Известия СПбГЭТУ «ЛЭТИ». 2026. Т. 19, № 8. С. 79–88
LETI Transactions on Electrical Engineering & Computer Science. 2026. Vol. 19, no. 8. P. 79–88

УДК 621.396.06 Научная статья

https://doi.org/10.32603/2071-8985-2026-19-8-79-88

Гибридная методика очистки данных проектного мониторинга на основе интегрального индекса качества и метода анализа иерархий

Д. Л. Тукеев[1], **О. В. Афанасьева**[1], **И. М. Новожилов**[2], **Е. А. Прокина**[1]✉, **О. С. Волгина**[1]

[1] Санкт-Петербургский горный университет императрицы Екатерины II, Санкт-Петербург, Россия

[2] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина), Санкт-Петербург, Россия

✉ elizvtprk@gmail.com

Аннотация. Цель исследований состоит в разработке новой методики очистки данных проектного мониторинга, обеспечивающей оценку качества записей до их загрузки в информационные системы. Для этого используются методы системного анализа, семантического анализа текстов, робастной статистики и анализа иерархий. Предложена модель очистки данных на основе интегрального индекса качества (IQI), которая объединяет синтаксические, семантические и логические критерии. Разработанный алгоритм рассчитывает индекс IQI и позволяет принимать обоснованные решения о допуске записей. Данные, полученные вследствие эксперимента, демонстрируют значительное повышение полноты обнаружения дефектов при сохранении высокой точности, что позволяет выявлять даже скрытые ошибки и эффективно применять наш подход в системах предварительной очистки данных.

Ключевые слова: очистка данных, качество данных, проектный мониторинг, интегральный индекс качества, анализ иерархий, семантический анализ, обнаружение аномалий

Для цитирования: Гибридная методика очистки данных проектного мониторинга на основе интегрального индекса качества и метода анализа иерархий / Д. Л. Тукеев, О. В. Афанасьева, И. М. Новожилов, Е. А. Прокина, О. С. Волгина // Изв. СПбГЭТУ «ЛЭТИ». 2026. Т. 19, № 8. С. 79–88. doi: 10.32603/2071-8985-2026-19-8-79-88.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Original article

Hybrid Data Cleaning Method for Project Monitoring Based on an Integrated Quality Index and the Analytic Hierarchy Process

D. L. Tukeev[1], **O. V. Afanasieva**[1], **I. M. Novozhilov**[2], **E. A. Prokina**[1]✉, **O. S. Volgina**[1]

[1] Empress Catherine II Saint Petersburg Mining University, Saint Petersburg, Russia

[2] Saint Petersburg Electrotechnical University, Saint Petersburg, Russia

✉ elizvtprk@gmail.com

Abstract. The aim of this study is to develop a novel data cleaning method for project monitoring that enables data quality assessment prior to loading into information systems. The approach is based on methods of system analysis, semantic text analysis, robust statistics, and the Analytic Hierarchy Process. A data cleaning model based on an integrated quality index (IQI) is proposed, combining syntactic, semantic, and logical criteria. The developed algorithm calculates the IQI and supports informed decision-making regarding data acceptance. Experimental results demonstrate a significant increase in defect detection recall while maintaining high precision, enabling the identification of hidden errors and effective application of the proposed approach in data pre-cleaning systems.

© Д. Л. Тукеев, О. В. Афанасьева, И. М. Новожилов, Е. А. Прокина, О. С. Волгина, 2026

Keywords: data cleaning, data quality, project monitoring, integrated quality index, analytic hierarchy process, semantic analysis, anomaly detection

For citation: Hybrid Data Cleaning Method for Project Monitoring Based on an Integrated Quality Index and the Analytic Hierarchy Process / D. L. Tukeev, O. V. Afanasieva, I. M. Novozhilov, E. A. Prokina, O. S. Volgina // LETI Transactions on Electrical Engineering & Computer Science. 2026. Vol. 19, no. 8. P. 79–88. doi: 10.32603/2071-8985-2026-19-8-79-88.

Conflict of interest. The authors declare no conflict of interests.

Введение. Объем данных в проектных организациях за последние годы заметно вырос. Он формируется в системах проектного мониторинга на различных этапах жизненного цикла работ. Данные поступают из распределенной сети источников, включая подрядные организации, внутренние подразделения и автоматизированные системы учета. Качество этих данных оказывает непосредственное влияние на достоверность аналитических отчетов и эффективность управленческих решений [1].

Большинство существующих подходов к контролю качества данных ориентировано либо на синтаксическую валидацию и контроль схем данных, либо на локальное выявление аномалий в числовых признаках, где семантические расхождения терминов и логические противоречия между атрибутами записи часто рассматриваются изолированно, без интеграции в единый механизм принятия решений. В среде проектного мониторинга данный разрыв – критический. Здесь данные проходят через несколько уровней агрегации, и даже небольшие искажения на входе способны масштабироваться. Значительная часть инцидентов, связанных с некорректной аналитикой, имеет корни именно в ранних дефектах исходных записей, причем речь идет не только о банальных опечатках, но даже и о более сложных случаях: несогласованности терминов между подразделениями, расхождениях в интерпретации показателей или логических нестыковках в календарно-сетевых моделях.

Отдельного внимания заслуживает феномен комплексно-дефектных данных – записей, в которых одновременно присутствуют несколько типов искажений. Это могут быть корректные по форме, но противоречивые по содержанию значения, или, наоборот, логически допустимые, но неверно закодированные параметры. Классические ETL-методы, как правило, не справляются с подобными случаями: они ориентированы на стандартизированные проверки и слабо учитывают предметный контекст, более того, очистка на завершающих этапах обработки часто сопровождается потерей части информации, которая могла бы быть полезна для объяснения отклонений. Алгоритм исключает значимые аномалии из выборки либо не обеспечивает их идентификацию [2], [3].

Существующие подходы к валидации данных недостаточно эффективны при выявлении семантических и логических нарушений – таких, как терминологические расхождения и противоречия в календарно-сетевых показателях. В результате значительная часть дефектов либо не выявляется, либо обнаруживается с запаздыванием, поэтому актуальна задача разработки методов очистки данных, обеспечивающих оценку качества записей на этапе их формирования [4], [5].

Цель работы – разработать многокритериальную модель оценки качества данных проектного мониторинга на основе интегрального индекса качества данных IQI, объединяющего синтаксические, семантические и логические критерии.

Новизна предложенного подхода заключается в переносе контроля качества на стадию формирования записи (пре-ETL). В отличие от существующих решений, здесь используется трехкомпонентный индекс, где веса критериев верифицированы методом Саати (AHP), что минимизирует субъективность оценки.

Постановка задачи. На основе анализа существующих подходов к валидации данных формулируется задача разработки метода очистки, обеспечивающего комплексную оценку качества данных проектного мониторинга до их загрузки в централизованные системы. Требуется разработать модель, позволяющую учитывать синтаксические, семантические и логические аспекты качества данных, а также формировать интегральную оценку, пригодную для автоматического принятия решений о допуске записей.

Формализация задачи и модель индекса качества данных. Данные проектного мониторинга рассматриваются как множество записей вида

$$R = \{r_1, r_2, \ldots, r_n\}.$$

Табл. 1. Типы дефектов данных проектного мониторинга
Tab. 1. Types of data defects in project monitoring

Тип дефекта	Пример	Метод обнаружения	Частный критерий
Синтаксический	Пустое обязательное поле, некорректная дата	Проверка формата и ограничений	K_{syn}
Семантический	Опечатка, неканоничный термин	Левенштейна, TF-IDF, cosine similarity	K_{sem}
Логический	Дата окончания раньше даты начала, аномальные трудозатраты	Бизнес-правила, robust Z-score	$K_{\log}$

Каждая запись r_i описывается вектором атрибутов $r_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, где x_{ij} – значение j-го признака, характеризующего параметры проектной деятельности.

Корректность данных задается системой ограничений

$$C = C_{\text{syn}} \cup C_{\text{sem}} \cup C_{\log}.$$

Соответственно, множество дефектов, типы которых расписаны в табл. 1, может быть представлено в виде $E = E_{\text{syn}} \cup E_{\text{sem}} \cup E_{\log}$.

Для оценки качества записи вводится интегральный индекс

$$\text{IQI}(r_i) = w_{\text{syn}} K_{\text{syn}}(r_i) + \\ + w_{\text{sem}} K_{\text{sem}}(r_i) + w_{\log} K_{\log}(r_i),$$

где $w_{\text{syn}} + w_{\text{sem}} + w_{\log} = 1$; w_{syn}, w_{sem} и $w_{\log}$ – весовые коэффициенты синтаксического, семантического и логического критериев соответственно, удовлетворяющие условию нормировки (сумма весов равна 1). Их значения (0.106; 0.260 и 0.633) определены методом анализа иерархий (см. далее) и отражают трехкратное превосходство значимости логического критерия над синтаксическим.

Все частные критерии нормированы в интервале [0, 1], где значение 1 соответствует полному выполнению требований качества, а значение 0 – их полному нарушению [6], [7].

Для практического применения модели вводится система пороговых значений:

$\text{IQI}(r_i) \tau_1 \Rightarrow$ запись принимается;

$\tau_2 \le \text{IQI}(r_i) < \tau_1 \Rightarrow$ запись принимается с предупреждением;

$\text{IQI}(r_i) < \tau_2 \Rightarrow$ запись отклоняется.

τ_1 и τ_2 ($\tau_1 > \tau_2$) – верхний и нижний пороги интегрального индекса. Значения порогов, установленные экспериментальной калибровкой на синтетическом датасете: $\tau_1 = 0.85$, $\tau_2 = 0.70$.

Синтаксический критерий отражает соответствие записи формальным требованиям и определяется как

$$K_{\text{syn}}(r_i) = \sum_{j=1}^{m} \lambda_j s_j(r_i),$$

где $s_j(r_i) \in [0,1]$ – результат проверки ограничений; λ_j – вес ограничения.

Семантический критерий оценивает терминологическое соответствие записи глоссарию предметной области [7], [8]. Для атрибута используется комбинированная оценка

$$s_{\text{sem}}(x_{ij}) = \alpha K_{\text{lev}}(x_{ij}, g^*) + (1-\alpha) K_{\text{tfidf}}(x_{ij}, g^*),$$

где $K_{\text{lev}}(x,g) = 1 - \dfrac{d_{\text{lev}}(x,g)}{\max(|x|,|g|)}$; $K_{\text{tfidf}}(x,g) = \dfrac{\mathbf{x}\cdot\mathbf{g}}{\|\mathbf{x}\|\|\mathbf{g}\|}$,

x – значение текстового атрибута записи (например, наименование работы); g – эталонный (канонический) термин глоссария предметной области; g^* – ближайший к x термин глоссария по выбранной мере сходства.

Обе меры близости – $K_{\text{lev}}(x, g*)$ и $K_{tfidf}(x, g*)$ – вычисляются относительно одного и того же ближайшего термина g^* и затем объединяются в комбинированную оценку через коэффициент $\alpha \in [0;1]$, где α – доля вклада лексической составляющей (Левенштейн), а $(1-\alpha)$ – доля вклада контекстной составляющей (косинусное сходство TF-IDF). В работе принято $\alpha = 0.45$, что обеспечивает преимущественную чувствительность к синонимическим заменам и переформулировкам при сохранении значимого вклада лексической близости для отлова опечаток.

Интегральная оценка имеет вид

$$K_{\text{sem}}(r_i) = \sum_{j=1}^{p} \mu_j s^{\text{sem}}(x_{ij}), \sum_{j=1}^{p} \mu_j = 1,$$

где p – число семантически значимых текстовых атрибутов записи (в модели $p = 2$: наименование работы и категориальный статус); μ_j – вес j-го

атрибута, причем $\mu_1 = 0.80$ и $\mu_2 = 0.20$. Веса отражают предметную мотивацию: содержательная корректность наименования работы существенно важнее формальной согласованности статуса для семантической интерпретации записи.

Логический критерий учитывает бизнес-правила и статистические отклонения следующим образом:

$$K_{\log}(r_i) = \beta K_{\text{rule}}(r_i) + (1-\beta) K_{\text{anom}}(r_i),$$

$$K_{\text{rule}}(r_i) = \sum_{k=1}^{q} \gamma_k \, \text{rule}_k(r_i)$$

– доля выполненных бизнес-правил (rule-based составляющая).

Это взвешенная сумма локальных бинарных оценок с весами γ_k по q учитываемым правилам (в реализации $q = 3$: непротиворечивость порядка дат, согласованность статуса и процента выполнения, правдоподобие длительности). K_{anom} – робастная оценка аномальности количественных атрибутов (anomaly-based-составляющая, определяется через медианное абсолютное отклонение MAD, см. далее); $\beta \in [0;1]$ – коэффициент баланса между двумя составляющими. В исследовании принято $\beta = 0.60$, т. е. приоритет отдан формальным бизнес-правилам над статистической аномальностью.

Компонента аномалий определяется через робастную оценку:

$$Z_{\text{rbst}} = \frac{|x - \text{median}|}{1.4826 \cdot \text{MAD}}.$$

Таким образом, модель объединяет формальные, терминологические и логические процедуры проверки качества данных и их результаты в единую интегральную оценку, что позволяет выявлять более широкий спектр дефектов по сравнению с традиционными методами валидации [9], [10].

Определение весовых коэффициентов методом анализа иерархий. В рамках модели интегрального индекса качества данных IQI частные критерии $K_{\text{syn}}, K_{\text{sem}}$ и $K_{\log}$ показывают различные аспекты корректности данных и, очевидно, обладают неодинаковой значимостью. Прямое задание весов в данном эксперименте приведет и к искажению итоговой оценки, поэтому требуется немного более формализованный подход к их определению [11], [12].

Для вычисления весов коэффициентов применяется метод анализа иерархий (AHP), который позволяет преобразовать экспертные оценки важности критериев в количественную форму.

С учетом специфики проектного мониторинга принимается следующая система приоритетов:

– логические дефекты оказывают наибольшее влияние на достоверность данных;

– семантические нарушения влияют на корректность интерпретации;

– синтаксические ошибки, как правило, менее критичны.

На основе принятой системы приоритетов матрица попарных сравнений принимает вид

$$A = \begin{pmatrix} 1 & \frac{1}{3} & \frac{1}{5} \\ 3 & 1 & \frac{1}{3} \\ 5 & 3 & 1 \end{pmatrix}.$$

Далее выполняется нормализация матрицы и вычисление средних значений по строкам, что позволяет получить вектор приоритетов

$$\mathbf{w} = \begin{pmatrix} w_{\text{syn}} \\ w_{\text{sem}} \\ w_{\log} \end{pmatrix} \approx \begin{pmatrix} 0.106 \\ 0.260 \\ 0.633 \end{pmatrix}.$$

Полученные значения удовлетворяют условию нормировки

$$w_{\text{syn}} + w_{\text{sem}} + w_{\log} = 1.$$

Для проверки согласованности экспертных оценок определяется максимальное собственное значение $\lambda_{\max}$, а также индекс согласованности:

$$\text{CI} = \frac{\lambda_{\max} - n}{n-1}; \ \lambda_{\max} \approx 3.041; \ \text{CI} \approx 0.0205,$$

где n – количество сравниваемых элементов (порядок матрицы).

Отношение согласованности рассчитывается как

$$\text{CR} = \frac{\text{CI}}{\text{RI}},$$

где RI – случайный индекс согласованности (для $n = 3$, $\text{RI} = 0.58$).

В результате $\text{CR} \approx 0.035$.

Поскольку выполняется условие $\text{CR} < 0.1$, матрица попарных сравнений считается согласованной, а полученные весовые коэффициенты – корректными. Таким образом, полученные веса критериев

$$w_{\text{syn}} = 0.106; \quad w_{\text{sem}} = 0.260; \quad w_{\log} = 0.633.$$

Наибольший вклад в итоговую оценку качества данных вносит логический критерий, что соответствует особенностям проектного мониторинга [13], [14].

Расчет индекса качества данных выполняется в виде последовательного алгоритма, изображенному при помощи блок-схемы на рис. 1.

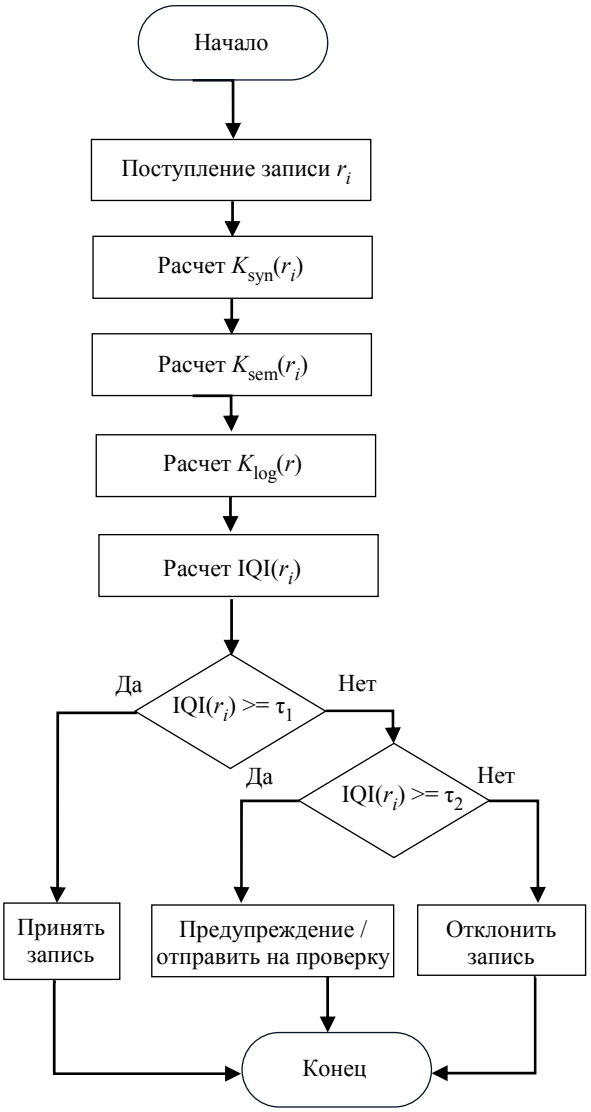


Рис. 1. Блок-схема алгоритма
Fig. 1. Flowchart of the algorithm

Данный алгоритм можно описать следующим образом:

1. На вход подается запись r_i.

2. Для каждой записи последовательно вычисляются значения критериев K_{syn}, K_{sem} и K_{log}.

3. На основе полученных значений определяется интегральный индекс:

$$\mathrm{IQI}(r_i) = w_{syn} K_{syn}(r_i) + \\ + w_{sem} K_{sem}(r_i) + w_{log} K_{log}(r_i).$$

4. Полученное значение индекса сравнивается с заданными порогами τ_1, τ_2.

5. В зависимости от результата сравнения принимается решение о дальнейшей обработке записи.

Представленный алгоритм реализует поэтапную процедуру оценки качества данных проектного мониторинга. Сначала осуществляется проверка на соответствие формальным ограничениям, после чего выполняется сопоставление значений с терминологическим глоссарием предметной области. На завершающем этапе анализируются логические зависимости и выявляются возможные аномалии.

Такой подход позволяет не только обнаруживать отдельные типы дефектов, но и формировать итоговую оценку качества записи в интегральной форме, учитывающую различные аспекты корректности данных, т. е. обнаруживать как явные, так и скрытые дефекты.

Синтаксические дефекты типичны на этапе ввода и первичной выгрузки данных: пропуск обязательного поля (например, project_code), некорректный формат даты («31.02.2025», «не_дата»), строковое значение вместо числового в трудозатратах, код раздела, отсутствующий в справочнике ({АС, ТХ, ГП, КЖ, ОВ, ВК}). Семантические дефекты – следствие свободного текстового ввода наименований работ: опечатка («бетон*о*рование» вместо «бетонирование»), неканоничная формулировка («монтаж задви*жки*»), подмена термина синонимом, не зафиксированным в словаре. Логические дефекты – наиболее опасный класс, поскольку запись формально корректна и терминологически согласована, но внутренне противоречива: дата окончания раньше даты начала; статус «завершено» при проценте выполнения 30 %; трудозатраты на порядок выше медианы по данному виду работ; нарушение последовательности этапов. Важно, что перечисленные классы не исключают друг друга: запись может одновременно содержать нарушения нескольких типов (например, сочетание семантических и логических дефектов), что и обосновывает необходимость интегральной, а не одноуровневой оценки качества.

Экспериментальная оценка эффективности предложенной модели. Для оценки предложенного подхода был сформирован синтетический набор данных, моделирующий типичные ситуации, возникающие при проектном мониторинге. В выборку включено 500 записей, из кото-

рых около 39 % содержат различные дефекты. Каждая запись описывает этап работ и включает текстовое описание, временные параметры, процент выполнения, количественные показатели (например, объем, трудозатраты); статус.

Структура тестового датасета:

1. Число записей – 500.

2. Доля дефектных записей – 39 % (195 записей); корректных – 61 % (305 записей).

3. Структура дефектных записей по категориям: только синтаксические – 50 (25.6 % от дефектных); только семантические – 60 (30.8 %); только логические – 60 (30.8 %); одновременно семантические и логические – 25 (12.8 %). Пересечения с синтаксическими дефектами в датасете намеренно не предусмотрены – для однозначного соответствия записи и уровня модели, на котором она должна обнаруживаться.

4. Итоговое число внедренных дефектов по типам (с учетом пересечений): синтаксических – 50, семантических – 85, логических – 85.

5. Базовый метод сравнения (Baseline) – rule-based-синтаксическая валидация.

6. Предлагаемый метод – многокритериальный индекс качества IQI.

В эксперименте сопоставлялись два подхода: базовая синтаксическая валидация (Baseline) и предложенная модель на основе интегрального индекса качества (IQI). Для сравнения используются метрики точности и полноты.

Точность вычислялась по формуле

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}};$$

полнота определялась как

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

где TP – корректно обнаруженные дефекты; FP – ложные срабатывания; FN – пропущенные дефекты [14]. В табл. 2 представлены результаты эксперимента.

Табл. 2. Результаты работы методов
Tab. 2. Results of the methods

Метод	Метрика			
	Precision	Recall	F1-score	Accuracy
Baseline	1.00	0.195	0.326	0.686
IQI	1.00	0.667	0.800	0.870

Обсуждение результатов. Полученные в ходе исследования результаты эксперимента с большой долей вероятности подтверждают первоначальную гипотезу о том, что синтаксическая валидация – необходимое, но недостаточное условие обеспечения качества данных.

Применение в одной плоскости двух уровней контроля – семантического и логического, позволяет существенным образом расширить область выявления дефектов.

Полученные экспериментальные данные показывают, что полнота выявления ошибок в этом случае увеличивается на 50–55 % по сравнению с общепринятыми, классическими схемами валидации данных.

Детализация результатов в рамках эксперимента подтверждает этот вывод на уровне отдельных типов дефектов и итоговых решений модели.

Полнота обнаружения по типам дефектов: на синтаксических дефектах (50 записей) Baseline и IQI показывают одинаковый результат – Recall = = 76.0 % (оба метода используют пересекающийся набор формальных проверок); на семантических дефектах (85 записей) Baseline дает Recall = = 0 %, IQI – 49.4 %; на логических дефектах (85 записей) Baseline дает Recall = 0 %, IQI – 88.2 %. Прирост общей полноты обеспечен исключительно подключением семантического и логического уровней контроля.

Распределение решений IQI-модели по итогам обработки 500 записей: accept – 403 записи (80.6 %); warning – 32 записи (6.4 %); return – 0 записей (0 %); block – 65 записей (13.0 %), что соответствует критическому логическому нарушению (нарушение порядка дат либо статус «завершено» при прогрессе ниже 95 %).

Дополнительно проведен анализ вклада частных критериев: средняя разница значений между корректными и дефектными записями составляет + 0.032 для K_{syn}, + 0.077 для K_{sem} и +0.144 для $K_{\log}$. Логический критерий обладает наибольшей разделительной способностью, что согласуется с его наибольшим весом AHP.

Базовый формальный поход, используемый отдельно от других, дает очевидное преимущество – высокую точность. При его использовании проверка типов данных, диапазонов знаний и форматов проводится достаточно строго и дает возможность исключить ложные срабатывания. Одновременно с этим такая точность применяемых алгоритмов приводит к тому, что значительная часть ошибок им игнорируется. К примеру, до 80 % дефектов при его использовании не фиксируются, если они не приводят к нарушениям синтаксических ограничений.

Интегральная оценка качества (модель IQI) более применима, так как она исследует запись как совокупность взаимосвязанных атрибутов и учитывает контекст данных. Ее особенность состоит в том, что она проверяет согласованность терминов, инспектирует противоречия между терминами, может уловить смысловые несостыковки, нелогичные последовательности и взаимосвязи. К примеру, правильно оформленная запись о проекте может содержать непоследовательные и/или несовместимые стадии или значения показателей, которые можно определить как «аномальные», т. е. те, что не выходят за формальные границы, но противоречат логике процесса. Такие случаи в большинстве своем проходят через базовые фильтры и могут искажать окончательный результат. Необходимо, чтобы увеличение полноты не приводил к лавинообразному росту ложных срабатываний. При подключении семантического слоя возникают «пограничные» ситуации, при которых выявляются записи, которые сложно однозначно отнести к корректным или ошибочным. По наблюдениям, обычно их доля находится в диапазонах 5–10 % и все они требуют дополнительной интерпретации [15]. Поэтому система начинает работать скорее как экспертная логика, а не как механический фильтр.

Главная причина различий между походами основывается на их базовой архитектуре. Упор на фиксированный набор правил и отсутствие анализа контекста – это база для классических методов. Традиционные методы, существенным образом, опираются на формальную проверку корректности и не учитывают семантическую достоверность полученных значений. В сложных же моделях появляется второй уровень исследования – анализ взаимосвязи и интерпретация значений. Именно применение сложных, двухуровневых моделей актуально для проектного мониторинга, где формально корректные данные, прошедшие «классический фильтр», приводят к искаженным выводам, которые, в свою очередь, ложатся в основу неверных, ошибочных управленческих решений.

Таким образом, переход от проверки по единственному критерию к многослойной оценке качества обладает более высокой эффективностью. Предложенный метод дает возможность идентифицировать латентные несоответствия, формально удовлетворяющих структурным ограничениям. Его применение позволяет достичь более полного и реалистичного представления о состоянии информационных потоков и формулировать на их основе верные управленческие решения.

Выводы и заключение. Предложенная в исследовании модель очистки данных проектного мониторинга, базирующаяся на применении интегрального индекса качества IQI, более применима для целей принятия решений. Объединяя формальные, смысловые и логические критерии, она рассматривает данные как единую взаимосвязанную систему, в комплексе, исключая анализ набора отдельных полей. В отличие от традиционных методов, ее применение позволяет перенести процедуру контроля на этап формирования данных. Это в значительной степени уменьшает риск распространения ошибок и позволяет сохранить истинный контекст.

Определенные с использованием метода анализа иерархий весовые коэффициенты критериев обеспечивают обоснованность модели и позволяют учесть различную значимость типов выявленных дефектов. Разработанный в исследовании алгоритм применим для автоматизированного принятия решений о допуске данных.

Результаты проведенного эксперимента на синтетическом датасете из 500 записей (см. раздел «Обсуждение результатов») убедительно доказывают увеличение полноты выявления дефектов примерно на 50–55 % при сохранении высокой точности данного метода. При его применении наибольший эффект достигается за счет выявления скрытых и контекстных ошибок, которые не фиксируются при использовании только синтаксических методов.

Основными направлениями продолжения исследований в части применимости предложенного метода являются тестирование его на массивах реальных данных, а также более углубленное исследование методики семантического анализа с учетом специфики предметной области.

Список литературы

1. Pervukhin D. A., Lisha T. A Multi-objective ε-Constraint optimization of coal supply chain performance considering customer satisfaction in multi-layer logistics networks // Intern. J. of Eng. Trans. B Appl. 2026. Vol. 39, no. 7. P. 1716–1729. doi: 10.5829/ije.2026.39.07a.15.

2. Квасов М. Н., Митряев Г. А., Прохоров М. А. Разработка математической модели семантического анализа и синтеза данных // Изв. ТулГУ. Техн. науки. 2023. № 11. С. 262–266. doi: 10.24412/2071-6168-2023-11-262-263.

3. Буркин Н. Г. Подходы, стандарты и методологии управления качеством данных // Совр. ИТ в образовании, науке и промышленности: XXV Междунар. конф. XXIII Междунар. конкурс науч. и науч.-метод. работ, IX Междунар. конкурс «Научное школьное сообщество». М.: ООО Изд-во «Экон-Информ», 2023. С. 29–33.

4. Ilyushin Y. V., Asadulagi M.-A. M. Development of a distributed control system for the hydrodynamic processes of aquifers, taking into account stochastic disturbing factors // Water Switzerland. 2023. Vol. 15, no. 4. Art. 770. doi: 10.3390/w15040770.

5. Скуковский П. А., Афанасьева О. В. Методы машинного обучения для анализа текстовых данных на примере автоматизации поиска типовых замечаний // Моя специальность: Тез. докл. конкурса науч. работ обучающихся. Санкт-Петербург, 2025. СПб.: СПб Горный ун-т императрицы Екатерины II, 2026. С. 114–116.

6. Любицын В. Н. Повышение качества данных в контексте современных аналитических технологий // Вестн. Юж-УрГУ. Сер.: Компьютерные технологии, управление, радиоэлектроника. 2012. № 23(282). Вып. 16. С. 83–86.

7. Игошева В. А., Первухин Д. А. Процесс автоматизации деятельности отдела в IT-компании / Анализ и прогнозирование систем управления в промышленности, на транспорте и в логистике: сб. тр. XXV Междунар. науч.-практ. конф. мол. уч., студ. и аспирантов. СПб.: Медиапапир, 2025. С. 145–150.

8. Глущенко А. Г. Разработка системы автоматизации обработки заявок на техническое обслуживание на основе NLP-методов // Изв. СПбГЭТУ «ЛЭТИ». 2026. Т. 19, № 2. С. 71–83. doi: 10.32603/ 2071-8985-2026-19-2-71-83.

9. Кочегаров М. Б. Метод матричного латентно-семантического анализа для структурирования массивов текстовых данных // Вологдинские чтения. 2006. № 57. С. 7–10.

10. Николаев А. В., Кычкин А. В. Разработка и интеграция имитационной модели процесса проветривания подземного горно-добывающего предприятия в сервис управления спросом на электроэнергию // Зап. Горного ин-та. 2025. Т. 275. С. 18–29.

11. Systematic study of structural divisions of industrial enterprise using queuing systems / V. A. Navatskaya, S. K. Neyrus, M. A. Skorobogatova, M. P. Afanasiev // J. of Technol. Management and Innovation. 2023. Vol. 18, no. 3. P. 51–59. doi: 10.4067/s0718-27242023000300051.

12. Нейрус С. К., Навацкая В. А., Афанасьев М. П. Исследование структурных подразделений промышленных предприятий с помощью систем массового обслуживания // Междунар. науч.-исслед. журн. 2022. № 12(126). doi: 10.23670/IRJ.2022.126.32.

13. Давлианидзе Я. С., Гребенюк К. А. Принятие решений на основе метода анализа иерархий // Вестн. Донецкого НУ. Сер. В. Экономика и право. 2019. № 2. С. 106–113.

14. Martyshov M. I., Nikitenko D. A. Preprocessing of system monitoring data for workload analysis of HPC systems // Num. Methods and Programming. 2021. Vol. 22, no. 3. P. 230–238. doi: 10.26089/NumMet.v22r314.

15. Захарова О. И., Коробейников В. С. Интегрированное управление качеством данных в организации // Вестн. СибГУТИ. 2025. Т. 19, № 4. С. 28–47. doi: 10.55648/1998-6920-2025-19-4-28-47.

Информация об авторах

Тукеев Дмитрий Леонидович – д-р техн. наук профессор кафедры системного анализа и управления. Санкт-Петербургский горный университет императрицы Екатерины II, 21-я линия д. 2, В. О., Санкт-Петербург, 199106, Россия.
E-mail: dimleo@inbox.ru
https://orcid.org/0009-0006-8103-9733

Афанасьева Ольга Владимировна – канд. техн. наук, доцент кафедры системного анализа и управления. Санкт-Петербургский горный университет императрицы Екатерины II, 21-я линия, д. 2, В. О., Санкт-Петербург, 199106, Россия.
E-mail: afanaseva_ov@pers.spmi.ru
https://orcid.org/0000-0003-3169-4781

Новожилов Игорь Михайлович – канд. техн. наук, доцент кафедры автоматики и процессов управления СПбГЭТУ «ЛЭТИ».
E-mail: novozhilovim@list.ru
https://orcid.org/0000-0002-2056-3930

Прокина Елизавета Андреевна – студентка 4-го курса, гр. САМ-22 кафедры системного анализа и управления. Санкт-Петербургский горный университет императрицы Екатерины II, 21-я линия, д. 2, В. О., Санкт-Петербург, 199106, Россия.
E-mail: elizvtprk@gmail.com

Волгина Ольга Сергеевна – студентка 2-го курса магистратуры, гр. СИСТ-24 кафедры информационных систем и вычислительной техники. Санкт-Петербургский горный университет императрицы Екатерины II, 21-я линия, д. 2, В. О., Санкт-Петербург, 199106, Россия.
E-mail: s242553@stud.spmi.ru

Вклад авторов:

Тукеев Д. Л. – методология исследования, построение концептуальной и математической модели.

Афанасьева О. В.– разработка и проектирование методологии исследования.

Новожилов И. М. – общее руководство и координация исследовательской деятельности.

Прокина Е. А. – экспериментальные исследования, сбор данных, подготовка и создание графиков и других визуальных материалов.

Волгина О. С. – применение статистических и вычислительных методов для анализа данных, проверка результатов на достоверность и воспроизводимость.

References

1. Pervukhin D. A., Lisha T. A Multi-objective ε-Constraint optimization of coal supply chain performance considering customer satisfaction in multi-layer logistics networks // Intern. J. of Eng. Transactions B Appl. 2026. Vol. 39, no. 7. P. 1716–1729. doi: 10.5829/ije.2026.39.07a.15.

2. Kvasov M. N., Mitrjaev G. A., Prohorov M. A. Razrabotka matematicheskoj modeli semanticheskogo analiza i sinteza dannyh // Izv. TulGU. Tehn. nauki. 2023. № 11. S. 262–266. doi: 10.24412/2071-6168-2023-11-262-263. (In Russ.).

3. Burkin N. G. Podhody, standarty i metodologii upravlenija kachestvom dannyh // Sovr. IT v obrazovanii, nauke i promyshlennosti: XHV Mezhdunar. konf. XHIII Mezhdunar. konkurs nauchnyh i nauch.-metod. rabot, IX Mezhdunar. konkurs «Nauchnoe shkol'noe soobshhestvo». Mytishhi, 2023. M.: OOO Izd-vo «Jekon-Inform», 2023. S. 29–33. (In Russ.).

4. Ilyushin Y. V., Asadulagi M.-A. M. Development of a distributed control system for the hydrodynamic processes of aquifers, taking into account stochastic disturbing factors // Water Switzerland. 2023. Vol. 15, no. 4. Art. 770. doi: 10.3390/w15040770. (In Russ.).

5. Skukovskij P. A., Afanas'eva O. V. Metody mashinnogo obuchenija dlja analiza tekstovyh dannyh na primere avtomatizacii poiska tipovyh zamechanij // Moja special'nost': Tez. dokl. konkursa nauch. rabot obuchajushhihsja. Sankt-Peterburg, 2025. SPb.: SPb Gornyj un-t imperatricy Ekateriny II, 2026. S. 114–116. (In Russ.).

6. Ljubicyn V. N. Povyshenie kachestva dannyh v kontekste sovremennyh analiticheskih tehnologij // Vestn. Juzh-UrGU. Ser.: Komp'juternye tehnologii, upravlenie, radiojelektronika. 2012. № 23(282). Vyp. 16. S. 83–86. (In Russ.).

7. Igosheva V. A., Pervuhin D. A. Process avtomatizacii dejatel'nosti otdela v IT-kompanii / Analiz i prognozirovanie sistem upravlenija v promyshlennosti, na transporte i v logistike: sb. tr. XXV Mezhdunar. nauch.-prakt. konf. mol. uch., stud. i aspirantov. SPb.: Mediapapir, 2025. S. 145–150. (In Russ.).

8. Glushhenko A. G. Razrabotka sistemy avtomatizacii obrabotki zajavok na tehnicheskoe obsluzhivanie na osnove NLP-metodov // Izv. SPbGJeTU «LJeTI». 2026. T. 19, № 2. S. 71–83. doi: 10.32603/ 2071-8985-2026-19-2-71-83. (In Russ.).

9. Kochegarov M. B. Metod matrichnogo latentno-semanticheskogo analiza dlja strukturirovanija massivov tekstovyh dannyh // Vologdinskie chtenija. 2006. № 57. S. 7–10. (In Russ.).

10. Nikolaev A. V., Kychkin A. V. Razrabotka i integracija imitacionnoj modeli processa provetrivanija podzemnogo gorno-dobyvajushhego predprijatija v servis upravlenija sprosom na jelektrojenergiju // Zap. Gornogo in-ta. 2025. T. 275. S. 18–29. (In Russ.).

11. Systematic study of structural divisions of industrial enterprise using queuing systems / V. A. Navatskaya, S. K. Neyrus, M. A. Skorobogatova, M. P. Afanasiev // J. of Technol. Management and Innovation. 2023. Vol. 18, no. 3. P. 51–59. doi: 10.4067/s0718-27242023000300051. (In Russ.).

12. Nejrus S. K., Navackaja V. A., Afanas'ev M. P. Issledovanie strukturnyh podrazdelenij promyshlennyh predprijatij s pomoshh'ju sistem massovogo obsluzhivanija // Mezhdunar. nauch.-issled. zhurn. 2022. № 12(126). doi: 10.23670/IRJ.2022.126.32. (In Russ.).

13. Davlianidze Ja. S., Grebenjuk K. A. Prinjatie reshenij na osnove metoda analiza ierarhij // Vestn. Doneckogo NU. Ser. V. Jekonomika i pravo. 2019. № 2. S. 106–113. (In Russ.).

14. Martyshov M. I., Nikitenko D. A. Preprocessing of system monitoring data for workload analysis of HPC systems // Numerical Methods and Programming. 2021. Vol. 22, no. 3. P. 230–238. doi: 10.26089/NumMet.v22r314. (In Russ.).

15. Zaharova O. I., Korobejnikov V. S. Integrirovannoe upravlenie kachestvom dannyh v organizacii // Vestn. SibGUTI. 2025. T. 19, № 4. S. 28–47. doi: 10.55648/1998-6920-2025-19-4-28-47. (In Russ.).

Information about the authors

Dmitry L. Tukeev – Dr Sci. (Eng.), Professor of the Department of System Analysis and Management, Empress Catherine II Saint Petersburg Mining University, 21st line, 2, Vasilievsky Island, Saint Petersburg, 199106, Russia.
E-mail: dimleo@inbox.ru
https://orcid.org/0009-0006-8103-9733

Olga V. Afanasieva – Cand. Sci. (Eng.), Associate Professor of the Department of System Analysis and Management, Empress Catherine II Saint Petersburg Mining University, 21st line, 2, Vasilievsky Island, Saint Petersburg, 199106, Russia.
E-mail: afanaseva_ov@pers.spmi.ru
https://orcid.org/0000-0003-3169-4781

Igor M. Novozhilov – Cand. Sci. (Eng.), Associate Professor of the Department of Automation and Control Processes, Saint Petersburg Electrotechnical University.
E-mail: novozhilovim@list.ru
https://orcid.org/0000-0002-2056-3930

Elizaveta A. Prokina – student, Empress Catherine II Saint Petersburg Mining University, 21st line, 2, Vasilievsky Island, Saint Petersburg, 199106, Russia.
E-mail: elizvtprk@gmail.com.

Olga S. Volgina – second-year master's student, Empress Catherine II Saint Petersburg Mining University, 21st line, 2, Vasilievsky Island, Saint Petersburg, 199106, Russia.
E-mail: s242553@stud.spmi.ru

Author contribution statement:
Tukeev D. L. – research methodology, development of the conceptual and mathematical model.
Afanasyeva O. V. – development and design of the research methodology.
Novozhilov I. M. – general management and coordination of research activities.
Prokina E. A. – experimental research, data collection, preparation and creation of graphs and other visual materials.
Volgina O. S. – application of statistical and computational methods for data analysis, verification of results for reliability and reproducibility.

Статья поступила в редакцию 02.04.2026; принята к публикации после рецензирования 19.05.2026; опубликована онлайн 23.10.2026.

Submitted 02.04.2026; accepted 19.05.2026; published online 23.10.2026.