УДК 004.89:004.855.5:550.385

https://doi.org/10.30730/gtrz.2026.0.mfi-384
https://www.elibrary.ru/xkpbmf
Опубликована online 17.04.2026

Методика восстановления пропусков в вариациях геомагнитного поля на основе алгоритмов kNN и MICE

С. А. Имашев

Научная станция РАН в г. Бишкеке, Бишкек, Киргизия

Резюме. В статье представлены методы восстановления пропусков в геомагнитных данных, которые основаны на алгоритмах k-ближайших соседей (k-Nearest Neighbors, kNN) и многократного заполнения пропусков методом цепных уравнений (Multiple Imputation by Chained Equations, MICE). Анализ эффективности алгоритмов проведен на данных сети геомагнитного мониторинга Научной станции РАН по двум типам событий: регулярным Sq-вариациям и геомагнитным бурям. Согласно полученным результатам, алгоритм kNN демонстрирует хорошую точность при восстановлении регулярных вариаций с показателем MAE ≤ 0.4 нТл (Mean Absolute Error – средняя абсолютная ошибка), но его точность значительно снижается в условиях магнитных бурь (MAE = 5.7 нТл). Алгоритм MICE лучше справляется с обработкой пропусков в таких сложных условиях, снижая MAE до 1.1 нТл за счет учета межстанционных корреляций. Комбинированный подход, в котором используется алгоритм kNN для предварительного восстановления пропусков, а MICE – для последующего уточнения, показал эффективность как при заполнении пропусков в данных самой удаленной станции, так и при устранении импульсных выбросов в данных. Кроме того, показано, что предложенный подход может применяться для анализа магнитных возмущений, регистрируемых на близлежащей станции и вызванных работой установки ЭРГУ-600. Полученные результаты подтверждают возможности предложенных методов машинного обучения для автоматизации анализа многомерной информации, что особенно актуально при обработке больших массивов геомагнитных данных.

Ключевые слова: геомагнитные данные, восстановление пропусков, машинное обучение, алгоритм kNN, алгоритм MICE, магнитные бури, устранение выбросов

A methodology for imputing missing values in geomagnetic field variations using kNN and MICE algorithms

Sanjar A. Imashev

Research Station of the Russian Academy of Sciences in Bishkek, Bishkek, Kyrgyzstan

Abstract. The study presents methods for filling missing values in geomagnetic data based on the k-Nearest Neighbors (kNN) algorithm and Multiple Imputation by Chained Equations (MICE). The effectiveness of these algorithms was analyzed using data from the geomagnetic monitoring network of the Research Station of the RAS, focusing on two types of events: regular Sq-variations and geomagnetic storms. According to the results, the kNN algorithm demonstrates high accuracy in reconstructing regular variations with a Mean Absolute Error (MAE) of ≤ 0.4 nT. However, its accuracy significantly decreases during geomagnetic storms (MAE = 5.7 nT). In contrast, the MICE algorithm performs better in these challenging conditions, reducing the MAE to 1.1 nT by leveraging correlations between monitoring stations. A combined approach, utilizing kNN for preliminary imputation followed by MICE for refinement, proved effective both for filling missing values at the remote Karagai-Bulak station and for addressing impulsive outliers in the data. Additionally, it was shown that the proposed approach can be applied to analyze magnetic disturbances recorded at a nearby sta-

tion caused by the operation of the ERGU-600 system. The results confirm the potential of the methods to automate the analysis of multidimensional data, which is particularly crucial when working with large volumes of geomagnetic data.

Keywords: geomagnetic data, missing value imputation, machine learning, kNN algorithm, MICE algorithm, magnetic storms, outlier removal

Финансирование и благодарности

Работа выполнена в рамках государственного задания Научной станции Российской академии наук в г. Бишкеке (тема № 1025032800066-3-1.5.1).

Автор благодарит уважаемых рецензентов за полезные замечания, которые помогли улучшить текст этой статьи.

Funding and Acknowledgements

The work was carried out within the framework of the state task of the Research Station of RAS in Bishkek (No. 1025032800066-3-1.5.1).

Author thanks the respected Reviewers for helpful comments that improved the quality of this manuscript.

***Для цитирования*:** Имашев С.А. Методика восстановления пропусков в вариациях геомагнитного поля на основе алгоритмов kNN и MICE. *Геосистемы переходных зон,* 2026, т. 10, № 2, 384. URL: http://journal.imgg.ru/web/full/f2026-0-384.pdf; https://doi.org/10.30730/gtrz.2026.0.mfi-384; https://www.elibrary.ru/xkpbmf

***For citation*:** Imashev S.A. A methodology for imputing missing values in geomagnetic field variations using KNN and MICE algorithms. *Geosistemy perehodnykh zon = Geosystems of Transition Zones*, 2026, vol. 10, No. 2, Article 384. (In Russ.). URL: http://journal.imgg.ru/web/full/f2026-0-384.pdf; https://doi.org/10.30730/gtrz.2026.0.mfi-384; https://www.elibrary.ru/xkpbmf

Введение

Пропуски в данных являются характерной особенностью реальных геофизических временных рядов [1, 2]. Основными причинами таких пропусков могут быть технические сбои оборудования или программного обеспечения, экстремальные погодные условия, нестабильность связи, внешние помехи и шумы, а также человеческий фактор [2, 3]. Эти проблемы создают дополнительные трудности для анализа, особенно при работе с геомагнитными временными рядами [4].

Необходимость восстановления пропусков данных, в частности геомагнитных, обусловлена несколькими причинами. Во-первых, непрерывность временных рядов важна для корректного анализа трендов, сезонных колебаний и других характеристик геомагнитного поля. Отсутствующие значения в данных могут привести к их искажению, что затрудняет анализ результатов и приводит к некорректной интерпретации. Во-вторых, многие методы обработки временных рядов, такие как спектральный анализ, фильтрация или моделирование временных рядов (например, сезонное разложение и ARIMA), требуют непрерывности данных для корректной работы. Пропуски также могут снижать точность моделей машинного обучения, включая искусственные нейронные сети, что негативно сказывается на прогнозировании и классификации различных аномалий.

Разрывы временного ряда не только искажают корреляции между характеристиками геомагнитного поля на различных станциях, но и усложняют анализ его пространственной структуры. Так, сложная структура суточных вариаций геомагнитного поля, особенно в условиях внешних возмущающих факторов, таких как геомагнитные бури, сильно ограничивает эффективность традиционных методов восстановления, включая интерполяцию и статистические модели, которые оказываются недостаточно точными при наличии продолжительных пропусков [5]. Эти ограничения показывают необходимость разработки более эффективных методов восстановления данных, учитывающих специфику геофизических наблюдений.

Развитие алгоритмов машинного обучения привело к появлению новых методов восстановления пропусков во временных рядах [6, 7]. Например, в работе [8] показано, что нейронные сети могут успешно восстанавливать данные отдельных магнитных обсерваторий, используя информацию от других станций. В работе [5] для заполнения пропусков

использовали корреляцию соседних фрагментов временного ряда с аналогичными данными за прошлые годы при условии, что магнитная активность (значения K_p-индекса) в восстанавливающих данных была идентичной наблюдаемой в исследуемые сутки. Однако эффективность этого метода имеет свои ограничения: например, при длительности пропусков более 120 мин среднеквадратичная ошибка (MSE) начинает превышать 1.2 нТл. Также важно отметить, что метод работает корректно в условиях спокойной магнитосферы. В условиях геомагнитной активности с $K_p > 3$ погрешность восстановления возрастает до 35 нТл, что снижает эффективность метода. Это подчеркивает необходимость разработки более универсальных подходов, способных выполнять восстановление данных при различных внешних условиях.

В работе [9] предложена модификация классического метода сезонного разложения для анализа и выявления аномалий различных масштабов в вариациях геомагнитного поля. В частности, для выделения магнитной бури с помощью остаточной компоненты вариаций геомагнитного поля участок, содержащий возмущение, сначала заполняется пропусками, а затем восстанавливается на основе усредненного суточного профиля, рассчитанного по фрагментам с обеих сторон пропусков. Однако данный метод корректен только при работе с регулярными Sq-вариациями в условиях отсутствия внешних возмущений. Это связано с тем, что сезонное разложение основывается на периодической составляющей временного ряда, которая может быть искажена в присутствии магнитных бурь [9].

Таким образом, задача восстановления пропусков в случае регулярных Sq-вариаций и в периоды магнитных бурь диктует необходимость разработки новой методики. Целью настоящей работы является оценка эффективности алгоритмов машинного обучения для восстановления пропущенных значений в данных сети геомагнитного мониторинга, проводимого Научной станцией РАН (НС РАН) [10, 11]. В частности, алгоритма k-ближайших соседей (kNN – k-Nearest Neighbors) и метода замены пропущенных данных с помощью цепных уравнений (MICE – Multiple Imputation by Chained Equations)

Данные

В настоящее время сеть геомагнитного мониторинга НС РАН включает 7 стационарных пунктов наблюдения (рис. 1), где данные регистрируются с временным шагом 20 с. Такое временное разрешение позволяет фиксировать как регулярные вариации геомагнитного поля, так и кратковременные аномалии, вызванные внешними возмущениями.

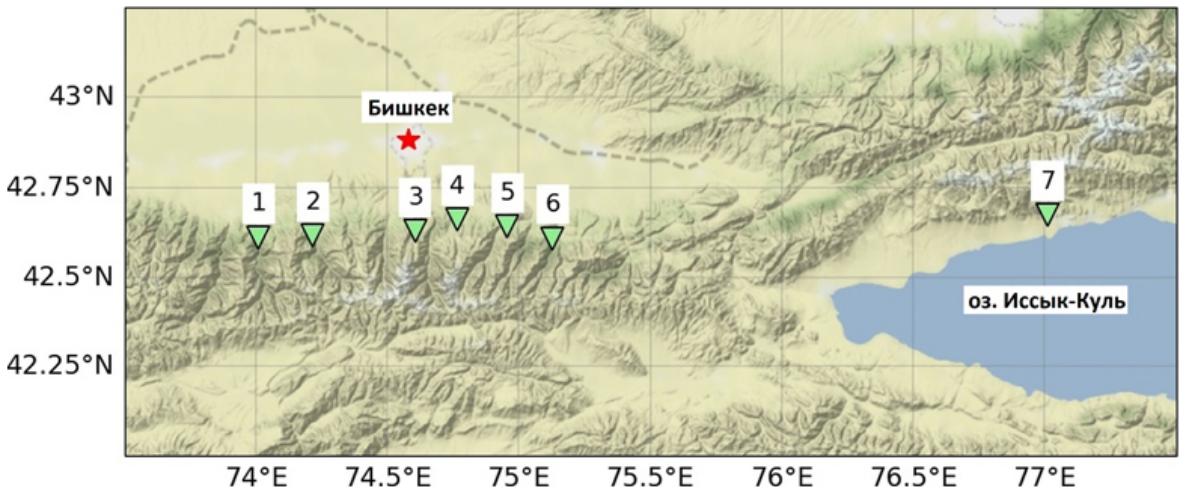

Рис. 1. Карта расположения стационарных пунктов сети геомагнитных наблюдений НС РАН: 1 – Ак-Суу, 2 – Шавай, 3 – Чункурчак, 4 – Таш-Башат, 5 – Иссык-Ата, 6 – Кегеты, 7 – Карагай-Булак.

Fig. 1. Location map of the stationary geomagnetic observation sites of the Research Station of the Russian Academy of Sciences (RS RAS): 1, Ak-Suu; 2, Shavai; 3, Chunkurchak; 4, Tash-Bashat; 5, Issyk-Ata; 6, Kegety; 7, Karagai-Bulak.

Основными причинами пропусков в рассматриваемых данных являются сбои в работе аппаратуры и длительное отсутствие электропитания вследствие аварий и веерных отключений. Кроме того, в процессе предварительной обработки удаляются техногенные помехи, которые возникают из-за воздействия внешних источников, таких как близлежащие электростанции или промышленное оборудование. Это, в свою очередь, создает дополнительные пропуски в данных.

Для анализа использованы данные пяти станций сети геомагнитного мониторинга НС РАН. Две станции, Чункурчак и Таш-Башат, были исключены из анализа, так как их близость к питающему диполю ЭРГУ-600 [12] (на расстояниях 8 и 10 км соответственно) вызывает значительные искажения данных – скачки уровня магнитного поля (ступени), возникающие во время каждого 12-минутного сеанса зондирования (рис. 2). Поскольку эти искусственные ступени не связаны с естественными вариациями поля и присутствуют только на отдельных пунктах, их использование в качестве донорских рядов привело бы к переносу техногенного сигнала в восстанавливаемые данные. Из всех данных предварительно удаляли импульсные выбросы с помощью алгоритма расширенного изолирующего леса [13] и фильтра Хампеля [14], для того чтобы выбросы из данных восстанавливающих станций не переносились в данные восстанавливаемой станции.

Для оценки эффективности алгоритмов были сформированы два набора данных. Первый включал наблюдения с типичными суточными Sq-вариациями [15, 16], характерными для спокойных геомагнитных условий. Второй набор состоял из данных, зарегистрированных во время магнитной бури. Искусственные пропуски с продолжительностью в одни сутки были добавлены в оба набора для имитации реальных длительных сбоев в работе станций. Отметим, что использование искусственно созданных пропусков необходимо для объективной оценки точности восстановления, поскольку в этом случае известны истинные значения данных. Такой подход широко применяется в задачах импутации временных рядов и позволяет корректно сравнивать эффективность различных алгоритмов. При этом длительность искусственно заданных пропусков (1 сутки) соответствует наиболее сложному, но относительно редкому сценарию, связанному с длительными сбоями в работе станций, например, при выходе из строя измерительного оборудования. В реальных условиях наблюдений типичные пропуски, как правило, не превышают нескольких часов и обусловлены профилактическими работами или удалением помех от внешних антропогенных источников. Таким образом, использование длительных пропусков позволяет

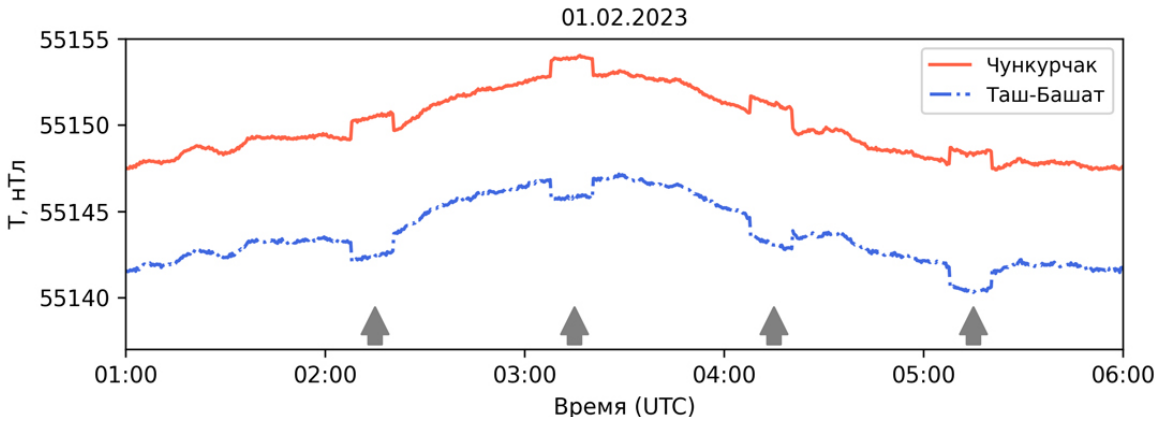

Рис. 2. Вариации величины геомагнитного поля на станциях Чункурчак и Таш-Башат 1 февраля 2023 г. с 01:00 по 06:00 UTC. Стрелками помечены моменты, соответствующие сеансам работы ЭРГУ-600. Для наглядности данные станций Чункурчак и Таш-Башат смещены на –170 нТл и +30 нТл соответственно.

Fig. 2. Variation of the geomagnetic field at the Chunkurchak and Tash-Bashat stations on February 1, 2023, during the period from 01:00 to 06:00 UTC. Arrows indicate the time intervals corresponding to the operation sessions of the ERGU-600 system. For clarity, the data from the Chunkurchak and Tash-Bashat stations are shifted by −170 nT and +30 nT, respectively.

оценить работоспособность алгоритмов в заведомо усложненных условиях и рассматривать полученные результаты как консервативную (пессимистическую) оценку их точности.

Данные магнитных бурь включались в анализ, поскольку аномальные вариации геомагнитного поля в этот период значительно превышают по длительности и амплитуде эффекты других источников возмущений. Кроме того, магнитные бури оказывают планетарное воздействие, что обеспечивает согласованный (синхронный) отклик на станциях сети геомагнитного мониторинга НС РАН.

Для дальнейшей обработки данные геомагнитного мониторинга были структурированы в виде таблицы. Каждая строка этой таблицы соответствует определенному моменту времени измерения, а столбцы содержат значения геомагнитного поля, измеренные в этот момент на различных станциях. Такое представление удобно для дальнейшего применения методов машинного обучения, где строки трактуются как наблюдения, а столбцы – как признаки. В этом случае каждое наблюдение будет представлять собой точку в многомерном пространстве признаков. Этот подход обеспечивает совместимость данных с большинством стандартных алгоритмов обработки временных рядов и позволяет эффективно анализировать межстанционные взаимосвязи.

На рис. 3 представлен первый набор данных, включающий регулярные Sq-вариации геомагнитного поля. Для 4 станций (Ак-Суу, Шавай, Иссык-Ата и Кегеты), расположенных близко друг к другу, были искусственно созданы пропуски длительностью одни сутки, обозначенные как А1, А2, А3 и А4. Аналогичные пропуски (Б1, Б2, Б3 и Б4) были созданы для удаленной станции Карагай-Булак, синхронно по времени с пропусками на других станциях. Такая схема позволяет оценить не только точность восстановления данных на станции Карагай-Булак, но и влияние пропусков в данных станций-доноров на качество работы алгоритмов. Весь набор данных включает 38 880 строк (9 дней × 24 ч × 60 мин × 3 измерения в минуту) и 5 столбцов, соответствующих количеству станций.

На рис. 4 показаны вариации геомагнитного поля для второго набора данных, содер-

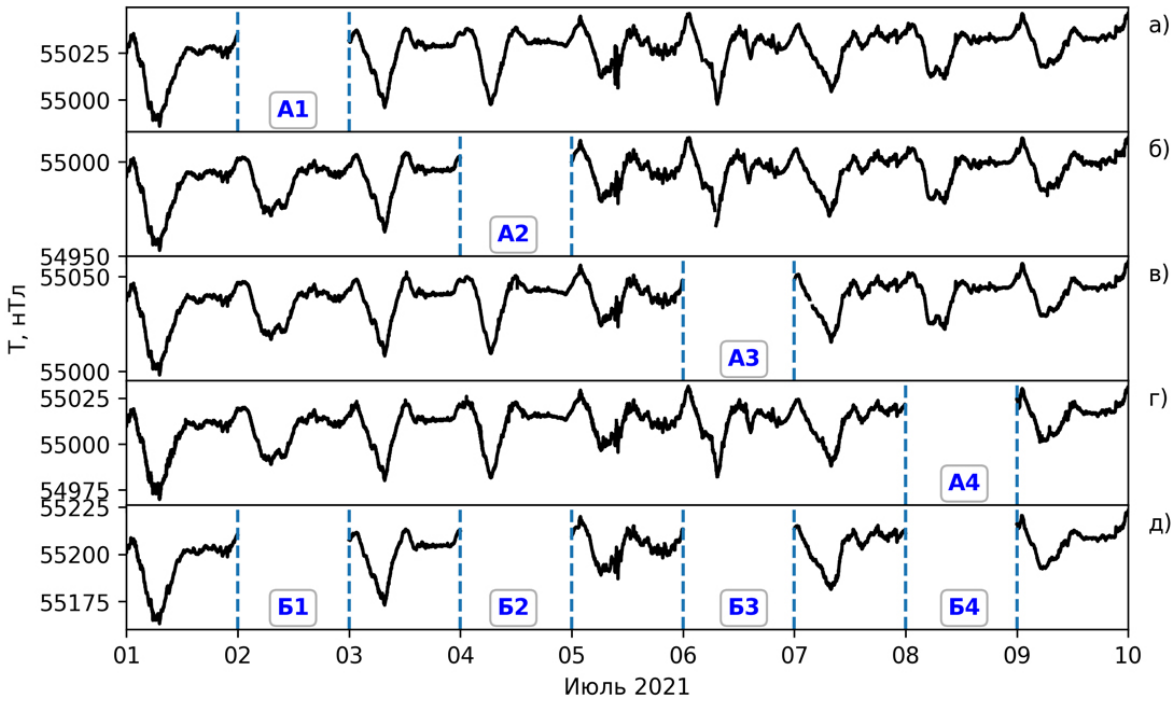

Рис. 3. Набор данных, содержащий типичные Sq-вариации величины геомагнитного поля на станциях Ак-Суу (а), Шавай (б), Иссык-Ата (в), Кегеты (г) и Карагай-Булак (д) и искусственно созданные пропуски в них (А1–А4 и Б1–Б4).

Fig. 3. Dataset containing typical Sq variations of the geomagnetic field at the Ak-Suu (а), Shavai (б), Issyk-Ata (в), Kegety (г), and Karagai-Bulak (д) stations, along with artificially generated gaps (A1–A4 and B1–B4).

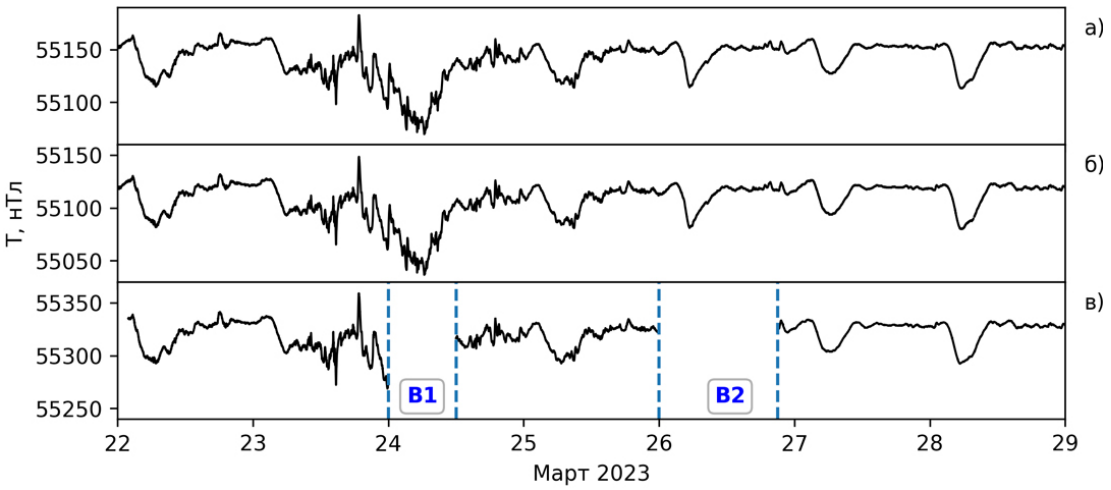

Рис. 4. Набор данных, содержащий магнитную бурю: Ак-Суу (а), Шавай (б), Карагай-Булак (в) – и искусственно созданные пропуски B1 и B2.

Fig. 4. Dataset containing a geomagnetic storm recorded at the Ak-Suu (а), Shavai (б), and Karagai-Bulak (в) stations, with artificially generated gaps (B1 and B2).

жащего магнитную бурю 24 марта 2023 г. с индексом A_p = 73, K_p = 6 (для краткости представлены данные только с трех станций). Для тестирования в этом наборе были созданы два искусственных пропуска. Пропуск B1 совпадал с периодом максимальной депрессии геомагнитного поля под воздействием бури, а B2 приходился на спокойный интервал, не затронутый возмущениями. Этот подход позволяет объективно сравнить точность восстановления данных в условиях геомагнитной активности и в спокойные периоды. Анализ подобных ситуаций важен, поскольку наличие геомагнитных бурь существенно влияет на качество работы алгоритмов и требует отдельного подхода для восстановления пропусков в такие периоды.

Методика

Алгоритм k-ближайших соседей

Алгоритм kNN является одним из наиболее простых и эффективных методов восстановления пропусков в коррелированных многомерных данных. Его широкое применение объясняется доступностью реализации и хорошими результатами, которые он демонстрирует на различных наборах данных [17]. Основная идея метода состоит в замене пропущенного значения средним значением k ближайших соседей, которые определяются по заданной метрике, такой как, например, евклидово расстояние. Такой подход применим для многомерных временных рядов, в частности к данным геомагнитного мониторинга, где измерения на станциях демонстрируют высокую корреляцию. Исследования [3, 18] показывают, что в задачах восстановления пропусков алгоритм kNN может быть эффективнее других методов машинного обучения. Это обусловлено способностью алгоритма учитывать локальные взаимосвязи между измерениями, что делает метод надежным для обработки данных с регулярными вариациями.

Процесс заполнения пропусков с использованием алгоритма kNN включает в себя несколько этапов.

Этап 1 – Выбор числа соседей k. Этот гиперпараметр играет ключевую роль в работе алгоритма. Если k слишком мало (например, k = 1), алгоритм становится чувствительным к случайным флуктуациям в данных и наличию шумовой компоненты, что снижает точность восстановления. Наоборот, слишком большое значение (например, близкое к размеру выборки n) приводит к чрезмерному сглаживанию результатов, что мешает учитывать локальные особенности данных.

Этап 2 – Расчет расстояний между наблюдениями. Для оценки сходства между данными в нашем случае используется модифи-

цированная евклидова метрика. Эта метрика подходит для данных геомагнитного мониторинга, поскольку значения измерений на всех станциях имеют одинаковый масштаб. Для работы с пропусками классическая формула расчета расстояния была адаптирована таким образом, чтобы игнорировать отсутствующие значения (Not a Number – NaN). Расстояние между двумя векторами наблюдений x_t и x_s (значения геомагнитного поля на станциях в моменты времени t и s) вычисляется только на основе тех признаков (величины геомагнитного поля на конкретной станции), которые доступны в обоих случаях:

$$R(x_t, x_s) = \sqrt{\frac{N}{m}\sum_{i=1}^{N} \delta_i \cdot (x_{t,i} - x_{s,i})^2}, \quad (1)$$

где N – размерность вектора признаков (число станций), m – количество доступных (непропущенных) признаков (где оба значения $x_{t,i}$ и $x_{s,i}$ не равны NaN), δ_i – индикатор наличия данных, где

$$\delta_i = \begin{cases} 1, & \text{если } x_{t,i} \neq NaN \text{ и } x_{s,i} \neq NaN \\ 0, & \text{иначе} \end{cases}$$

Масштабирование на коэффициент N/m компенсирует различное число доступных координат и предотвращает занижение расстояния при наличии пропусков. Поскольку значения геомагнитного поля на всех станциях имеют одинаковую размерность (нТл), дополнительная нормализация признаков не проводилась. На этом этапе создается симметричная матрица расстояний размерностью M×M, где M – количество временных отметок, а главная диагональ матрицы заполнена нулями (расстояние от точки до самой себя).

Этап 3 – Сортировка соседей и расчет заполняемого значения. После вычисления расстояний наблюдения упорядочиваются по возрастанию значения метрики. Для каждого наблюдения x_t, содержащего пропуски, выбираются k ближайших соседей среди других наблюдений. Заполнение осуществляется одновременно для всех отсутствующих признаков в данном наблюдении, а именно, каждое пропущенное значение заменяется средним арифметическим соответствующего признака у выбранных соседей. Отсутствующие у соседей в восстанавливаемом признаке значения исключаются из усреднения. Процедура повторяется для всех строк таблицы, после чего формируется полностью заполненная матрица наблюдений. Другими словами, алгоритм выполняет многомерное заполнение пропусков, используя межстанционные корреляции внутри каждого временного среза.

В такой реализации алгоритм kNN обладает рядом преимуществ при восстановлении пропусков в данных. Прежде всего, этот алгоритм прост в реализации и адаптивен к различным типам данных, что особенно важно для многомерных временных рядов, где значения коррелированы между измерениями. Еще одно преимущество заключается в том, что для восстановления пропусков kNN использует данные самой станции, что устраняет необходимость дополнительного масштабирования или нормализации. Кроме того, метод хорошо работает в условиях регулярных вариаций, таких как регулярные Sq-вариации геомагнитного поля, где отсутствуют резкие аномалии.

Тем не менее надо отметить ряд ограничений. Во-первых, при наличии выбросов в исходных данных алгоритм kNN может использовать такие значения при выборе ближайших соседей, что способно приводить к снижению точности восстановления. В связи с этим корректная предварительная фильтрация выбросов является необходимым этапом обработки данных перед применением алгоритма. Во-вторых, алгоритм требует значительных вычислительных ресурсов из-за квадратичной сложности $O(n^2)$, что особенно проявляется при обработке больших объемов данных. При этом увеличение объема выборки приводит к росту требований к оперативной памяти при расчете матрицы расстояний. Следует отметить, что на практике вычисления могут быть существенно оптимизированы за счет ограничения выборки во времени. Учитывая, что значимая корреляция геомагнитных вариаций сохраняется в пределах ограниченного временного интервала (как правило, нескольких суток), расчет расстояний может выполняться не для всего временного ряда, а в скользящем окне, что позволяет значительно снизить вычислительную нагрузку. Нако-

нец, существенным недостатком заполнения пропусков с помощью алгоритма kNN является его ограниченная эффективность при обработке экстремальных возмущений, таких как магнитные бури. В таких ситуациях в наборе данных отсутствуют аналогичные значения, которые могли бы использоваться для восстановления, что приводит к значительным ошибкам восстановления. Эти ограничения показывают необходимость комбинирования kNN с другими подходами для улучшения итоговых результатов.

Алгоритм MICE

MICE – это метод многократного заполнения пропусков, который строит вероятностную модель для каждого признака, основываясь на значениях остальных признаков [19, 20]. Алгоритм MICE выполняется итеративно, что обеспечивает постепенное улучшение качества заполнения пропусков на каждом этапе. Это проявляется наиболее эффективно при работе с данными, где взаимозависимости между признаками играют ключевую роль. Например, в рассматриваемых данных геомагнитного мониторинга MICE позволяет учитывать корреляции между различными станциями, что может повысить точность восстановления пропусков по сравнению с методами, где такие связи игнорируются.

Процесс заполнения пропусков с использованием алгоритма MICE включает несколько этапов.

Этап 1 – Предварительное заполнение пропусков. Пропуски заполняются с помощью простых методов, таких как среднее значение по столбцу, линейная интерполяция или метод kNN. Это необходимо для формирования полной начальной матрицы, необходимой для запуска итеративной процедуры MICE. Также отметим, что, поскольку значения на всех станциях имеют одинаковую физическую размерность (нТл), дополнительная стандартизация признаков не выполнялась.

Этап 2 – Построение модели для каждого признака. Для каждой станции j строится регрессионная модель:

$$x_{t,j} = f_j(x_{t,-j}) + \varepsilon, \qquad (2)$$

где $x_{t,j}$ – вектор значений на остальных станциях в момент времени t.

Например, алгоритм MissForest [21] использует подход итеративного заполнения, но вместо линейной или логистической регрессии применяет алгоритм случайного леса для предсказания пропущенных значений. Использование случайного леса в качестве предсказательной модели оправдано в случае, когда данные содержат как числовые, так и категориальные переменные, а также когда они имеют сложные нелинейные зависимости между переменными. В рассматриваемой задаче геомагнитные данные являются числовыми и характеризуются высокой межстанционной корреляцией, что делает линейную регрессию предпочтительным выбором вследствие простой реализации и вычислительной эффективности. В настоящей работе функция f_j реализована в виде множественной линейной регрессии. Временные зависимости в данной постановке не учитывались, и моделирование выполнялось только на основе межстанционных корреляций внутри каждого временного среза.

Этап 3 – Циклическое заполнение пропусков. На данном этапе для каждого признака (станции) последовательно выполняется предсказание только отсутствующих значений с использованием регрессионной модели, построенной на предыдущем этапе. После восстановления пропусков для одного признака обновленные значения используются при обработке следующего признака. Таким образом выполняется один полный проход по всем признакам.

Этап 4 – Итеративное уточнение. В следующем итерационном цикле используются обновленные значения, полученные на предыдущем этапе. Итерационный процесс продолжался до достижения заданного числа циклов K, которое по результатам предварительных численных экспериментов было выбрано равным 10 (дальнейшее увеличение числа итераций не приводило к существенному снижению ошибки восстановления).

Следует подчеркнуть, что в данной постановке задача не сводится к решению одной системы линейных уравнений, несмотря на

использование линейной регрессии в качестве модели f_j. Это связано с тем, что в алгоритме MICE рассматривается последовательность условных моделей для различных признаков (станций), каждая из которых строится с учетом текущих оценок пропущенных значений в других признаках.

В отличие от kNN, алгоритм MICE может учитывать сложные взаимозависимости между признаками, что делает его эффективным при работе с многомерными временными рядами. Итеративный характер метода позволяет на каждом этапе уточнять восстановленные значения, обеспечивая корректные результаты в сложных условиях, таких как магнитные бури. Также алгоритм обладает гибкостью, которая заключается в возможности выбора предсказательной модели для заполнения пропусков. Это могут быть как простые методы, например линейная регрессия, так и более сложные, включая деревья решений или случайные леса, что позволяет адаптировать MICE к различным типам данных в зависимости от задачи.

Тем не менее данный алгоритм имеет определенные ограничения. Так, его вычислительная сложность и значительное потребление оперативной памяти могут создавать трудности при работе с большими наборами данных. Кроме того, скорость сходимости алгоритма MICE зависит от качества начального заполнения пропусков, поскольку оно определяет начальную точку итерационного процесса и влияет на стабильность и скорость уточнения оценок на последующих шагах. Использование слишком простых методов, например среднего или медианного значения, на этом этапе может замедлить процесс. Еще одной проблемой является перенос артефактов, таких как выбросы или ступени, из данных станций-доноров в восстанавливаемые данные. Чтобы избежать этого, важно применять предварительную фильтрацию [13, 14], которая удаляет нежелательные артефакты перед началом работы алгоритма. При этом ключевым моментом является то, что использование алгоритма MICE в данной задаче основывается на предположении о стабильности межстанционных корреляций во времени, что может и не соблюдаться для других задач подобного рода. В общем случае выбор между kNN и MICE во многом будет зависеть от специфики самого исследования и характеристик данных. В условиях регулярных Sq-вариаций, характеризующихся высокой повторяемостью и ограниченным диапазоном значений, предпочтительным является использование алгоритма kNN. В то же время при наличии экстремальных возмущений, когда значения выходят за пределы типичного диапазона, более эффективным оказывается алгоритм MICE, учитывающий глобальные межстанционные зависимости. В текущей реализации выбор метода осуществляется на основе анализа данных и выполняется исследователем, что делает процедуру автоматизированной. При этом сам выбор конкретного метода может быть формализован. В частности, в качестве критерия может использоваться уровень вариабельности временного ряда или внешние индексы геомагнитной активности (K_p, A_p). В случае превышения заданного порогового значения будет использован алгоритм MICE, в противном случае – kNN. Пороговое значение при этом может быть определено эмпирически на основе обучающей выборки или задано в соответствии с классификацией геомагнитной активности.

Надо отметить, что, в отличие от классических задач машинного обучения, ориентированных на построение обобщающей предиктивной модели с последующим применением к новым данным, в данном случае восстановление пропусков выполняется непосредственно на основе имеющихся наблюдений с использованием межстанционных зависимостей. Иными словами, алгоритмы kNN и MICE применяются не как инструменты построения универсальной модели, а как методы локальной и глобальной реконструкции пропущенных значений. Так, алгоритм kNN реализует восстановление на основе поиска ближайших по структуре наблюдений, тогда как MICE последовательно восстанавливает пропуски, используя регрессионные зависимости между временными рядами различных станций. Таким образом, предложенный подход ориентирован на реконструкцию данных в пределах рассматриваемой сети наблюдений и не предполагает экстраполяции или прогнозирования за пределами имеющегося набора данных.

Результаты

Заполнение пропусков с помощью алгоритма kNN

В качестве основной метрики оценки качества заполнения пропусков использовалась средняя абсолютная ошибка MAE (Mean Absolute Error), рассчитанная как среднее значение модуля разности между исходными и восстановленными данными. Эта величина является устойчивой к выбросам и обеспечивает интерпретируемую оценку отклонения восстановленных значений от исходных данных в физических единицах (нТл). В отличие от квадратичных метрик, таких как MSE (Mean Squared Error), MAE менее чувствительна к отдельным аномальным значениям, что особенно важно при анализе геомагнитных данных, которые могут содержать выбросы и экстремальные вариации.

Заполнение пропусков с использованием алгоритма kNN было протестировано на двух наборах данных для оценки его возможностей. При этом необходимо отметить, что оптимальный выбор гиперпараметра k является ключевым при использовании алгоритма kNN. На практике часто применяется эмпирическое правило $k \leq \sqrt{n}$, где n – количество наблюдений [22]. Для первого набора данных (n = 38 880) это правило определяет значение k на уровне ~200. Для более точного анализа зависимости MAE от количества соседей k были проведены эксперименты с варьированием значения этого параметра в диапазоне от 30 до 1500 (рис. 5). Минимальные значения MAE достигаются при $k \approx 150–500$, после чего величина ошибки постепенно увеличивается вследствие включения менее коррелированных наблюдений. Этот диапазон можно считать оптимальным для восстановления пропусков в первом наборе данных, поскольку он обеспечивает баланс между точностью и вычислительными затратами. Малые значения k увеличивают MAE из-за недостаточной устойчивости к локальным выбросам, а слишком большие значения включают данные станций с низкой корреляцией, что также снижает итоговую точность восстановления. Таким образом, хотя правило $k \leq \sqrt{n}$ является удобным ориентиром, оптимальный выбор k должен зависеть от специфики данных и поставленной задачи. В данной работе значение гиперпараметра k было выбрано равным 150.

На рис. 6 представлены результаты восстановления искусственно созданных пропусков для типичных Sq-вариаций на разных станциях.

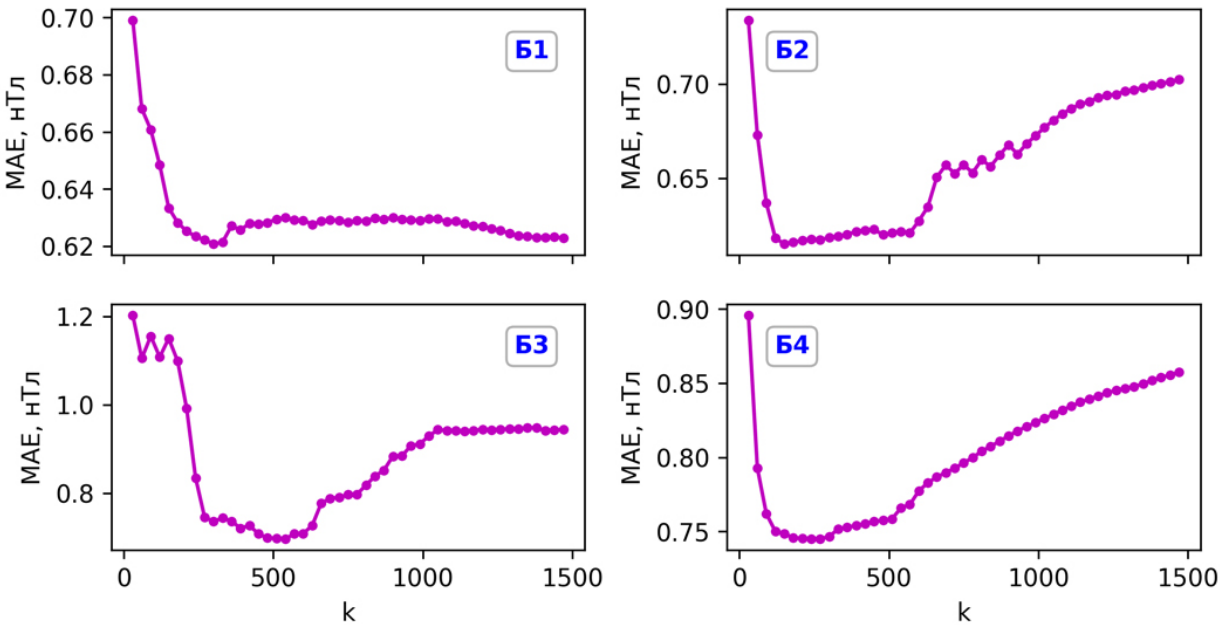

Рис. 5. Зависимость ошибки восстановления пропусков, MAE, от параметра k для участков Б1–Б4 на основе алгоритма kNN.
Fig. 5. Variation of the MAE of gap reconstruction as a function of the parameter k for segments B1–B4 using the kNN algorithm.

На рис. 7 показаны графики исходных и восстановленных временных рядов для искусственно созданных пропусков. Для наглядности восстановленные кривые смещены на 5 нТл относительно исходных данных. Результаты показывают, что на станциях, расположенных близко друг к другу (Ак-Суу, Шавай, Иссык-Ата, Кегеты – рис. 7, слева), пропуски были восстановлены с хорошей точностью (MAE ≤ 0.36 нТл). Однако на удаленной станции Карагай-Булак точность восстановления оказалась ниже (MAE варьирует от 0.62 до 1.15 нТл). Это связано с тем, что расстояние до ближайшей станции (Кегеты) составляет около 155 км, что снижает корреляцию между данными. Следует отметить, что полное устранение погрешности восстановления невозможно, так как каждая станция обладает уникальными локальными особенностями, включая географические, геомагнитные и техногенные факторы, которые создают различия в данных между станциями. Высокие значения коэффициента корреляции ($r \geq 0.99$) подтверждают сохранение общей формы сигнала, однако основная оценка точности проводилась по метрике MAE, которая более чувствительна к ошибкам восстановления.

Качество восстановления пропусков было дополнительно проанализировано на данных, зарегистрированных во время геомагнитной бури и в период после ее окончания (рис. 8). Для оценки использовалась метрика MAE аналогично процедуре, примененной для Sq-вариаций. Как видно из рисунка, геомагнитная буря создает значительные изменения в магнитном поле, что существенно усложняет задачу восстановления пропусков.

Результаты показали, что регулярные Sq-вариации восстанавливаются с высокой точностью, как и в предыдущем примере. Однако эффективность алгоритма резко снижается на участках, где присутствует магнитная буря (рис. 9). В условиях экстремальных значений во время бури алгоритм kNN не находит адекватных аналогов в выборке и склонен выбирать значения из диапазона спокойных условий, что

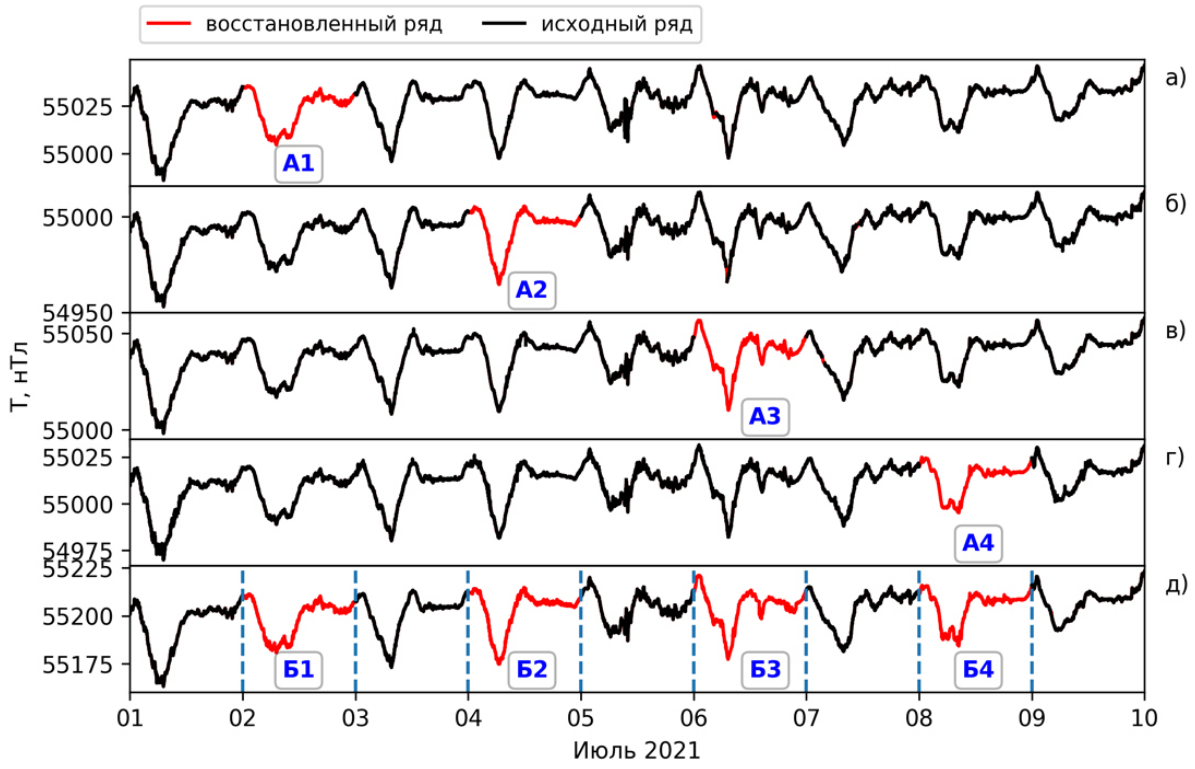

Рис. 6. Результат заполнения искусственно созданных пропусков А1–А4 и Б1–Б4 на основе алгоритма kNN для набора данных, содержащих типичные Sq-вариации величины геомагнитного поля на станциях Ак-Суу (а), Шавай (б), Иссык-Ата (в), Кегеты (г) и Карагай-Булак (д)

Fig. 6. Reconstruction of the artificially generated gaps A1–A4 and B1–B4 using the kNN algorithm for the dataset containing typical Sq variations of the geomagnetic field at the Ak-Suu (a), Shavai (б), Issyk-Ata (в), Kegety (г), and Karagai-Bulak (д) stations

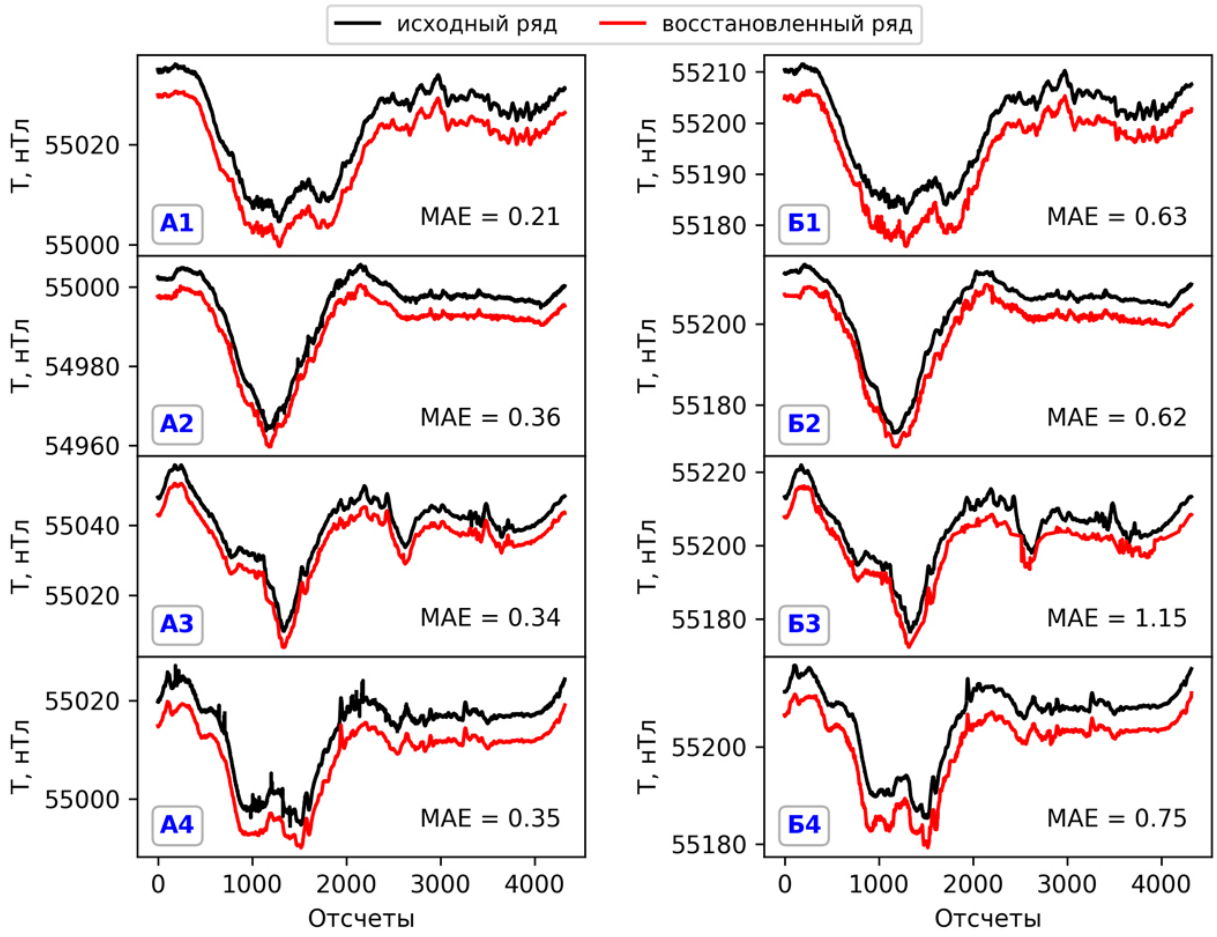

Рис. 7. Исходные и восстановленные временные ряды для участков А1–А4 и Б1–Б4 на основе алгоритма kNN. Слева – для станций Ак-Суу, Шавай, Иссык-Ата, Кегеты; справа – станция Карагай-Булак.

Fig. 7. Original and reconstructed time series for segments A1–A4 and B1–B4 using the kNN algorithm. Left panels: Ak-Suu, Shavai, Issyk-Ata, and Kegety stations; right panels: Karagai-Bulak station.

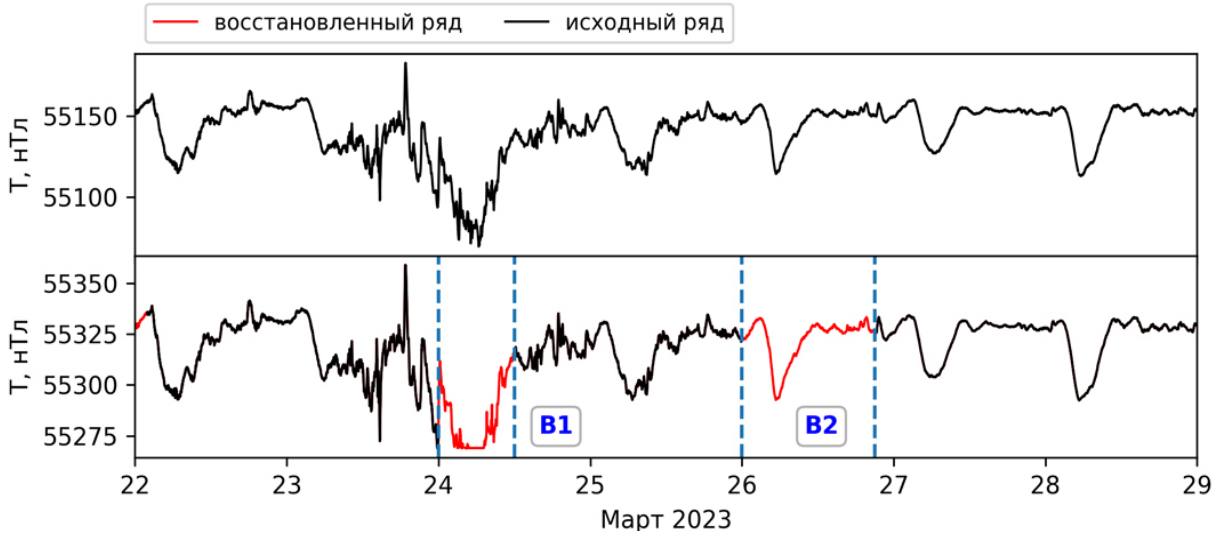

Рис. 8. Результат заполнения искусственно созданных пропусков В1, В2 на основе алгоритма kNN для набора данных, содержащих геомагнитную бурю и Sq-вариации. Вверху – станция Ак-Суу (для сравнения); внизу – станция Карагай-Булак.

Fig. 8. Reconstruction of the artificially generated gaps B1, B2 using the kNN algorithm for the dataset containing a geomagnetic storm and Sq variations. Top panel: Ak-Suu station (reference case); bottom panel: Karagai-Bulak station.

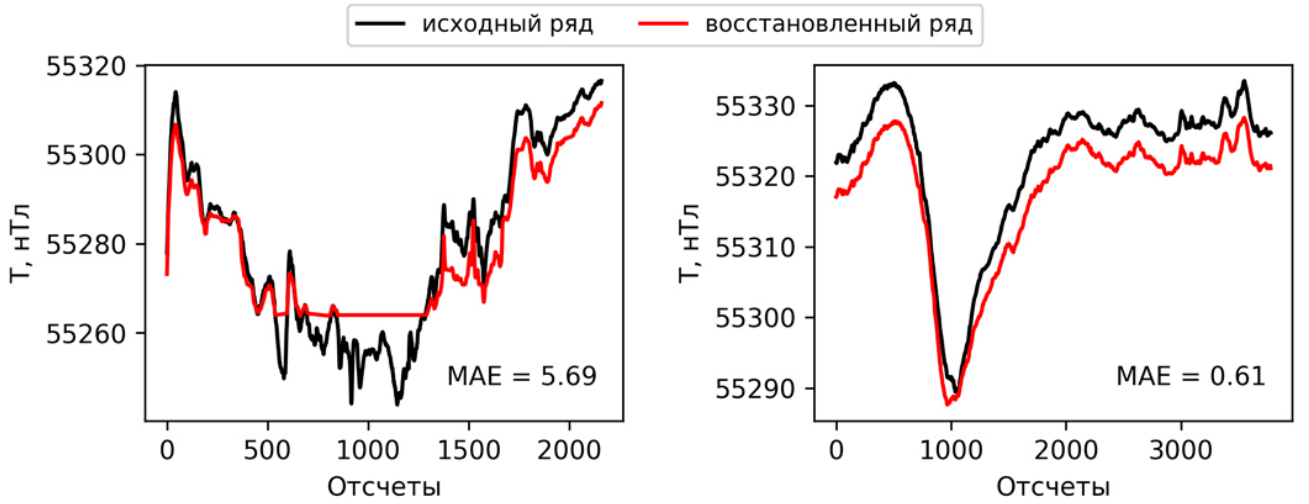

Рис. 9. Исходные и восстановленные временные ряды для участков B1 и B2 на основе алгоритма kNN для станции Карагай-Булак. Слева в случае геомагнитной бури; справа в случае типичных Sq-вариаций.

Fig. 9. Original and reconstructed time series for segments B1 and B2 at the Karagai-Bulak station obtained using the kNN algorithm. Left panel: geomagnetic storm conditions; right panel: typical Sq variations.

приводит к систематическому занижению амплитуды возмущения. Это подтверждается высоким значением MAE = 5.69 нТл, в то время как для Sq-вариаций MAE составляет всего 0.61 нТл.

Таким образом, kNN плохо подходит для восстановления данных в условиях экстремальных возмущений, когда значения выходят за пределы диапазонов нормальных вариаций. В таких случаях приходится использовать более сложные методы, способные учитывать глобальные корреляции и межстанционные связи.

Заполнение пропусков с помощью алгоритма MICE

Анализ работы алгоритма MICE был выполнен на тех же наборах данных, что и для kNN, включая искусственно созданные пропуски (А1–А4 и Б1–Б4). На рис. 10 представлены результаты восстановления данных с регулярными Sq-вариациями, зарегистрированными на различных станциях. Для наглядности восстановленные кривые смещены на 5 нТл относительно исходных данных.

Мы видим, что алгоритм MICE обеспечивает точность восстановления, сопоставимую с kNN. Для близкорасположенных станций значение MAE в среднем составило менее 0.37 нТл, а для удаленной станции Карагай-Булак – 0.73 нТл. Однако алгоритм MICE имеет склонность к переносу мелких выбросов из данных восстанавливающих станций в данные станции с пропусками. Это объясняется тем, что регрессионная модель воспроизводит локальные аномалии, присутствующие в донорских рядах. Это особенно заметно на участках А1, А2 и Б2 (рис. 10). В отличие от MICE, алгоритм kNN обеспечивает более сглаженные восстановленные данные за счет осреднения по k-ближайшим соседям, что снижает влияние выбросов. Эти различия подчеркивают, что выбор подходящего метода зависит от характеристик данных. Если в данных отсутствуют выбросы, то kNN может быть предпочтительнее из-за своей простоты.

В случаях, когда в данных присутствуют экстремальные вариации (магнитные бури), алгоритм MICE демонстрирует более устойчивое поведение по сравнению с kNN. Значение MAE в 5 раз ниже полученного при применении kNN (1.06 против 5.69 нТл) (рис. 11 А, слева), поскольку в MICE учитываются глобальные зависимости между станциями, которые важны в период экстремальных событий. Следовательно, MICE более эффективен в условиях, когда значения данных выходят за пределы нормального диапазона вариаций, характерного для спокойных периодов.

Изначально для предварительного заполнения пропусков использовался простой метод расчета среднего значения по всему времен-

ному ряду. Однако этот подход оказался неэффективным, так как он игнорирует корреляции между данными станций, что особенно критично для продолжительных пропусков. Для повышения точности восстановления было решено провести предварительное заполнение с помощью алгоритма kNN. Это позволяет учесть локальные особенности данных и получить более точное начальное состояние. Далее применялся итеративный алгоритм MICE, который уточняет восстановленные значения, учитывая взаимозависимости между станциями. На рис. 11 Б представлены результаты восстановления временных рядов для пропусков B1 и B2 на основе алгоритма MICE с предварительным заполнением методом kNN. Данный подход позволил снизить значение MAE для пропуска во время магнитной бури до 0.88 нТл.

Следует отметить, что эффективность предложенной методики определяется рядом условий, связанных с особенностями исходных данных. Ключевым фактором является степень коррелированности временных рядов между станциями. Метод показывает наилучшие результаты (MAE $\leq$ 0.37 нТл) при высокой межстанционной корреляции ($r \geq 0.9–0.99$), характерной для близкорасположенных пунктов наблюдений. При увеличении расстояния между станциями, как показано на примере станции Карагай-Булак, точность восстановления снижается (MAE $\geq$ 0.55 нТл) вследствие ослабления пространственной согласованности вариаций геомагнитного поля. При этом 6 близкорасположенных станций (среднее расстояние между двумя соседними станциями ~19 км) составляют вытянутую в широтном направлении область протяженностью ~92 км, тогда как удаленная станция Карагай-Булак находится на расстоянии 155 км от этой области.

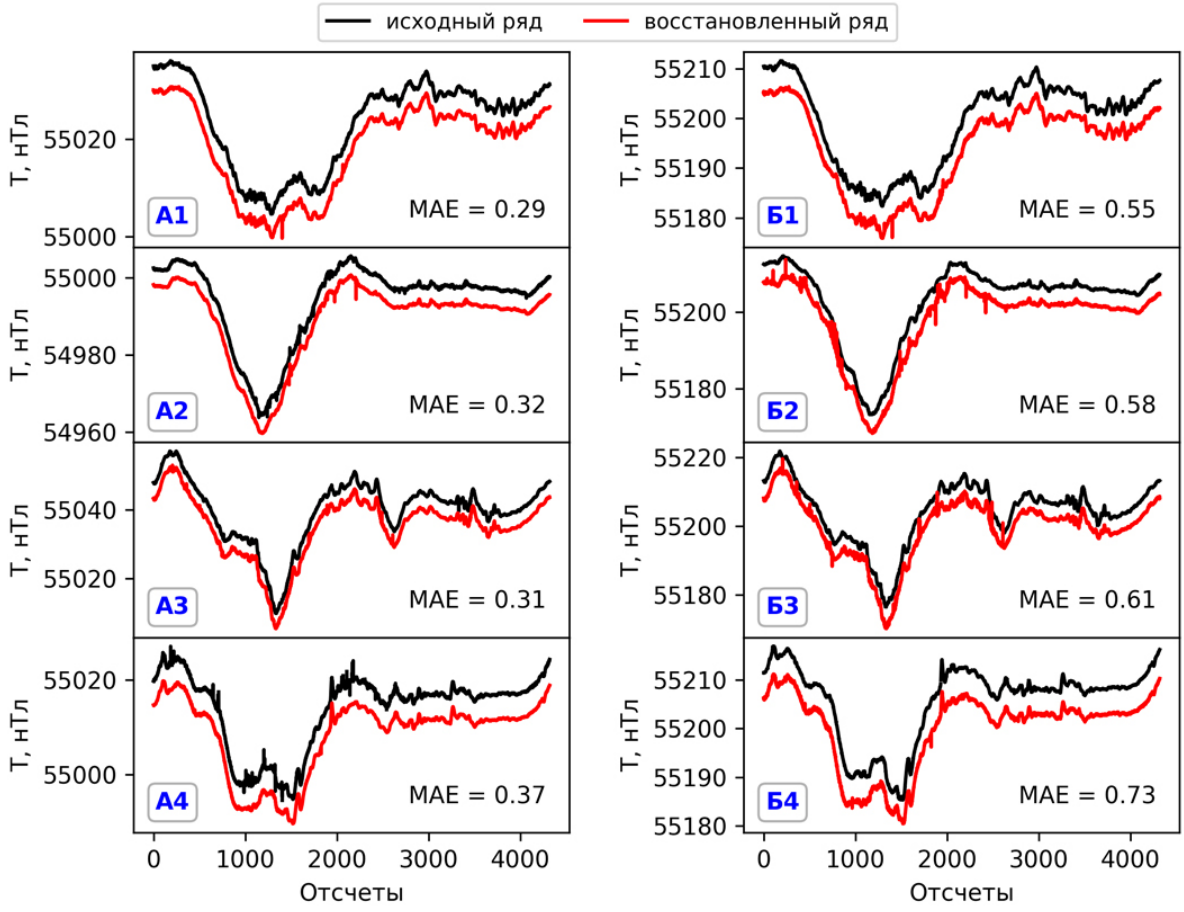

Рис. 10. Исходные и восстановленные временные ряды с регулярными Sq-вариациями для участков А1–А4 и Б1–Б4 на основе алгоритма MICE. Слева для станций Ак-Суу, Шавай, Иссык-Ата, Кегеты; справа для станции Карагай-Булак.

Fig. 10. Original and reconstructed time series for segments A1–A4 and B1–B4 obtained using the MICE algorithm. Left panels: Ak-Suu, Shavai, Issyk-Ata, and Kegety stations; right panels: Karagai-Bulak station.

Кроме того, как было показано выше, алгоритм kNN достаточно хорошо справляется с восстановлением данных на регулярных участках, где отсутствуют резкие аномалии, однако демонстрирует ограниченную применимость в условиях экстремальных возмущений, таких как магнитные бури, когда значения выходят за пределы диапазона, представленного в обучающей выборке. В этих условиях более предпочтительно использование алгоритма MICE, учитывающего глобальные межстанционные зависимости. Таким образом, предложенный комбинированный подход наиболее эффективен для сетей геомагнитного мониторинга с устойчивыми пространственными корреляциями и требует осторожного применения на данных с низкой степенью согласованности.

Гибридный подход, при котором kNN используется для предварительного заполнения, а MICE уточняет данные на сложных участках, был выбран в качестве основного метода восстановления пропусков в задачах, рассматриваемых далее.

Удаление выбросов

Восстановление значений на одной станции на основе данных других станций может быть использовано для моделирования хода геомагнитного поля с учетом глобальных источников, таких как тренд, суточные Sq-вариации и магнитные бури. При этом расчет модельного поведения на основе данных других станций автоматически исключает нерегулярные компоненты, например помехи и выбросы, характерные исключительно для рассматриваемой станции. Сравнение временного ряда станции, содержащего аномалии, с моделью, построенной на данных других станций, позволяет эффективно выявлять эти аномалии.

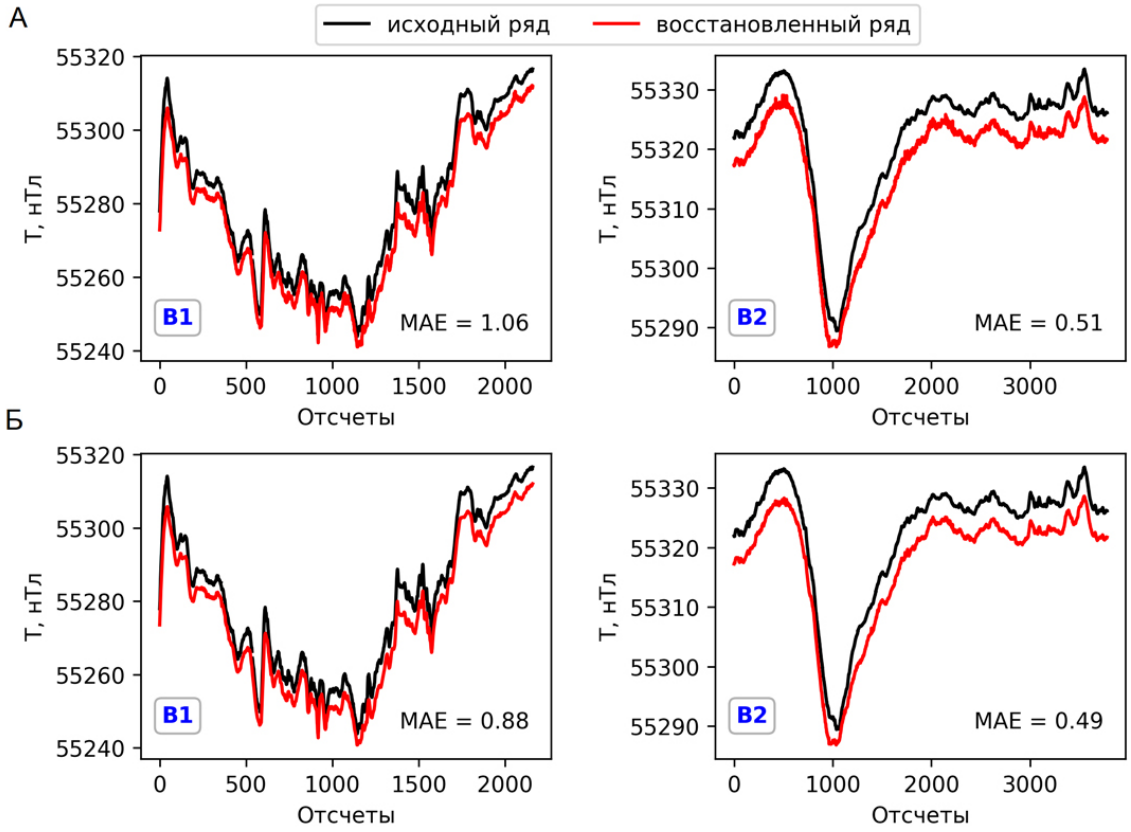

Рис. 11. Исходные и восстановленные временные ряды для участков B1 и B2 для станции Карагай-Булак на основе алгоритма MICE (А) и на основе алгоритма MICE с предварительным заполнением пропусков методом kNN (Б). Слева в случае геомагнитной бури; справа в случае типичных Sq-вариаций.

Fig. 11. Original and reconstructed time series for segments B1 and B2 at the Karagai-Bulak station obtained using the MICE algorithm (A) and using the MICE algorithm with preliminary gap filling by the kNN algorithm (Б). Left panel: geomagnetic storm conditions; right panel: typical Sq variations.

Такой подход можно использовать для мониторинга геомагнитного поля и обнаружения отклонений, вызванных локальными событиями или помехами.

В рамках данной работы под выбросами понимаются кратковременные аномальные отклонения значений геомагнитного поля, не связанные с естественными вариациями (такими как Sq-вариации или магнитные бури) и обусловленные, как правило, техногенными помехами, сбоями измерительного оборудования или внешними электромагнитными воздействиями (например, грозовой деятельностью) [14]. Такие выбросы, как правило, имеют импульсный характер и существенно отличаются по амплитуде от фоновых вариаций [13]. На рис. 12 представлен пример выбросов, наблюдаемых в реальных данных геомагнитного мониторинга (станция Карагай-Булак).

Такие выбросы, как правило, характеризуются малой длительностью и отсутствием пространственной согласованности между станциями (в отличие от геомагнитных бурь). Также следует отметить, что возмущения, возникающие при работе установки ЭРГУ-600, имеют иную природу по сравнению с импульсными выбросами. Они проявляются в виде ступенчатых изменений уровня сигнала, сохраняющихся в течение определенного интервала времени (рис. 2), и поэтому не могут рассматриваться как классические выбросы. Такие аномалии анализируются отдельно и требуют применения специализированного подхода, описанного в следующем разделе.

Для тестирования алгоритмов был выбран временной ряд станции Карагай-Булак (рис. 13 а), содержащий геомагнитную бурю и Sq-вариации. Далее к этому ряду были искусственно добавлены 200 выбросов, со случайным местоположением, знаком и амплитудой в диапазоне 30–100 нТл (рис. 13 б). Использование искусственно сгенерированных выбросов в данном случае обеспечивает контролируемую оценку эффективности алгоритма, поскольку их точное положение и параметры заранее известны. При этом диапазон амплитуд и их количество выбирались на основе анализа реальных данных, они соответствуют техногенным помехам, наблюдаемым на станциях геомагнитного мониторинга.

Выбранное количество выбросов для такого промежутка времени значительно превышает норму для рассматриваемых станций (превышение среднего количества выбросов в 4–10 раз). Добавление выбросов в таком нетипичном количестве позволяет протестировать алгоритм в условиях, когда другие методы не дают значимых результатов [13, 14]. Для того чтобы смоделировать временной ряд на одной из станций по данным других пунктов наблюдений, необходимо искусственно создать пропуски и оставить несколько значений, чтобы алгоритм рассчитал связь между данными станций. Для этого из зашумленного выбросами временного ряда $T_ш$ было выбрано случайным образом N = 150 значений (рис. 13 в). Значение N определялось эмпирически на основе предварительных численных экспериментов и

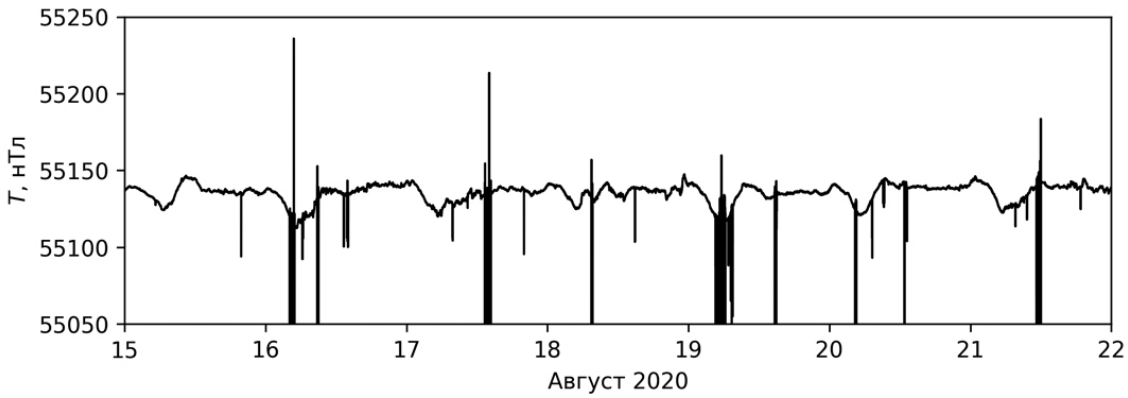

Рис. 12. Пример импульсных выбросов в данных вариации геомагнитного поля на станции Карагай-Булак.
Fig. 12. Example of impulsive outliers in geomagnetic field variation data recorded at the Karagai-Bulak station

оказалось достаточным для устойчивого восстановления целевого сигнала при заданной длине ряда и количестве искусственных выбросов. Анализ чувствительности к параметру N показал, что диапазон 100–200 обеспечивает сопоставимые результаты.

Полученная разреженная последовательность T_N использовалась для восстановления целевого временного ряда с помощью комбинации алгоритмов kNN и MICE. Нужно отметить, что среди выбранных наугад N точек есть вероятность наличия выбросов (рис. 13 в, отмечено стрелкой), которые могут исказить результаты восстановления пропусков. В этом случае необходимо провести несколько серий отборов N точек с последующим восстановлением целевого временного ряда, используя накопление и последующую робастную оценку для устранения влияния подобных выбросов.

Для устранения выбросов было выполнено L = 100 итераций, в каждой из которых рассчитывался временной ряд на основе разреженных данных. Получено 100 различных временных рядов, каждый из которых мог содержать отдельные выбросы. Для итогового восстановления временного ряда использовался расчет медианы в качестве устойчивой оценки среднего значения, что дает возможность нивелировать влияние выбросов, попавших в подвыборки на отдельных итерациях.

Важно отметить, что восстановленный временной ряд представляет собой модель данных, которая не является точной копией исходного ряда. Поэтому он не может использоваться напрямую. Для удаления выбросов из исходных данных рассчитывалась разность ΔT между исходным загрязненным временным рядом и модельным. Затем применялся метод выделения выбросов на основе анализа распределения ΔT (рис. 14, слева). Чтобы избежать искажений, вызванных большим числом выбросов высокой амплитуды (3σ = 17 нТл), вместо стандартного отклонения использовался межквартильный критерий (рис. 14, справа), согласно которому границы выбросов рассчитывались по следующим формулам:

$$L_1 = Q1 - W \cdot IQR,$$

$$L_2 = Q3 + W \cdot IQR,$$

где Q1, Q3 – первый и третий квартили, IQR = Q3 – Q1 – межквартильный размах (InterQuartile Range). Коэффициент W = 4 выбран эмпирически, исходя из характера распределения ΔT и доли устраняемых выбросов. В нашем случае этот подход обеспечивает гибкую фильтрацию аномалий с минимизацией

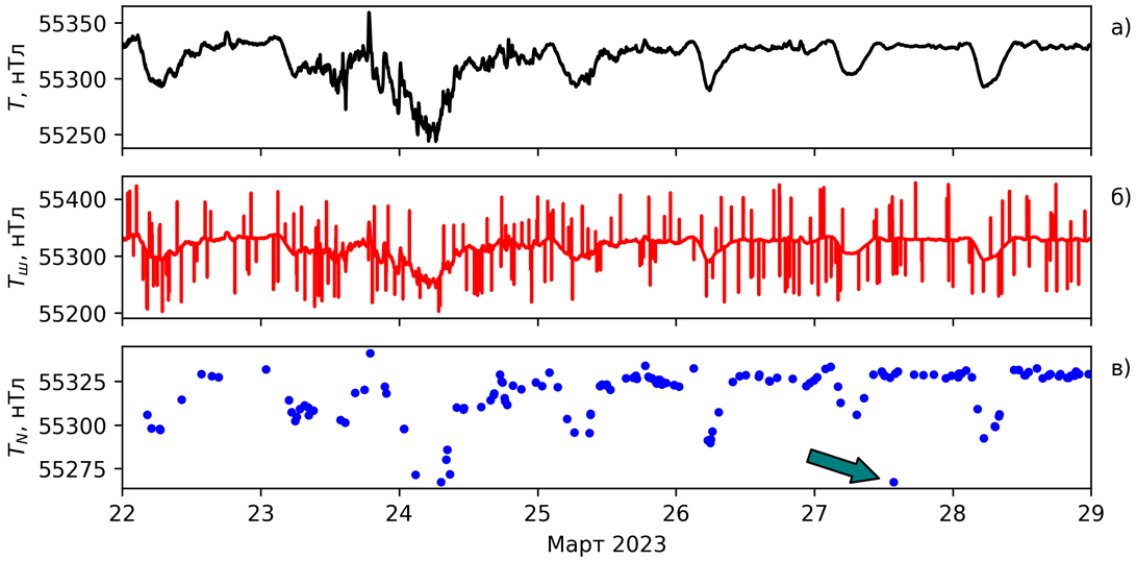

Рис. 13. Временные ряды вариаций величины геомагнитного поля: а) исходный, б) загрязненный выбросами, в) случайная выборка из N = 150 точек. Стрелкой указан выброс, который попал в ряд T_N.

Fig. 13. Time series of geomagnetic field variations: (a) original, (б) contaminated with outliers, and (в) a random sample of N = 150 points. The arrow marks the outlier that was included in the T_N series.

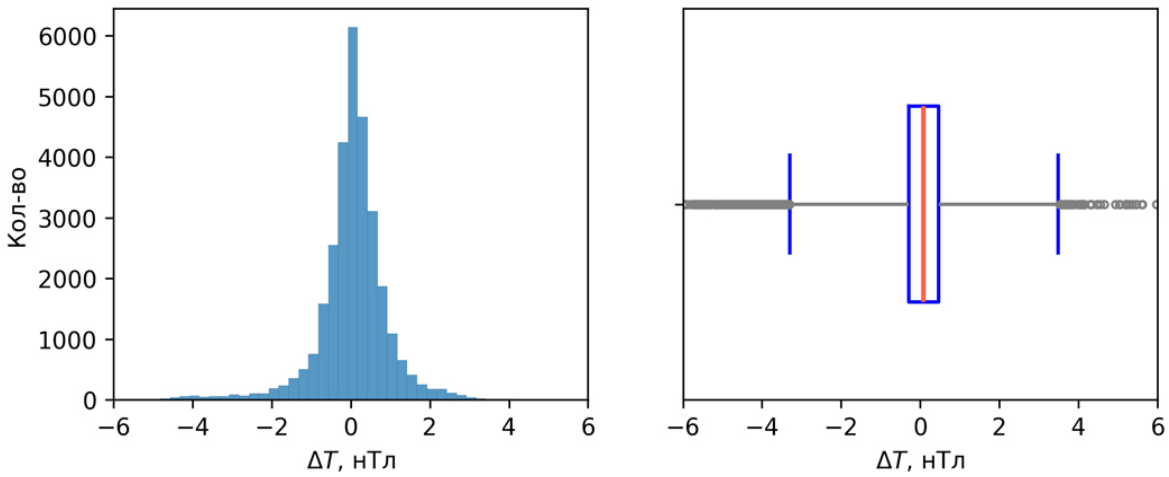

Рис. 14. Распределение разности ΔT (нТл) между исходным загрязненным временным рядом и модельным после применения комбинированного алгоритма kNN+MICE: слева – гистограмма распределения; справа – диаграмма размаха (boxplot), где выделены основные составляющие: медиана (красная вертикальная линия), межквартильный диапазон IQR (синие вертикальные линии) и выбросы (кружки серого цвета по краям).

Fig. 14. Distribution of the difference ΔT (nT) between the original contaminated time series and the modeled series after applying the combined kNN+MICE algorithm. Left panel: histogram of the distribution; right panel: boxplot highlighting the main statistical components – median (red vertical line), interquartile range (IQR; blue vertical lines), and outliers (gray circles at the extremes).

влияния выбросов на восстановленные данные.

Согласно межквартильному критерию, все значения ΔT, которые превышают пределы L_1 и L_2, классифицируются как выбросы. На рис. 15 представлены исходный временной ряд, модельный ряд и их разность ΔT, где аномальные значения выбросов отмечены отдельно. Данный подход позволяет выделить выбросы, которые существенно отклоняются от типичного (модельного) распределения ΔT. Количество выявленных выбросов, их амплитуда и распределение визуально подтверждают корректность работы алгоритма. В частности, в результате такой многоэтапной процедуры были правильно выделены все 200 искусствен-

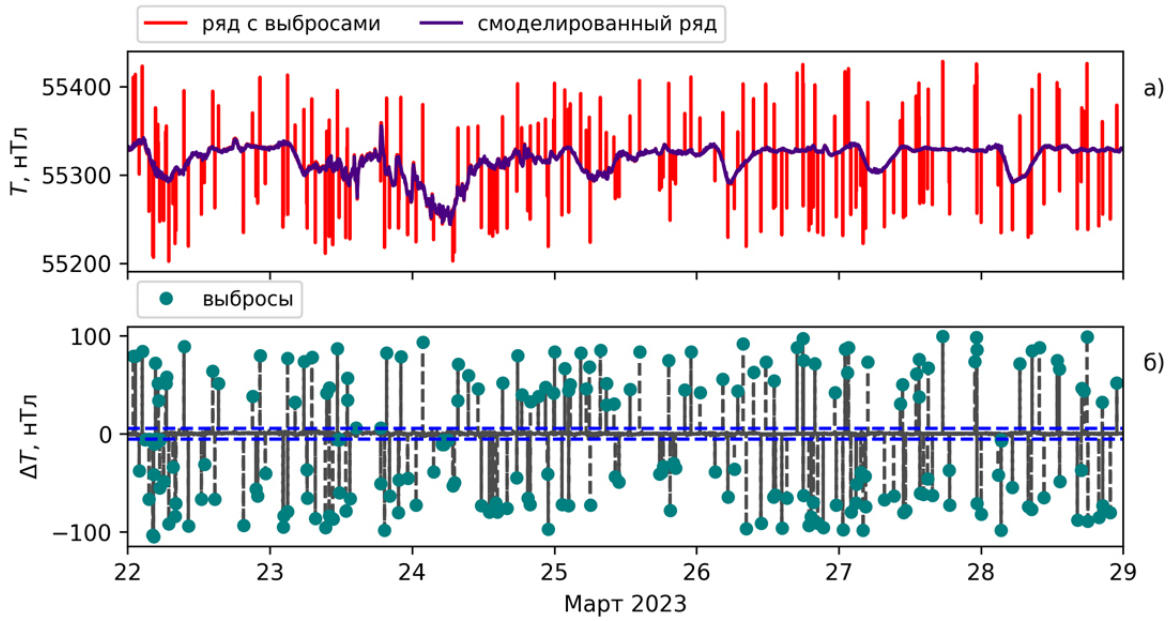

Рис. 15. Загрязненный выбросами и смоделированный временные ряды (а), их разность ΔT с выделенными выбросами (б). Горизонтальные штриховые линии – границы отсечки выбросов L_1 и L_2.

Fig. 15. Contaminated and reconstructed time series (a) and their difference ΔT with highlighted outliers (б). The horizontal dashed lines indicate the threshold levels L_1 and L_2 used for outlier detection.

но созданных выбросов, однако здесь нужно упомянуть чувствительность к выбору параметра W и возможные трудности при наличии большого числа пересекающихся выбросов.

Следует отметить, что в качестве отфильтрованного временного ряда используется исходный временной ряд, из которого удалены значения, классифицированные как выбросы. Такой подход позволяет сохранить максимум информации об исходных данных, исключая только те точки, которые существенно отклоняются от распределения.

Еще раз подчеркнем, что смоделированный временной ряд служит исключительно для выявления местоположений выбросов и не используется в качестве итогового результата. Такая методика позволяет сохранить структуру временного ряда, минимизируя потери информации при удалении выбросов.

Оценка возмущений, индуцированных ЭРГУ-600

Как отмечалось ранее, на станциях Таш-Башат и Чункурчак, расположенных ближе всего к питающему диполю установки ЭРГУ-600, наблюдаются изменения магнитного поля во время сеансов электромагнитного зондирования. Для оценки величины индуцированного возмущения в магнитном поле («магнитного эффекта») традиционно используется метод вычисления разности между значениями геомагнитного поля на исследуемой станции и базовой (Ак-Суу) [23, 24]. Однако данный подход имеет свои ограничения. Локальные особенности исследуемых станций (географическое положение, техногенные факторы, магнитные аномалии) могут существенно влиять на результаты измерений. Это затрудняет точное определение амплитуды аномалий, вызванных непосредственно сеансами электромагнитного зондирования (рис. 16), из-за присутствия сложного низкочастотного тренда. Для решения этой проблемы требуется разработка более универсальных методов анализа, которые учитывали бы как глобальные закономерности, так и локальные особенности.

Для более точной оценки возмущений, вызванных работой ЭРГУ-600, был применен многоэтапный подход. Сначала участки временного ряда, соответствующие сеансам зондирования, заменялись пропусками на основе данных из журнала пусков ЭРГУ-600. Эти пропуски заполнялись с использованием комбинации алгоритмов kNN + MICE. После этого восстановленные участки записывались обратно в копию исходного временного ряда. Далее вычислялась разность между исходным рядом и его копией, в которой восстановление проводилось на основе данных со станций Ак-Суу, Шавай, Кегеты, Иссык-Ата и Карагай-Булак. Эти станции, находясь далеко от питающего диполя, не фиксировали изменений, вызванных работой ЭРГУ-600. Такой метод позволяет исключить влияние локальных особенностей станций и сосредоточиться на выявлении аномалий, вызванных исключительно электромагнитным зондированием (рис. 17), что потенциально расширяет возможности его применения на других объектах мониторинга.

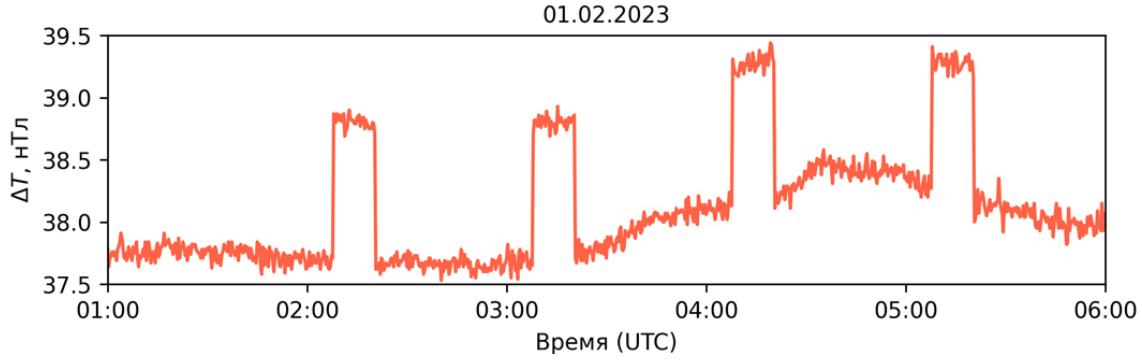

Рис. 16. Разность между временными рядами вариаций величины геомагнитного поля на станциях Таш-Башат и Ак-Суу во время сеансов электромагнитного зондирования с помощью установки ЭРГУ-600.

Fig. 16. Difference between time series of the geomagnetic field variation recorded at the Tash-Bashat and Ak-Suu stations during electromagnetic sounding sessions with the ERGU-600 system.

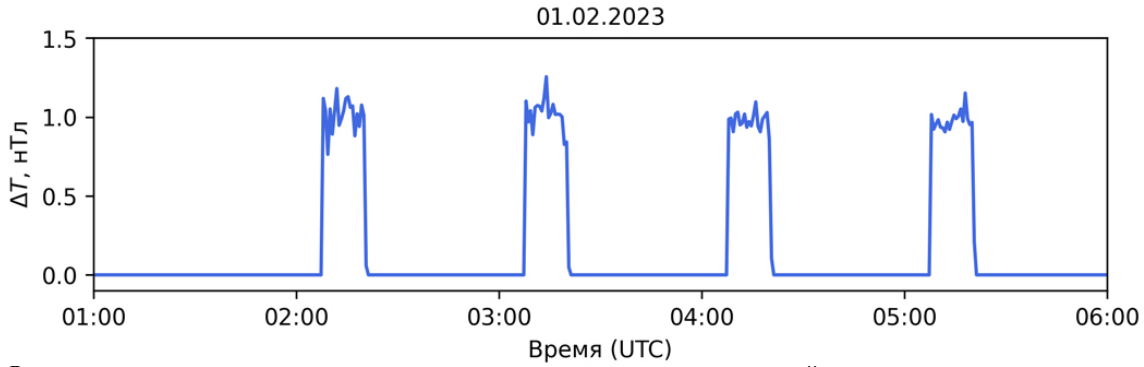

Рис. 17. Разность между исходным и восстановленным временными рядами вариаций величины геомагнитного поля на станции Таш-Башат во время сеансов электромагнитного зондирования с помощью установки ЭРГУ-600.

Fig. 17. Difference between the original and reconstructed time series of the geomagnetic field variation at the Tash-Bashat station during electromagnetic sounding sessions with the ERGU-600 system.

Результаты, представленные на рис. 17, демонстрируют, что предложенная комбинация алгоритмов kNN и MICE позволяет более точно оценить влияние ЭРГУ-600 на изменения геомагнитного поля, зарегистрированные на ближайших станциях. Нужно отметить, что при таком подходе минимизируется влияние человеческого фактора, что способствует воспроизводимости результатов. Также автоматизация значительно снижает трудозатраты на обработку больших объемов данных, что важно для анализа результатов многолетнего мониторинга.

Заключение

В данной работе рассмотрены методы восстановления пропусков в данных геомагнитного поля с использованием алгоритмов kNN и MICE. Проведен детальный анализ их эффективности на наблюдениях с различными типами вариаций, включая типичные Sq-вариации и магнитные бури, а также на данных, дополненных искусственными выбросами. Результаты исследования показали, что алгоритм kNN хорошо восстанавливает регулярные Sq-вариации (MAE $\leq$ 0.4 нТл), однако его эффективность значительно снижается во время магнитных бурь (MAE = 5.7 нТл) вследствие отсутствия аналогичных значений в диапазоне данных. В то же время алгоритм MICE, благодаря учету корреляций между станциями, дает существенно меньшую ошибку восстановления в условиях магнитной бури (MAE = 1.1 нТл).

Установлено, что комбинированное использование kNN для предварительного заполнения и MICE для уточнения восстановленных данных позволяет дополнительно повысить точность восстановления пропусков, особенно на участках с выбросами и магнитными бурями. Следует отметить, что предложенный подход наиболее эффективен для плотных сетей геомагнитного мониторинга с устойчивыми пространственно-временными связями и требует дополнительной валидации при применении к разреженным сетям.

Разработанная методика была адаптирована для оценки магнитных возмущений, создаваемых на ближайших к питающему диполю станциях Таш-Башат и Чункурчак во время работы установки ЭРГУ-600. Предложенный подход включает замену участков данных, полученных во время электромагнитных зондирований, на пропуски с последующим их восстановлением с использованием алгоритма kNN + MICE. Это позволяет снизить влияние локальных особенностей станций и обеспечить более точное выделение индуцированных магнитных возмущений. Предлагаемая методика восстановления пропусков на основе комбинации алгоритмов kNN и MICE также может быть использована для удаления выбросов во временных рядах вариации геомагнитного поля. Этот подход позволяет сохранить структуру временного ряда, минимизируя при этом потери информации при удалении выбросов.

Использование интерпретируемых методов (kNN и MICE) обеспечивает прозрач-

ность процедуры восстановления и позволяет анализировать влияние исходных данных на результат. Дальнейшее развитие исследований предполагает применение предложенных алгоритмов для выявления аномалий в геомагнитном поле, потенциально связанных с подготовкой и развитием сейсмических процессов на территории Бишкекского геодинамического полигона.

Список литературы

1. Schneider T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*. 2001,14(5):853-871. https://doi.org/10.1175/1520-0442(2001)014<0853:aoicde>2.0.co;2
2. Little R.J.A., Rubin D.B. Statistical analysis with missing data. Third ed. Hoboken, NJ: Wiley, 2020, 449 p. https://doi.org/10.1002/9781119482260
3. Jadhav A., Pramod D., Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*. 2019,33(10):913-933. https://doi.org/10.1080/08839514.2019.1637138
4. Love J.J. Missing data and the accuracy of magnetic observatory hour means. *Annals of Geophysics*. 2001,27:3601-3610. https://doi.org/10.5194/angeo-27-3601-2009
5. Воробьева Г.Р. Подход к восстановлению геомагнитных данных путем сопоставления суточных фрагментов временного ряда с равной геомагнитной активностью. *Компьютерная оптика.* 2019,43(6):1053-1063. https://doi.org/10.18287/2412-6179-2019-43-6-1053-1063
6. Richman M.B., Trafalis T.B., Adrianto I. Missing data imputation through machine learning algorithms. In: Haupt S.E., Pasini A., Marzban C. (eds) *Artificial intelligence methods in the environmental sciences*. Dordrecht: Springer, 2009. https://doi.org/10.1007/978-1-4020-9119-3_7
7. Abidin N.Z., Ismail A.R., Emran N.A. Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications*. 2018,9(6):442-447.
8. Бархатов Н.А., Левитин А.Е., Сахаров С.Ю. Метод искусственных нейронных сетей как способ восстановления пробелов в записях отдельных магнитных обсерваторий по данным других станций. *Геомагнетизм и аэрономия.* 2002,42(2):195-198.
9. Имашев С.А., Паров С.В. Модифицированное сезонное разложение вариаций модуля индукции магнитного поля Земли. *Информационные технологии.* 2024,30(2):59-67. https://doi.org/10.17587/it.30.59-67
10. Мухамадеева В.А., Воронцова Е.В., Лазарева Е.А. Опыт проведения геомагнитных наблюдений на Бишкекском геодинамическом полигоне. *Вестник Кыргызско-Российского Славянского университета*. 2015,15(3):130-133.
11. Имашев С.А., Лазарева Е.А. Пространственное распределение составляющих главного геомагнитного поля на основе модели IGRF-13 для территории Кыргызстана. *Вестник Кыргызско-Российского Славянского университета.* 2022,22(4):192-198. https://doi.org/10.36979/1694-500X-2022-22-4-192-198
12. Имашев С.А., Рыбин А.К. Сейсмические и геоакустические отклики земной коры на зондирования мощными электрическими импульсами на территории Бишкекского геодинамического полигона. *Наука и технологические разработки.* 2023,102(2-3):63-88. https://doi.org/10.21455/std2023.2-3-3
13. Imashev S.A. Extended isolation forest – Application to outlier detection in geomagnetic data. *IOP Conference Series: Earth and Environmental Science*. 2021,012022. https://doi.org/10.1088/1755-1315/929/1/012022
14. Имашев С.А., Лазарева Е.А. Удаление выбросов во временных рядах геомагнитного поля на основе фильтра Хампеля. *Информационные технологии.* 2025,31(4):191-198. https://doi.org/10.17587/it.31.191-198
15. Campbell W.H. *Introduction to geomagnetic fields.* Cambridge University Press, 2003, 337 p. https://doi.org/10.1017/cbo9781139165136
16. Imashev S.A. Method for detecting anomalies in geomagnetic field variations based on artificial neural network. *Geosystems of Transition Zones*. 2024,8(4):343-356. https://doi.org/10.30730/gtrz.2024.8.4.343-356
17. Beretta L., Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*. 2016,16(S3):74. https://doi.org/10.1186/s12911-016-0318-z
18. Batista G.E.A.P.A., Monard M.C. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*. 2003,17(5–6):519-533. https://doi.org/10.1080/713827181
19. White I.R., Royston P., Wood A.M. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. 2011,30(4):377-399. https://doi.org/10.1002/sim.4067
20. Huque M.H., Carlin J.B., Simpson J.A., Lee K.J. A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*. 2018,18:168. https://doi.org/10.1186/s12874-018-0615-6.
21. Stekhoven D.J., Bühlmann P. MissForest – nonparametric missing value imputation for mixed-type data. *Bioinformatics*. 2012,28(1):112-118. https://doi.org/10.1093/bioinformatics/btr597

22. Hassanat A.B., Abbadi M.A., Altarawneh G.A., Alhasanat A.A. Solving the problem of the k parameter in the kNN classifier using an ensemble learning approach. *International Journal of Computer Science and Information Security*. 2014,12(8):33-39. https://doi.org/10.48550/arXiv.1409.0919

23. Лазарева Е.А., Имашев С.А. Вариации полного вектора геомагнитного поля во время пусков электроразведочной генераторной установки (ЭРГУ-600-2). *Современные техника и технологии в научных исследованиях*: сб. материалов XIII Междунар. конф. молодых ученых и студентов. Бишкек, 2021, с. 107-112.

24. Сорокин В.М., Ященко А.К., Новиков В.А., Имашев С.А., Лазарева Е.А. Распространение электромагнитного сигнала в ионосферу от излучающего заземленного диполя электроразведочной генераторной установки ЭРГУ-600-2 (Северный Тянь-Шань). *Динамические процессы в геосферах*. 2025,17(2):41-53. https://doi.org/10.26006/29490995_2025_17_2_41

References

1. Schneider T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*. 2001,14(5):853-871. https://doi.org/10.1175/1520-0442(2001)014<0853:aoicde>2.0.co;2

2. Little R.J.A., Rubin D.B. Statistical analysis with missing data. Third ed. Hoboken, NJ: Wiley, 2020, 449 p. https://doi.org/10.1002/9781119482260

3. Jadhav A., Pramod D., Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*. 2019,33(10):913-933. https://doi.org/10.1080/08839514.2019.1637138

4. Love J.J. Missing data and the accuracy of magnetic observatory hour means. *Annals of Geophysics*. 2001,27:3601-3610. https://doi.org/10.5194/angeo-27-3601-2009

5. Vorobyeva G.R. Approach to the recovery of geomagnetic data by comparing daily fragments of a time series with equal geomagnetic activity. *Computer Optics*. 2019,43(6):1053-1063. (In Russ.). https://doi.org/10.18287/2412-6179-2019-43-6-1053-1063

6. Richman M.B., Trafalis T.B., Adrianto I. Missing data imputation through machine learning algorithms. In: Haupt S.E., Pasini A., Marzban C. (eds) *Artificial intelligence methods in the environmental sciences*. Dordrecht: Springer, 2009. https://doi.org/10.1007/978-1-4020-9119-3_7

7. Abidin N.Z., Ismail A.R., Emran N.A. Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications*. 2018,9(6):442-447

8. Barkhatov N.A., Levitin A.E., Sakharov S.Yu. The method of artificial neuron networks as a procedure for reconstructing gaps in records of individual magnetic observatories from the data of other stations. *Geomagnetism and Aeronomy*. 2002,42(2):184-186.

9. Imashev S.A., Parov S.V. Modified seasonal decomposition variations of earth magnetic field induction module. *Information Technologies*. 2024,30(2):59-67. (In Russ.). https://doi.org/10.17587/it.30.59-67

10. Mukhamadeeva V.A., Vorontsova E.V., Lazareva E.A. Experience of geomagnetic observations at the geodynamic test ground in Bishkek. *Vestnik of KRSU = Herald of KRSU*. 2015,15(3):130-133. (In Russ.).

11. Imashev S.A., Lazareva E.A. Spatial distribution of the main geomagnetic field components based on IGRF-13 model for Kyrgyzstan territory. *Vestnik of KRSU = Herald of KRSU*. 2022,22(4):192-198. (In Russ.). https://doi.org/10.36979/1694-500X-2022-22-4-192-198

12. Imashev S.A., Rybin A.K. Seismic and geoacoustic responses of the earth's crust to sensing with high energy electric pulses at the territory of the Bishkek geodynamic polygon. *Nauka i tekhnologicheskiye razrabotki*. 2023,102(2-3):63-88. (In Russ.). https://doi.org/10.21455/std2023.2-3-3

13. Imashev S.A. Extended isolation forest – Application to outlier detection in geomagnetic data. *IOP Conference Series: Earth and Environmental Science*. 2021,012022. https://doi.org/10.1088/1755-1315/929/1/012022

14. Imashev S.A., Lazareva E.A. Removal of outliers in geomagnetic field time series using the Hampel filter. *Information Technologies*. 2025,31(4):191-198. (In Russ.). https://doi.org/10.17587/it.31.191-198

15. Campbell W.H. *Introduction to geomagnetic fields*. Cambridge University Press, 2003, 337 p. https://doi.org/10.1017/cbo9781139165136

16. Imashev S.A. Method for detecting anomalies in geomagnetic field variations based on artificial neural network. *Geosystems of Transition Zones*. 2024,8(4):343-356. https://doi.org/10.30730/gtrz.2024.8.4.343-356

17. Beretta L., Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*. 2016,16(S3):74. https://doi.org/10.1186/s12911-016-0318-z

18. Batista G.E.A.P.A., Monard M.C. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*. 2003,17(5–6):519-533. https://doi.org/10.1080/713827181

19. White I.R., Royston P., Wood A.M. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. 2011,30(4):377-399. https://doi.org/10.1002/sim.4067

20. Huque M.H., Carlin J.B., Simpson J.A., Lee K.J. A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*. 2018,18:168. https://doi.org/10.1186/s12874-018-0615-6.

21. Stekhoven D.J., Bühlmann P. MissForest – nonparametric missing value imputation for mixed-type data. *Bioinformatics*. 2012,28(1):112-118. https://doi.org/10.1093/bioinformatics/btr597

22. Hassanat A.B., Abbadi M.A., Altarawneh G.A., Alhasanat A.A. Solving the problem of the k parameter in the kNN classifier using an ensemble learning approach. *International Journal of Computer Science and Information Security*. 2014,12(8):33-39. https://doi.org/10.48550/arXiv.1409.0919

23. Lazareva E.A., Imashev S.A. [Variations of the full vector of the geomagnetic field during the launch of the electrical exploration generator unit (ERGU-600-2)]. *Sovremennyye tekhnika i tekhnologii v nauchnykh issledovaniyakh:* sb. materialov XIII Mezhdunar. konf. molodykh uchenykh i studentov. Bishkek, 2021, p. 107-112. (In Russ.).

24. Sorokin V.M., Yaschenko A.K., Novikov A.V., Imashev S.A., Lazareva E.A. Electromagnetic signal propagation into ionosphere from the radiating grounded dipole of the ERGU-600-2 electric prospecting generator facility (Northern Tien-Shan). *Dynamic Processes in Geospheres*. 2025,17(2):41-53. (In Russ.). https://doi.org/10.26006/29490995_2025_17_2_41

Об авторе

Имашев Санжар Абылбекович, кандидат физико-математических наук, ведущий научный сотрудник, Научная станция РАН в г. Бишкеке, Бишкек, Киргизия, sanzhar.imashev@gmail.com. https://orcid.org/0000-0003-3293-3764

About the Author

Imashev, Sanjar A., Cand. Sci. (Phys. and Math.), Leading Researcher, Research Station of the Russian Academy of Sciences in Bishkek, Bishkek city, Kyrgyzstan, sanzhar.imashev@gmail.com. https://orcid.org/0000-0003-3293-3764

Поступила 18.02.2026
Принята к публикации 23.03.2026

Received 18 February 2026
Accepted 23 March 2026