УДК 519.25;519.248;004.6;005.6 DOI: http://doi.org/10.25728/pu.2019.3.6

# ПОИСК АНОМАЛИЙ В ЗАДАЧЕ ПОВЫШЕНИЯ КАЧЕСТВА ОТКРЫТЫХ ДАННЫХ

М.Ю. Чесноков

**Аннотация.** Отмечено, что в настоящее время растет число проектов открытых данных (ОД) — публикации в свободном доступе информации государственных органов или частных компаний для последующего использования. Один из барьеров в получении выгод от ОД состоит в наличии проблемы качества публикуемых данных. Проанализирована указанная проблема, причины ее появления, рассмотрены метрики и стратегии повышения качества ОД, предложена общая стратегия и ее имплементация для случаев наличия временного и категориального контекстов, предполагающие применение методов поиска аномалий.

**Ключевые слова:** открытые данные, качество данных, поиск аномалий.

## ВВЕДЕНИЕ

Концепция открытых данных позволяет увеличить доступность информации, прозрачность деятельности государственных и частных структур, стимулировать создание новых сервисов и услуг, вовлекать общественность в решение важных вопросов [1]. Важность концепции ОД подтверждают количественные оценки экономического эффекта от их публикации: согласно исследованию [2], общемировой совокупный эффект от использования ОД в различных областях составляет около $3—5 трлн. в год. Тем не менее, для получения максимальных эффектов от использования ОД недостаточно только их публикации, необходимо соблюдение определенных юридических, управленческих и технических условий, которые помогут преодолеть существующие барьеры в получении выгод от использования ОД [1]. Одним из них является проблема качества ОД, которая может приводить к неверным решениям, дополнительным издержкам, затруднять применение ОД. Основная причина появления ОД низкого качества заключается в незаинтересованности лиц, публикующих данные, в улучшении ситуации, что можно показать при рассмотрении ОД в аспекте теории общественных благ — в случае ОД возникает так называемая проблема «безбилетника», решение которой возможно путем формирования требований качества и стандартов публикации ОД.

Существует множество исследований, посвященных разработке методологий качества данных [3], однако они не фокусируются на общей методологии и не рассматривают важную группу метрик качества — контекстуальную корректность (см. далее табл. 1). В настоящей работе предложена общая стратегия повышения качества ОД и рассмотрена ее реализация для случаев наличия временного и категориального контекстов, предполагающая применение методов поиска аномалий [4] для контроля контекстуальной корректности.

Структура статьи: § 1 посвящен описанию метрик и проблемы качества ОД; в § 2 проанализированы причины проблем качества ОД; в § 3 рассмотрены стратегии повышения качества ОД, предложена общая стратегия и ее реализация; в § 4 рассмотрена предложенная имплементация стратегии на конкретном примере наборов опубликованных ОД.

## 1. КАЧЕСТВО ОТКРЫТЫХ ДАННЫХ

На рис. 1 представлен типичный жизненный цикл данных в проекте ОД (в котором участвует владелец ОД и конечные пользователи): сбор и подготовка данных на внутренних системах владельца; публикация на платформе ОД (портал в интернете); анализ и применение ОД конечными пользователями. Сбор данных включает в себя интеграцию информации из различных источников, подготовка — обработку и агрегацию собранных

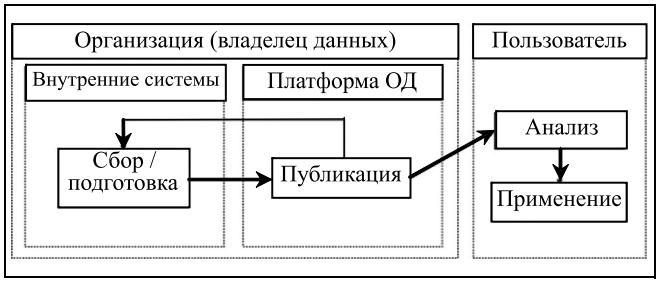


**Рис. 1. Типичный жизненный цикл открытых данных**

данных в конечный вид. Публикация открывает доступ к данным для внешних пользователей. Анализ данных пользователем может происходить на платформе ОД в случае наличия соответствующих инструментов и инфраструктуры.

На практике положительные эффекты проектов ОД ограничивают ряд проблем, среди которых одной из важнейших является проблема качества данных [1, 5—7], исследуемая в ряде работ: рассматривается качество ОД национальных порталов Чехии [8], Италии [9]; в публикациях [6, 7] авторы разрабатывают фреймворк для автоматического сбора, проверки и мониторинга качества метаданных; в работе [10] с помощью ряда методов поиска аномалий без учителя производится обнаружение ошибочных данных в проекте связных открытых данных (Linked Open Data, LOD) DBpedia [11]; в работе [12] поиск потенциально некорректных записей в LOD производится путем кластеризации с помощью метода ближайших соседей и выделения аномальных объектов из полученных кластеров.

Качество данных в общем случае определяется как «степень удовлетворения заявленных и подразумеваемых требований при использовании в определенных условиях» [8, 9]. Для измерения качества необходимо введение метрик, при этом создание универсальных метрик представляется практически сложным [3, 7]. Из анализа литературы можно выделить наиболее распространенные группы метрик качества ОД, представленные в табл. 1.

Отметим, что любые прикладные проекты всегда содержат «грязные» данные (неточные, противоречивые, неполные и т. п.), обнаружение которых является важной задачей [13], так как они могут значительно уменьшать положительные эффекты от реализации проектов ОД из-за: заключения некорректных выводов, принятия неверных решений; построения искаженных прогнозных моделей; дополнительных издержек на повторный сбор и обработку данных; снижения доверия к данным и их владельцам; сложности поиска релевантной информации, интеграции с другими источниками.

## 2. ПРИЧИНЫ ПОЯВЛЕНИЯ ПРОБЛЕМ КАЧЕСТВА ОТКРЫТЫХ ДАННЫХ

Среди основных причин появления проблем качества ОД [1, 7]: отсутствие стандартов публикации ОД и метаданных; нерепрезентативные данные в источниках (малая или неправильно собранная выборка и др.); разнородные источники, требующие интеграции; намеренное ухудшение качества данных; отсутствие ресурсов для публикации и незаинтересованность владельцев ОД в повышении качества ОД. Последняя из указанных причин становится понятной при рассмотрении ОД в рамках теории общественных благ.

Общественное благо — благо, потребление которого одним лицом не делает его недоступным для потребления другими лицами (свойство неконкурентности), и из потребления которого, невозможно в силу физических или организационных причин устранить какое-либо лицо (свойство неисключаемости) [14]. Открытые данные — общественное благо, так как они обладают свойствами

*Таблица 1*

**Метрики качества открытых данных**

| Группа метрик | Описание |
|---|---|
| Синтаксическая корректность | Близость значения к соответствующей области определения: e-mail, URL-адрес, дата |
| Непротиворечивость | Отсутствие формальных/логических несоответствий: дубликатов, соответствие типов данных |
| Полнота | Отсутствие пропусков в данных |
| Контекстуальная корректность | Соответствие контекстным ограничениям, времени, региону, категории и т. п. |
| Своевременность | Актуальность информации на текущую дату |
| Свойства ОД | Открытый формат (csv, json, xml) и лицензия, машиночитаемость, контактная информация |

неконкурентности (при использовании одним лицом ОД не становятся меньше для других) и неисключаемости (данные публикуются в открытом доступе в сети Интернет). Важно обратить внимание, что свойство неисключаемости ведет к проблеме, известной как «проблема безбилетника» [15] — потребитель осознано желает получить выгоду от общественного блага, не внося платы за него. Проблема безбилетника для ОД заключается в вопросе выделения ресурсов для их создания и обновления — и потребители, и владельцы данных заинтересованы в росте доступной информации и повышении качества публикуемых ОД, но это требует дополнительных издержек.

Существует несколько вариантов решения проблемы безбилетника [16]: 1) экономические (а) и институциональные (б) формы принуждения (например, принудительная вакцинация населения, налоги); 2) преодоление свойства неисключаемости (платная дорога); 3) формирование социальных установок («субботник»). Отметим, что вариант 1 (а) решения проблемы безбилетника в виде налогов за пользование ОД (экономическая форма принуждения) не реалистичен в силу сложности измерения получаемых выгод конкретным лицом. Всеобщая доступность — основной принцип ОД, что ведет к невозможности применения варианта 2 решения — преодоление неисключаемости, не выйдя за рамки концепции ОД (например, данные «по подписке»). Учитывая сложность формирования социальных установок в соответствии с принципами открытости в краткосрочной перспективе (вариант 3), единственный реалистичный вариант решения проблемы безбилетника для ОД заключается в институциональной форме принуждения (вариант 1 (б)), а именно введение обязательства для владельцев данных публикации определенной информации по установленным стандартам — они будут вынуждены следить за требованиями качества ОД.

## 3. РАЗРАБОТКА СТРАТЕГИИ ПОВЫШЕНИЯ КАЧЕСТВА ОТКРЫТЫХ ДАННЫХ

Отметим, что данные низкого качества могут быть следствием плохо организованных процессов создания/обновления данных. В связи с этим подходы к повышению качества данных (табл. 2) можно разделить на две большие группы [3, 9]. В первую из них входят методы, модифицирующие непосредственно данные, а во вторую — методы, модифицирующие процесс создания/обновления данных. Отметим, что в долгосрочной перспективе методы из второй группы более эффективны, так как устраняют причины появления «плохих» данных, в отличие от методов первой группы, направленных на исправление следствий «плохих» процессов создания/обновления данных. Применение методов, модифицирующих процессы, может быть очень трудоемким по сравнению с методами, модифицирующими данные [3].

*Таблица 2*

**Методы повышения качества данных**

| Модификация данных | Модификация процесса |
|---|---|
| получение новых данных;<br>стандартизация / нормализация данных;<br>создание / добавление метаданных;<br>объединение различных представлений одних и тех же данных, удаление дубликатов;<br>обнаружение ошибочных данных и их корректировка: заполнение пропусков, обработка выбросов | формирование и соблюдение требований к качеству ОД;<br>добавление ручных / автоматических процедур валидации и контроля при создании/ обновлении данных;<br>интеграция разнородных источников;<br>выбор надежных источников;<br>обратная связь с пользователями ОД;<br>версионирование данных |

В работе [3] проведено детальное сравнение и систематизация 13-ти методологий оценки и повышения качества данных. Авторы называют наиболее полными, универсальными и в то же время простыми для применения на практике две из них: TIQM (Total Information Quality Management) [17] и CDQ (Comprehensive Data Quality) [18]. Методология TIQM предполагает сбор и консолидацию всех данных в одну интегрированную базу данных, при этом устраняя ошибки в источниках. Она подразумевает наличие сформулированных требований качества и состоит из трех фаз:

— оценка: анализ требований качества, измерение качества, оценка затрат;

— улучшение: идентификация причин ошибок, разработка улучшений (стандартизация, исправление, заполнение), изменение процессов;

— управление изменениями и мониторинг: исследование удовлетворенности клиентов, анализ систематических барьеров изменений, проверка эффективности изменений.

Методология CDQ уделяет внимание выявлению процессов работы с данными, формулировке требований качества:

— определение текущего состояния: идентификация внешних и внутренних источников данных, их потребителей, описание их взаимодействия, формирование требований;

— оценка: анализ требований качества данных, измерение качества, установка целевых показателей;

— улучшение: идентификация причин ошибок, оценка затрат, выбор методов улучшения.

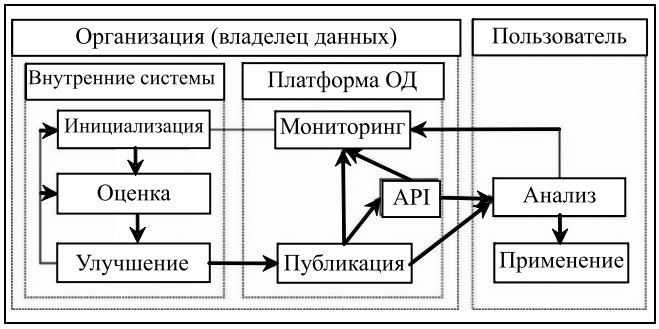


**Рис. 2. Жизненный цикл открытых данных по стратегии повышения качества**

Объединим и применим приведенные методологии к ОД, разбив их на фазы:

— инициализация: определение необходимых данных, их источников, формирование требований качества;

— оценка: сбор и подготовка данных, установка целевых показателей качества, оценка качества;

— улучшение: идентификация причин ошибок, оценка затрат на исправление данных/процесса, выбор и применение методов повышения качества (см. табл. 2), изменение требований при необходимости, повторная оценка качества;

— мониторинг: обратная связь с пользователями, измерение популярности наборов, поиск новых потребностей в данных.

Формулировка требований качества возможна на основании групп метрик (см. табл. 1) и присутствующих данных в наборах ОД. Важность формулировки требований как части методологии подчеркивается: определением «качества данных»; решением проблемы безбилетника для ОД; обобщением прочих методологий. В результате получена общая стратегия повышения качества ОД.

Приведенный ранее жизненный цикл ОД (см. рис. 1) может быть улучшен в соответствии с описанной стратегией (рис. 2), а также благодаря доступу к данным через API (англ. *Application Programming Interface*) и представление данных как в машино-, так и в человекочитаемом (визуализированном) формате.

Сравним работы, рассматривающие проблему качества данных применительно к ОД. Исследования [6, 7] сфокусированы на оценке, сравнении, мониторинге и визуализации метрик метаданных и порталов ОД, не затрагивая стратегии повышения качества самих наборов ОД. В работе [8] приводятся расплывчатые требования качества наборов ОД (например, «все записи в наборе должны быть корректные», однако для разных наборов это требование может отличаться), а также общие рекомендации по применению некоторых из методов повышения качества (см. табл. 2) без конкретных примеров. Статья [9] сфокусирована на оценке и сравнении метрик качества ОД, в ней рассматривается контекстуальная корректность и не затрагивается соответствие ОД требованиям качества. Заметим, что в литературе большинство работ сфокусированы на исследовании синтаксической корректности, непротиворечивости и полноты, контекстуальная корректность для ОД не рассматривается. Хотя именно соответствие контекстным ограничениям позволяет говорить об «адекватности» данных. В качестве контекста может выступать дата, регион и прочие категориальные переменные, присутствующие в данных. Ближе всего к этой проблематике работы [10, 12], в которых проверяется контекстуальная корректность, но не для ОД, а для связных открытых данных (для LOD — не табличной формы данных), и в их фокусе методы выявления конкретных ошибок, а не повышение качества данных в целом.

Восполняя этот пробел, реализуем общую стратегию повышения качества ОД для случая наличия контекста (категориального/временного). Для контроля выполнения требований контекстуальной корректности ОД предлагается применять методы поиска аномалий, изучаемые для разнообразных типов данных и прикладных задач [4]. Аномалиями называют объекты, не похожие на остальные, не подчиняющиеся «нормальным» паттернам поведения [4, 19]. Отметим, что недостаток всех методов поиска аномалий заключается в зависимости результатов от входных параметров и структуры данных (разная структура для различных наборов при рассмотрении сразу нескольких из них). Для нивелирования указанных недостатков целесообразно воспользоваться ансамблированием методов поиска аномалий [19]. Сравнение ансамблирования методов поиска аномалий в режиме отсутствия разметки норма/аномалия для случая временных рядов (временного контекста) представлено в работе [20]. Один из методов — метод медианного абсолютного отклонения (Median Absolute Deviation, MAD) [21] — показывает хорошие экспериментальные результаты (не всегда максимально лучшие). Его преимущества: простота применения и интерпретации, универсальность — возможность модификации под соответствующие нужды (например, наличие категориального или временного контекста).

В случае контекста по времени целесообразно применение метода MAD по скользящим окнам, т. е. для каждой точки временного ряда рассчитывается отношение ее абсолютного отклонения от медианы к MAD по окну из предшествующих во времени точек и сравнивается с заранее выбранным порогом. Рассмотрим метод подробнее. Пусть имеется временной ряд из $N$ точек $X = \{x_1, ..., x_N\}$ — последовательность значений. Тогда абсолютным

медианным отклонением от медианы по $X$ (т. е. $i = \overline{1, N}$) будет называться $MAD_X = \underset{i \in X}{median}(|x_i - \bar{x}_X|)$, $\bar{x}_X = \underset{i \in X}{median}(x_i)$. Окном размера $w$ для точки $j$ назовем набор $w$ значений, предшествующих $j$, т. е. $X_w^j = \{x_i | i \in (j - w, ..., j)\}$. Аналогично можно задать $MAD_{X_w^j}$ и $\bar{x}_{X_w^j}$ по окну $X_w^j$ и для точки $j + 1$ вычислить ее относительное отклонение от медианы на этом окне:

$$z_{j+1} = \frac{\left| x_{j+1} - \bar{x}_{X_w^j} \right|}{MAD_{X_w^j}}.$$

Далее каждой точки $i \in (w + 1, ..., N)$ присваивается метка аномальности $l_i \in \{0, 1\}$ (0 — норма, 1 — аномалия) в зависимости от того, отклоняется ли она от медианы по окну на заранее заданное $r$ — пороговое число MAD:

$$l_i = \begin{cases} 0, \text{ если } z_i \le r, \\ 1, \text{ если } z_i > r, \end{cases} \quad i \in (w + 1, ..., N).$$

В связи с указанными выше недостатками методов поиска аномалий предлагается модификация метода MAD для временных рядов: ансамблирование по пересекающимся скользящим окнам различного размера $\{w_1, ..., w_Q\}$ и различным порогам MAD $\{r_1, ..., r_P\}$. Для каждой пары $\{w_q, r_p\}$ рассчитаем бинарные метки аномальности точек $L^{qp} = \{l_{w_q+1}^{qp}, ..., l_N^{qp}\}$, $l_i^{qp} \in \{0, 1\}$, $q \in (1, ..., Q\}$, $p \in (1, ..., P)$, с помощью MAD по окну из $w_q$ предшествующих точек. Всего получим $QP$ меток для каждой точки, начиная с $\max\limits_{q \in (1, ..., Q)} w_q + 1$. Оценку аномальности $a_i$ для наблюдения $i$ получаем усреднением по $QP$ меткам:

$$a_i = \frac{1}{QP} \sum_{j=1}^{QP} l_i^j, \quad i \in (\max_{q \in (1, ..., Q)} w_q + 1, ..., N),$$

где $a_i \in (0, 1)$ интерпретируется как вероятность принадлежности наблюдения множеству аномалий. В результате аномалиями будем считать точки, для которых оценка аномальности превышает заранее заданный порог ансамблирования $a_i > a_{ens}$.

## 4. ПРИМЕРЫ ПРИМЕНЕНИЯ СТРАТЕГИИ

Применим предложенную реализацию общей стратегии для проверки уже опубликованных наборов ОД. Фазы, имплементированные из общей схемы:

— инициализация (формирование требований качества);

— оценка (установка целевых показателей качества, оценка качества);

— улучшение (корректировка записей, удаление дублей, заполнение пропусков, обработка выбросов).

Рассмотрим наборы с портала ОД правительства г. Москвы [22]. Целевыми показателями качества будем считать выполнение всех приведенных требований.

### 4.1. Набор «Нарушения ПДД, выявляемые системой фото-видео-фиксации» [23]

В наборе представлено число нарушений ПДД на дату в разрезе по группам. Помимо основных полей («Общее число нарушений по скорости» и т. п.) в наборе присутствует поле «Дата» (временной контекст). Формулировка требований качества приведена в табл. 3. Набор содержит только вре-

*Таблица 3*

**Требования качества к набору примера п. 4.1**

| Группа | № | Формулировка | Выполнено |
|---|---|---|---|
| Синтаксическая корректность | T1.1 | Даты должны соответствовать формату «дд.мм.гггг» | Да |
| | T1.2 | Основные поля должны иметь целочисленный тип данных | Да |
| Непротиворечивость | T2.1 | Основные поля должны содержать значения ≥ 0 | Да |
| | T2.2 | Даты должны соответствовать логическим ограничениям и временному периоду: дни от 1 до 31, месяцы от 1 до 12, годы от 2013 до 2018 | Нет |
| | T2.3 | В наборе должны отсутствовать дубли по датам | Да |
| Полнота | T3.1 | В наборе должны присутствовать записи на все даты с 01.01.2013 по 30.04.2018 | Нет |
| Контекстуальная корректность | T4.1 | Основные поля должны соответствовать контекстным ограничениям (времени) | Нет |

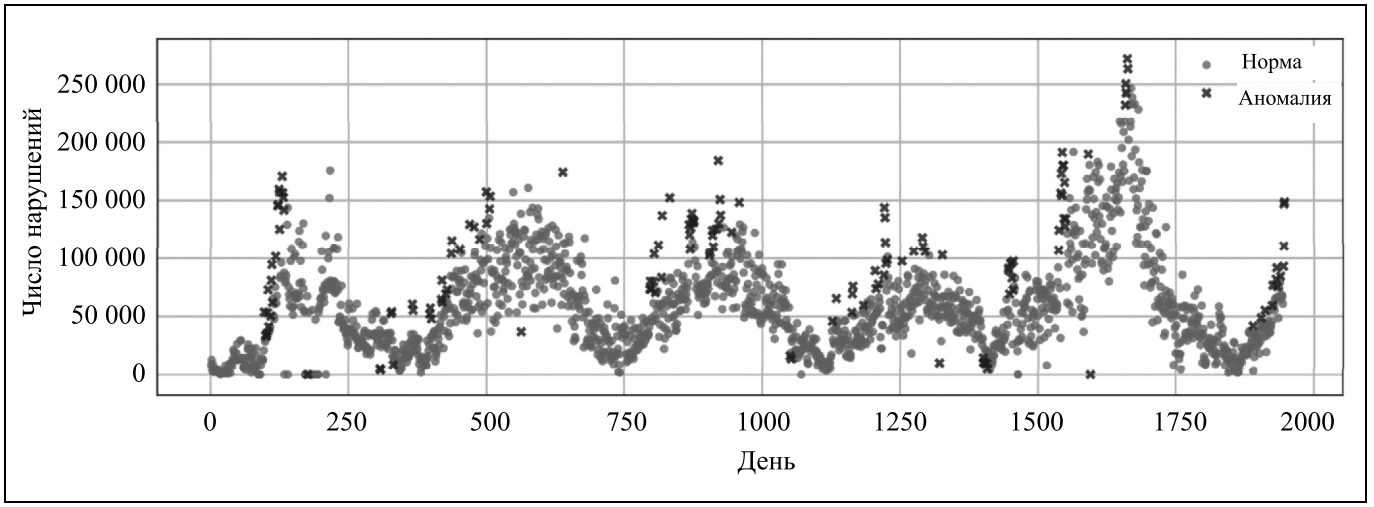


**Рис. 3. Число нарушений скоростного режима в день**

менной контекст (в разрезе дней), поэтому для проверки требования Т4.1 будем применять приведенный в § 3 метод ансамблирования MAD по скользящим окнам. Исходя из числа наблюдений в наборе (1947 дней), наличия годовой сезонности (рис. 3) и более важной ошибки пропуска аномалии, чем ошибки ложного срабатывания (так как можно сделать дополнительную проверку найденных выбросов) для ансамблирования будем использовать параметры порогов MAD: {1,5; 2,0; 2,5} и размеров окон:{30, 60, 90} с общим порогом ансамблирования 0,7. На рис. 3 представлено применение метода для ряда «Число нарушений скоростного режима».

### 4.2. Набор «Данные вызовов пожарной службы по административным округам г. Москвы» [24]

В наборе отражено число вызовов пожарно-спасательной службы помесячно в разрезе административных округов (АО) Москвы (поля: «Год», «Месяц» (текст), АО (категориальный контекст), «Число вызовов»). Соответствующие требования качества приведены в табл. 4. Набор содержит временной (в разрезе месяцев) и категориальный контекст (АО), поэтому для проверки выполнения требования Т4.1 будем пользоваться версией MAD по всем АО со стандартным порогом 3, т. е. для каждого месяца берем распределение по АО (масштабируя значения АО) и для каждого АО прове-

*Таблица 4*

**Требования качества к набору примера п. 4.2**

| Группа | № | Формулировка | Выполнено |
|---|---|---|---|
| Синтаксическая корректность | Т1.1 | Месяцы должны соответствовать шаблону написания | Нет |
| | Т1.2 | АО должны соответствовать шаблону написания | Нет |
| | Т1.3 | Поле «Число вызовов» должно иметь целочисленный тип данных | Да |
| Непротиворечивость | Т2.1 | Поле «Число вызовов» должно содержать значения ≥ 0 | Да |
| | Т2.2 | Годы должны быть от 2015 до 2018 | Да |
| | Т2.3 | В наборе должны отсутствовать дубли по времени и АО | Да |
| Полнота | Т3.1 | В наборе должны присутствовать записи на все АО на все месяцы с января 2015 по апрель 2018 | Да |
| Контекстуальная корректность | Т4.1 | Поле «Число вызовов» должно соответствовать контекстным ограничениям (времени, категории АО) | Нет |

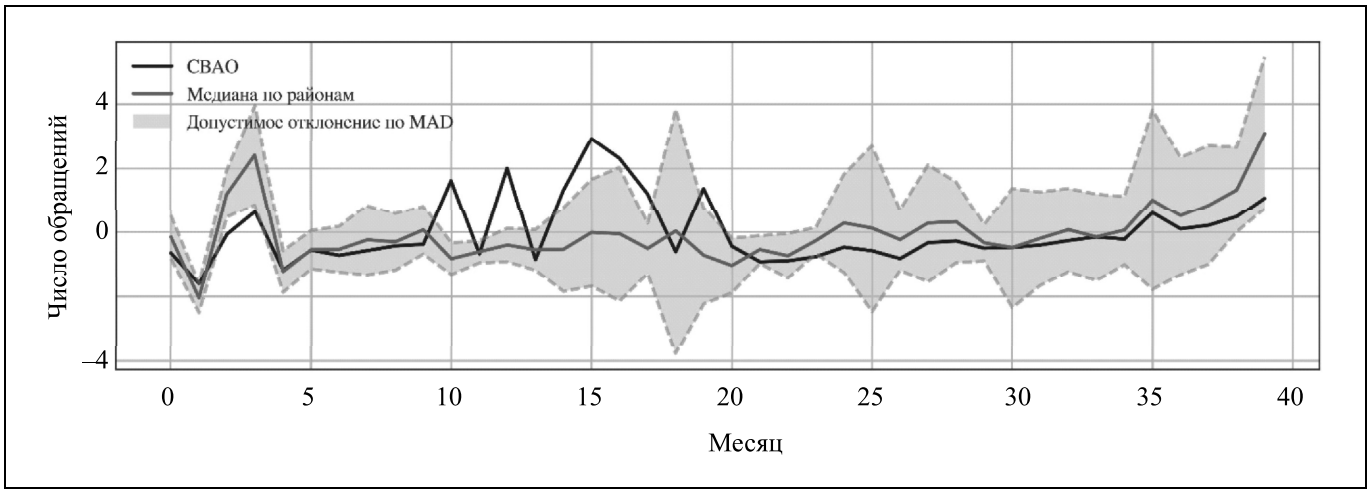


**Рис. 4. Число вызовов в месяц (масштабированное) для Северо-Восточного административного округа (СВАО)**

ряем, находится ли значение в пределах 3MAD от медианы по АО в этом месяце. Визуализация метода представлена на рис. 4 для Северо-Восточного АО (рис. 4 отличается от рис. 3, так как учитываются временной и категориальный контексты).

### 4.3. Обсуждение результатов

В результате проверки набора из п. 4.1 можно обнаружить даты, нарушающие требование Т2.2[1]. Также не удовлетворяется требование Т3.1, так как в наборе 1917 записей, при этом число дней с 01.01.2013 по 30.04.2018 составляет 1947. Кроме этого, нашлись записи, не удовлетворяющие требованию Т4.1, что отражено на рис. 3 (черные крестики). По итогам проверки набора из п. 4.2 можно заключить, что требование Т1.1 не удовлетворяется из-за различного написания месяцев. Требование Т1.2 нарушается в силу присутствия различного написания АО. Также в наборе присутствуют точки, нарушающие требование Т4.1, что показано на рис. 4, — крайние точки отрезков ломаной, выступающие за границы допустимого отклонения (серая заливка). Выявленные потенциально некорректные записи, нарушающие рассмотренные требования качества, могут быть дополнительно обработаны на фазе «улучшение».

В соответствии с выбранными целевыми показателями качества его количественной оценкой может служить относительное число выполненных требований: для набора п. 4.1 выполнены 4 из 7 (или 57 %) требований, а для набора п. 4.2 — 5 из 8 (или 62 %) требований. Обнаружение потенциально некорректных записей и относительный показатель свидетельствуют о невысоком уровне качества рассмотренных наборов ОД, что подчеркивает актуальность рассматриваемой проблематики качества ОД.

## ЗАКЛЮЧЕНИЕ

Для полноценного использования положительных эффектов открытых данных (ОД) необходима публикация качественных данных. В работе проанализированы причины появления проблемы качества ОД с точки зрения теории общественных благ («проблема безбилетника»), показано, что единственный реалистичный вариант решения данной проблемы состоит во введении стандартов публикации ОД (соблюдение требований качества). На основании методологий TIQM и CDQ предложена общая стратегия повышения качества ОД и ее имплементация для случая наличия категориального/временного контекстов, предусматривающая применение методов поиска аномалий для контроля контекстуальной корректности, в том числе модификацию метода MAD для временных рядов с ансамблированием по пересекающимся скользящим окнам различного размера и различным порогам. Применение предложенной общей стратегии позволило обнаружить ряд потенциально некорректных записей в двух опубликованных наборах ОД, что свидетельствует о невысоком уровне их качества и актуальности рассматриваемой проблемы, а также о целесообразности ее применения.

В дальнейшем предполагается применить предложенную стратегию для большего числа наборов ОД и сравнить ее с прочими методологиями повышения качества данных.

[1] Не соответствующие отчетному периоду «03.01.0016», «15.09.2105», «24.12.2018».

## ЛИТЕРАТУРА


1. *Janssen, M., Charalabidis, Y., Zuiderwijk, A.* Benefits, adoption barriers and myths of open data and open government // Information systems management. — 2012. — Vol. 29, No. 4. — P. 258—268.
2. *Manyika, J., Chui, M., Farrell, D.* Open data: Unlocking innovation and performance with liquid information. — McKinsey Global Institute. — October 2013. — Vol. 21. — 116 p.
3. *Batini, C., Cappielo, C., Francalanci, C., Maurino, A.* Methodologies for data quality assessment and improvement // ACM Computing Surveys. — 2009. — Vol. 41, No. 3. — P. 16.
4. *Chandola, V., Banerjee, A., Kumar, V.* Anomaly detection: A survey // ACM Computing Surveys. — 2009. — Vol. 41, No. 3. — Article No. 15.
5. *Волков А.И., Рейнгольд Л.А.* Открытые данные: проблемы и решения // Прикладная информатика. — 2014. — № 3 (51). — С. 5—12. [*Volkov, A., Reingold, L.* Open data: problems and solutions // Journal of Applied Informatics. — 2014. — No. 3 (51). — P. 5—12. (In Russian)]
6. *Umbrich, J., Neumaier, S., Polleres, A.* Quality assessment and evolution of open data portals. — URL: https://aic.ai.wu.ac.at/~polleres/publications/umbr-etal-2015OBD.pdf.
7. *Neumaier, S., Umbrich, J., Polleres, A.* Automated quality assessment of metadata across open data portals // Journal of Data and Information Quality. — 2016. — Vol. 8, No. 1. — Article No. 2.
8. *Kučera, J., Chlapek, D., Nečaský, M.* Open government data catalogs: Current approaches and quality perspective // Inter. Conf. on Electronic Government and the Information Systems Perspective EGOVIS/EDEM — Springer-Verlag, Berlin, Heidelberg, 2013. — P. 152—166.
9. *Vetrò, A., Canova, L., Torchiano, M.,* et al. Open data quality measurement framework: Definition and application to Open Government Data // Government Information Quarterly. — 2016. — Vol. 33, No. 2. — P. 325—337.
10. *Wienand, D., Paulheim, H.* Detecting incorrect numerical data in DBpedia // European Semantic Web Conference ESWC 2014. — Springer, Cham, 2014. — P. 504—518.
11. *URL:* http://wiki.dbpedia.org/about.
12. *Debattista, J., Lange, C., Auer, S.* A Preliminary Investigation Towards Improving Linked Data Quality Using Distance-Based Outlier Detection // Joint Inter. Semantic Technology Conf. JIST 2016. — Springer, Cham, 2016. — P. 116—124.
13. *Chu, X., Ilyas, I.F., Krishnan, S., Wang, J.* Data cleaning: Overview and emerging challenges // Proc. of the 2016 Int. Conf. on Management of Data. — P. 2201—2206.
14. *Бусыгин В.П., Желободько Е.В., Цыплаков А.А.* Микроэкономика — третий уровень: учеб. пособие. — Новосибирск: Изд-во СО РАН. — 2005. — 704 с. [*Busygin, V.P., Zhelobodko, E.V., Tsyplakov, A.A.* Microeconomics. — Novosibirsk: Publishing house SB RAS. — 2005. — 704 p. (In Russian)]
15. *Baumol, W.J.* Welfare Economics and the Theory of the State / The encyclopedia of public choice. — Boston, MA: Springer, 2004. — P. 937—940.
16. *Гринберг Р.С., Рубинштейн А.Я.* Индивидуум & Государство: экономическая дилемма. — М.: Весь Мир, 2014. — 480 с. [*Grinberg, R.S., Rubinshtein, A. Ya.* Individuum & Gosudarstvo: ekonomicheskaya dilemma. — Moscow: Ves' Mir, 2014. — 480 p. (In Russian)]
17. *English, L.* Total information quality management: A complete methodology for IQ management // Dm Review. — 2003. — Vol. 9, No. 3.
18. *Batini, C., Cappielo, C., Francalanci, C., Maurino, A.* A comprehensive data quality methodology for web and structured data // Int. J. of Innovative Computing and Applications. — 2008. — Vol. 1, No. 3. — P. 205—218.
19. *Aggarwal, C.C., Sathe, S.* Outlier Ensembles: An Introduction. — Springer, 2017. — 276 p.
20. *Чесноков М.Ю.* Поиск аномалий во временных рядах на основе ансамблей алгоритмов DBSCAN // Искусственный интеллект и принятие решений. — 2018. — № 1. — С. 99—107. [*Chesnokov, M. Yu.* Poisk anomalii vo vremennykh ryadakh na osnove ansamblei algoritmov DBSCAN // Iskusstvennyi intellekt i prinyatie reshenii. — 2018. — No. 1. — P. 99—107. (In Russian)]
21. *Leys, C., Ley, C., Klein, O.*, et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median // J. of E. Social Psychology. — 2013. — Vol. 49, No. 4. — P. 764—766.
22. *URL:* https://data.mos.ru/.
23. *URL:* https://data.mos.ru/opendata/7704786030-narusheniya-pdd-vyyavlyaemye-s-ispolzovaniem-avtomaticheskoy-sistemy-fotovideofiksatsii-narusheniy-pdd (version 3.30 от 07.05.2018).
24. *URL:* https://data.mos.ru/opendata/7710474791-dannye-vyzovov-pojarnoy-slujby-po-ao-goroda-moskvy (version 1.34 от 8.05.2018).


*Статья представлена к публикации членом редколлегии В.М. Вишневским.*

**Чесноков Михаил Юрьевич** — Московский физико-технический институт (государственный университет), г. Долгопрудный, ✉ mikhail.chesnokov@phystech.edu.

*Поступила в редакцию 17.07.2018, после доработки 24.12.2018. Принята к публикации 6.02.2019.*

# ANOMALY DETECTION FOR OPEN DATA QUALITY IMPROVEMENT

M. Yu. Chesnokov

Moscow Institute of Physics and Technology, Russia,
✉ mikhail.chesnokov@phystech.edu

**Abstract.** It is noted that an increasing number of Open Data (OD) projects are making governmental and corporate data available to public with free access and reuse. One of the barriers of getting benefits from OD is the quality of published data. This problem and its causes are analyzed, metrics and strategies of improvement of the quality of OD are considered, the general strategy using anomaly detection techniques and its' implementation for cases of time and categorical contexts are proposed.

**Keywords**: open data, data quality, anomaly detection.