedings*. 2018;2017:1080-1089. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC5977591/pdf/2731442.pdf (accessed 13.02.2024).



[24] Bian J., et al. 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
[25] 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
[26] 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
[27] 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
[28] 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: IEEE Press; 1993. p. 670-677. https://doi.org/10.1109/ICDE.1993.344012
[29] Jaya I., et al. Systematic review of data quality research. *Journal of Theoretical and Applied Information Technology.* 2019;97(21):3043-3068. Available at: https://www.jatit.org/volumes/Vol97No21/13Vol97No21.pdf (accessed 13.02.2024).
[30] Yang Y., Yuan Y., Li B. Data Quality Evaluation: Methodology and Key Factors. In: Qiu M. (eds.) Smart Computing and Communication. SmartCom 2017. *Lecture Notes in Computer Science*. Vol. 10699. Cham: Springer; 2018. p. 222-230. https://doi.org/10.1007/978-3-319-73830-7_22
[31] Zhgun T.V. Data transformations when constructing a composite system quality index. *Journal of Physics: Conference Series.* 2021;2052:012058. https://doi.org/10.1088/1742-6596/2052/1/012058
[32] Zhgun T.V. Evaluation of Statistical Data Quality in the Problem of Calculating the Integral 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
[33] 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(1):012082. https://doi.org/10.1088/1742-6596/1658/1/012082
[34] 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. (In Russ., abstract in Eng.) https://doi.org/10.25559/SITITO.019.202302.393-402

*Поступила 13.02.2024; одобрена после рецензирования 18.04.2024; принята к публикации 21.05.2024.*
*Submitted 13.02.2024; approved after reviewing 18.04.2024; accepted for publication 21.05.2024.*

Об авторах:

**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
**Жгун Антонина Александровна,** 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 технических наук, **ORCID: https://orcid.org/0009-0008-7006-7960,** Antonina.Zhgun@novsu.ru
**Проузи Давид Кристофер,** ассистент 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), **ORCID: https://orcid.org/0009-0002-4659-6260,** prowsedavid@yandex.ru

*Все авторы прочитали и одобрили окончательный вариант рукописи.*

About the authors:

**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
**Antonina A. 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. (Eng.), **ORCID: https://orcid.org/0009-0008-7006-7960,** Antonina.Zhgun@novsu.ru
**David C. Prowse,** Assistant 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), **ORCID: https://orcid.org/0009-0002-4659-6260,** prowsedavid@yandex.ru

*All authors have read and approved the final manuscript.*