УДК 005.63

ПРОПУСКИ В ЦЕЛЕВЫХ ПРИЗНАКАХ КАК ИСТОЧНИК ИНФОРМАЦИИ О КАЧЕСТВЕ: МЕТОД ДВУМЕРНОГО РЕЙТИНГА ФАКТОРОВ

© 2026 В.Г. Мосин, М.Л. Рауш, В.Н. Козловский, А.С. Демкин

Самарский государственный технический университет, Самара, Россия

Статья поступила в редакцию 23.06.2026

В статье представлен метод двумерного рейтинга факторов, позволяющий эффективно использовать информацию о пропусках в данных для управления качеством процессов. Подход базируется на расчете пессимистических и оптимистических оценок доли аномалий, что дает интегральный показатель подозрительности каждого фактора. Метод успешно апробирован на задачах машиностроения и сервисного обслуживания, обеспечивая приоритизацию управленческих решений в условиях неполноты данных.

Ключевые слова: Анализ данных, машинное обучение, управление качеством, статистическое управление качеством.

DOI: 10.37313/1990-5378-2026-28-4-35-40
EDN: FELOEY

1. ВВЕДЕНИЕ

В задачах управления качеством исходные данные редко бывают идеальными. Пропуски в записях о результатах контроля, неполные логи операций, отсутствующие метки о проверках — все это традиционно воспринимается исследователями и практиками как дефект данных, подлежащий устранению. Наиболее распространённые стратегии работы с пропусками сводятся либо к их удалению (listwise deletion), либо к восстановлению пропущенных значений методами импутации — от простого заполнения средним до применения более сложных моделей машинного обучения [1, 2]. Такой подход, безусловно, оправдан, когда требуется получить «чистый» датасет для построения прогностической модели или корректного статистического вывода.

Однако во многих прикладных задачах управления качеством сам факт пропуска может нести важную содержательную информацию [3, 4]. Отсутствие оценки качества может быть следствием не только технического сбоя системы регистрации, но и сознательного или подсознательного уклонения от фиксации неудовлетворительного результата. Оператор, допустивший брак, может «забыть» внести запись; контролёр, закрывающий глаза на систематические дефекты, — пропустить отметку о проверке; супервайзер, не желающий портить статистику своему сотруднику, — не выставить низкую оценку. В таких случаях пропуск превращается из статистической помехи в ценный диагностический признак, указывающий на потенциально проблемные зоны бизнес-процесса.

Настоящая работа лежит на стыке двух научно-прикладных областей — управления качеством и анализа данных. С одной стороны, мы опираемся на классические представления о контрольных картах, факторном анализе и управлении процессами; с другой стороны, предлагаемый метод является результатом осмысления пропусков не как «мусора», который нужно выбросить, а как источника управленческой информации. Такой подход соответствует современному тренду Data Quality Management, в рамках которого качество данных рассматривается не как статистическая характеристика, а как объект целенаправленного управления [4].

В настоящей работе разрабатывается подход, позволяющий использовать информацию о пропусках для ранжирования факторов по степени их влияния на целевой показатель качества.

2. ПРОБЛЕМА: ИНФОРМАЦИОННЫЙ ПОТЕНЦИАЛ ПРОПУСКОВ

Современные производственные и сервисные процессы оставляют после себя развёрнутый цифровой след. Системы учёта фиксируют параметры продукции, логи диалогов, временные метки операций, идентификаторы исполнителей и контролёров. При достаточном уровне автоматизации задача состоит уже не в том, чтобы собрать данные, а в том, чтобы корректно их проанализировать и извлечь из них управленчески значимую информацию [5].

Мосин Владимир Геннадьевич, кандидат физико-математических наук, доцент. E-mail: yanbacha@yandex.ru
Рауш Мария Леонидовна. E-mail: mary.raush@mail.ru
Козловский Владимир Николаевич, доктор технических наук, профессор, заведующий кафедрой. E-mail: Kozlovskiy-76@mail.ru,
Демкин Александр Сергеевич, аспирант. E-mail: etf3gtu@mail.ru

2.1. Пропуски данных

Пропуски в таких массивах данных — явление практически неизбежное. Природа их может быть различной: технические сбои (отказ датчика, потеря пакета данных при передаче), человеческий фактор (оператор забыл внести запись, контролёр пропустил проверку) или умышленные действия (сознательное сокрытие нежелательных результатов).

2.2. Целевые признаки

Под целевым признаком мы понимаем показатель, на основе которого принимается решение о качестве результата процесса. Природа этого показателя может быть различной: в машиностроении — отклонение геометрического размера детали от номинала.

2.3. Аномалии и пропуски — бинарные маркеры качества

Для унификации анализа мы вводим понятие бинарного маркера аномалии — out_marker. Если значение целевого признака выходит за установленные границы нормы (допуск, пороговое значение, минимально приемлемый балл), маркер принимает значение 1; в противном случае — 0.

Так, для машиностроения out_marker = 1, если диаметр вала вышел за поле допуска, для муниципального управления — если время ответа на обращение превысило регламент, а для здравоохранения — если наступило осложнение. Бинарный маркер аномалии делает анализ принципиально инвариантным к природе целевого признака и позволяет применять единый математический аппарат к самым разным задачам.

Наряду с маркером аномалии вводится бинарный маркер пропуска — nan_marker. Он принимает значение 1, если для данной записи отсутствует значение целевого признака, и 0 в противном случае. Именно этот простейший индикатор, как будет показано ниже, позволяет существенно дополнить картину анализа.

Можно предположить, что факторы, ассоциированные с повышенной долей пропусков, — это те же факторы, которые ассоциированы с повышенной долей явных аномалий, однако это не всегда так, так как пропуски могут указывать на серую зону: ситуации, в которых аномалия формально не зафиксирована, но и норма не подтверждена.

Но в любом случае — учёт nan_marker наряду с out_marker даёт более полное представление о влиянии факторов на качество процесса.

3. МЕТОД: ДВУМЕРНЫЙ РЕЙТИНГ ФАКТОРОВ

На входе имеется таблица данных, содержащая записи о результатах функционирования технологической системы или бизнес-процесса. В обязательном порядке в таблице присутствует столбец, отражающий целевой показатель качества (далее — target). Именно по его значению определяется, находится ли результат в пределах установленной нормы.

3.1. Шаг 1 — создание маркеров

Для каждой записи вводится два бинарных маркера.

- Маркер пропуска nan_marker принимает значение 1, если значение target отсутствует (пропуск), и 0 в противном случае.
- Маркер аномалии out_marker принимает значение 1, если значение target выходит за установленные границы нормы (зафиксирована аномалия), и 0 во всех остальных случаях.

Важно, что если target пропущен, то out_marker не присваивается неопределённое состояние — в коде он получает 0. Это означает, что при расчётах в оптимистичном сценарии любой пропуск интерпретируется как норма. Данный подход является осознанным допущением: мы не подменяем неизвестность произвольным значением, а откладываем её учёт на следующий шаг.

Таким образом, после первого шага в таблице появляются два столбца: nan_marker и out_marker. Пропуски не удаляются, не заполняются средними, медианами или предсказанными значениями. Они остаются пропусками. Их информационный потенциал будет раскрыт на этапе расчёта двух сценариев (раздел 3.3).

3.2. Шаг 2 — группировка и агрегация

Второй шаг заключается в переходе к анализу на уровне факторов качества. Фактором качества будем называть любую категориальную величину, значения которой (далее — градации фактора) предположительно влияют на целевой показатель. Примерами факторов являются: единица оборудования (станок), персонал (оператор), смена, режим работы, поставщик, тип сырья и т.д. Полный перечень факторов определяется спецификой решаемой задачи.

Для каждого категориального признака из заданного перечня выполняется группировка записей по его градациям. В пределах каждой группы рассчитываются три скалярные агрегаты:

1. count — общее количество записей в группе;
2. sum_nan — количество записей, для которых nan_marker = 1 (пропуски целевого показателя);
3. sum_out — количество записей, для которых out_marker = 1 (зафиксированные аномалии).

Результатом второго шага является таблица grouped_factors, где каждая строка соответствует одной градации фактора. Столбцы: factor (наименование градации), count, nan_sum, out_sum. Название исходного категориального признака на данном этапе не фиксируется — все градации из всех факторов сводятся в общий перечень для последующего ранжирования.

3.3. Шаг 3 — две оценки доли аномалий

Пессимистичная оценка (out_prop_pes) предполагает, что каждый пропуск целевого показателя фактически является аномалией. Данный сценарий соответствует наихудшему предположению: если данные отсутствуют, то причина — выход показателя за границы нормы. Формула расчёта:

$$\text{out_prop_pes} = \frac{\text{sum_nan} + \text{sum_out}}{\text{count}} .$$

Оптимистичная оценка (out_prop_opt), напротив, трактует все пропуски как отсутствие аномалии. Иными словами, если значение отсутствует, предполагается, что оно находилось в пределах нормы. Формула расчёта:

$$\text{out_prop_pes} = \frac{\text{sum_out}}{\text{count}} .$$

Таким образом, для каждой градации фактора получается пара значений (pes, opt), где pes — пессимистичная доля аномалий, opt — оптимистичная.

Разность между пессимистичной и оптимистичной оценками представляет собой долю пропусков в группе.

3.4. Шаг 4 — переход к новому базису

Рассмотрим пару значений (pes, opt), полученную на предыдущем шаге, как координаты точки на декартовой плоскости. Область возможных значений ограничена единичным квадратом: $pes \in [0, 1]$, $opt \in [0, 1]$, причём всегда выполняется условие $pes \geqslant opt$, поскольку pes учитывает пропуски, а opt — нет. Таким образом, все точки лежат в треугольной области единичного квадрата, расположенной на и ниже главной диагонали.

Введём поворот координатных осей на угол 45° против часовой стрелки. Матрица поворота имеет вид:

$$R = \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}.$$

Умножая вектор исходных координат (pes, opt) на матрицу поворота, получаем координаты точки в новом базисе. Первая координата повёрнутой системы соответствует проекции исходной точки на направление правого верхнего угла единичного квадрата. Выражение для этой координаты после упрощения принимает вид:

$$\text{rating} = \frac{\text{pes} + \text{opt}}{\sqrt{2}} .$$

Величина rating тем выше, чем ближе точка (pes, opt) расположена к правому верхнему углу единичного квадрата, то есть к максимальным значениям обеих оценок доли аномалий. Фактор, имеющий высокий рейтинг, обладает одновременно высокой пессимистичной и высокой оптимистичной оценкой доли аномалий, что указывает на систематический характер отклонений.

3.5. Шаг 5 — ранжирование

На завершающем шаге метода выполняется упорядочивание всех градаций факторов по убыванию значения rating, рассчитанного на предыдущем шаге. Результатом является ранжированный список, в котором первые позиции занимают факторы с наибольшей склонностью к аномалиям.

Каждая градация получает интегральную оценку, учитывающую как зафиксированный брак, так и подозрительные пропуски.

Ранжированный список позволяет ответить на практический вопрос управления качеством: какие станки, операторы, смены или режимы требуют первоочередного внимания.

3.6. Обсуждение ограничений метода

Предложенный метод двумерного рейтинга факторов обладает рядом ограничений, которые необходимо учитывать при интерпретации результатов.

Эффект малых групп. Основное ограничение метода связано с ситуацией, когда градация фактора содержит малое количество записей (малый объём группы). Малые группы могут неоправданно занимать верхние позиции рейтинга, создавая ложное впечатление о высокой проблемности фактора. Для смягчения данного эффекта могут быть предложены два подхода. Первый подход заключается во введении порогового значения по минимальному объёму группы count_min. Града-

ции, объём которых ниже установленного порога, исключаются из ранжирования. Второй подход предполагает взвешивание rating на долю группы в общем объёме данных. Пусть w_i = count_i / N, где count_i — объём i-й группы, N — общее число записей. Тогда взвешенный рейтинг rating_weighted = rating_i * w_i.

Неразличение причин пропусков. Метод не различает природу пропусков: технические сбои, человеческий фактор или умышленное сокрытие нежелательных результатов.

Зависимость от корректности границ нормы. Метод опирается на заданные границы целевого показателя, отделяющие норму от аномалии.

4. ПРИМЕР: АНАЛИЗ БРАКА ПРИ МЕХАНИЧЕСКОЙ ОБРАБОТКЕ ВАЛОВ

Управление качеством в машиностроении относится к числу наиболее ресурсоёмких и одновременно критически значимых бизнес-процессов.

4.1. Постановка задачи и описание данных

Рассматривается технологическая операция чистовой токарной обработки вала. Контролируемый параметр — наружный диаметр вала (далее — «диаметр»), нормативное значение которого составляет 50 мм с полем допуска ±0,05 мм (от 49,95 мм до 50,05 мм). Выход диаметра за границы поля допуска классифицируется как брак.

Исходный датасет содержит 10 000 записей, каждая из которых соответствует результату обработки одного вала. Состав признаков (таблица 1) включает переменные, потенциально способные влиять на возникновение брака: оборудование (станок), оператор, смена и технологический режим обработки.

- диаметр — наружный диаметр вала, мм (float),
- станок — идентификатор станка (A, B, C) (str),
- оператор — сотрудник (Контсантин_С, Артем_Т, Рашид_Ф) (str),
- смена — номер смены (1, 2) (int),
- режим — скоростной режим обработки (нормальный, ускоренный) (str).

Целевой бинарный признак — out_marker (0 — годный вал, 1 — брак). Следует подчеркнуть, что датасет не является эталонным с точки зрения полноты: примерно 8% записей в поле диаметр содержат пропуски (NaN), что типично в реальных условиям сбора производственных данных, где пропуски могут быть вызваны сбоями измерительного оборудования, человеческим фактором при внесении записей или локальными проблемами информационной системы цеха.

Таким образом, перед аналитиком встаёт классическая задача, осложнённая неполнотой данных: необходимо выявить факторы, в наибольшей степени связанные с браком, и при этом корректно учесть неопределённость, вносимую пропусками. Именно поэтому к этим данным был применён метод, описанный в разделе 3 настоящей статьи.

4.2. Результаты применения метода

По исходным данным на основе алгоритма, подробно описанного в разделе 3 (а именно: этапы 1–5), для каждого фактора были рассчитаны доли аномалий в пессимистичном (pes) и оптимистичном (opt) сценариях, а также итоговый рейтинг rating — первая координата точек (pes, opt) в системе координат, повёрнутой на 45°. Результаты ранжирования факторов представлены в табл. 1.

Как видно из табл. 1, с существенным отрывом лидирует ускоренный режим обработки: его рейтинг примерно на 32% выше, чем у ближайшего «преследователя» — станка B. Важно отметить, что, хотя станок B и уступает ускоренному режиму по абсолютному значению rating, он всё же заметно опережает остальные факторы по пессимистичной оценке. Разрыв между значениями rating ускоренного режима и станка B высок. Однако при интерпретации данного результата следует учитывать, что, во-первых, сам по себе рейтинг не является вероятностью, а во-вторых — выявленная дифференциация указывает на необходимость первоочередных управленческих решений.

Визуальное представление результатов ранжирования приведено на рис. 1 (см. раздел «Результаты и обсуждение»): на графике в координатах (pes, opt) каждый фактор представлен точкой, причём положение точки относительно диагонали иллюстрирует степень неопределённости, связанной с пропусками данных. Ярко выраженное смещение в правую верхнюю область для ускоренного режима и станка B подтверждает их наибольшую ассоциацию с дефектами.

4.3. Интерпретация результатов

Полученный рейтинг факторов (табл. 1) представляет собой не столько статистическую оценку, сколько инструмент для принятия решений. В отличие от классических метрик (p-value, коэффициент корреляции), значение rating прямо указывает на приоритетность фактора с точки зрения его связи с браком, причём с учётом неопределённости, обусловленной пропусками данных.

Первое и наиболее очевидное следствие: ускоренный режим обработки является доминирующим фактором риска. Второй по значимости фактор — станок B. Здесь интерпретация менее одно-

Таблица 1. Рейтинг факторов, влияющих на качество механической обработки валов

Место	Фактор	Значение	pes	opt	rating
1	режим	ускоренный	0.191781	0.123288	0.222787
2	станок	B	0.157459	0.055249	0.150407
3	смена	2	0.105263	0.065263	0.120580
4	оператор	Рашид_Ф	0.095238	0.056689	0.107429
5	оператор	Константин_С	0.096852	0.048426	0.102727
6	станок	A	0.068862	0.068862	0.097386
7	смена	1	0.085714	0.030476	0.082159
8	оператор	Артем_Т	0.089041	0.013699	0.072648
9	режим	нормальный	0.055085	0.015537	0.049937
10	станок	C	0.049342	0.013158	0.044194

значна: рейтинг станка B может быть обусловлен как его техническим состоянием (износ направляющих, люфт осей вращения), так и спецификой операций, которые на нём выполняются. Плотная группа остальных факторов (операторы, смены, иные станки) со сравнимыми значениями rating говорит о том, что их влияние на брак носит сопоставимый характер. Это не означает, что данные факторы можно игнорировать — скорее, их следует рассматривать как «фоновый шум»..

Особого внимания заслуживает положение точек на диаграмме (pes, opt) относительно друг друга (рис. 1). Для ускоренного режима характерно смещение в область высоких значений как по оси pes, так и по оси opt» — это говорит о том, что их высокая ассоциация с браком является устойчивой и не зависит от предположений о пропущенных данных. Напротив, станок B имеет высокий pes, но низкий opt, его рейтинг может быть обусловлен исключительно пропусками — и доверие к такому результату ниже.

Таким образом, по итогам анализа можно сформулировать следующие рекомендации:

1. Незамедлительно провести технологический аудит ускоренного режима обработки: проанализировать карты наладок, фактические режимы резания и стойкость инструмента.

2. В краткосрочной перспективе организовать внеплановую проверку станка B (геометрия, точность позиционирования и т. д.).

3. В долгосрочной перспективе — рассмотреть возможность внедрения автоматизированного сбора данных о работе оборудования (для сокращения доли пропусков) и мониторинга режимов обработки в реальном времени.

5. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Применение метода двумерного рейтинга демонстрирует универсальность предложенного подхода. Несмотря на различие в физической сути «аномалий» (отклонение диаметра вала в первом случае и низкая оценка качества диалога во втором), в обоих примерах метод позволил эффективно выделить ключевые факторы риска. На рис. 1 представлены графики распределения групп факторов в координатах (pes, opt).

В качестве перспективного направления развития метода видится введение весовых коэффициентов для коррекции эффекта малых групп. В текущей версии факторы с малым количеством наблюдений могут демонстрировать аномальные значения рейтинга из-за высокой чувствительности к единичным инцидентам. Использование взвешенного рейтинга (rating_weighted = rating × w_i, где w_i — доля группы в общем объ-

Рисунок 1. Рейтинги факторов качества в машиностроении

еме данных) позволит сгладить влияние групп с низкой статистической значимостью и сделать приоритизацию управленческих воздействий более устойчивой к шуму.

6. ВЫВОДЫ И ЗАКЛЮЧЕНИЕ

В данной работе был предложен концептуально новый взгляд на проблему пропусков в данных прикладных бизнес-процессов. Мы показали, что пропуски — это не просто дефект или статистическая помеха, подлежащая удалению или замене, а полноценный источник управленческой информации. В условиях, когда автоматизированные системы контроля качества работают не идеально, именно пропуски часто указывают на «слепые зоны» процесса, где скрываются системные сбои или нежелательные результаты деятельности.

Разработанный метод двумерного рейтинга позволяет эффективно перевести качественные представления о «подозрительности» данных в количественный инструмент поддержки принятия решений.

СПИСОК ЛИТЕРАТУРЫ

1. *Бурков, А.* Инженерия машинного обучения / А. Бурков. – М: ДМК Пресс, 2022. – 306 с.
2. *Вьюгин, В. В.* Математические основы теории машинного обучения и прогнозирования / В. В. Вьюгин. – М.: МЦИМО. – 2013. – 387 с.
3. *Гембрис, С.* Управление качеством / С. Гембрис, Й. Геррманн. – М.: СмартБук, 2013. – 128 с.
4. *Харрингтон, Д. Х.* Управление качеством в американских корпорациях / Д. Х. Харрингтон. – М.: Экономика, 2013. – 272 с.
5. *Мосин, В. Г.* Обоснование методологии детекции аномалий в качестве нового цифрового инструмента управления качеством / В. Г. Мосин, В. Н. Козловский // Известия Самарского научного центра Российской академии наук. – 2024. – Т. 26. – № 3. – С. 16-22.

MISSING VALUES IN TARGET FEATURES AS A SOURCE OF QUALITY INFORMATION: TWO-DIMENSIONAL FACTOR RATING METHOD

© 2026 V.G. Mosin, M.L. Raush, V.N. Kozlovsky, A.S. Demkin

Samara State Technical University, Samara, Russia

The article presents a two-dimensional factor rating method that allows for the effective use of information about data gaps in process quality management. The approach is based on calculating pessimistic and optimistic estimates of the proportion of anomalies, which provides an integral measure of the suspiciousness of each factor. The method has been successfully tested on engineering and service tasks, enabling the prioritization of management decisions in the face of incomplete data.

Key words: Data science, machine learning, quality management, statistical quality management.

DOI: 10.37313/1990-5378-2026-28-4-35-40
EDN: FELOEY

REFERENCES

1. *Burkov, A.* Inzheneriya mashinnogo obucheniya / A. Burkov. – M: DMK Press, 2022. – 306 s.
2. *V'yugin, V. V.* Matematicheskie osnovy teorii mashinnogo obucheniya i prognozirovaniya / V. V. V'yugin. – M.: MCIMO. – 2013. – 387 s.
3. *Gembris S.* Upravlenie kachestvom / S. Gembris, J. Gerrmann. – M.: SmartBuk, 2013. – 128 c.
4. *Harrington, D. H.* Upravlenie kachestvom v amerikanskih korporaciyah / D. H. Harrington. – M.: Ekonomika, 2013. – 272 s.
5. *Mosin, V. G.* Obosnovanie metodologii detekcii anomalij v kachestve novogo cifrovogo instrumenta upravleniya kachestvom / V. G. Mosin, V. N. Kozlovskij // Izvestiya Samarskogo nauchnogo centra Rossijskoj akademii nauk. – 2024. – T. 26. – № 3. – S. 16-22.

Vladimir Mosin, Samara State Technical University, Associate Professor, PhD in Physics and Mathematics. E-mail: yanbacha@yandex.ru.
Maria Raush, Student. E-mail: mary.raush@mail.ru
Vladimir Kozlovsky, Doctor of Engineering Sciences, Professor, Head of Department. E-mail: kozlovskiy-76@mail.ru,
Aleksandr Demkin, Postgraduate Student. E-mail: etf3gtu@mail.ru