

Ермаков Александр Валентинович

кандидат технических наук, старший преподаватель Инженерно-экономического института, Национальный исследовательский университет «МЭИ», Москва. Author ID: 208495

Электронный адрес: YermakovAIV@mpei.ru

Aleksandr V. Ermakov

Ph.D. of Engineering Sciences, Senior Lecturer at the Institute of economics and management, National Research University "MPEI", Moscow. Author ID: 208495

E-mail address: YermakovAIV@mpei.ru

Федосов Артемий Николаевич

студент, Национальный исследовательский университет «МЭИ», Москва. ORCID: 0009-0004-1397-3351, SPIN-код: 2157-1210, AuthorID: 1300374

Электронный адрес: artemfedosov.00@mail.ru

Artemiy N. Fedosov

Student, National Research University "MPEI", Moscow. ORCID: 0009-0004-1397-3351, SPIN-code: 2157-1210, AuthorID: 1300374

E-mail address: artemfedosov.00@mail.ru

Сало Анна Андреевна

студент, Национальный исследовательский университет «МЭИ», Москва. ORCID: 0009-0009-2640-8614

Электронный адрес: anyasalo@yandex.ru

Anna A. Salo

Student, National Research University "MPEI", Moscow. ORCID: 0009-0009-2640-8614

E-mail address: anyasalo@yandex.ru

МЕТОДИКА ПРЕДВАРИТЕЛЬНОЙ ЭКСПЕРТНОЙ ОЦЕНКИ КАЧЕСТВА ДАННЫХ ВРЕМЕННЫХ РЯДОВ ДЛЯ ЦЕЛЕЙ ПРОГНОЗИРОВАНИЯ

Аннотация. В статье исследуется проблема зависимости точности прогнозов в сложных системах от качества исходных данных. Анализируются ключевые аспекты, определяющие пригодность исторических данных для целей прогнозирования: горизонт прогнозирования, объем и репрезентативность выборки, консистентность, стационарность, наличие аномальных наблюдений, релевантность, частота и регулярность поступления данных. Особое внимание уделяется методологическим ограничениям классических подходов к оценке качества данных. Предлагается оригинальная методика предварительной экспертной оценки, основанная на принципах системного анализа и включающая три последовательных этапа: 1) оценка достаточности объема данных с использованием модифицированного принципа Парето; 2) верификация регулярности данных через коэффициент вариации; 3) идентификация аномалий с применением концепции «труба прогноза». Для количественной оценки значимости факторов разработана система взвешивания на базе критериев теории принятия решений (Лапласа, Вальда, Сэвиджа, Гурвица). Методика позволяет систематизировать процесс предварительного анализа, минимизировать риски использования нерепрезентативных данных и обоснованно выбирать методы прогнозирования. Практическая значимость заключается в сокращении временных затрат на подготовку данных и повышении достоверности прогнозных моделей в условиях неопределенности.

Методика предварительной экспертной оценки качества данных временных рядов
для целей прогнозирования

Ключевые слова: качество данных, прогнозирование, временные ряды, системный анализ, экспертная оценка, теория принятия решений, стационарность, релевантность данных.

Для цитирования: Ермаков А.В., Федосов А.Н., Сало А.А. Методика предварительной экспертной оценки качества данных временных рядов для целей прогнозирования // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ, управление. 2025. № 4. С. 88 – 95. DOI: 10.18137/RNUV9187.25.04.P.88

THE METHODOLOGY OF PRELIMINARY EXPERT ASSESSMENT
OF THE QUALITY OF TIME SERIES DATA FOR FORECASTING PURPOSES

Abstract. The article examines the problem of the dependence of the accuracy of forecasts in complex systems on the quality of the source data. The key aspects determining the suitability of historical data for forecasting purposes are analyzed: the forecasting horizon, the volume and representativeness of the sample, consistency, stationarity, the presence of anomalous observations, relevance, frequency and regularity of data receipt. Special attention is paid to the methodological limitations of classical approaches to data quality assessment. An original method of preliminary expert assessment is proposed, based on the principles of system analysis and including three successive stages: 1) assessment of the sufficiency of the data volume using the modified Pareto principle, 2) verification of the regularity of data through the coefficient of variation, 3) identification of anomalies using the “forecast tube” concept. To quantify the importance of factors, a weighting system has been developed based on the criteria of decision theory (Laplace, Wald, Savage, Hurwitz). The methodology makes it possible to systematize the process of preliminary analysis, minimize the risks of using unrepresentative data, and reasonably choose forecasting methods. The practical significance lies in reducing the time spent on data preparation and increasing the reliability of predictive models in conditions of uncertainty.

Keywords: data quality, forecasting, time series, system analysis, expert assessment, decision theory, stationarity, data relevance.

For citation: Ermakov A.V., Fedosov A.N., Salo A.A. (2025) The Methodology of Preliminary Expert Assessment of the Quality of Time Series Data for Forecasting Purposes. *Vestnik of Russian New University. Series: Complex Systems: Models, analysis, management*. No. 4. Pp. 88 – 95. DOI: 10.18137/RNUV9187.25.04.P.88 (In Russian).

Введение

В сложных системах точность прогнозов во многом зависит от качества входных данных. Неполные, противоречивые или сильно зашумленные данные могут привести к ложным сигналам и снижению точности моделей. Исторические данные используются для анализа, прогнозирования будущих событий, обучения моделей и понимания развития чего-либо. Они собирают за определенный промежуток времени информацию о прошлых событиях, значениях, тенденциях [1].

Понятие «хорошие данные» является зависимым от контекста. То, что является ценной информацией для одного класса задач, может оказаться информационным шумом для другого.

Зависимость от горизонта прогнозирования. По С.А. Касперовичу, «горизонт прогнозирования – это число периодов в будущем, которые охватывает прогноз» [2, с. 34]. Бывает три вида горизонта прогнозирования:

- краткосрочное;
- среднесрочное;
- долгосрочное [2].

Для долгосрочного прогноза требуется гораздо больше данных, чем для краткосрочного, и они должны быть более обобщенными. Краткосрочные прогнозы более точны, поскольку учитывают недавние события, в то время как долгосрочные прогнозы менее точны из-за большого количества переменных, которые могут измениться за длительный период времени [3].

Зависимость от объема выборки и полноты данных. Как правило, чем больше объем данных, тем надежнее модели, иначе может возникнуть ее недообучение. Причиной недообучения становится излишняя простота модели – алгоритму не удастся построить достаточно сложную аппроксимирующую функцию, позволяющую принимать решения, которые хорошо подходили бы для текущей задачи [4]. Поэтому любые *дополнительные* релевантные данные будут полезны, будь то новые данные или исторические наблюдения. Вместе с тем не менее важна и *полнота* данных. Недостающие значения, нерепрезентативная выборка или отсутствие ключевых признаков могут исказить модель независимо от её сложности или объема обучающей выборки. В такой ситуации модель может сделать некорректные выводы, которые не будут работать в реальных условиях. Однако не стоит забывать о переобучении модели. Переобучение модели возникает, когда алгоритм хорошо аппроксимирует обучающую выборку, теряя при этом обобщающую способность и возможность делать прогнозы для новых данных. Риск переобучения возрастает с увеличением сложности модели [5].

Зависимость от консистентности данных. Данные могут быть несогласованными, что может привести к проблемам при интеграции и анализе данных. Консистентность является важнейшим понятием теории управления данными.

Зависимость от стационарности. Временной ряд называется стационарным, когда наблюдения колеблются относительно постоянного уровня или среднего значения [6]. Большинство классических моделей прогнозирования требуют стационарности ряда для построения надежных прогнозов. Для стационарных рядов применяются относительно простые и интерпретируемые модели, такие как ARIMA. В случае нестационарности (когда ряд имеет тренд, сезонность или изменяющуюся дисперсию) подход к прогнозированию меняется, и тогда используются другие модели прогнозирования: LSTM, XGBoost, Holt-Winters. Поэтому в зависимости от стационарности и модели прогнозирования качество данных может быть как «хорошее», так и «плохое».

Стационарность ряда проверяется с помощью статистических критериев. Классическим методом служит расширенный тест Дики – Фуллера, который тестирует гипотезу о единичном корне. Опровержение этой гипотезы на заданном уровне значимости означает, что ряд можно считать стационарным [7].

Зависимость от шумов и выбросов. Выбросы называют точки данных, которые существенно отклоняются от характерных закономерностей в массиве информации. Такие объекты могут существовать в виде единичных экземпляров или формировать локальные кластеры [8]. Если выбросы и шумы являются информативной частью данных, используются методы и инструменты для их устранения.

Различные методы машинного обучения имеют разную чувствительность к выбросам, и это нужно учитывать при выборе метода прогнозирования данных.

Методика предварительной экспертной оценки качества данных временных рядов
для целей прогнозирования

Зависимость от релевантности данных. Исторические данные должны не только быть технически качественными, но и содержательно соответствовать решаемой задаче. Релевантность подразумевает, что данные отражают именно те процессы и факторы, которые модель должна прогнозировать. Например, данные о продажах зонтов за последний год могут быть нерелевантны для прогноза на следующий год, если в текущем году произошло резкое изменение климата.

Зависимость от частоты и регулярности данных. Для временных рядов критическую роль играет также частота дискретизации и регулярность поступления данных. Высоочастотные данные (например, показания датчиков каждую секунду) позволяют уловить краткосрочные аномалии и быстрые процессы, но могут требовать сложной обработки и агрегации для долгосрочного прогноза. Низкочастотные данные (например, ежемесячные отчеты), напротив, скрывают краткосрочные колебания, но лучше подходят для выявления общих трендов. Не менее важна регулярность: пропуски во времени, непостоянная частота или изменяющиеся временные метки вносят дополнительную нестационарность и требуют применения специальных методов предобработки [9].

Методика предварительной экспертной оценки качества данных

Предлагается дополнить существующий инструментарий статистических и визуальных методов оценки качества данных временных рядов этапом предварительной экспертной оценки, основанной на принципах системного анализа. Разработанная методика, несмотря на эмпирический характер, позволяет существенно сократить временные затраты на предварительный анализ качества данных. Исходной предпосылкой методики является предположение о свойстве историчности прогнозируемой системы, означающее, что ее текущее и будущее состояние полностью определяются предыдущим состоянием.

Этап 1. Оценка достаточности объема и репрезентативности данных с учетом горизонта прогнозирования. Для количественной оценки данного параметра предлагается использовать модифицированный принцип Парето: существующий объем данных должен составлять не менее 80 % от общего требуемого объема, при этом прогнозируемые значения не должны превышать 20 % от общего количества наблюдений. Например, при наличии данных за 8 месяцев допустимый горизонт прогнозирования составляет не более двух месяцев.

Этап 2. Верификация регулярности данных и оценка критичности пропущенных значений. Базовое ограничение заключается в том, что минимальный период прогнозирования не может быть меньше временного интервала между наблюдениями в исходной зависимости, поскольку в противном случае возникает ситуация полной неопределенности, описываемая равномерным законом распределения вероятностей. Достоверное прогнозирование возможно только при условии, что минимальный период прогнозирования содержит не менее трех контрольных точек, что позволяет применять метод PERT и другие вероятностные модели.

Для количественной оценки регулярности временных интервалов предлагается рассчитывать среднее время поступления данных и среднеквадратическое отклонение для каждого интервала прогнозирования. Качественным индикатором регулярности служит коэффициент вариации, определяемый как отношение среднеквадратического отклонения к среднему значению. При значениях коэффициента вариации, не превышающих 0,1, отклонения в регулярности данных считаются статистически незначимыми. Критичность пропусков данных классифицируется следующим образом: при изменении коэффициента

вариации от 0,1 до 0,2 критичность средняя; при значениях от 0,2 до 0,33 – высокая, что требует исключения соответствующего интервала из анализа.

Этап 3. Идентификация аномальных наблюдений. В рамках данного этапа исходят из предположения, что после исключения трендовой компоненты значения прогнозируемой величины должны подчиняться нормальному закону распределения на всех отрезках исследуемой зависимости. Для верификации этого предположения зависимость разбивается на интервалы, соответствующие планируемому горизонту прогнозирования.

Вводится концепция «трубы прогноза», определяемая потенциалом системы (максимальные значения при оптимальных условиях) и минимумом системы (минимальные значения при наихудших условиях). Согласно данной концепции, исторические значения системы должны находиться в пределах указанного диапазона. Для нормального распределения выборка считается репрезентативной, если 68,2 % данных укладываются в одно стандартное отклонение от среднего значения, 95,4 % – в два стандартных отклонения и 99,7 % – в три стандартных отклонения. Наблюдения, выходящие за указанные пределы, классифицируются как выбросы, при этом необходима дополнительная проверка на предмет систематического характера таких отклонений.

Методика взвешивания функции полезности

Для количественной оценки значимости факторов предлагается использовать аппарат теории принятия решений.

Критерий Лапласа применяется при отсутствии априорной информации о значимости факторов, предполагая равномерное распределение весов.

Критерий Вальда используется при известном негативном влиянии определенных факторов с распределением 80 % весового коэффициента среди наиболее критичных факторов в соответствии с принципом Парето.

Критерий максимального оптимизма применяется при идентификации доминирующего позитивного фактора, которому присваивается 80 % веса.

Критерий крайнего пессимизма применяется при доминировании негативного фактора с аналогичным распределением весов.

Критерий Сэвиджа эффективен при множественных негативных факторах с неопределенным соотношением влияния и предполагает равномерное распределение 80 % весов между ними.

Критерий Гурвица позволяет учитывать экспертные оценки через введение поправочного коэффициента, регулирующего баланс между оптимистичными и пессимистичными сценариями, при этом максимальный вес любого фактора не должен превышать 80 %.

Практическая реализация оценки

Методика оценки критериев качества данных была реализована на языке программирования Python в среде разработки Visual Studio Code. Для апробации методики были отобраны два набора исторических данных¹, визуальное представление которых приведено на Рисунке 1. Первый временной ряд демонстрирует свойства стационарности – колебания вокруг постоянного уровня без выраженного тренда. Дополнительный анализ подтвердил отсутствие пропущенных значений, дубликатов записей, а также соответствие требований регулярности и информативной частоты наблюдений. Второй набор данных,

¹ Загрузка исторических данных о погоде // Meteoblue. URL: <https://www.meteoblue.com/ru/погода/archive/export> (дата обращения: 25.10.2025).

Методика предварительной экспертной оценки качества данных временных рядов
для целей прогнозирования

напротив, содержит аномальное наблюдение (выброс), а также характеризуется наличием пропущенных значений и дублирующихся записей, что позволяет проверить устойчивость разработанной методики к различным типам несовершенств данных.

Таким образом, на основе критерия Гурвица была проведена комплексная оценка пригодности исторических данных и получены следующие оценки.

1. Достаточность объема и охвата – соответствие горизонта данных целям прогноза (кратко-, средне- или долгосрочному). Оценка критерия 0,05.

2. Полнота – минимизация пропущенных значений и систематических смещений в данных. Оценка критерия 0,15.

3. Непротиворечивость (консистентность) – отсутствие внутренних противоречий и соответствие единым стандартам. Оценка критерия 0,15.

4. Стационарность – соответствие статистических свойств ряда требованиям выбранного класса моделей. Оценка критерия 0,2.

5. Информативная чистота – уровень шума и количество выбросов должны быть совместимы с выбранным алгоритмом машинного обучения. Оценка критерия 0,15.

6. Релевантность – содержательное соответствие данных текущему контексту и будущим сценариям работы системы. Оценка критерия 0,15.

7. Частота и регулярность данных – временные интервалы между наблюдениями должны быть согласованы с динамикой прогнозируемых процессов в системе. Оценка критерия 0,1.

Выведенный программой результат соответствует визуальному представлению о наборах исторических данных (см. Рисунок 2).

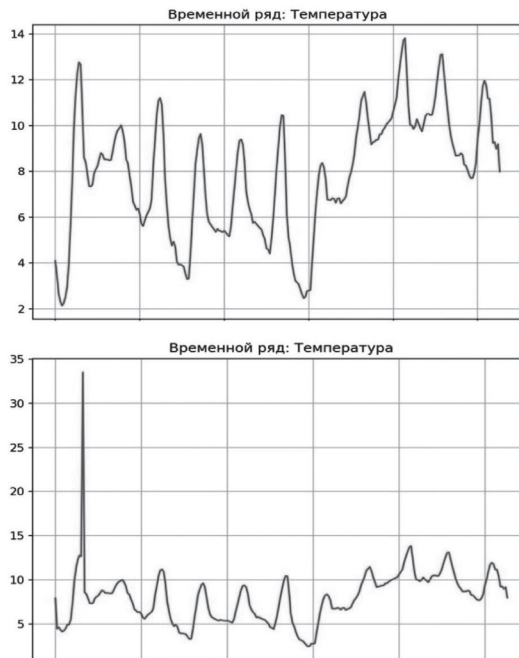


Рисунок 1. Наборы исторических данных

Источник: здесь и далее рисунки выполнены авторами.

ИНТЕГРАЛЬНЫЙ ПОКАЗАТЕЛЬ ПРИГОДНОСТИ: 0.8627
 СТАТУС: Исторические данные являются хорошими для использования прогнозирования

ИНТЕГРАЛЬНЫЙ ПОКАЗАТЕЛЬ ПРИГОДНОСТИ: 0.6136
 СТАТУС: Исторические данные можно использовать для прогнозирования, однако, есть некоторые недочеты: зафиксирована аномалия и несколько дубликатов

Рисунок 2. Результат оценки критериев

Только совокупная оценка по всем этим критериям позволяет сделать обоснованный вывод о том, являются ли имеющиеся исторические данные надежным фундаментом для построения прогнозных моделей.

Заключение

Предложенная методика позволяет систематизировать процесс предварительного анализа данных, минимизировать риски использования нерепрезентативных исторических данных и обоснованно выбирать методы прогнозирования, соответствующие выявленным свойствам данных.

В качестве перспективных направлений дальнейших исследований целесообразно рассматривать развитие оценки критериев весов.

Литература

1. Литвинчук С.Ю. Информационные технологии в экономике. Анализ и прогнозирование временных рядов с помощью Excel : Учебное пособие. Нижний Новгород : ННГАСУ, 2010. 78 с.
2. Касперович С.А. Прогнозирование и планирование экономики : Курс лекций. Минск : Белорусский государственный технологический университет, 2015. 170 с. EDN GITNKO.
3. Капитанова О.В. Прогнозирование социально-экономических процессов : учебно-методическое пособие. Нижний Новгород : Нижегородский госуниверситет, 2016. 74 с. URL: http://www.unn.ru/books/met_files/Kapitanova.pdf (дата обращения: 25.10.2025).
4. Пател А. Прикладное машинное обучение без учителя с использованием Python / Пер. с англ. А.Г. Гузикевича. СПб. : Диалектика, Вильямс, 2020. 432 с. ISBN 978-5-907144-99-6.
5. Бринк Х., Ричардс Дж., Феверолф М. Машинное обучение. СПб. : Питер, 2023. 336 с. ISBN 978-5-496-02989-6.
6. Крутиков В.К., Дорожкина Т.В., Зайцев Ю.В., Федорова О.В. Макроэкономическое планирование и прогнозирование : Учебно-методическое пособие. Калуга : Эйдос, 2014. 112 с. ISBN 978-5-905697-99-9. EDN TNGYGB.
7. Нильсен Э. Практический анализ временных рядов: прогнозирование со статистикой и машинное обучение / Пер. с англ. Д.И. Ключина. СПб. : Вильямс, 2021. 544 с. ISBN 978-5-907365-04-9.
8. Вольфсон М.Б. Анализ данных : учебное пособие. СПб. : СПбГУТ, 2015. 82 с.
9. Шорохова И.С., Кисляк Н.В., Мариев О.С. Статистические методы анализа : учебное пособие. Екатеринбург : Изд-во Уральского университета, 2015. 300 с. ISBN 978-5-7996-1633-5. EDN VGKLAL.

References

1. Litvinchuk S.Yu. (2010) *Informatsionnye tekhnologii v ekonomike. Analiz i prognozirovaniye vremennykh ryadov s pomoshch'yu Excel* [Information Technologies in Economics. Time Series Analysis and Forecast-

- ing with Excel] : Tutorial. Nizhny Novgorod : Nizhny Novgorod State University of Architecture and Construction Publ. 78 p. (In Russian).
2. Kasperovich S.A. (2015) *Prognozirovanie i planirovanie ekonomiki* [Forecasting and Planning the Economy] : Lecture Course. Minsk : Belarusian State Technological University Publ. 170 p. (In Russian).
3. Kapitanova O.V. (2016) *Prognozirovanie sotsial'no-ekonomicheskikh protsessov* [Forecasting Socioeconomic Processes] : Tutorial. Nizhny Novgorod : Lobachevsky State University of Nizhni Novgorod Publ. 74 p. URL: http://www.unn.ru/books/met_files/Kapitanova.pdf (accessed 25.10.2025). (In Russian).
4. Patel A.A. (2020) *Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data*. O'Reilly Media. 362 p. ISBN 1492035610. (Russian edition: transl. by A.G. Guzikovich. St. Petersburg : Dialektika, Williams Publ. 2019. 432 p.).
5. Brink H., Richards J., Fetherolf M. (2016) *Real-World Machine Learning*. Manning Publications. 264 p. ISBN 1617291927. (Russian edition: transl. by Ruzmaikina I. St. Petersburg : Piter Publ., 2023. 336 p.).
6. Krutikov V.K., Dorozhkina T.V., Zaitsev Yu.V., Fedorova O.V. (2014) *Makroekonomicheskoe planirovanie i prognozirovanie* [Macroeconomic Planning and Forecasting] : Tutorial. Kaluga : Eidos Publ. 112 p. ISBN 978-5-905697-99-9. (In Russian).
7. Nielsen E. (2021) *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*. O'Reilly Media. 504 p. ISBN 1492041602. (Russian edition: transl. by D.I. Klyushin. St. Petersburg : Williams. 2021. 544 p.).
8. Wolfson M.B. (2015) *Analiz dannykh* [Data Analysis] : Tutorial. St. Petersburg : Bonch-Bruевич Saint Petersburg State University of Telecommunications Publ. 82 p. (In Russian).
9. Shorokhova I.S., Kislyak N.V., Mariev O.S. (2015) *Statisticheskie metody analiza* [Statistical Methods of Analysis] : Tutorial. Ekaterinburg : Ural State University Publ. 300 p. ISBN 978-5-7996-1633-5. (In Russian).

Поступила в редакцию: 27.10.2025

Received: 27.10.2025

Поступила после рецензирования: 26.11.2025

Revised: 26.11.2025

Принята к публикации: 12.12.2025

Accepted: 12.12.2025