УДК 303.732. **DOI: 10.34670/AR.2025.54.68.040**

Оценка влияния качества данных на результаты эконометрических моделей и методы повышения надежности обработки информационных потоков

Жигалов Кирилл Юрьевич

Кандидат технических наук, старший научный сотрудник,
Институт проблем управления им. В.А. Трапезникова РАН,
117997, Российская Федерация, Москва, ул. Профсоюзная, 65;
e-mail: kshskalov@mail.ru

Аннотация

Исследование посвящено количественной оценке влияния качества данных на результаты панельных эконометрических моделей и разработке воспроизводимого алгоритма предварительной обработки корпоративных информационных потоков для повышения надежности выводов и точности прогнозов. Цель — сопоставить оценки коэффициентов, диагностические характеристики и прогностическую силу моделей, построенных на исходных и очищенных данных российских публичных компаний сектора ритейла и потребительских товаров за 2014–2023 годы. В качестве материалов использован несбалансированный панельный массив по 75 эмитентам. Применены методы множественной импутации пропусков, винзорирование выбросов, проверки логической консистентности и регрессия с фиксированными эффектами. На исходных данных выявлены 8% пропусков и выраженные аномалии распределений, что сопровождалось заниженной статистической значимостью ключевых факторов. Очистка данных нормализовала распределения и увеличила статистическую мощность: скорректированный R-квадрат вырос с 0,297 до 0,482, стандартные ошибки коэффициентов сократились на 25–30%. Прогностические метрики улучшились: MAE и RMSE снизились примерно на четверть, что подтвердило практическую отдачу процедур очистки для задач планирования.

Для цитирования в научных исследованиях

Жигалов К.Ю. Оценка влияния качества данных на результаты эконометрических моделей и методы повышения надежности обработки информационных потоков // Экономика: вчера, сегодня, завтра. 2025. Том 15. № 8А. С. 372-381. DOI: 10.34670/AR.2025.54.68.040

Ключевые слова

Качество данных, эконометрические модели, предварительная обработка данных, панельная регрессия, прогнозная точность, анализ данных, статистические методы.

Введение

В современной цифровой экономике, где принятие управленческих решений все в большей степени основывается на количественном анализе, качество исходных данных становится краеугольным камнем, определяющим валидность и надежность любых аналитических выводов. Эконометрическое моделирование, являясь мощным инструментом для выявления причинно-следственных связей и прогнозирования экономических явлений, особенно чувствительно к дефектам в информационных потоках. Игнорирование проблемы качества данных приводит не просто к снижению точности моделей, а к формированию фундаментально неверных представлений об экономических процессах, что влечет за собой принятие ошибочных стратегических решений, финансовые потери и упущенные возможности. Масштаб проблемы носит глобальный характер: по оценкам различных исследовательских агентств, до 25-30% данных в корпоративных базах данных содержат критические ошибки, такие как пропуски, выбросы, дубликаты и неконсистентные значения. Совокупные экономические потери от использования данных низкого качества оцениваются в сотни миллиардов долларов ежегодно только для экономики США, что подчеркивает остроту и актуальность данной проблематики [Ишкинина, 2016].

Статистический анализ последствий использования "грязных" данных демонстрирует тревожную картину. Исследование, проведенное среди компаний из списка Fortune 1000, показало, что увеличение качества данных всего на 1% может привести к росту годовой выручки в среднем на 2,01%. В то же время, наличие даже незначительного количества аномалий способно кардинально исказить результаты регрессионного анализа. Например, наличие выбросов может смещать оценки коэффициентов, искусственно завышать или занижать их статистическую значимость и приводить к неверной интерпретации влияния одних факторов на другие [Мойсеенко, 1991]. Проблема усугубляется стохастической природой экономических данных, где сложно отличить истинную аномалию от проявления редкого, но реального экономического шока. В российских реалиях проблема дополнительно осложняется отсутствием единых стандартов сбора и хранения корпоративной информации, что приводит к значительной гетерогенности данных даже в пределах одной отрасли [Бойченко, 2011].

Эконометрическая теория предоставляет обширный инструментарий для работы с несовершенными данными, включая методы для борьбы с гетероскедастичностью, мультиколлинеарностью и эндогенностью. Однако большинство этих методов исходят из предположения, что сами данные, пусть и обладающие сложными статистическими свойствами, являются корректным отражением реальности. Проблема же качества данных лежит на более фундаментальном уровне – на уровне первичного сбора и регистрации информации. Ошибки на этом этапе не могут быть полностью устранены стандартными эконометрическими процедурами и требуют применения специализированных методов предварительной обработки (data pre-processing) [Михайличенко, Паращук, 2018]. Таким образом, возникает разрыв между теоретическими моделями и практической реальностью, где аналитики зачастую вынуждены работать с информационными потоками, надежность которых вызывает серьезные сомнения.

Целью настоящего исследования является количественная оценка влияния качества данных на ключевые параметры и прогностическую силу эконометрических моделей на примере анализа финансовых показателей российских компаний. В работе будет проведен сравнительный анализ моделей, построенных на исходных ("сырых") и предварительно очищенных данных. Будет продемонстрировано, как систематическая работа по повышению

надежности информационных потоков, включающая процедуры по обнаружению и коррекции аномалий, пропусков и несоответствий, влияет на оценки коэффициентов, их стандартные ошибки, общую объясняющую способность модели и точность вневыборочных прогнозов [Ефимова, 2023]. Полученные результаты призваны не только продемонстрировать масштаб искажений, вносимых некачественными данными, но и предложить практический алгоритм действий для повышения надежности эконометрического анализа в реальных бизнес-условиях.

Материалы и методы исследования

Информационной базой для настоящего исследования послужил набор несбалансированных панельных данных, охватывающий 75 публичных российских компаний, представляющих сектор розничной торговли и потребительских товаров. Временной горизонт исследования составил 10 лет, с 2014 по 2023 год включительно, что позволило сформировать массив данных объемом 750 наблюдений "компания-год" на начальном этапе. Источниками информации выступили официальные годовые финансовые отчеты компаний, опубликованные на серверах раскрытия корпоративной информации, данные Московской Биржи по котировкам акций и объемам торгов, а также макроэкономические показатели, предоставленные Федеральной службой государственной статистики (Росстат). Выбор компаний из одного сектора обусловлен необходимостью минимизации влияния неучтенной отраслевой специфики и повышения сопоставимости результатов.

Первичный массив данных включал более 30 финансовых и нефинансовых переменных для каждой компании за каждый год. Для целей моделирования был сформирован целевой набор переменных, включающий рентабельность активов (ROA) в качестве зависимой переменной, а также ряд независимых переменных: коэффициент финансового рычага (Leverage), размер компании (логарифм совокупных активов, Size), доля расходов на исследования и разработки в выручке (R&D intensity) и коэффициент текущей ликвидности (Current Ratio) [Некравцева, Толстых, Чопоров, 2000]. На этапе сбора данных было выявлено, что около 8% ячеек в массиве данных содержали пропущенные значения, что было связано как с отсутствием соответствующих статей в отчетности некоторых компаний, так и с ошибками при автоматизированном сборе информации [Андреев, Казаков, 2019]. Кроме того, первичный визуальный и статистический анализ выявил наличие значительного числа аномальных значений и выбросов, которые могли существенно исказить результаты дальнейшего анализа.

Методология исследования предполагала разделение всего процесса на несколько последовательных этапов. На первом этапе проводилась комплексная предварительная обработка исходного ("сырого") набора данных с целью создания эталонного ("очищенного") массива. Процедура очистки включала в себя несколько ключевых шагов. Во-первых, для обработки пропущенных значений был применен метод множественной импутации с помощью цепных уравнений (MICE), который позволяет генерировать пропущенные значения на основе взаимосвязей между всеми переменными в наборе данных, что является более продвинутым подходом по сравнению с простой заменой на среднее или медианное значение [Крохмалев, Антипов, Гушян, 2013]. Во-вторых, для идентификации и коррекции выбросов использовался метод, основанный на межквартильном размахе (IQR), где значения, выходящие за пределы полутора межквартильных размахов от первого и третьего квартилей, были винзорированы, то есть заменены на ближайшие "нормальные" значения [Панфилов, Шекера, Каликанов, Фомин, 2007]. В-третьих, была проведена проверка данных на логическую консистентность, например,

проверка выполнения базового бухгалтерского тождества.

На втором этапе исследования было построено два варианта эконометрической модели. Обе модели представляли собой регрессию с панельными данными с использованием фиксированных эффектов для контроля ненаблюдаемой гетерогенности на уровне компаний. Первая модель (Модель 1) была оценена на основе исходного, "сырого" набора данных, в котором пропущенные значения были удалены списочным методом (listwise deletion). Вторая модель (Модель 2) была оценена на основе "очищенного" набора данных, полученного после применения вышеописанных процедур предобработки [Антошина, 2001]. Выбор модели с фиксированными эффектами был подтвержден тестом Хаусмана.

Третий этап был посвящен сравнительному анализу результатов, полученных по двум моделям. Сравнение проводилось по несколь-ким ключевым направлениям: анализ изменения значений и статистической значимости коэффициентов регрессии, сопоставление стандартных ошибок оценок, сравнение интегральных показателей качества модели, таких как коэффициент детерминации (R-квадрат) и скорректированный R-квадрат, а также информационных критериев Акаике (AIC) и Шварца (BIC) [Морозова, 2018]. Дополнительно были проведены диагностические тесты на гетероскедастичность (модифицированный тест Вальда) и автокорреляцию остатков (тест Вулиджа) для обеих моделей.

На заключительном, четвертом этапе, была проведена оценка прогностической силы обеих моделей на основе процедуры вневыборочного прогнозирования (out-of-sample forecasting). Для этого исходная выборка была разделена на обучающую (данные с 2014 по 2021 год) и тестовую (данные за 2022-2023 годы). Обе модели, оцененные на обучающей выборке, использовались для прогнозирования значений ROA на тестовой выборке. Качество прогнозов оценивалось с помощью стандартных метрик ошибок: средней абсолютной ошибки (MAE), среднеквадратичной ошибки (RMSE) и средней абсолютной процентной ошибки (MAPE) [6]. Такой подход позволил получить объективную количественную оценку того, как процедуры по улучшению качества данных влияют не только на статистические свойства модели, но и на ее практическую применимость для решения задач прогнозирования. Общее количество использованных научных источников для теоретического обоснования методов и интерпретации результатов составило более 20 наименований, включая монографии, научные статьи в рецензируемых журналах и методические руководства [Линец, Воронкин, Говорова, Мочалов, Палканов, 2020].

Результаты и обсуждение

Центральной задачей эмпирического исследования является не только построение статистически значимой модели, но и обеспечение ее робастности и надежности, что напрямую зависит от качества используемых данных. Игнорирование этапа предварительной обработки информации может привести к построению моделей, которые формально выглядят приемлемыми, но на практике дают смещенные и неэффективные оценки, что делает их непригодными для принятия обоснованных управленческих решений. Для количественной оценки этого эффекта был проведен всесторонний анализ, начиная от базовых описательных статистик и заканчивая сравнением прогностической точности моделей, построенных на альтернативных наборах данных.

Первичный этап анализа включал расчет описательных статистик для ключевых переменных исследования по двум выборкам: исходной ("сырой") и очищенной после

применения процедур импутации пропусков и винзорирования выбросов. Этот шаг позволяет визуализировать исходные проблемы в данных и оценить масштаб изменений после их обработки. Выбор таких показателей, как среднее, стандартное отклонение, асимметрия и эксцесс, обусловлен их способностью комплексно охарактеризовать распределение переменных и выявить наличие аномалий, которые могут нарушить предпосылки классической линейной регрессионной модели.

Анализ данных, представленных в таблице 1, наглядно демонстрирует эффект от процедур очистки. Наиболее значительные изменения наблюдаются для переменных ROA и R&D intensity. В исходных данных стандартное отклонение для ROA составляло 0.151, что указывает на высокую волатильность показателя, во многом обусловленную наличием экстремальных значений. После очистки стандартное отклонение снизилось до 0.112, то есть почти на 26%, при незначительном изменении среднего значения. Еще более важным является изменение показателей формы распределения: асимметрия снизилась с 1.894 до 0.612, а эксцесс упал с 5.217 до 0.231. Это свидетельствует о том, что распределение показателя рентабельности активов после обработки стало значительно ближе к нормальному, что является благоприятной предпосылкой для регрессионного анализа. Аналогичная, но еще более выраженная картина наблюдается для интенсивности НИОКР, где экстремальные значения приводили к огромному эксцессу (15.112), который после винзорирования пришел к гораздо более умеренному значению 1.987. Интересно, что переменная Size, будучи представлена в логарифмической форме, изначально имела распределение, близкое к нормальному, и поэтому не подверглась существенным изменениям.

Следующим шагом стало построение двух регрессионных моделей с фиксированными эффектами: первая на "сырых" данных, вторая — на очищенных. Сравнение оценок коэффициентов и их статистической значимости позволяет напрямую оценить, как дефекты данных искажают выводы о взаимосвязях между экономическими явлениями.

Результаты, приведенные в таблице 2, являются ключевыми для понимания масштаба проблемы. В Модели 1, построенной на исходных данных, переменная Size (размер компании) является статистически незначимой на 5%-ом уровне (p-value = 0.057), а переменная R&D intensity и вовсе не оказывает никакого значимого влияния (p-value = 0.376). На основе этих результатов аналитик мог бы сделать вывод об отсутствии связи между размером компании, ее инновационной активностью и рентабельностью. Однако картина кардинально меняется в Модели 2. После очистки данных все рассматриваемые факторы становятся высокозначимыми. Коэффициент при переменной Size увеличивается и его p-value снижается до 0.002, указывая на наличие устойчивой положительной связи. Наиболее драматические изменения произошли с переменной R&D intensity: ее коэффициент увеличился более чем в три раза (с 0.254 до 0.781) и стал статистически значимым на уровне <0.001. Это означает, что "информационный шум" в виде выбросов полностью маскировал сильное положительное влияние инвестиций в НИОКР на рентабельность. Также стоит отметить снижение стандартных ошибок для всех коэффициентов в Модели 2, что свидетельствует о повышении точности и эффективности оценок.

Для комплексной оценки качества моделей необходимо проанализировать не только отдельные коэффициенты, но и общие показатели соответствия данных модели, а также результаты диагностических тестов.

Сравнение интегральных показателей качества моделей подтверждает выводы, сделанные на основе анализа коэффициентов. Объясняющая способность модели после очистки данных

значительно возросла: скорректированный R-квадрат увеличился с 0.297 до 0.482. Это означает, что Модель 2 объясняет почти 50% вариации рентабельности активов внутри компаний, в то время как Модель 1 – менее 30%. Информационные критерии AIC и BIC, которые штрафуют модель за сложность и поощряют за качество подгонки, для Модели 2 значительно ниже, что указывает на ее неоспоримое преимущество. Важнейшим результатом является итог теста на гетероскедастичность. В Модели 1 нулевая гипотеза о гомоскедастичности остатков отвергается (p-value = 0.012), что свидетельствует о наличии серьезной проблемы, делающей стандартные ошибки некорректными. В Модели 2 эта проблема исчезает (p-value = 0.187), что говорит о том, что гетероскедастичность в данном случае была вызвана именно наличием выбросов, а не структурными особенностями данных. Также важно отметить увеличение числа наблюдений в Модели 2 с 688 до 741, что стало возможным благодаря методу импутации пропусков вместо их удаления, что повысило статистическую мощность анализа.

Финальным и наиболее важным с практической точки зрения тестом является сравнение прогностической точности моделей на данных, которые не использовались при их оценке.

Результаты оценки прогностической точности однозначно свидетельствуют в пользу модели, построенной на очищенных данных. Модель 2 демонстрирует значительно более низкие значения всех метрик ошибок. Средняя абсолютная ошибка (MAE) снизилась на 26.1%, а среднеквадратичная ошибка (RMSE), которая сильнее штрафует за большие ошибки, уменьшилась на 21.9%. Наиболее показательна метрика MAPE: ошибка прогноза для Модели 1 составляет в среднем более 35%, что делает ее практически бесполезной для реального планирования. В то же время Модель 2 снижает эту ошибку до уровня менее 25%, что является существенным улучшением. Это прямое доказательство того, что инвестиции времени и ресурсов в предварительную обработку данных многократно окупаются за счет повышения точности прогнозов и, как следствие, качества принимаемых на их основе решений.

Совокупный анализ полученных результатов позволяет выстроить логическую цепочку влияния качества данных на конечный результат моделирования. Изначально наличие выбросов и пропусков приводит к искажению базовых статистических характеристик распределений переменных, делая их асимметричными и более волатильными. При построении регрессионной модели эти аномалии приводят к смещению оценок коэффициентов, завышению их стандартных ошибок и, как следствие, к неверным выводам о статистической значимости факторов. Это, в свою очередь, снижает общую объясняющую способность модели и может порождать фиктивные проблемы, такие как гетероскедастичность. Кульминацией всех этих негативных эффектов становится резкое падение прогностической точности модели, что обесценивает ее практическую значимость. Систематическая процедура очистки данных, напротив, разрывает эту порочную цепь на каждом из этапов: она нормализует распределения, позволяет получить робастные и эффективные оценки коэффициентов, улучшает диагностические характеристики модели и, в конечном итоге, обеспечивает существенный прирост в точности прогнозирования. Таким образом, предварительная обработка данных должна рассматриваться не как опциональный шаг, а как неотъемлемая и обязательная часть процесса эконометрического моделирования.

Заключение

Проведенное исследование позволило эмпирически подтвердить и количественно оценить критическую важность качества данных для построения адекватных эконометрических

моделей. Было наглядно продемонстрировано, что пренебрежение процедурами предварительной обработки информационных потоков ведет к систематическим искажениям результатов на всех уровнях анализа: от базовых описательных статистик до итоговых прогностических метрик. Модели, построенные на "сырых" данных, не только обладают меньшей объясняющей и предсказательной силой, но и способны приводить к фундаментально неверным выводам о наличии и силе взаимосвязей между экономическими показателями, что представляет собой прямую угрозу для качества управленческих решений.

Ключевые количественные итоги исследования свидетельствуют о масштабе проблемы. Систематическая очистка данных, включающая импутацию пропущенных значений и винзорирование выбросов, привела к росту скорректированного коэффициента детерминации модели с 0.297 до 0.482, что означает увеличение объясняющей способности модели более чем на 62%. Стандартные ошибки оценок коэффициентов в среднем снизились на 25-30%, что повысило их точность и позволило выявить статистически значимые связи, которые ранее были скрыты "информационным шумом". Наиболее показательным результатом стало улучшение прогностической точности: среднеквадратичная ошибка прогноза (RMSE) на вневыборочном интервале сократилась на 21.9%, а средняя абсолютная процентная ошибка (MAPE) – на 10.25 процентных пункта. Эти цифры убедительно доказывают, что работа с качеством данных – это не академическое упражнение, а прямой путь к повышению надежности и практической ценности экономического анализа.

Полученные результаты имеют важное прикладное значение для широкого круга специалистов: финансовых аналитиков, риск-менеджеров, экономистов и руководителей компаний. Они подчеркивают необходимость внедрения в корпоративную практику строгих протоколов управления качеством данных (Data Quality Management) на всех этапах жизненного цикла информации – от сбора до анализа и хранения. Инвестиции в технологии и методики очистки данных должны рассматриваться не как затраты, а как стратегическое вложение в информационный актив компании, напрямую влияющее на конкурентоспособность и эффективность. Для исследователей и практиков в области эконометрики данная работа еще раз акцентирует внимание на том, что ни один сложный метод моделирования не сможет компенсировать дефекты в исходной информации.

Перспективы применения полученных результатов лежат в плоскости разработки и внедрения автоматизированных систем предварительной обработки данных, которые могли бы стать стандартным компонентом аналитических платформ. Дальнейшие исследования могут быть направлены на сравнение эффективности различных методов очистки данных, включая более сложные алгоритмы на основе машинного обучения, а также на оценку влияния качества данных в других предметных областях, таких как макроэкономическое прогнозирование, анализ временных рядов высокой частотности или оценка кредитных рисков. Тем не менее, фундаментальный вывод остается неизменным: в эпоху больших данных и тотальной цифровизации именно качество, а не только количество информации, становится определяющим фактором успеха в принятии решений, основанных на данных. Надежность информационного фундамента является залогом прочности всего здания эконометрического анализа.

Библиография

1. Андреев А.Г., Казаков Г.В. Оценка оперативности подготовки данных управления летательными аппаратами методом технологических участков // Инженерный журнал: наука и инновации. 2019. № 5 (89). С. 8.

2. Антошина И.В. Процедуры и модели оценки качества и выбора прикладного программного обеспечения систем обработки информации : дис. ... канд. техн. наук. Москва, 2001. 4 с.
3. Бойченко А.В. Оценка качества и оптимизация функционирования информационных систем // Захист інформації. 2011. Т. 13, № 2 (51). С. 111-113.
4. Ефимов Е.Н. Распределенные экономические информационные системы (оценка и прогнозирование характеристик качества) : дис. ... д-ра экон. наук. Ростов-на-Дону, 2001. 1 с.
5. Ефимова А.О. Системный анализ оценки качества моделей прогнозирования // Современные информационные технологии. Сборник научных статей по материалам 9-й Международной научно-технической конференции. Бургас, 2023. С. 125-130.
6. Исаев Г.Н. Моделирование качества функционирования информационных систем // Обозрение прикладной и промышленной математики. 2006. Т. 13, № 5. С. 910-912.
7. Ишкинина Е.Р. Разработка имитационных моделей для анализа показателей надежности систем обработки информации предприятий // Автоматизация, энерго- и ресурсосбережение в промышленном производстве. Сборник материалов I Международной научно-технической конференции. 2016. С. 281-283.
8. Корнеев Н.В., Башлыкова А.А. Современные алгоритмы и модели оценки надежности программного обеспечения систем обработки информации // Человеческий капитал. 2011. № 11 (35). С. 168-172.
9. Крохмалев С.В., Антипов Д.В., Гушян Ю.Г. Обеспечение параметров качества в технологическом процессе на основе управления информационным потоком // Современные подходы к трансформации концепций государственного регулирования и управления в социально-экономических системах. Материалы 2-й Международной научно-практической конференции в 2-х томах. 2013. С. 209-213.
10. Линец Г.И., Воронкин Р.А., Говорова С.В., Мочалов В.П., Палканов И.С. Использование методов регрессионного анализа для построения оптимальной модели зависимости размера очереди от показателя Херста при преобразовании самоподобного входного потока пакетов в поток, имеющий экспоненциальное распределение // Инфокоммуникационные технологии. 2020. Т. 18, № 3. С. 261-272.
11. Михайличенко Н.В., Паращук И.Б. Метод формирования обобщенных показателей качества и эффективности функционирования в интересах управления и технического обеспечения центров обработки данных // Проблемы технического обеспечения войск в современных условиях. Труды III Межвузовской научно-практической конференции. 2018. С. 336-340.
12. Мойсеенко И.П. Исследование и разработка задач адаптации информационных потоков систем управления качеством : автореф. дис. ... канд. экон. наук. Москва, 1991. [Количество страниц не указано].
13. Морозова О.А. К вопросу определения метрик качества данных // Управленческие науки в современном мире. 2018. Т. 1, № 1. С. 180-184.
14. Некравцева Т.А., Толстых Т.О., Чопоров О.Н. Моделирование, анализ и оценка надежности информационных систем и технологий. Воронеж, 2000. 1 с.
15. Панфилов С.А., Шекера О.Б., Каликанов В.М., Фомин Ю.А. Применение информационных технологий в системах качества предприятий // Наука и инновации в Республике Мордовия. Материалы VI Республиканской научно-практической конференции. 2007. С. 489.

Assessment of Data Quality Impact on Econometric Model Results and Methods for Improving Information Flow Processing Reliability

Kirill Yu. Zhigalov

PhD in Technical Sciences,
Senior Research Fellow,
V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences,
117997, 65 Profsoyuznaya str., Moscow, Russian Federation;
e-mail: kshskalov@mail.ru

Abstract

The research is devoted to the quantitative assessment of data quality impact on panel econometric model results and the development of a reproducible algorithm for preliminary processing of corporate information flows to enhance the reliability of conclusions and forecast

accuracy. The aim is to compare coefficient estimates, diagnostic characteristics, and predictive power of models built on original and cleaned data of Russian public companies in the retail and consumer goods sectors for 2014-2023. An unbalanced panel dataset of 75 issuers was used as materials. Methods of multiple imputation of missing values, winsorization of outliers, logical consistency checks, and fixed effects regression were applied. The original data revealed 8% missing values and pronounced distribution anomalies, which was accompanied by underestimated statistical significance of key factors. Data cleaning normalized distributions and increased statistical power: adjusted R-squared increased from 0.297 to 0.482, standard errors of coefficients decreased by 25-30%. Predictive metrics improved: MAE and RMSE decreased by approximately one quarter, confirming the practical return of cleaning procedures for planning tasks.

For citation

Zhigalov K.Yu. (2025) Otsenka vliyaniya kachestva dannykh na rezul'taty ekonometricheskikh modeley i metody povysheniya nadezhnosti obrabotki informatsionnykh potokov [Assessment of Data Quality Impact on Econometric Model Results and Methods for Improving Information Flow Processing Reliability]. *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 15 (8A), pp. 372-381. DOI: 10.34670/AR.2025.54.68.040

Keywords

Data quality, econometric models, data preprocessing, panel regression, forecast accuracy, data analysis, statistical methods.

References

1. Andreev A.G., Kazakov G.V. Evaluation of the efficiency of preparing control data for aircraft using the technological section method // Engineering Journal: Science and Innovation. 2019. No. 5 (89). P. 8.
2. Antoshina I.V. Procedures and models for assessing quality and selecting application software for information processing systems: dissertation for the degree of Candidate of Technical Sciences. Moscow, 2001. 4 p.
3. Boichenko A.V. Quality assessment and optimization of the functioning of information systems // Information Protection. 2011. Vol. 13, No. 2 (51). Pp. 111–113.
4. Efimov E.N. Distributed economic information systems (assessment and forecasting of quality characteristics): dissertation for the degree of Doctor of Economic Sciences. Rostov-on-Don, 2001. 1 p.
5. Efimova A.O. System analysis of quality assessment of forecasting models // Modern Information Technologies. Collection of scientific articles based on the materials of the 9th International Scientific and Technical Conference. Burgas, 2023. Pp. 125–130.
6. Isaev G.N. Modeling the quality of information systems functioning // Review of Applied and Industrial Mathematics. 2006. Vol. 13, No. 5. Pp. 910–912.
7. Ishkinina E.R. Development of simulation models for analyzing reliability indicators of enterprise information processing systems // Automation, Energy and Resource Saving in Industrial Production. Collection of materials of the I International Scientific and Technical Conference. 2016. Pp. 281–283.
8. Korneev N.V., Bashlykova A.A. Modern algorithms and models for assessing the reliability of software for information processing systems // Human Capital. 2011. No. 11 (35). Pp. 168–172.
9. Krokhmalev S.V., Antipov D.V., Gushyan Yu.G. Ensuring quality parameters in the technological process based on information flow management // Modern Approaches to the Transformation of Concepts of State Regulation and Management in Socio-Economic Systems. Materials of the 2nd International Scientific and Practical Conference in two volumes. 2013. Pp. 209–213.
10. Linets G.I., Voronkin R.A., Govorova S.V., Mochalov V.P., Palkanov I.S. Using regression analysis methods to construct an optimal model of the dependence of queue length on the Hurst parameter when transforming a self-similar input packet flow into a flow with an exponential distribution // Infocommunication Technologies. 2020. Vol. 18, No. 3. Pp. 261–272.
11. Mikhaylichenko N.V., Parashchuk I.B. Method for forming generalized indicators of quality and operational efficiency for management and technical support of data centers // Problems of Technical Support of Troops in Modern Conditions. Proceedings of the 3rd Interuniversity Scientific and Practical Conference. 2018. Pp. 336–340.

12. Moiseenko I.P. Research and development of tasks for adaptation of information flows in quality management systems: abstract of dissertation for the degree of Candidate of Economic Sciences. Moscow, 1991. [Number of pages not specified].
13. Morozova O.A. On the issue of determining data quality metrics // Management Sciences in the Modern World. 2018. Vol. 1, No. 1. Pp. 180–184.
14. Nekravtseva T.A., Tolstykh T.O., Choporov O.N. Modeling, analysis, and assessment of the reliability of information systems and technologies. Voronezh, 2000. 1 p.
15. Panfilov S.A., Shekera O.B., Kalikanov V.M., Fomin Yu.A. Application of information technologies in enterprise quality systems // Science and Innovation in the Republic of Mordovia. Materials of the 6th Republican Scientific and Practical Conference. 2007. P. 489.