DOI: 10.25728/mlsd.2023.1535

ЭФФЕКТИВНЫЕ МЕТРИКИ КАЧЕСТВА СТАТИСТИЧЕСКИХ ДАННЫХ

Жгун Т.В., Жгун А.А., Проузи Д.К.
Новгородский государственный университет имени Ярослава Мудрого,
Великий Новгород, Россия
Tatyana.Zhgun@novsu.ru

Аннотация. Предлагается методика оценки качества совокупности данных с использованием аппарата конечных разностей. Для предлагаемых характеристик качества определены соответствующие меры, модели их оценки и референсные значения. Предлагаемая методика применена для анализа совокупности статистических данных, характеризующих смертность населения России, Великобритании, Швеции и Японии за 2013-2020 годы.

Ключевые слова: количественный математико-статистический анализ, измерения качества данных, конкретные показатели качества данных ошибки данных, статистические данные, метод конечных разностей.

Введение

На протяжении длительного времени государственная статистика фиксирует множество параметров большого количества различных социальных и экономических систем. При синтезе адаптивной системы управления этими 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], Интернета вещей [2], социальных сетей [3], искусственного интеллекта [4] промышленности [4,5], информационных системах [6]. Наибольшее количество публикаций посвящено проблемам качества данных в области больших данных [7] и здравоохранения [8-10].

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 [11-13].

Разногласие мнений относительно того, какие именно параметры определяют качество данных, определяется сложной и неоднородной природой данных и областью их применения [14]. 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 [15]. 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, основанная на принципе максимального удовлетворения потребностей пользователей. Исходя из этого принципа и в соответствии с международными рекомендациями и стандартами Приказ Росстата от 07.12.2018 N 732 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, производящих оценки компонентов модели. Для сравнения при оценке качества программного обеспечения стандарт ISO/IEC 9126-2 качество программного обеспечения измеряют мерой размера ПО, мерой времени выполнения компонента, мерой усилий (производительность труда, трудоемкость и др.), мера учета (количество ошибок, число отказов, ответов системы и др.). Введенные метрики позволяют оценить совокупные качество программного продукта.

Результаты оценивания качества государственной статистики неизвестны, и вопрос об уровне качества публикуемых статистических данных остается открытым. Поэтому разработка показателей качества, позволяющих однозначно характеризовать рассматриваемую совокупность данных 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 иметь ответ «положительно» или «отрицательно».

1. Введение метрик качества данных с применением конечных разностей

1.1. Различия понятий меры и метрики

Дискуссия о количественных характеристиках порождает использования понятий меры и метрика, которые полезно различать. Известно, что мера является количественной характеристикой какого-либо свойства объекта.

Метрика же является мерой «расстояния» в том смысле, что метрика вычисляется по значениям опорных характеристик и позволяет оценить, в какой степени объект обладает заданными свойствами. Понятие «расстояние», используемое в метрике, не всегда соответствует обычному физическому понятию расстояния. Скорее, это мера непохожести, различия или разделения по некоторым измерениям, определяемым опорными характеристиками. Этими характеристиками могут быть любые признаки или свойства, которые имеют числовые характеристики и которые имеют отношение к рассматриваемой задаче.

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.

1.2. Стандартные меры качества данных

В справочнике [15] 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 мер 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. Приведенные характеристики не имеют эффективных метрик для измерения.

Единственным критерием качества статистических данных 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 (SE, standard error) 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 (RSE, relative standard error) – 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% [34]. 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 (LE, limit error) 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 (RLE, relative limit error) 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 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.

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 общего пользования (км дорог на 1000 км^2)» по всей выборке за 8 лет для России составляет 53,0%, а для Японии 80,9%. Вычисленные значения коэффициента вариации не позволяют дать качественную характеристику оцениваемого объекта в сравнении лучше – хуже, а говорят лишь о сильной неравномерности протяженности дорог в разных субъектах этих стран. Также как и коэффициент вариации показателя «Среднедушевые денежные доходы населения» для России 46,2%, а для Японии 55,6% говорит исключительно большом разрыве в доходах населения регионов стран, а не о самих данных.

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

2. Метрики качества данных

Задачей исследования является определение качества набора данных, описывающих функционирование некоторой слабо формализованной системы. Качество данных определяет система регистрации данных, в данном случае это система государственной статистики. Очевидно, качество может быть определено только в сравнении, поэтому следует рассмотреть функционирование нескольких систем государственной статистики, предоставляющих данные по регионам на некотором временном интервале. В общем случае, оцениваться будет трехмерный куб данных.

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. Качественны те данные, которые точно представляют конкретную систему [11-13] 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 ***достоверность 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 $y_0, y_1, \ldots, y_k$ 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 Δ^k_i– значения приближенных конечных разностей Δ^{*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 ε. 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 $\left|\Delta_i^{*k}\right| \le \left|\Delta_i^k\right| + 2^k\cdot\varepsilon$.

Если функцию, описывающую измеряемый параметр, можно аппроксимировать полиномом степени менее *k*, то значения точных конечных разностей Δ_i^k 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 $\varepsilon \ge \left|\Delta_i^{*k}\right| / 2^k$.

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 $x^*_{ij}(0), x_{ij}{}^*(1), ..., x^*_{ij}(k)$,: $x^*_{ij}(t) = x_{ij}(t) + \varepsilon_{ij}(t)$, $\varepsilon_{ij} = \max_t \left|\varepsilon_{ij}(t)\right|$. Вычисленная оценка представления данных для параметра *j* объекта *i* на промежутке наблюдения определяется соотношением: $\varepsilon^*_{ij} = |\Delta^{*k}_{ij}|/2^k$ Вычисленное значение ε^*_{ij} является оценкой снизу возможной ошибки регистрации *j*-го параметра для *i*-го объекта. Характеристикой параметра *j* будет максимальная из наблюдаемых ошибок

$$\varepsilon^*_j = \max_i \left|\varepsilon^*_{ij}\right| = \max_i \left|\Delta^{*k}_{ij}\right| / 2^k . \qquad (6)$$

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

$$M(\left|\Delta^k\right|) \le = \frac{a}{3}\cdot 2^{k-1} = \frac{a}{6}\cdot 2^k . \qquad (7)$$

Сравнение поведения абсолютных величин вычисленных приближенных конечных разностей с оценкой (7) 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

$$\mu_j = \frac{6\cdot M(\left|\Delta^{*k}_{ij}\right|)}{a\cdot 2^k}\cdot 100\%\cdot \qquad (8)$$

Отношение (8) характеризует относительный вклад случайного компонента в исследуемый процесс для переменной *j*.

Если данные рассматриваются на едином интервале [1,100] [17], то мера 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. Следовательно, соотношения (6) и (8) определяют метрики.

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. Также вычисленные оценки позволяют определить единую метрику качества каждой переменной и оценить среднее качество выборки одной величиной:

$$Q = \left(\sum_{j=1=}^{n} (\varepsilon_j^2 + \mu_j^2)^{1/2} \right) / n\cdot \qquad (9)$$

3. Применение метрик точности и достоверности для оценки качества статистических данных Оценка качества статистических данных на примере анализа показателей смертности

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 [18]. 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. Исследование включает анализ данных государственной статистики за 8-летний период (2013–2020) для административных подразделений 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» [19], для Великобритании – государственный агрегатор «nomis» [20], для Японии – государственный агрегатор «e-stat» [21] и данные, предоставленные Японским национальным институтом исследований народонаселения и социального обеспечения (IPSS) [22], для Швеции – Национальный совет по вопросам здравоохранения и социального обеспечения Швеции (Socialstyrelsen) [23].

Далее рассматриваются показатели смертности за 8-летний период (2013–2020) государственной статистики России, Великобритании, Японии и Швеции:

В таблице 1 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 точность.

Наименее точно зафиксированы специфические показатели смертности в Великобритании, включают 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. Необходимо помнить, что метрика

точности представляет собой нижнюю оценку и определяется самым слабым элементом системы регистрации.

Таблица 1. Оценки точности данных

Причины смертности	Метрика точности			
	Россия	Великобритания	Швеция	Япония
Общая смертность	3.1	11.2	4.0	1.0
Младенческая смертность	6.9	20.6		11.0
Детская смертность	3.9	24.1		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		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

Источник: разработки авторов

В таблице 2 представлены рассчитанные согласно формуле (8) оценки достоверности данных для отдельных 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 являются 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. утопления и смертности от болезни органов дыхания. Это свидетельствует о различиях в методологии и подходах к регистрации данных, применяемых в различных странах.

Таблица 2. Оценки достоверности данных

Причины смертности	Метрика достоверности			
	Россия	Великобритания	Швеция	Япония
Общая смертность	3.6	4.9	6.5	1.4
Младенческая смертность	8.1	17.9		13.9
Детская смертность	6.4	22.5		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		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

Источник: разработки авторов

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 вероятно 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 u Salford.*

В таблице 3 представлена общая оценка качества рассматриваемых выборок. Оценка системы регистрации данных в России близка к выбранному значению эталона: при пятипроцентных значениях точности и достоверности общее качество оценивается величиной 7,1. Система регистрации данных Японии также близка к этой оценке. Системы регистрации данных Великобритании и Швеции предоставляют данные значительно менее точные.

Таблица 3. Суммарная оценка систем регистрации данных

Страна	Россия	Великобритания	Швеция	Япония
Показатель качества	7.9	22.3	22.1	9.6

Источник: разработки авторов

4. Заключение

Критически важным компонентом управления качеством данных является разработка метрик, информирующих потребителей о таких характеристиках качества данных, которые наиболее важны для оценки степени пригодности данных к использованию. Измеримых параметров качества данных всегда имеется в избытке, но далеко не все из них актуальны и стоят времени и труда, затрачиваемых на их измерение и учет. Эффективные для контура управления метрики качества данных должны поддаваться количественному определению (измеримость), быть полезными для контура управления (значимость).

Предлагаемая методика предоставляет формализованный и вычислительно несложный алгоритм оценки качества системы регистрации данных. Предлагаемая методика применена для анализа совокупности статистических данных, характеризующих смертность населения субъектов Российской Федерации, Великобритании, Швеции и Японии за 2013-2020 годы.

Анализ 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.

Введенные измерения качества обладают свойствами эффективной метрики качества: погрешность и достоверность данных измерима, значима и контролируема. Введенные метрики характеризуют разные стороны наблюдаемого процесса. Также вычисленные оценки позволяют оценить качество выборки одной величиной. Чем меньше значение вычисляемой характеристики качества Q, тем качественнее рассматриваемая выборка. Вычисленные оценки качества пула данных свидетельствуют о более высоком качестве функционирования органов регистрации данных Российской Федерации.

Литература

1. *To A., Meymandpour R., Davis J. G., Jourjon G., Chan J.* A linked data quality assessment framework for network data. // Proc. of the 2nd Joint International Workshop on Graph Data Management Experiences Systems (GRADES) and Network Data Analytics (NDA). – 2019. P. 1-8
2. *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 Journ. – 2019. –Vol. 6(3), – P.5651-5664.
3. *Salvatore C., Biffignandi S., Bianchi A.* Social media and twitter data quality for new social indicators. // Social Indicators Research. – 2021. –Vol. 156(2). – P. 601-630. doi.org/10.1007/s11205-020-02296
4. *Benedick P.L., Robert J., Le Traon Y. A* Systematic Approach for Evaluating Artificial Intelligence Models in Industrial Settings. // Sensors. – 2021. – Vol. 21 (18), – P. 6195. doi: 10.3390/s21186195
5. *Xiao Q., Shan M., Xiao X., Rao C.* Evaluation model of industrial operation quality under multi-source heterogeneous data information // International Journal of Fuzzy Systems. – 2020. – Vol, 22(2). – P.522-547.
6. *Timmerman Y., Bronselae, A.* Measuring data quality in information systems research // Decision Support Systems. – 2019. – Vol. 126. – P.1-7, 113138.

7. *Ramasamy A., Chowdhury S.* Big data quality dimensions: a systematic literature review.// JISTEM-Journal of Information Systems and Technology Management. – 2020. – Vol.17.
8. *Pezoulas V. C., Kourou K. D., Kalatzis F., Exarchos T. P., Venetsanopoulou A., Zampeli E., ... Fotiadis D. I.*. Medical data quality assessment: On the development of an automated framework for medical data curation.. //. Computers in biology and medicine. – 2019. –Vol. 107. – P.270-283.
9. *Terry A. L., Stewart M., Cejic S., Marshall J. N., de Lusignan S., Chesworth B. M., ... Thind A.* A basic model for assessing primary health care electronic medical record data quality //. BMC medical informatics and decision making. – 2019. – Vol. 19(1). P.1-11.
10. *Lee K., Weiskopf N., Pathak J.* A framework for data quality assessment in clinical research datasets // American Medical Informatics Association Annual Symposium Proceedings. – 2017.– Vol. 2017. –P. 1080).
11. *Fürber C.* "3. Data Quality". Data Quality Management with Semantic Technologies. Springer. – 2015. – P. 20–55.
12. *Batini C., Scannapieca M*. Data quality. Springer-Verlag, Berlin, Germany. – 2006. – P.19-31
13. *Herzog Thomas N., Scheuren Fritz J., Winkler William E.* What is Data Quality and Why Should We Care? Data Quality and Record Linkage Techniques. New York: Springer New York. – 2007.– P.7-15.
14. *Izham Jaya, Fatimah Sidi, Lilly Suriani Affendy, Marzanah Jabar, Iskandar Ishak* Systematic review of data quality research // Journal of Theoretical and Applied Information Technology. – 2019. – Vol. 97 (21). – P. 3043-3068
15. Data Management Body of Knowledge: 2nd Edition. Technics Publications; New Jersey. – 2017. –590 p
16. *Жгун Т. В*. Оценка качества статистических данных в задаче вычисления интегральной характеристики системы по ряду наблюдений // Современные информационные технологии и ИТ-образование. – 2020. –Т. 16. –.№2. – С.295-303
17. *Zhgun T.V*. Data transformations when constructing a composite system quality index // *J. Phys.: Conf. Ser.* , 2021/ 2052 012058 Doi: 10.1088/1742-6596/2052/1/012058
18. Семенова В.Г. 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. 11-18.
19. Бюллетень «Естественное движение населения Российской Федерации» URL: https://rosstat.gov.ru/folder/11110/document/13269 (дата обращения: 15.06.2023)
20. Официальный агрегатор британской статистики «nomis» URL: https://www.nomisweb.co.uk (дата обращения: 15.06.2023)
21. Официальный агрегатор японской статистики «e-stat» URL: https://www.e-stat.go.jp/en (дата обращения: 15.06.2023)
22. Японский национальный институт исследований народонаселения и социального обеспечения (IPSS) URL: https://www.ipss.go.jp/index-e.asp (дата обращения: 15.06.2023)
23. Национальный совет по вопросам здравоохранения и социального обеспечения Швеции (Socialstyrelsen) URL: https://www.socialstyrelsen.se/en (дата обращения: 15.06.2023)