

МЕТОДОЛОГИЯ СОЗДАНИЯ ЭМПИРИЧЕСКОЙ БАЗЫ ДАННЫХ ДЛЯ ЭКОНОМЕТРИЧЕСКОЙ ОЦЕНКИ ЭФФЕКТИВНОСТИ УПРАВЛЕНИЯ В СЕКТОРЕ УСЛУГ

Турсунова Шахло Сайфиллаевна

Независимый исследователь из Государственного университета Карши.

Электронная почта: shahlotursunova84@gmail.com

orcid:0009-0005-9349-3063

<https://doi.org/10.5281/zenodo.22811429>

Абстрактный. В статье разработана полная методология формирования эмпирической базы данных для эконометрической оценки эффективности управления сектором услуг. На основе шести показателей качества данных определены процедуры их проверки. В результате исследования предложены шестиступенчатый алгоритм формирования базы данных, критерии выбора методов восстановления пропущенных значений, процедура выявления аномальных наблюдений и индекс оценки готовности базы данных к анализу. Отдельно обоснованы требования к структуре панельных данных.

Ключевые слова: эмпирическая база данных, панельные данные, качество данных, выпавшие значения, аномальные наблюдения, импутация, нормализация данных, эконометрическая оценка, индекс готовности к анализу.

Введение

Надежность любого эконометрического исследования зависит от двух факторов: правильности используемого метода и качества используемых данных. Из этих двух факторов последний часто является решающим. Даже самая современная и сложная модель даст ненадежные результаты на данных низкого качества. В статистической практике эта ситуация выражается коротким правилом: если на входе мусор, то и на выходе будет мусор.

Однако в научных исследованиях этапу формирования базы данных обычно уделяется мало внимания. В статьях подробно описывается спецификация модели и результаты оценки, а данные ограничиваются коротким предложением: «На основе данных Статистического агентства». Однако именно на этом этапе принимаемые решения — какой показатель выбрать, как заполнить пропущенные значения, что делать с аномальными наблюдениями — оказывают существенное влияние на конечный результат.



Формирование базы данных в секторе услуг сопряжено с рядом специфических трудностей. Во-первых, сектор структурно разнообразен, и данные по различным видам деятельности собираются не с одинаковой полнотой. Во-вторых, в секторе высока доля нематериальных результатов, и их количественное измерение ограничено. В-третьих, ряды данных в территориальном разделе невелики, что ограничивает выбор эконометрических методов.

Цель данной статьи — разработка систематической методологии формирования эмпирической базы данных для эконометрической оценки эффективности управления в секторе услуг и обоснование конкретных процедур для каждого из ее этапов.

Литературный анализ по теме

Теоретическая основа концепции качества данных была создана Р.Ю. Ваном и Д.М. Стронгом. Они определили качество данных как многомерную концепцию, не ограничивающуюся точностью, и систематизировали показатели качества с точки зрения пользователя [5]. Этот подход послужил основой для различения показателей качества в данном исследовании.

Фундаментальное исследование проблемы пропущенных значений было проведено Р. Дж. А. Литтлом и Д. Б. Рубиным. Они выделили три типа механизмов пропущенных значений: пропущенные значения случайным образом (MCAR), пропущенные значения условно случайным образом (MAR) и пропущенные значения неслучайным образом (MNAR) [3]. Эта классификация имеет фундаментальное значение: для каждого механизма используется свой метод восстановления.

ДБРубин разработал метод множественной импутации, который предложил идею учета неопределенности путем замены пропущенных значений несколько раз, а не один раз, и объединения результатов [4]. Этот метод дает более надежные результаты, чем однократная импутация.

Требования к структурам панельных данных были изложены в основополагающих работах Дж. М. Вулдриджа [1] и Б. Х. Балтаги [2]. Они установили различие между сбалансированными и несбалансированными панелями, а также влияние длины и ширины панели на метод оценки.

Методика построения составных показателей и нормализации данных систематизирована в Руководстве ОЭСР [6]. Стандарты расчета показателей производительности изложены в другом методическом документе ОЭСР [7].



М. Д. Уилкинсон и др. разработали принципы FAIR в управлении научными данными — доступность, открытость, интероперабельность и возможность повторного использования [10]. Дж. Ф. Хэйр и др. описали процедуры подготовки данных для многомерного анализа, включая методы выявления аномальных наблюдений [9]. А. Чарнес и др. определили требования к структуре данных для метода DEA [8].

ВНТютюрюков уделил особое внимание проблеме сопоставимости данных в региональных исследованиях, показав, что различия в методологии расчета показателей в разных регионах искажают результаты анализа [14]. Этот вывод имеет непосредственное отношение к данному исследованию.

СВБрунов представил опыт формирования территориальной базы данных для оценки эффективности государственных услуг и показал способы решения проблемы полноты данных [13]. ААФирсова и соавторы описали процедуру отбора и подготовки данных для анализа DEA [12]. ВАПлотников проанализировал специфику статистики сектора услуг [11].

А. Турсинбаев представил опыт использования национальных статистических данных при проведении эконометрического анализа сектора услуг [15]. А. Шералиев проанализировал показатели сектора и отметил ограничения базы данных [16]. Г. Ерназаров определил состав показателей, необходимых для оценки [17]. Правовая основа национальной статистической системы устанавливается соответствующим законом [19].

Методы исследования

В исследовании использовались методы систематического анализа, классификации, статистического анализа и алгоритмизации.

Многомерный подход, предложенный Р.Ю. Ваном и Д.М. Стронгом [5], был взят за основу для разделения мер качества данных и адаптирован из требований эконометрического анализа.

При разработке критериев выбора методов восстановления утраченных значений использовалась классификация механизмов потерь на основе Р. Дж. А. Литтла и Д. Б. Рубина [3].

Были сопоставлены три метода выявления выбросов: метод межквартильного размаха (IQR), метод стандартизированных остатков и метод расстояния Махаланобиса [9]. Была определена применимость каждого из них.



В качестве основы для выбора методов нормализации данных использовалась методология Руководства по составным показателям ОЭСР [6].

Методология была протестирована на региональной панельной базе данных за период 2019–2024 гг.; источником были официальные данные Статистического агентства [18].

Результаты и обсуждение

Первый результат исследования заключается в определении шести аспектов качества данных для эконометрического анализа и разработке процедуры проверки для каждого из них.

Таблица 1

Меры оценки качества эмпирических данных и процедуры проверки

Измерение	Содержание	Процедура проверки	Последствия нарушения
Полнота	Наличие всех необходимых наблюдений	Расчет процента выпавших значений	Размер выборки уменьшается; происходит сдвиг.
Точность	Соответствие ценностей реальной ситуации	Сравнение с альтернативным источником; логический контроль	Шансы меняются.
Сравнимость	Единообразие методики расчета для всех наблюдений	Анализ методологических замечаний	Региональные различия создаются искусственно.
Последовательность	Непрерывность и отсутствие перерывов во времени.	Визуальный и статистический осмотр строк	Тенденция нарушена; прогноз оказался неверным.
Своевременность	Релевантность информации	Проверьте дату публикации.	Выводы будут актуальны в контексте устаревшей ситуации.
Внутренняя согласованность	Сохранение логических связей между показателями	Уравнения для проверки баланса	Спецификация модели не работает.

Третье измерение, сопоставимость, является наиболее важным для регионального анализа. Если показатель рассчитывается с использованием одной методологии в одном регионе, а другой — в другом, то разница между ними отражает методологическое различие, а не реальное



экономическое различие. Эта проблема, отмеченная Тютюрюковым [14], выделена в качестве отдельного этапа проверки в данной методологии.

2. Шестиэтапный алгоритм формирования базы данных

Разработан полный алгоритм формирования базы данных. Каждый этап алгоритма завершается определенным результатом, и только после этого происходит переход к следующему этапу.

Таблица 2

Алгоритм формирования эмпирической базы данных

Этап	Содержание	Чем заняться	Результат выполнения
I	Выбор показателей	Составьте список переменных на основе теоретической модели; обоснуйте каждую из них.	Список поддерживаемых индикаторов и их определения.
II	Сбор данных	Определите источники; приведите данные в единый формат.	Исходный массив данных
III	Гарантия качества	Шестимерная диагностика; проблемы с записью	Отчет о диагностике качества
IV	Чистка и реставрация	Восстановление утраченных значений; работа с аномальными наблюдениями; исправление ошибок.	Очищенный массив данных
V	Трансформация	Приведение к сопоставимым ценам; нормализация; логарифмизация; расчет производных показателей.	База данных, готовая к анализу
VI	Документация	Подготовьте метаданные; зафиксируйте все изменения.	База данных паспортов

Шестой этап — документирование — часто упускается из виду на практике, но является ключевым условием обеспечения проверяемости научных исследований. Паспорт базы данных должен содержать: четкое определение и источник каждого показателя; период наблюдения и территориальный охват; долю пропущенных значений и метод их



восстановления; порядок обработки аномальных наблюдений; использованные преобразования.

При работе с пропущенными значениями выбор метода должен зависеть от механизма пропущенных значений и доли пропущенных значений [3].

Таблица 3

Критерии выбора метода восстановления утраченных значений

процент отсева	Механизм	Рекомендуемый метод	Ограничения
До 5%	Любой	Исключить наблюдение или заполнить средним значением.	Размер выборки уменьшен; дисперсия искусственно занижена.
5–15%	МКАР или МАР	Импутация на основе регрессии; интерполяция по смежным периодам.	Силу связи можно искусственно увеличить.
15–30%	МАР	Множественная импутация [4]	Вычислительная сложность возрастает
Более 30%	Любой	Удаление индикатора из модели или замена его альтернативным	Требуется пересмотр спецификации модели.

Важно отметить, что в случаях, когда отсев происходит по неслучайному механизму (MNAR) — например, когда компании с низкими показателями не предоставляют отчетность, — ни один метод импутации не может полностью устранить смещение. В таких случаях причина отсева должна быть проанализирована и четко указана в качестве ограничения исследования.

Выбросы — это наблюдения, которые значительно отклоняются от общего распределения выборки. Они могут оказывать непропорционально большое влияние на результаты эконометрического анализа. Для их выявления используются три метода:

– Метод межквартильного размаха (IQR): значения меньше $Q1 - 1,5 \cdot IQR$ или больше $Q3 + 1,5 \cdot IQR$ считаются аномальными. Подходит для одномерного анализа.

– Метод стандартизированных остатков: анализируются стандартизированные остатки со значением модуля больше 3. Используется в контексте регрессии.

– Метод расстояния Махаланобиса: выявляет многомерные аномалии, то есть находит наблюдения, которые являются нормальными для каждой переменной, но необычными для их комбинации [9].

При обнаружении аномального наблюдения автоматически исключать его нецелесообразно. Сначала необходимо определить его природу: является ли это технической ошибкой (например, неправильно указана единица измерения) или это реальный, но необычный случай? В первом случае значение корректируется, во втором — наблюдение сохраняется, но проверяется чувствительность результатов к нему.

Структура панельных данных напрямую определяет применимые эконометрические методы. Основными параметрами являются количество поперечных единиц (N) и количество временных периодов (T).

Размер панели: $NT = N \times T$, требование для DEA: $N \geq 3 \times (m + s)$ (1)

где m — количество входных переменных; s — количество выходных переменных. Это правило обеспечивает дискриминационную способность метода DEA: если количество переменных больше количества наблюдений, почти все единицы считаются эффективными, и анализ теряет смысл [8].

Для динамических панельных моделей (GMM) существует дополнительное требование: желательно, чтобы T было небольшим, а N — большим. В случае большого T количество инструментов чрезмерно возрастает, и качество оценки снижается. Из-за ограниченного числа регионов в национальном контексте ($N \approx 14$) основной способ увеличения размера панели — это увеличение временной глубины или сокращение единицы анализа до уровня предприятия, а не региона.

Предлагается составной индекс для оценки уровня готовности базы данных к анализу:

$$DRI = 0,25 Q1 + 0,20 Q2 + 0,20 Q3 + 0,15 Q4 + 0,10 Q5 + 0,10 Q6 \quad (2)$$

где $Q1...Q6$ — нормализованные значения шести показателей качества, перечисленных в таблице 1. Значение индекса варьируется от 0 до 1.

Целесообразно придавать наибольший вес показателям полноты и точности: если эти два показателя нарушены, даже высокие баллы по остальным показателям не будут иметь смысла. База данных с $DRI \geq 0,80$ считается готовой к анализу; в диапазоне 0,60–0,79 ее можно использовать с осторожностью; если она ниже 0,60, базу данных необходимо переформатировать.

Научная значимость предложенной методологии заключается в формализации процесса формирования базы данных и превращении его в



полноценный этап исследования. На практике этот этап часто рассматривается как техническая задача и не описывается, что делает невозможным повторение и проверку результатов исследования.

Практическая ценность данной методологии заключается в трех аспектах: она обеспечивает четкие процедуры проверки качества данных; снижает субъективность при выборе метода восстановления; и позволяет количественно оценить готовность базы данных.

Следует отметить следующие ограничения. Во-первых, весовые коэффициенты индекса DRI основаны на экспертной оценке. Во-вторых, некоторые показатели качества — например, точность — требуют альтернативного источника оценки, который не всегда доступен. В-третьих, ограниченное количество регионов в национальных условиях является объективным фактором, который не может быть преодолен методологическим решением — это требует изменения единицы анализа.

Выводы и предложения

В результате исследования были сделаны следующие выводы.

1. Надежность эконометрических исследований в большей степени зависит от качества данных, чем от качества модели; поэтому формирование базы данных следует формализовать как полноценный этап исследования.

2. Качество данных оценивается по шести параметрам: полнота, точность, сопоставимость, согласованность, актуальность и внутренняя согласованность; сопоставимость является наиболее важным параметром для регионального анализа.

3. Метод восстановления пропущенных значений следует выбирать в зависимости от механизма и доли пропущенных данных; ни один метод импутации не сможет полностью устранить сдвиг при механизме MNAR.

4. Неправильно автоматически исключать аномальные наблюдения: сначала необходимо определить, являются ли они технической ошибкой или подлинной аномалией.

5. Структура панельных данных определяет используемые методы; для DEA количество наблюдений должно быть как минимум в три раза больше суммы переменных.

Практические рекомендации:

1. Внедрить практику составления паспорта базы данных в научных исследованиях и его представления в качестве приложения к статье; это обеспечит проверяемость результатов.



2. Стандартизация методологии расчета региональных показателей Статистическим агентством и публичная публикация методических пояснений; это позволит решить проблему сопоставимости.

3. Расширить временную глубину региональных рядов данных по сектору услуг — сформировать и архивировать непрерывные ряды данных как минимум за десять лет.

4. Предоставить исследователям доступ к анонимизированным данным корпоративного уровня для расширения масштабов анализа.

5. Внедрить практику расчета Индекса готовности базы данных (DRI) и использования его в качестве критерия оценки научных исследований.

Список использованных литератур

1. Wooldridge J.M. Econometric Analysis of Cross Section and Panel Data. 2nd ed. – Cambridge, MA: The MIT Press, 2010. – 1064 p.
2. Baltagi B.H. Econometric Analysis of Panel Data. 6th ed. – Cham: Springer, 2021. – 424 p.
3. Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data. 3rd ed. – Hoboken: John Wiley & Sons, 2019. – 449 p.
4. Rubin D.B. Multiple Imputation for Nonresponse in Surveys. – New York: John Wiley & Sons, 1987. – 258 p.
5. Wang R.Y., Strong D.M. Beyond Accuracy: What Data Quality Means to Data Consumers // Journal of Management Information Systems. – 1996. – Vol. 12, No. 4. – P. 5–33.
6. OECD. Handbook on Constructing Composite Indicators: Methodology and User Guide. – Paris: OECD Publishing, 2008. – 162 p.
7. OECD. Measuring Productivity: OECD Manual. – Paris: OECD Publishing, 2001. – 156 p.
8. Charnes A., Cooper W.W., Rhodes E. Measuring the Efficiency of Decision Making Units // European Journal of Operational Research. – 1978. – Vol. 2, No. 6. – P. 429–444.
9. Hair J.F., Black W.C., Babin B.J., Anderson R.E. Multivariate Data Analysis. 8th ed. – Andover: Cengage Learning, 2019. – 813 p.
10. Wilkinson M.D. et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship // Scientific Data. – 2016. – Vol. 3, Article 160018. – P. 1–9.
11. Плотников В.А. Экономика услуг как элемент современной экономической системы // Известия Санкт-Петербургского государственного экономического университета. – 2019. – № 4. – С. 12–19.



12. Фирсова А.А., Челнокова О.Ю. Оценка эффективности региональных инновационных систем методом DEA // Известия Саратовского университета. Серия: Экономика. Управление. Право. – 2019. – Т. 19, вып. 2. – С. 145–153.
13. Брунов С.В. Эконометрическая оценка эффективности государственных услуг в регионах России // Вопросы государственного и муниципального управления. – 2020. – № 3. – С. 88–107.
14. Тютюрюков В.Н. Оценка эффективности региональных социальных услуг: методологические подходы // Региональная экономика: теория и практика. – 2021. – № 5. – С. 44–58.
15. Tursinbayev A. Xizmatlar sohasining yalpi ichki mahsulotga ta'sirini ekonometrik baholash // Iqtisodiyot va innovatsion texnologiyalar. – Toshkent, 2021. – № 4. – B. 121–130.
16. Sheraliyev A. Xizmatlar sohasini rivojlantirishning zamonaviy tendensiyalari va omillari // O'zbekiston iqtisodiy axborotnomasi. – Toshkent, 2022. – № 6. – B. 55–63.
17. Ernazarov G'. Xizmat ko'rsatish sohasi samaradorligini baholash ko'rsatkichlari tizimi // Iqtisodiyot va ta'lim. – Toshkent, 2022. – № 3. – B. 78–86.
18. O'zbekiston Respublikasi Statistika agentligi ma'lumotlari. Xizmatlar sohasi ko'rsatkichlari. – Toshkent, 2025. // www.stat.uz.
19. O'zbekiston Respublikasining «Rasmiy statistika to'g'risida»gi Qonuni, O'RQ-575-son, 2019-yil 11-noyabr. // www.lex.uz.
20. Tursunova Sh.S. Xizmatlar sektorini boshqarish samaradorligini oshirish mexanizmini takomillashtirish: PhD dissertatsiyasi.

