УДК 004.95 **DOI** 10.15622/ia.25.2.4

А.В. Воробьев, Г.Р. Воробьева

КОНТЕКСТНО-ЗАВИСИМЫЙ МЕТОД АДАПТИВНОЙ НАСТРОЙКИ ПАРАМЕТРОВ АВТОРЕГРЕССИОННЫХ МОДЕЛЕЙ ДЛЯ НЕСТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДОВ

Воробьев А.В., Воробьева Г.Р. **Контекстно-зависимый метод адаптивной настройки параметров авторегрессионных моделей для нестационарных временных рядов.**

Аннотация. Предлагается метод контекстно-зависимой настройки параметров авторегрессионных моделей для восстановления пропусков в нестационарных временных рядах. Ключевая особенность метода заключается в адаптивном выборе параметров модели ARIMA (p, d, q) на основе двух факторов контекста: длительности пропуска и уровня внешних возмущений в соответствующий период. В отличие от стандартных подходов автоматического подбора, ориентированных на глобальную оптимизацию для прогнозирования, разработанный алгоритм сужает пространство поиска моделей и осуществляет выбор оптимальной конфигурации с помощью локальной кросс-валидации, что позволяет учитывать специфические условия в области пропуска. Метод реализован в виде программного модуля на языке Python с модульной архитектурой, обеспечивающей вычислительную эффективность за счет кеширования и параллельных вычислений. Эффективность метода проверена в ходе эксперимента на реальных геомагнитных данных (компонента DBE_NEZ обсерватории Ловозеро). Результаты демонстрируют, что в условиях спокойной и слабовозмущенной геомагнитной обстановки (индекс SME = 50–200 нТл) метод обеспечивает высокую точность восстановления (R^2 = 0.71–0.85) для пропусков длиной от 5 до 120 минут. При этом показано, что точность закономерно снижается с ростом уровня возмущений, что отражает фундаментальное ограничение, связанное с возрастающей стохастичностью исходного сигнала. Предложенный подход обеспечивает интерпретируемость и адаптивность, открывая перспективы для создания инструментов восстановления данных в различных прикладных областях.

Ключевые слова: авторегрессионные модели, восстановление пропусков, нестационарные временные ряды, геомагнитные данные.

1. Введение. Автоматический анализ и обработка нестационарных временных рядов представляет собой фундаментальную задачу в области информатики, машинного обучения и прикладной математики. Особую сложность в этом контексте вызывает проблема восстановления пропусков в данных, которые неизбежно возникают в реальных системах мониторинга вследствие сбоев оборудования, потери сигнала или иных технологических причин. Традиционным и хорошо зарекомендовавшим себя аппаратом для работы с такими рядами являются авторегрессионные модели, в частности, класс ARIMA-моделей, который за счет операции дифференцирования явно предназначен для анализа нестационарных процессов [1]. Однако эффективность этих моделей для задач интерполяции критически зависит от корректного выбора их

параметров – порядков авторегрессии (p), дифференцирования (d) и скользящего среднего (q) в модели ARIMA (p, d, q).

Существующие стандартные методы автоматического подбора параметров, такие как широко используемый алгоритм, реализованный в пакете «forecast» для R [2], ориентированы, в первую очередь, на задачи прогнозирования и основаны на минимизации информационных критериев (Акаике, Байеса) и анализе автокорреляционных функций. Несмотря на свою распространенность, эти методы обладают существенным ограничением при использовании для восстановления данных: они осуществляют выбор параметров, ориентируясь исключительно на внутреннюю статистическую структуру всего временного ряда, игнорируя локальные условия вокруг пропуска. В реальных же условиях, особенно при работе с данными физического мониторинга, финансовой аналитики или телеметрии, характер ряда в области пропуска может существенно зависеть от внешних, зачастую измеримых факторов. Длительный пропуск, возникший в период высокой внешней нагрузки на систему, принципиально отличается от пропуска аналогичной длины в период ее спокойной работы. Применение стандартных параметров, оптимизированных для глобального прогноза, к локальной задаче интерполяции может приводить к значительным погрешностям восстановления.

Таким образом, актуальной научно-технической задачей является разработка методов настройки моделей восстановления данных, которые бы учитывали не только внутренние глобальные закономерности ряда, но и локальный контекст, в котором наблюдаются пропуски. Подобный контекстно-зависимый подход позволяет перейти от универсальных, но зачастую излишне общих решений к адаптивным, которые тонко подстраиваются под конкретные условия восстанавливаемого сегмента данных. Это особенно важно для работы с нестационарными рядами, характер которых может резко меняться под воздействием внешних факторов [3].

В данной работе предлагается контекстно-зависимый метод адаптивной настройки параметров авторегрессионных моделей для задач интерполяции. Его основное отличие заключается в том, что алгоритм выбора параметров (p, d, q) использует два ключевых контекстных фактора, специфичных для области пропуска: длительность пропуска в данных и количественную оценку интенсивности внешних возмущений в соответствующий период. В качестве демонстрации эффективности предложенного метода и валидации подхода он применен к задаче восстановления реальных геомагнитных данных, характеризующихся выраженной

нестационарностью и подверженных влиянию солнечной активности, что моделируется через соответствующий геомагнитный индекс. Целью исследования является демонстрация принципиальной возможности и преимущества включения локальной контекстной информации в процесс параметризации моделей для интерполяции временных рядов [4], что открывает путь к созданию более надежных и точных аналитических инструментов для восстановления данных в различных предметных областях.

2. Состояние вопроса. Развитие методов автоматического анализа временных рядов непосредственно связано с потребностью в эффективных инструментах для обработки неполных данных, возникающих в реальных системах мониторинга. Исторически проблема восстановления пропусков решалась либо простыми статистическими методами (линейная интерполяция, заполнение средним), либо с помощью скользящих средних. Однако эти подходы не учитывали автокорреляционную структуру данных, что приводило к значительным искажениям при работе с нестационарными рядами, характерными для геофизических, экономических и инженерных измерений.

Переломным моментом стало широкое внедрение в практику методологии Бокса–Дженкинса и класса авторегрессионных интегрированных моделей скользящего среднего (ARIMA), которые за счет оператора дифференцирования позволяют работать с нестационарностью, а компоненты авторегрессии и скользящего среднего эффективно идентифицируют внутреннюю динамику процесса [5]. Это создало теоретическую основу для использования данных моделей не только для прогнозирования, но и для интерполяции пропусков.

Следующим логическим шагом стала автоматизация трудоемкого процесса идентификации и оценивания параметров ARIMA-моделей. Алгоритмы, подобные реализованному в пакете forecast для R [1], стали промышленным стандартом, демонстрируя высокую эффективность в условиях, когда временной ряд может быть описан единой структурой на всем периоде наблюдений. Эти алгоритмы, основанные на комбинации статистических тестов и информационных критериев, минимизируют необходимость экспертного вмешательства. Однако их фундаментальное ограничение проистекает из самой цели их создания – они оптимизированы для глобального описания ряда в целях прогнозирования. При этом задача локального восстановления пропуска, особенно в условиях изменяющейся внешней среды, предъявляет иные требования к модели.

Параметры, обеспечивающие минимальную ошибку на всей исторической выборке, могут оказаться субоптимальными для аппроксимации поведения системы в конкретный, потенциально аномальный, интервал времени, отмеченный пропуском данных.

Расширение ARIMA до моделей с экзогенными переменными (ARIMAX) стало попыткой учесть влияние внешних факторов [6]. Этот подход показал свою эффективность в ситуациях, когда существует четко измеряемый внешний драйвер, коррелирующий с основным рядом. Тем не менее, его применимость для восстановления пропусков ограничена необходимостью наличия полных данных по экзогенным переменным за весь период интерполяции, что на практике часто невыполнимо. Более того, ARIMAX не решает проблему адаптивного выбора структурных параметров (p, d, q) – модель получает внешний вход, но ее архитектура остается фиксированной и подобранной глобально.

В последние годы исследовательский интерес сместился в сторону машинного обучения и гибридных моделей. Работы, подобные [4], демонстрируют потенциал комбинирования линейных авторегрессионных моделей с нелинейными аппроксиматорами, такими как искусственные нейронные сети, для учета сложных паттернов. Однако возрастающая сложность таких моделей часто требует больших объемов данных для обучения, может вести к переобучению и снижает интерпретируемость результатов. Кроме того, вопросы адаптивного выбора гиперпараметров и архитектуры гибридных моделей в зависимости от локального контекста пропуска остаются открытыми.

Параллельно в отдельных работах (например, [7]) отмечается, что характеристики самого пропуска (механизм возникновения, длина, расположение) являются критически важной мета-информацией. Эмпирически установлено, что точность большинства методов интерполяции снижается с увеличением длины пропуска. Тем не менее, эта зависимость редко формализуется в виде явного алгоритмического правила для перенастройки модели. Существующие решения, как классические, так и современные, действуют в парадигме, где алгоритм либо не принимает во внимание эту информацию, либо пассивно наблюдает ухудшение качества, но не меняет свою стратегию для компенсации данного эффекта.

Таким образом, обзор современных подходов выявляет существующую методологическую проблему. Недостаточно разработанными остаются методы, которые активно использовали бы контекстную информацию о пропуске и внешних условиях для динамической адаптации структурных параметров базовой модели, а не

только ее коэффициентов. Требуется переход от принципа «одна модель для всего ряда» к концепции «адаптивная модель для конкретного пропуска».

3. Описание предлагаемого метода. В основе предлагаемого подхода лежит принцип контекстно-зависимой адаптации [8], согласно которому параметры авторегрессионной модели, используемой для интерполяции пропуска, должны определяться не глобальными свойствами всего временного ряда, а локальными условиями, характерными для данного конкретного пропуска.

Данный принцип реализуется в виде формального алгоритма, который интегрирует два ключевых фактора контекста: количественную характеристику самого пропуска и меру интенсивности внешних возмущений в соответствующий период. Это позволяет перейти от модели с фиксированной структурой к адаптивной процедуре, где архитектура модели является функцией от контекста:

$$M = f(L, I), \tag{1}$$

где M – выбираемая модель ARIMA(p, d, q), L – длительность пропуска, I – индекс внешней активности.

Первый контекстный фактор, длительность пропуска L (измеряемая в количестве отсчетов временного ряда, где для минутных данных 1 отсчет = 1 минута), непосредственно влияет на сложность задачи восстановления. Эмпирически установлено, что с увеличением L авторегрессионная составляющая теряет предсказательную силу из-за ослабления корреляционной связи с граничными известными значениями. Для учета этого эффекта в предложенном подходе вводится эвристическое правило, связывающее максимальный допустимый порядок авторегрессии $p_{\max}$ с длиной пропуска:

$$p_{\max} = \max(1, \lfloor k / L \rfloor), \tag{2}$$

где k – эмпирический коэффициент, характеризующий типичный временной масштаб (лаг) значимой автокорреляции ряда. Он оценивается, например, как наибольший лаг, на котором модуль выборочной автокорреляционной функции (АКФ) ряда в окрестности пропуска превышает заданный порог значимости. Конструкция формулы обеспечивает, что $p_{\max}$ является целым положительным числом: обозначение $\lfloor k / L \rfloor$ означает взятие целой части от деления, а функция max(1, ...) гарантирует, что порядок авторегрессии будет не менее 1, даже если $k < L$ (т.е. при очень длинных пропусках или слабой

автокорреляции ряда). Это предотвращает попытку построения модели с $p = 0$, которая была бы бессмысленна в данном контексте.

Данное ограничение предотвращает переобучение модели на малом объеме релевантных данных, доступных для обучения в локальном окне.

Второй и наиболее значимый с точки зрения новизны фактор – это индекс внешней активности I, характеризующий интенсивность возмущений в системе в период возникновения пропуска. В качестве индекса внешней активности I может использоваться любой количественный дескриптор состояния внешней среды, коррелирующий с динамикой исследуемого процесса (например, индекс солнечной активности для геофизических данных, объем торгов для финансовых рядов, показатель нагрузки для инженерных систем).

Предполагается, что уровень I коррелирует с характером нестационарности ряда. В периоды высокой внешней активности (высокое I) процесс может демонстрировать поведение, близкое к белому шуму с резкими скачками, что требует увеличения веса компоненты скользящего среднего (MA) в модели.

Для формализации указанной зависимости предлагается использовать нормализованный индекс

$$I_{norm} = \frac{I - I_{\min}}{I_{\max} - I_{\min}}, \tag{3}$$

на основе которого вычисляется эвристический весовой коэффициент ω для смещения баланса между AR и MA компонентами в пространстве параметров.

Коэффициент ω определяется логистической функцией:

$$\omega(I_{norm}) = \frac{1}{1 + \exp(-\alpha(I_{norm} - \beta))}, \tag{4}$$

где параметры α и β калибруются на валидационной выборке. В логистической функции параметр $\beta \in (0, 1)$ задает пороговое значение I_{norm}, при котором $\omega = 0.5$, а параметр $\alpha > 0$ определяет скорость перехода функции от 0 к 1. Их конкретные оптимальные значения находятся методом поиска по сетке в процессе калибровки метода для конкретного типа данных.

Высокое значение ω (близкое к 1) указывает на предпочтительность моделей с повышенным порядком q.

Используя входные параметры L и I, алгоритм формирует ограниченное пространство допустимых моделей C. Это пространство представляет собой подмножество всевозможных троек (p, d, q), отобранное по правилам:

$$p \in [1, p_{\max}], \tag{5}$$

где $d \in \{0, 1, 2\}$ (определяется предварительным тестом Дики-Фуллера на стационарность остатков в локальном окне [9, 10]), q выбирается из диапазона, смещенного в зависимости от ω(I_{norm}):

$$q \in [q_{\min}, q_{\max}], \tag{6}$$

где $q_{\min} = \max(1, \lfloor \omega Q \rfloor)$, $q_{\max}$ – константа, задающая верхнюю границу. Здесь ω – весовой коэффициент из (4), Q – положительная масштабирующая константа, определяющая чувствительность порядка скользящего среднего q к изменению коэффициента ω. Конкретное значение Q, как и $q_{\max}$, выбирается на этапе калибровки метода и зависит от типичного диапазона порядков q, адекватных для моделирования данного типа рядов.

Таким образом, контекстные факторы не предписывают жестко единственную модель, а сужают область поиска до наиболее правдоподобных с точки зрения текущих условий допустимых моделей.

Ключевым этапом подхода является процедура кросс-валидации на смежных данных для выбора окончательной модели из множества C. Для этого в окрестностях пропуска, на известных данных, искусственно создается валидационный пропуск той же длины L. Для каждой модели из C производится ее обучение на усеченном ряду (с искусственным пропуском) и последующая интерполяция этого пропуска. Качество интерполяции оценивается на известных значениях, которые были искусственно скрыты. В качестве целевой функции оптимизации $Q(M_i)$ используется взвешенная комбинация среднеквадратичной ошибки (RMSE) и информационного критерия Акаике (AIC) [11], которая позволяет учитывать как точность аппроксимации, так и сложность модели, предотвращая излишнее усложнение:

$$Q(M_i) = \gamma RMSE_{norm} + (1 - \gamma) AIC_{norm}, \tag{7}$$

где $RMSE_{norm}$ и AIC_{norm} – нормализованные значения метрик, γ – весовой коэффициент.

Модель M_{opt}, доставляющая минимум функции Q, выбирается для финального восстановления целевого пропуска. Оптимальная модель M_{opt} (p_{opt}, d_{opt}, q_{opt}), доставляющая минимум функции Q, выбирается для финального восстановления целевого пропуска.

Представленные эмпирические правила (2) и (4) не являются теоретически выведенными, а представляют собой содержательную эвристику, направленную на решение двух ключевых практических проблем при восстановлении пропусков в реальных нестационарных рядах. Правило (2), связывающее максимальный порядок авторегрессии p_{max} с длиной пропуска L, ограничивает структурную сложность модели при дефиците релевантных данных для обучения, предотвращая переобучение. Правило (4), определяющее весовой коэффициент ω через индекс внешней активности I, адаптирует баланс между авторегрессионной и скользящей средней компонентами модели к изменяющемуся уровню стохастических внешних возмущений.

Калибровка гиперпараметров алгоритма – коэффициентов α и β логистической функции (4) и весового коэффициента γ целевой функции (7) – выполнялась на выделенной валидационной выборке исторических данных. Для оптимизации использовался метод полного перебора (Grid Search) [8] по предопределенным сеткам значений: параметр α варьировался в диапазоне от 5 до 20 с шагом 1, параметр β – в диапазоне от 0.3 до 0.7 с шагом 0.05. Данные диапазоны были определены эмпирически на основе предварительных экспериментов и охватывают область, в которой логистическая функция (4) демонстрирует плавный, но отчетливый переход выходного значения ω от состояния, близкого к 0 (для AR-компоненты), к состоянию, близкому к 1 (для MA-компоненты), на всем диапазоне нормализованного индекса I_{norm} от 0 до 1. Критерием оптимизации служило максимальное значение среднего коэффициента детерминации (R^2), достигнутое на множестве валидационных пропусков. Весовой коэффициент γ, управляющий балансом между ошибкой (RMSE) и сложностью модели (AIC) в выражении (7), был подобран аналогичным образом; его оптимальное значение составило 0.7. Этот процесс обеспечил объективный и воспроизводимый выбор гиперпараметров [12], адаптированных к специфике анализируемых геомагнитных рядов.

Подобный подход к адаптивному управлению сложностью модели согласуется с общей практикой в прикладном анализе данных и машинном обучении, где строгий теоретический вывод часто

дополняется или заменяется эмпирически обоснованными правилами для работы с конкретными данными. Обоснованность и адекватность данных эвристик проверяются в рамках процедуры кросс-валидации, описанной выше (например, выбор модели на основе минимизации функции $Q(M_i)$ в (7)), где они непосредственно влияют на формирование пространства множества допустимых моделей C и, как следствие, на итоговое качество восстановления. Правило (2) основывается на статистическом принципе, согласно которому для устойчивой оценки параметров авторегрессии порядка p требуется объем выборки, существенно превышающий p. Отношение k/L является упрощенной оценкой количества доступных независимых наблюдений на один оцениваемый параметр, что напрямую связано с проблемой переобучения. Ключевые параметры правил – коэффициент k в (2), а также α и β в (4) – не задаются априори, а калибруются на отдельной валидационной выборке для конкретного типа данных, что делает метод принципиально адаптируемым к различным предметным областям и условиям наблюдений.

После выбора оптимальной модели $M_{\text{opt}}(p_{\text{opt}}, d_{\text{opt}}, q_{\text{opt}})$ производится заключительный этап – непосредственная интерполяция исходного пропуска длины L. Модель обучается на всем доступном сегменте данных, окружающем пропуск, который включает N точек до и после него. Обученная модель ARIMA затем используется для получения оценок пропущенных значений.

Рассмотрим модель ARIMA(p, d, q), заданную разностным уравнением:

$$(1-B)^d y_t = c + \sum_{i=1}^{p} \varphi_i (1-B)^d y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j}, \qquad (8)$$

или в эквивалентной операторной форме:

$$\nabla^d y_t = c + \sum_{i=1}^{p} \varphi_i \nabla^d y_{t-i} \, \varepsilon_t + \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j}, \qquad (9)$$

где ∇^d – оператор дифференцирования порядка d, y_t – значение ряда в момент t, ε_t – белый шум, φ_i (i = 1,…, p) и θ_j (j = 1,…, q) – параметры авторегрессии и скользящего среднего, c – константа модели.

Для (9) восстановление значений внутри пропуска осуществляется последовательным вычислением условного

математического ожидания. Этот процесс эквивалентен односторонней фильтрации, при которой каждое последующее восстанавливаемое значение вычисляется с учетом как ранее предсказанных значений внутри пропуска, так и известных исторических данных и оценок ошибок.

Алгоритмическая реализация формальных правил, заданных выражениями (1)-(7), требует практической адаптации к особенностям реальных данных, которые часто демонстрируют сложные нестационарные паттерны, не укладывающиеся в строгие теоретические предположения. Для устойчивой работы алгоритма на этапе предварительной обработки входной временной ряд подвергается процедуре мягкого сглаживания и очистки от выбросов. Это необходимо для стабилизации оценок локальной статистики, таких как автокорреляционная функция и дисперсия, которые критически важны для формул (2) и (3). В частности, значение эмпирического коэффициента k в уравнении (2) корректируется с учетом локальной волатильности ряда в окрестности пропуска. На спокойных участках с низкой дисперсией можно допустить использование более высокого порядка p, так как даже слабые корреляционные связи могут быть информативными. Напротив, в турбулентных сегментах, где шумовая компонента велика, соотношение (2) ужесточается, чтобы избежать попыток моделирования случайных флуктуаций, что предотвращает переобучение. Эта динамическая корректировка делает адаптацию к первому контекстному фактору (L) не механической, а учитывающей качество доступной для обучения информации.

Калибровка параметров логистической функции (4), связывающей нормализованный индекс активности I_{norm} с весовым коэффициентом ω, представляет собой отдельную задачу оптимизации [13]. Для ее решения создается специальная валидационная выборка, состоящая из множества исторических пропусков с известными истинными значениями. На этой выборке методом поиска по сетке определяются значения α и β, которые максимизируют общее качество восстановления. При этом качество понимается не только как минимизация среднеквадратичной ошибки, но и как способность модели сохранять важные структурные особенности сигнала, такие как экстремумы и точки перегиба, которые часто теряются при неудачном выборе параметров. Этот итеративный процесс настройки превращает уравнение (4) из абстрактной зависимости в конкретный инструмент, настроенный на специфику предметной области, будь то геомагнитные данные или другой тип сигналов. В результате, для спокойных периодов [14] ($I_{\text{norm}} \rightarrow 0$)

алгоритм склоняется к выбору моделей с более выраженной авторегрессионной компонентой, что хорошо согласуется с теорией о возможности более длинной памяти у стабильных процессов. В периоды высокой возмущенности ($I_{\text{norm}} \to 1$) возрастающая ω смещает предпочтения алгоритма в сторону моделей с высоким порядком q, что позволяет более гибко адаптироваться к резким, похожим на шум изменениям, характерным для таких условий.

4. Программная реализация и архитектура предлагаемого решения. Разработанный метод был реализован в виде специализированного программного модуля на языке Python, выбранного ввиду его распространенности в научных вычислениях и наличию развитой экосистемы библиотек для анализа данных.

Ядро модуля построено по объектно-ориентированной архитектуре [15], центральным элементом которой является класс ContextAwareImputer. Такой дизайн инкапсулирует всю логику контекстно-зависимой настройки, предоставляя пользователю простой программный интерфейс для восстановления пропусков в виде метода impute(data, gap_indices, context_I).

Для обеспечения вычислительной эффективности и воспроизводимости результатов в реализации активно задействованы библиотеки NumPy, pandas и statsmodels [16 – 18]. Последняя предоставляет надежную реализацию оценки и прогнозирования для моделей ARIMA, которая была интегрирована и расширена в рамках предлагаемого подхода. В частности, стандартный класс ARIMA из statsmodels был обернут в процедуру, которая динамически меняет его параметры order в соответствии с алгоритмом, описанным в разделе 3.

Критически важным аспектом реализации является управление вычислительной сложностью. Полный перебор всех возможных троек (p, d, q) даже в суженном пространстве моделей C, определенном выражениями (5) и (6), может стать ресурсоемкой операцией при обработке длинных рядов или большого количества пропусков. Для оптимизации этого процесса в алгоритм внедрен механизм кеширования.

Результаты предварительного теста Дики–Фуллера на стационарность, а также вычисленные автокорреляционные функции для стандартных сегментов данных сохраняются и повторно используются, что позволяет избежать дублирующих вычислений. Кроме того, процедура кросс-валидации для оценки возможных моделей, заданная выражением (7), была распараллелена. Используя возможности библиотеки joblib, оценка качества каждой модели из множества C производится асинхронно на доступных ядрах процессора,

что приводит к почти линейному ускорению вычислений на многоядерных системах. Этот подход делает метод практичным для работы с данными большого объема.

Входными данными для модуля являются одномерный массив числовых значений (временной ряд), массив индексов или меток времени, определяющих начало и конец каждого пропуска, и соответствующий массив значений контекстного индекса I для каждого пропуска.

Программный конвейер обработки начинается с этапа валидации и предварительной очистки входных данных. На этом этапе проверяется согласованность длин массивов, осуществляется мягкое сглаживание ряда для подавления высокочастотного шума, не несущего информации для структурного анализа, и выполняется первичное обнаружение и обработка выбросов. Затем для каждого пропуска определяется его локальное окно – сегмент данных, окружающий пропуск, размер которого пропорционален длине пропуска L, но ограничен сверху для сохранения локальности анализа. Именно в пределах этого окна вычисляются все необходимые статистики:

– оценка автокорреляционной функции для расчета коэффициента k в соответствии с выражением (2),

– локальная дисперсия для корректировки коэффициента k.

Кроме того проводится тест Дики–Фуллера для принятия решения о порядке дифференцирования d.

Архитектурно модуль разделен на несколько логических компонентов (рисунок 1). Компонент ContextAnalyzer отвечает за прием исходных данных, их предобработку и вычисление параметров L и I_{norm} для каждого пропуска. ParameterSpaceConstructor реализует правила, заданные уравнениями (2), (4), (5) и (6), преобразуя пару (L, I_{norm}) в конкретное множество моделей C. ModelSelector выполняет процедуру кросс-валидации, описанную в разделе 3, включая создание искусственного валидационного пропуска, обучение моделей и вычисление целевой функции качества $Q(M_i)$. Наконец, компонент ImputationEngine, получив от ModelSelector оптимальные параметры (p_{opt}, d_{opt}, q_{opt}), осуществляет финальное обучение модели ARIMA на полном локальном окне и выполняет последовательную интерполяцию пропущенных значений, используя механизм условного прогноза, основанный на выражении (9).

Такая модульная архитектура не только улучшает читаемость и поддерживаемость кода, но и облегчает его возможное расширение, например, для интеграции других типов прогнозных моделей или дополнительных контекстных факторов.

Для обеспечения удобства использования модуль снабжен подробной документацией, сгенерированной с помощью Sphinx, включающей описание API, примеры запуска и руководство по калибровке гиперпараметров алгоритма, таких как коэффициенты α, β и γ.

Расширение модуля для обработки различных типов временных рядов потребовало создания универсального механизма адаптации базовых гиперпараметров. Исходные значения коэффициентов α и β в логистической функции (4), а также веса γ в целевой функции (7) были получены в ходе калибровки на тестовых данных.

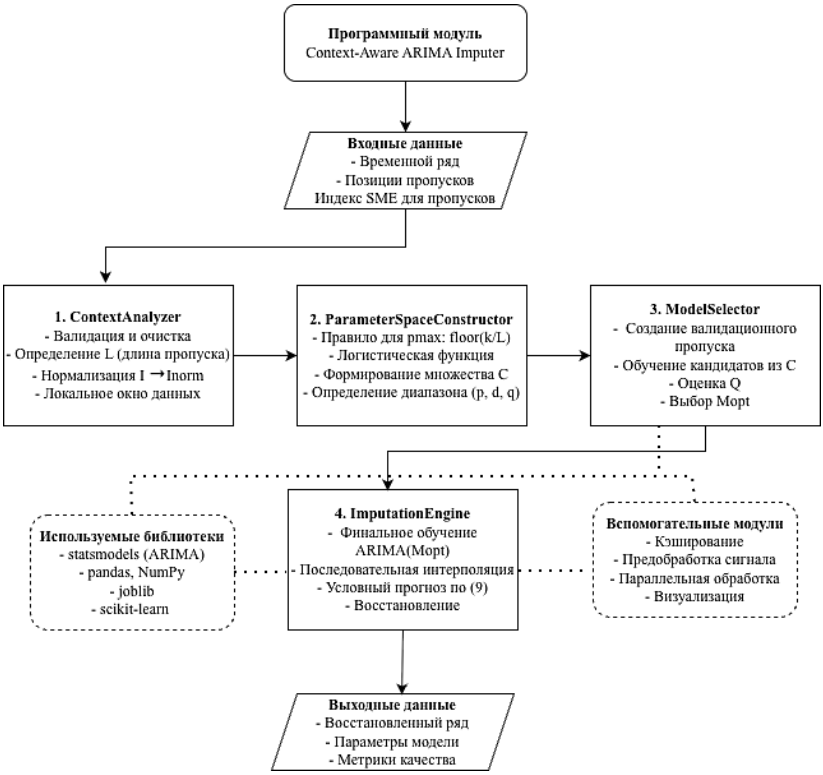

Рис. 1. Взаимодействие компонентов в методе адаптивной параметризации. Сплошные стрелки обозначают основной поток данных и управления между ядром алгоритма. Пунктирные стрелки обозначают зависимость компонентов от внешних библиотек

Однако для применения метода к рядам из других предметных областей необходима возможность их тонкой настройки. В архитектуру модуля был включен специализированный класс HyperparameterOptimizer, который через конфигурационный файл позволяет пользователю задавать границы для автоматизированного поиска оптимальных значений этих параметров для своего набора данных. Процедура поиска основана на минимизации средней ошибки восстановления на выделенной калибровочной выборке с использованием метода случайного поиска по сетке, что обеспечивает баланс между эффективностью и вычислительными затратами. Этот механизм формализует процесс адаптации ядра алгоритма к новой предметной области, сохраняя при этом его основную логику неизменной.

Одной из ключевых задач стала разработка устойчивых стратегий обработки пограничных случаев и аномальных сценариев, возникающих при работе с реальными данными. В компонентах ContextAnalyzer и ParameterSpaceConstructor реализованы многоуровневые проверки корректности входных данных и промежуточных вычислений. Например, если алгоритм обнаруживает, что локальное окно данных содержит недостаточное количество наблюдений для надежной оценки автокорреляционной функции или дисперсии, он автоматически переключается на использование глобальных оценок, вычисленных для всего ряда, или консервативных значений по умолчанию.

Аналогичным образом, в ситуациях, когда вычисленное множество моделей C оказывается пустым из-за чрезмерно жестких ограничений, правила ослабляются в итеративном режиме до тех пор, пока не будет сформировано хотя бы несколько допустимых моделей. Эти стратегии гарантируют, что алгоритм не завершит работу с фатальной ошибкой, а выдаст осмысленный, хотя и, возможно, субоптимальный результат даже в нестандартных условиях.

Обеспечение численной устойчивости и точности вычислений при оценке и прогнозировании моделей ARIMA потребовало особого внимания к настройке параметров оптимизации и обработке численных ошибок. Библиотека statsmodels, лежащая в основе вычислений, предоставляет различные алгоритмы оптимизации для подбора коэффициентов моделей (метод максимального правдоподобия). В ходе экспериментов было установлено, что выбор конкретного решателя и его гиперпараметров (допуск, максимальное число итераций) может существенно влиять на сходимость и точность оценки, особенно для моделей с высокими порядками p и q. Поэтому в компонентах

ModelSelector и ImputationEngine реализована обертка, которая отслеживает предупреждения и ошибки в процессе оптимизации. В случае сбоя или отсутствия сходимости для сложной модели, алгоритм автоматически переключается на более простой решатель или уменьшает максимальный порядок модели в множестве моделей C. Это обеспечивает надежность итоговых вычислений, предотвращая ситуации, когда весь процесс восстановления прерывается из-за неудачи на этапе обучения одной из множества проверяемых моделей.

Валидация корректности численных расчетов и логики работы всех компонентов потребовала создания комплексной системы модульных и интеграционных тестов. Тестовый набор включает синтетические данные, сгенерированные процессами ARIMA с известными параметрами, что позволяет напрямую проверять способность алгоритма восстанавливать пропуски в условиях, когда истинная модель ряда априори известна.

Модульные тесты проверяют, что каждый компонент для фиксированных входных данных выдает строго ожидаемый результат, например, что ParameterSpaceConstructor для заданных L и I_norm формирует определенное множество моделей C. Интеграционные тесты проверяют согласованность работы всего конвейера: восстановленные значения для искусственного пропуска в синтетическом ряду должны с заданной точностью совпадать с заведомо известными истинными значениями. Эта система тестов служит не только для обеспечения надежности, но и как формальная спецификация ожидаемого поведения системы в различных условиях, что критически важно для долгосрочной поддержки и развития кодовой базы.

Для обеспечения прозрачности и интерпретируемости работы сложного адаптивного алгоритма была разработана система детального логирования и генерации отчетов. Каждый основной компонент модуля записывает в структурированный лог ключевые промежуточные результаты и принятые решения: вычисленные значения L и I_norm, построенное множество моделей C, значения целевой функции $Q(M_i)$ для каждой модели, результаты проверки на сходимость и финальный выбор параметров модели. Эти данные могут быть автоматически агрегированы в текстовый или визуальный отчет, который позволяет исследователю реконструировать весь процесс обработки конкретного пропуска. Такая прозрачность превращает алгоритм из «черного ящика» в инструмент, работа которого может быть проанализирована, объяснена и, при необходимости, скорректирована, что является важным требованием для научных и инженерных приложений, где

понимание причинно-следственных связей не менее важно, чем итоговый результат.

5. Экспериментальное исследование на геомагнитных данных. Для практической проверки предложенного метода и демонстрации его работоспособности было выполнено экспериментальное исследование на реальных геомагнитных временных рядах.

Выбор геомагнитных данных в качестве тестового стенда обусловлен комплексом причин, делающих их уникальным и требовательным полигоном для алгоритмов восстановления нестационарных процессов. Во-первых, геомагнитное поле является непрерывно наблюдаемым физическим полем, чья динамика формируется под воздействием как внутренних процессов ядра Земли, так и внешних воздействий со стороны солнечного ветра и магнитосферной активности. Эта двойственная природа приводит к формированию сложного сигнала, сочетающего в себе относительно плавные суточные вариации, обусловленные вращением Земли, и резкие, импульсные возмущения (суббури, бури), связанные с солнечной активностью. Таким образом, геомагнитный ряд по своей сути является ярко выраженным нестационарным процессом со смешанным спектром, где спокойные периоды сменяются интервалами высокой турбулентности. Такое поведение представляет собой полезный, но крайне сложный случай для проверки способности алгоритма адаптироваться к фундаментальным изменениям в характере данных.

Во-вторых, для геомагнитных данных существует хорошо разработанная система количественных индексов, объективно описывающих уровень внешней возмущенности. Индекс SME (Substorm Magnetospheric Index), используемый в данном исследовании, рассчитывается по глобальной сети наземных обсерваторий и служит надежной интегральной мерой энергии, вкладываемой в магнитосферу во время суббуревых событий. Наличие такого независимого, непрерывного и количественного дескриптора внешнего контекста (фактор I в предлагаемом методе) является редким преимуществом. Во многих других областях (финансы, медицина, техника) подобные внешние факторы либо ненаблюдаемы, либо имеют качественный характер, что затрудняет их формальное использование в алгоритмах. Таким образом, геомагнитные данные предоставляют возможность проверить гипотезу о полезности интеграции внешней контекстной информации в процесс параметризации модели, поскольку здесь эта

информация доступна, измерима и имеет четкую физическую интерпретацию.

В качестве конкретного объекта исследования использовались минутные данные вариаций магнитного поля Земли, предоставляемые международной сетью обсерваторий SuperMAG [19, 20]. Эта сеть обеспечивает стандартизованную предобработку сырых измерений, включая удаление основного поля и приведение к единой системе координат, что гарантирует высокое качество и сопоставимость данных из различных источников.

Для проведения вычислительных экспериментов был использован ряд северной компоненты (DBE_NEZ), полученный на высокоширотной обсерватории Ловозеро (LOZ) Полярного геофизического института (ПГИ) [21]. Выборка была ограничена периодом 2015 года, что позволяет обеспечить сопоставимость условий наблюдений и полную доступность синхронных индексов геомагнитной активности SME. Целевым параметром, подвергаемым восстановлению, была выбрана северная компонента возмущенного магнитного поля (DBE_NEZ). Данная компонента является одним из ключевых индикаторов суббуревых процессов в высоких широтах, демонстрируя особенно сильный отклик на возмущения в магнитосферных токах. Ее динамика характеризуется значительной изменчивостью амплитуды и частоты колебаний, что создает дополнительную сложность для задач интерполяции по сравнению с более плавными компонентами.

Для моделирования условий пропусков из исходного непрерывного ряда были искусственно удалены сегменты различной длины, что позволило иметь точный эталон для объективной оценки качества интерполяции. Диапазон длин пропусков L был выбран от 5 до 120 минут с дискретным шагом (5, 10, 15, 30, 45, 60, 90, 120 минут), что соответствует длинам от 5 до 120 отсчетов при минутном разрешении данных. Этот выбор не является произвольным; он соответствует типичным временным масштабам потери данных в реальных системах мониторинга, которые могут быть вызваны кратковременными сбоями в передаче (минуты), плановым техническим обслуживанием (десятки минут) или более длительными отказами оборудования. Исследование именно таких диапазонов длин представляет наибольший практический интерес, поскольку методы простой экстраполяции или прогноза на один шаг оказываются неадекватными, а необходимость в сложной модели, учитывающей структуру ряда, становится критической.

Ключевой особенностью экспериментального плана являлась неслучайная, а систематическая привязка каждого искусственно

созданного пропуска к конкретному уровню геомагнитной активности, характеризуемому индексом SME. Вместо того чтобы случайным образом распределять пропуски по всему временному ряду, они целенаправленно размещались внутри заранее определенных временных интервалов, классифицированных по уровням SME: спокойные условия (SME < 100 нТл), слабые возмущения (100–300 нТл), умеренные бури (300–600 нТл) и сильные возмущения (SME > 600 нТл). Такой подход позволил сформировать сбалансированную тестовую выборку, равномерно покрывающую весь спектр возможных условий, в которых может работать алгоритм восстановления.

Подготовка данных к эксперименту включала несколько обязательных этапов. Исходный минутный ряд, как и любой реальный геофизический сигнал, содержал технические артефакты и выбросы, не связанные с геофизическими процессами. Для их подавления применялся мягкий фильтр на основе алгоритма Савицкого–Голея, который эффективно удаляет высокочастотный шум, минимально искажая форму основного сигнала. Важно отметить, что эта предобработка применялась только к данным, используемым для обучения и тестирования моделей; исходные значения сохранялись в качестве эталона для расчета финальных метрик ошибки. Это позволяет оценить, насколько хорошо восстановленные значения соответствуют реальным наблюдаемым данным, а не их сглаженной версии. После очистки проводилась проверка ряда на стационарность с помощью расширенного теста Дики–Фуллера (ADF), которая подтвердила наличие единичного корня, то есть нестационарность ряда. Это обосновывает необходимость использования именно класса ARIMA-моделей.

Планирование эксперимента дало возможность проверить не только общую эффективность метода в терминах средней ошибки, но и его ключевую концептуальную гипотезу – способность адаптивно и осмысленно менять параметры модели в зависимости от контекста, определяемого парой (L, SME). Для этого в ходе работы алгоритма для каждого тестового пропуска протоколировался финальный выбор оптимальной модели $M_{\text{opt}}(p_{\text{opt}}, d_{\text{opt}}, q_{\text{opt}})$. Последующий статистический анализ этих выборов позволил выявить устойчивые паттерны: например, тенденцию к выбору моделей с более высоким порядком q (скользящее среднее) в периоды высокой активности SME, что соответствует физическому ожиданию о возрастающей роли стохастических, похожих на шум возмущений. Или тенденцию к уменьшению допустимого порядка p (авторегрессии) с ростом длины пропуска L, что свидетельствует о корректной работе эвристического

правила, ограничивающего сложность модели при дефиците релевантной информации.

Для каждого тестового случая работа предложенного контекстно-зависимого метода сравнивалась с двумя базовыми подходами, представляющими разные философии восстановления данных. Первый базовый метод – классическая кусочно-линейная интерполяция – представляет собой простейший детерминированный подход, полностью игнорирующий как автокорреляционную структуру ряда, так и внешний контекст. Он служит нижним примером, демонстрирующим минимальный ожидаемый уровень качества, который должен быть превзойден любым более сложным методом.

Второй метод для сравнения – интерполяция с помощью модели ARIMA, параметры которой (p, d, q) были однократно подобраны глобальным алгоритмом auto.arima по всему доступному временному ряду (за исключением самого пропуска). Этот метод представляет собой современный стандарт де-факто в автоматическом анализе временных рядов. Однако, будучи примененным к задаче интерполяции, он воплощает философию «одна модель для всех случаев»: структура модели, выбранная на основе глобальных свойств всего ряда, используется для восстановления любого пропуска независимо от его локальных особенностей и условий внешней среды. Сравнение с этим методом позволяет количественно оценить ценность, которую добавляет контекстно-зависимая адаптация, предлагаемая в данной работе.

Качество восстановления оценивалось по двум взаимодополняющим метрикам: коэффициенту детерминации (R^2) и среднеквадратичной ошибке (RMSE), вычисленным путем сравнения восстановленных значений с исходными, необработанными («сырыми») данными внутри каждого искусственного пропуска. Использование R^2 в качестве основной метрики позволяет оценить, какая доля дисперсии исходного сигнала объясняется восстановленными значениями, в то время как RMSE дает понимание типичной величины отклонения в абсолютных единицах (нТл). Качественная иллюстрация работы метода в сложных условиях представлена на рисунке 2, который демонстрирует восстановление 60-минутного пропуска в спокойной обстановке (SME $\approx$ 80 нТл) и позволяет визуально сравнить подходы.

Количественные результаты эксперимента позволяют сделать следующие выводы об эффективности предлагаемого метода в зависимости от контекста. В спокойных и слабовозмущенных условиях (SME = 50–200 нТл) метод демонстрирует стабильно высокое качество

восстановления, с коэффициентом детерминации R^2 = 0.71–0.85 для всего диапазона длин пропусков от 5 до 120 мин. Это подтверждает его надежность в режимах, где динамика ряда относительно предсказуема.

В условиях умеренных возмущений (SME ≈ 400 нТл) точность заметно снижается, достигая значений R^2 = 0.23–0.49, что отражает возрастающую стохастичность сигнала. Наиболее сложным случаем для восстановления оказались периоды сильных и экстремальных бурь (SME ≥ 800 нТл), где качество падает до R^2 = 0.15–0.41, что является фундаментальным ограничением, связанным с приближением исходного ряда к шуму при высокой внешней возмущенности. Интересно отметить, что для самых длинных пропусков (90–120 мин) в этих экстремальных условиях метод иногда показывает сопоставимую или даже чуть более высокую точность, чем для коротких пропусков, что может указывать на сглаживающий эффект модели на длинных интервалах при очень хаотичном сигнале. Эти результаты наглядно иллюстрируют как сильные стороны метода – его адаптивность и надежность в широком диапазоне условий, так и объективные границы его применимости, определяемые природой исходных данных.

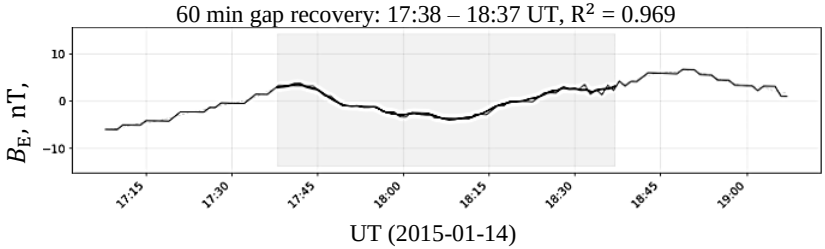

Рис. 2. Пример восстановления 60-минутного пропуска dbe_nez при SME = 80 нТл

Для каждого тестового случая работа предложенного контекстно-зависимого метода сравнивалась с двумя базовыми подходами, реализующими разные принципы восстановления данных. Первый, кусочно-линейная интерполяция, служил простейшим детерминированным ориентиром и, как и ожидалось, показывал резкое снижение точности с ростом длины пропуска (например, с $R^2 \approx 0.91$ для 5-минутного до $R^2 < 0.20$ для 30-минутного интервала). Второй, более совершенный подход – интерполяция с помощью модели ARIMA, параметры которой (p, d, q) были однократно подобраны глобальным алгоритмом `auto.arima` по всему ряду, – воплощал принцип «одна модель для всех случаев». Однако в условиях умеренных геомагнитных

возмущений (SME ≈ 400 нТл) предложенный адаптивный метод для коротких пропусков (5–15 мин) более чем вдвое превосходил эту глобальную ARIMA по точности (R^2 = 0.79–0.85 против R^2 = 0.39–0.46), а в экстремальных условиях (SME ≥ 800 нТл) сохранял преимущество в 30–60 процентных пунктов по R^2. Даже в спокойной обстановке (SME 50–200 нТл) адаптивный метод поддерживал стабильно высокое качество (R^2 = 0.83–0.94) на длинных интервалах (30–60 мин), где линейная интерполяция уже не работала. Эти результаты количественно подтверждают, что переход от универсальной к адаптивной, контекстно-зависимой параметризации модели обеспечивает существенный и устойчивый выигрыш в точности восстановления данных, особенно в условиях внешних возмущений.

6. Обсуждение результатов. Результаты экспериментального исследования демонстрируют, что предложенный контекстно-зависимый метод обеспечивает новый подход к задаче интерполяции пропусков в нестационарных временных рядах по сравнению с классическими решениями. Полученные данные не просто подтверждают его работоспособность, но и позволяют глубоко проанализировать механизмы его работы и границы эффективности. Ключевым практическим выводом является доказанная способность алгоритма поддерживать высокое качество восстановления ($R^2 > 0.7$) в широком диапазоне длин пропусков (5–120 мин) при условии спокойной или слабовозмущенной геомагнитной обстановки. Это указывает на то, что метод успешно решает выявленную проблему: он эффективно использует локальную автокорреляционную структуру данных, адаптивно ограничивая сложность модели в зависимости от доступного для анализа объема информации. Стабильность результатов на длинных пропусках особенно важна, так как именно в таких сценариях традиционные методы, основанные на экстраполяции, терпят неудачу.

Вместе с тем, экспериментальные данные ясно показывают фундаментальную зависимость качества восстановления от уровня внешней возмущенности, характеризуемого индексом SME. Резкое снижение коэффициента R^2 при SME > 300 нТл является не недостатком алгоритма, а отражением объективного физического ограничения. В периоды высокой геомагнитной активности динамика параметра DBE_NEZ становится крайне турбулентной, приближаясь к поведению окрашенного шума с резкими, плохо прогнозируемыми скачками.

Полученные результаты также позволяют четко очертить область наиболее эффективного применения метода. Его сильные стороны максимально раскрываются при работе с нестационарными рядами,

которые, однако, демонстрируют относительно устойчивую автокорреляционную структуру в пределах локального окна анализа. Это характерно для данных мониторинга физических процессов в штатных режимах их работы. В этих условиях метод обеспечивает точное, структуро-сохраняющее восстановление пропусков различной длины. С другой стороны, в периоды экстремальных возмущений, когда ряд теряет выраженную автокорреляцию, метод, как и любой другой, основанный на линейном прогнозе, достигает своего теоретического предела точности. Однако важно отметить, что даже в этих условиях он не «ломается», а выдает консервативный результат, часто превосходящий по точности простую линейную интерполяцию.

Стоит подчеркнуть, что предложенный метод носит универсальный характер и не ограничивается геомагнитными данными. Данный конкретный случай был выбран для экспериментальной проверки в первую очередь в силу его доступности для авторов и наличия в нем всех ключевых атрибутов, необходимых для валидации метода: выраженной нестационарности и наличия формализованного внешнего индекса активности (SME). Успешное применение алгоритма в этих сложных условиях служит убедительным доказательством его работоспособности. Модульная архитектура алгоритма и его ключевые принципы (учет длительности пропуска L и внешнего контекста I) заведомо готовы к адаптации для работы с временными рядами из других предметных областей (например, финансовой аналитики, мониторинга технологических процессов), где могут быть определены аналогичные контекстные факторы.

Представляется возможным несколько перспективных направлений для дальнейшего развития метода. Во-первых, интеграция более сложных моделей, учитывающих нелинейные зависимости, например, в рамках гибридного подхода, где контекстно-зависимый механизм выбирал бы не только параметры, но и класс модели (например, между линейной ARIMA и нелинейной моделью на основе деревьев). Во-вторых, использованием не одного, а нескольких контекстных индексов, которые могли бы более тонко описывать состояние системы. Наконец, разработанная модульная архитектура программной реализации позволяет относительно легко адаптировать ядро алгоритма для работы с данными из других предметных областей, таких как финансовая аналитика или мониторинг промышленного оборудования, где также существуют проблемы пропусков и доступны внешние индикаторы состояния рынка или технологического процесса. В этом смысле предлагаемое решение закладывает основу для создания универсального адаптивного инструментария восстановления данных,

способного учитывать специфику контекста в самых разных приложениях.

Авторы благодарят рецензентов за внимательное прочтение работы и конструктивную критику, которая позволила существенно улучшить изложение и методологическую строгость представленных результатов.

Литература

1. Januschowski T., Gasthaus J., Wang Y., Salinas D., Flunkert V., Bohlke-Schneider M., Callot L. Criteria for classifying forecasting methods // International Journal of Forecasting. 2020. vol. 36. no. 1. pp. 167–177. DOI: 10.1016/j.ijforecast.2019.05.008.
2. pmdarima: Arima estimators for python. Online Code Repos. Available at: http://www.alkaline-ml.com/pmdarima. (accessed 26.02.2026).
3. Hamilton J.D. Time Series Analysis. Princeton: Princeton University Press, 2020. 816 p. DOI: 10.2307/j.ctv14jx6sm.
4. Lama A., Ray S., Biswas T., Narsimhaiah L., Raghav Y.S., Kapoor P., Singh K.N., Mishra P., Gurung B. Python code for modeling ARIMA-LSTM architecture with random forest algorithm // Software Impacts. 2024. vol. 20. DOI: 10.1016/j.simpa.2024.100650.
5. Jiang Y., Ning K., Pan Z., Shen X., Ni J., Yu W., Schneider A., Chen H., Nevmyvaka Y., Song D. Multi-modal time series analysis: A tutorial and survey. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2025. vol. 2. pp. 6043–6053. DOI: 10.1145/3711896.3736567.
6. Kowal D. Dynamic Regression Models for Time-Ordered Functional Data // Bayesian Analysis. 2021. vol. 16(2). pp. 459–487. DOI: 10.1214/20-BA1213.
7. Bokde N.D., Yaseen Z.M., Andersen G.B. ForecastTB – An R Package as a Test-Bench for Time Series Forecasting – Application of Wind Speed and Solar Radiation Modeling. Energies 2020. vol. 13. no. 10. DOI: 10.3390/en13102578.
8. Vorobev A.V., Vorobeva G.R. An approach to dynamic visualization of heterogeneous geospatial vector images // Computer Optics. 2024. vol. 48(1). pp. 123–138. DOI: 10.18287/2412-6179-CO1279.
9. Maitra S., Politis D.N. Prepivoted Augmented Dickey-Fuller Test with Bootstrap-Assisted Lag Length Selection. Stats. 2024. vol. 7(4). pp. 1226–1243. DOI: 10.3390/stats7040072.
10. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974. vol. 19. no. 6. pp. 716–723. DOI: 10.1109/TAC.1974.1100705.
11. Hill C., Du L., Johnson M., McCullough B.D. Comparing programming languages for data analytics: Accuracy of estimation in Python and R. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2024. vol. 14(3). DOI: 10.1002/widm.1531.
12. Kataoka R. Extreme geomagnetic activities: a statistical study. Earth Planets Space. 2020. vol. 72(1). DOI: 10.1186/s40623-020-01261-8.
13. Boroyev R.N., Vasiliev M.S. The auroral activity during the main phase of magnetic storms. Advances in Space Research. 2023. vol. 71(1). pp. 1137–1145. DOI: 10.1016/j.asr.2022.10.034.
14. Newell P.T., Gjerloev J.W. Evaluation of SuperMAG auroral electrojet indices as indicators of substorms and auroral power. Journal of Geophysical Research: Space Physics. 2011. vol. 116. DOI: 10.1029/2011JA016779.

15. Chu X., Ma D., Bortnik J., Tobiska W.K., Cruz A., Bouwer S.D., et al. Relativistic electron model in the outer radiation belt using a neural network approach. Space Weather. 2021. vol. 19. pp. 1–18. DOI: 10.1029/2021SW002808.
16. Gupta P., Bagchi A. Data Manipulation with Pandas. Essentials of Python for Artificial Intelligence and Machine Learning. Synthesis Lectures on Engineering, Science, and Technology. 2024. pp. 197-235. DOI: 10.1007/978-3-031-43725-0_6.
17. Sundaram J., Gowri K., Devaraju S., Gokuldev S., Jayaprakash S., Anandaram H., Manivasagan C., Thenmozhi M. An Exploration of Python Libraries in Machine Learning Models for Data Science. 2023. pp. 1–31. DOI: 10.4018/978-1-6684-8696-2.ch001.
18. Faraway J.J. Linear Models with Python. Boca Raton, FL: Chapman and Hall/CRC, 2021. 308 p.
19. Gjerloev J.W. A Global Ground-Based Magnetometer Initiative. Eos, Transactions American Geophysical Union. 2009. vol. 90(27). pp. 230–231. DOI: 10.1029/2009EO270002.
20. Gjerloev J.W. The SuperMAG data processing technique. Journal of Geophysical Research: Space Physics. 2012. vol. 117. DOI: 10.1029/2012JA017683.
21. PGI Geophysical data. January, February, March 2013. Murmansk, Apatity: PGI KSC RAS, 2013.

Воробьев Андрей Владимирович — д-р техн. наук, профессор, заведующий кафедрой, кафедра информатики, Уфимский университет науки и технологий. Область научных интересов: геоинформационные технологии, цифровая обработка сигналов. Число научных публикаций — 200. geomagnet@list.ru; улица Карла Маркса, 12, 450000, Уфа, Россия; р.т.: +7(917)345-2299.

Воробьева Гульнара Равилевна — д-р техн. наук, профессор кафедры, кафедра вычислительной математики и кибернетики, Уфимский университет науки и технологий. Область научных интересов: геоинформационные и веб-технологии, системы хранения и обработки информации. Число научных публикаций — 163. gulnara.vorobeva@gmail.com; улица Карла Маркса, 12, 450000, Уфа, Россия; р.т.: +7(917)417-4111.

Поддержка исследований. Работа выполнена при поддержке Российского научного фонда (проект № 21-77-30010-П).

DOI 10.15622/ia.25.2.4

A. VOROBEV, G. VOROBEVA

A CONTEXT-DEPENDENT METHOD FOR ADAPTIVE TUNING OF PARAMETERS OF AUTOREGRESSIVE MODELS FOR NON-STATIONARY TIME SERIES

Vorobev A., Vorobeva G. **A Context-Dependent Method for Adaptive Tuning of Parameters of Autoregressive Models for Non-Stationary Time Series.**

Abstract. A method for context-sensitive tuning of autoregressive model parameters for gap reconstruction in nonstationary time series is proposed. A key feature of the method is the adaptive selection of ARIMA (p, d, q) model parameters based on two context factors: the gap duration and the level of external disturbances during the corresponding period. Unlike standard automatic model selection approaches focused on global optimization for forecasting, the developed algorithm narrows the model search space and selects the optimal configuration using local cross-validation, allowing for consideration of specific conditions in the gap region. The method is implemented as a Python software module with a modular architecture that ensures computational efficiency through caching and parallel computing. The effectiveness of the method was tested experimentally on real geomagnetic data (the DBE_NEZ component of the Lovozero Observatory). The results demonstrate that under calm and weakly disturbed geomagnetic conditions (SME index = 50–200 nT), the method provides high reconstruction accuracy (R^2 = 0.71–0.85) for gaps ranging from 5 to 120 minutes in length. However, accuracy is shown to decrease consistently with increasing disturbance level, reflecting a fundamental limitation associated with the increasing stochasticity of the original signal. The proposed approach ensures interpretability and adaptability, opening up prospects for the development of data reconstruction tools in various application areas.

Keywords: autoregressive models, gap reconstruction, nonstationary time series, geomagnetic data.

References

1. Januschowski T., Gasthaus J., Wang Y., Salinas D., Flunkert V., Bohlke-Schneider M., Callot L. Criteria for classifying forecasting methods. International Journal of Forecasting. 2020. vol. 36. no. 1. pp. 167–177. DOI: 10.1016/j.ijforecast.2019.05.008.
2. pmdarima: Arima estimators for python. Online Code Repos. Available at: http://www.alkaline-ml.com/pmdarima. (accessed 26.02.2026).
3. Hamilton J.D. Time Series Analysis. Princeton: Princeton University Press, 2020. 816 p. DOI: 10.2307/j.ctv14jx6sm.
4. Lama A., Ray S., Biswas T., Narsimhaiah L., Raghav Y.S., Kapoor P., Singh K.N., Mishra P., Gurung B. Python code for modeling ARIMA-LSTM architecture with random forest algorithm. Software Impacts. 2024. vol. 20. DOI: 10.1016/j.simpa.2024.100650.
5. Jiang Y., Ning K., Pan Z., Shen X., Ni J., Yu W., Schneider A., Chen H., Nevmyvaka Y., Song D. Multi-modal time series analysis: A tutorial and survey. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2025. vol. 2. pp. 6043–6053. DOI: 10.1145/3711896.3736567.
6. Kowal D. Dynamic Regression Models for Time-Ordered Functional Data. Bayesian Analysis. 2021. vol. 16(2). pp. 459–487. DOI: 10.1214/20-BA1213.
7. Bokde N.D., Yaseen Z.M., Andersen G.B. ForecastTB – An R Package as a Test-Bench for Time Series Forecasting – Application of Wind Speed and Solar Radiation Modeling. Energies 2020. vol. 13. no. 10. DOI: 10.3390/en13102578.

8. Vorobev A.V., Vorobeva G.R. An approach to dynamic visualization of heterogeneous geospatial vector images. Computer Optics. 2024. vol. 48(1). pp. 123–138. DOI: 10.18287/2412-6179-CO1279.
9. Maitra S., Politis D.N. Prepivoted Augmented Dickey-Fuller Test with Bootstrap-Assisted Lag Length Selection. Stats. 2024. vol. 7(4). pp. 1226–1243. DOI: 10.3390/stats7040072.
10. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974. vol. 19. no. 6. pp. 716–723. DOI: 10.1109/TAC.1974.1100705.
11. Hill C., Du L., Johnson M., McCullough B.D. Comparing programming languages for data analytics: Accuracy of estimation in Python and R. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2024. vol. 14(3). DOI: 10.1002/widm.1531.
12. Kataoka R. Extreme geomagnetic activities: a statistical study. Earth Planets Space. 2020. vol. 72(1). DOI: 10.1186/s40623-020-01261-8.
13. Boroyev R.N., Vasiliev M.S. The auroral activity during the main phase of magnetic storms. Advances in Space Research. 2023. vol. 71(1). pp. 1137–1145. DOI: 10.1016/j.asr.2022.10.034.
14. Newell P.T., Gjerloev J.W. Evaluation of SuperMAG auroral electrojet indices as indicators of substorms and auroral power. Journal of Geophysical Research: Space Physics. 2011. vol. 116. DOI: 10.1029/2011JA016779.
15. Chu X., Ma D., Bortnik J., Tobiska W.K., Cruz A., Bouwer S.D., et al. Relativistic electron model in the outer radiation belt using a neural network approach. Space Weather. 2021. vol. 19. pp. 1–18. DOI: 10.1029/2021SW002808.
16. Gupta P., Bagchi A. Data Manipulation with Pandas. Essentials of Python for Artificial Intelligence and Machine Learning. Synthesis Lectures on Engineering, Science, and Technology. 2024. pp. 197-235. DOI: 10.1007/978-3-031-43725-0_6.
17. Sundaram J., Gowri K., Devaraju S., Gokuldev S., Jayaprakash S., Anandaram H., Manivasagan C., Thenmozhi M. An Exploration of Python Libraries in Machine Learning Models for Data Science. 2023. pp. 1–31. DOI: 10.4018/978-1-6684-8696-2.ch001.
18. Faraway J.J. Linear Models with Python. Boca Raton, FL: Chapman and Hall/CRC, 2021. 308 p.
19. Gjerloev J.W. A Global Ground-Based Magnetometer Initiative. Eos, Transactions American Geophysical Union. 2009. vol. 90(27). pp. 230–231. DOI: 10.1029/2009EO270002.
20. Gjerloev J.W. The SuperMAG data processing technique. Journal of Geophysical Research: Space Physics. 2012. vol. 117. DOI: 10.1029/2012JA017683.
21. PGI Geophysical data. January, February, March 2013. Murmansk, Apatity: PGI KSC RAS, 2013.

Vorobev Andrei — Ph.D., Dr.Sci., Professor, Head of the department, Informatics Department, Ufa University of Science and Technology. Research interests: geoinformation technologies, digital signal processing. The number of publications — 200. geomagnet@list.ru; 12, Karl Marx St., 450000, Ufa, Russia; office phone: +7(917)345-2299.

Vorobeva Gulnara — Ph.D., Dr.Sci., Professor of the department, Computational Mathematics and Cybernetics Department, Ufa University of Science and Technology. Research interests: geoinformation and web technologies, systems of information storing and processing. The number of publications — 163. gulnara.vorobeva@gmail.com; 12, Karl Marx St., 450000, Ufa, Russia; office phone: +7(917)417-4111.

Acknowledgements. This work was funded by the Russian Science Foundation (project No. 21-77-30010-P).