Информационные процессы, Том 19, № 1, 2018, стр. 1–18

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ

Методы резервирования в задачах восстановления временных рядов геомагнитных данных

А.В. Воробьев, Г.Р. Воробьева

ФГБОУ ВО Уфимский государственный авиационный технический университет, Уфа, Россия

Поступила в редколлегию 6.02.2018

Аннотация—В работе предложены два подхода к восстановлению временных рядов геомагнитных данных, в основе которых лежат принципы информационного резервирования технических объектов и технологии машинного обучения. Методы резервной станции и прецедентного резервирования являются соответственно пространственным и временным типами информационного резервирования, при котором необходимая для повышения надежности измерительных систем магнитных станций информационная избыточность обеспечивается магнитной станцией, обладающей наиболее сильной корреляционной связью с анализируемой — в первом случае, и статистическими данными, накопленными самой магнитной обсерваторией, — во втором случае. В статье приведены результаты проведенного эксперимента по восстановлению временного ряда реальных геомагнитных данных с использованием предложенных методов резервирования и показана их эффективность в условиях возмущенной магнитосферы. Так, проведенный эксперимент показал, что, к примеру, метод прецедентного резервирования позволяет в среднем на 79.54 % повысить точность восстановления временного ряда геомагнитных данных в условиях возбужденной магнитосферы по сравнению с известными методами импутации геомагнитных данных.

КЛЮЧЕВЫЕ СЛОВА: временные ряды, геомагнитные данные, восстановление временных рядов, надежность технических систем, информационное резервирование, метод резервной станции, прецедентное резервирование.

1. ВВЕДЕНИЕ

В современном мире сфера функционирования отраслей народного хозяйства вследствие научно-технического прогресса распространилась в область околоземного космического пространства как значимого стратегического фактора обеспечения национальных интересов страны. На сегодняшний день околоземное пространство представляет собой сложную многоуровневую систему с непрерывно меняющимися малоизученными параметрами, динамика которых во многом определяет эволюционные или иные процессы в био- и экосфере. Экологический мониторинг этого пространства (внесено в закон Российской Федерации «Об охране окружающей среды» как объект охраны), в свою очередь, представляет собой комплексную проблему на естественно-научном, техническом и правовом уровнях.

Одной из важнейших компонент околоземного космического пространства является магнитосфера Земли, в которой под воздействием естественных и антропогенных факторов могут возникать опасные для био- и техносферы вариации и аномалии. Достоверно известно, что геомагнитные вариации и магнитные аномалии с высокой степенью вероятности могут спровоцировать кратковременные перестройки вегетативно-гуморальной и сердечно-сосудистой систем человека, что чревато серьезными губительными последствиями. Экстремальные геомагнитные явления приводят к полному выходу из строя сетей электропитания, появлению сильных токов в трубопроводах, практически полному прекращению радиосвязи на всех частотах, так

называемому «магнитному торможению» искусственных спутников Земли. Такое воздействие на техносферу наносит зачастую непоправимый экономический ущерб. И это далеко не полный перечень тех последствий, к которым могут привести изменения нормального состояния геомагнитосферы.

Главным способом профилактики неблагоприятного воздействия геомагнитных факторов на био- и техносферу является наблюдение и анализ геомагнитной обстановки. Значимость решения проблемы мониторинга геомагнитных вариаций и аномалий при этом сложно переоценить. С одной стороны, это система наблюдений и оценки текущей геомагнитной обстановки, а с другой — средство информационного обеспечения процесса подготовки и принятия управленческих решений в соответствующей прикладной области (биология, медицина, геофизика, геология, метеорология и пр.).

В настоящее время наиболее распространенным, достоверным и доступным для большинства ученых и специалистов методом наблюдения параметров геомагнитного поля и его вариаций являются наземные высокотехнологичные магнитные станции, объединенные в единую мировую информационную сеть ИНТЕРМАГНЕТ (INTERMAGNET — International Real-Time Magnetic Observatory Network). Глобальная сеть ИНТЕРМАГНЕТ объединяет более 120 постоянных цифровых магнитных станций, измеряющих значения комплекса параметров магнитного поля Земли и передающих в режиме квазиреального времени полученные (отчетные) данные (посекундные и поминутные) в необработанном виде в один из пяти мировых специализированных информационных центров по сбору и распространению геомагнитных данных. Доступ к данным осуществляется по одному из двух протоколов: HTTP для загрузки конечными пользователями и FTP для обращения к ним отдельными программными системами по интерфейсу взаимодействия «программа-программа».

Одной из важнейших задач магнитных станций ИНТЕРМАГНЕТ является обеспечение непрерывности регистрации данных об измеряемых параметрах геомагнитного поля и его вариаций. Однако несовершенство применяемой аппаратуры и задействованных каналов передачи информации обуславливает наличие пропусков во временных рядах зарегистрированных данных, что наряду с пространственной анизотропией создает серьезное препятствие для обработки геомагнитных данных при решении прикладных задач.

Усложняет сложившуюся ситуацию и тот факт, что одной из особенностей временных рядов геомагнитных данных является недетерминированная зависимость характера изменения их уровней от состояния магнитосферы в соответствующий момент времени. Сложность восстановления геомагнитных данных в условиях неспокойной магнитосферы обусловлена возникающими при этом вариациями параметров геомагнитного поля, которые, в свою очередь, приводят к сложным скачкообразным изменениям уровней временного ряда и разрыву линий тренда, нарушению их цикличности и периодичности [1].

В этой связи совершенствование методов и алгоритмов эффективной обработки больших объемов геомагнитных данных, включая способов восстановления пропущенных значений (в том числе в условиях возбужденной магнитосферы), входит в число первостепенных проблем современной геофизики.

2. АНАЛИЗ ИЗВЕСТНЫХ МЕТОДОВ ВОССТАНОВЛЕНИЯ ГЕОМАГНИТНЫХ ДАННЫХ

Для критического анализа известных методов восстановления временных рядов применительно к наборам геомагнитных данных в работе в качестве экспериментального временного ряда авторами использованы геомагнитные данные, зарегистрированные магнитной обсерваторией DOUrbes (50.1180° N, 4.6180° E) 1.01.2017 г. (по данным www.intermagnet.org). Согласно

теории анализа временных рядов, представленный ряд данных является моментным, изолированным, равномерным, неполным и скалярным.

Для оценки точности анализируемых методов восстановления данных в середину экспериментального временного ряда (Рис. 1, *а*) был искусственно введен серийный пропуск длиной в 20 значений (Рис. 1, *б*). Дальнейший анализ эффективности известных методов импутации пропущенных значений временного ряда геомагнитных данных был выполнен на основании оценки величины среднеквадратической ошибки, допущенной при восстановлении синтезированного пропуска.

Для анализа были выбраны наиболее распространенные методы восстановления временных рядов, одни из которых основаны на сглаживании, а другие базируются на прогностическом подходе к анализу данных. Кроме того, была рассмотрена эффективность методов, используемых в настоящее время российскими и зарубежными специалистами для восстановления временных рядов геомагнитных данных, зарегистрированных магнитными станциями сети ИНТЕРМАГНЕТ [2].

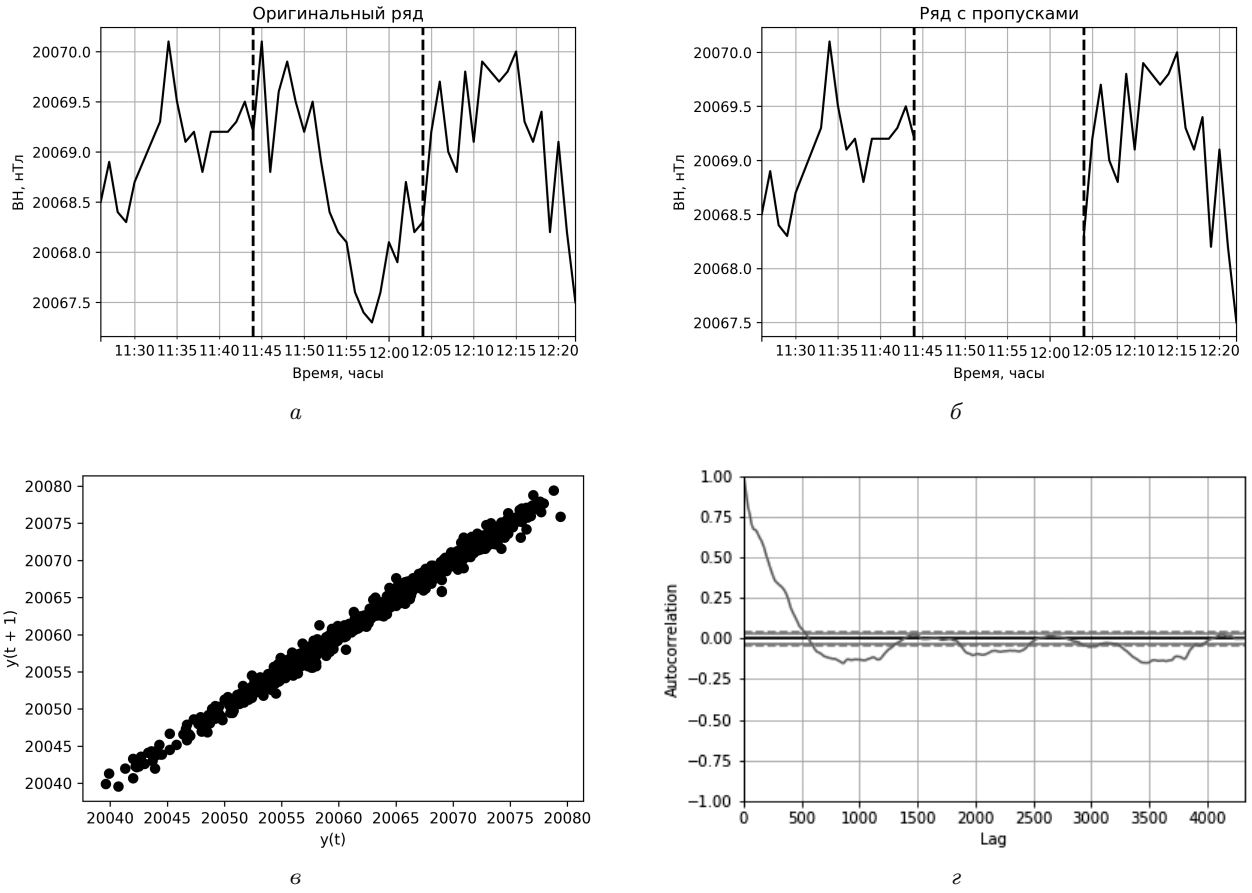

Рис. 1: Диаграмма временного распределения геомагнитных данных, зарегистрированных магнитной станцией DOUrbes 3 ноября 2017 года (*а*, *б*) и соответствующая ему коррелограмма уровней ряда (*в*, *г*)

Так, на сегодняшний день задача заполнения пропусков геомагнитных данных решается сетью ИНТЕРМАГНЕТ простейшим, но малоэффективным ad-hock методом — заменой пропусков на зарезервированные значения [2]. Так, стандарт IAGA-2002 [3], [4] определяет последовательности «99999.00» и «88888.00» в качестве индикатора отсутствующего значения

параметра геомагнитного поля, что при отсутствии предварительной обработки данных может существенно исказить результаты их интерпретации и анализа [5], [6].

Другой известный и широко практикуемый в России и за рубежом подход основан на линейной интерполяции временных рядов геомагнитных данных, содержащих пропущенные значения [7]. Суть метода состоит в том, что крайние точки пропущенного фрагмента временного ряда соединяются друг с другом прямой линией, т. е. составляется полином первой степени, поиск коэффициентов которого и выполняется в ходе интерполяции [8]. Формально метод может быть выражен следующим образом [9]:

$$\frac{(y_2 - y_1)}{(x_2 - x_1)} = \frac{(y - y_1)}{(x - x_1)};$$
$$x_1 \le x \le x_2; \quad y = ax + b;$$
$$a = \frac{(y_2 - y_1)}{(x_2 - x_1)}; \quad b = y_1 - ax_1,$$

где x_1 и y_1 — первая крайняя точка пропуска и значение уровня в ней, x_2 и y_2 — вторая крайняя точка пропуска и значение уровня в ней, x и y — пропущенная точка и значение уровня в ней, a, b — коэффициенты построенной прямой.

На Рис. 2 приведена схема применения метода линейной интерполяции для восстановления временного ряда геомагнитных данных. Верхний график показывает временное распределение уровней восстанавливаемого временного ряда, где S_0 — отсутствующий фрагмент временного ряда, S_- и S_+ — соответственно предшествующий и следующий за ним фрагменты оригинального ряда той же размерности. Нижний график иллюстрирует временной ряд геомагнитных данных, уже восстановленный методом линейной интерполяции, где S'_0 — восстановленный фрагмент временного ряда, S'_- и S'_+ — соответственно предшествующий и следующий за ним фрагменты восстановленного ряда той же размерности.

Метод линейной интерполяции обеспечивает наибольшую эффективность при восстановлении единичных пропусков, поскольку интервал, в который попадает искомое значение, в данном случае минимален. Увеличение длины пропущенного фрагмента приводит к пропорциональному увеличению значения среднеквадратической ошибки. Так, применение указанного метода для восстановления пропуска в экспериментальном временном ряду характеризуется среднеквадратической ошибкой величиной 0.509 нТл, что существенно выше допустимой для геомагнитных данных по стандарту IAGA-2002 погрешности измерений (0.1 нТл) [2] (Рис. 3, *б*).

Из теории анализ временных рядов [10] известно, что одним из распространенных подходов к их восстановлению является применение методов сглаживания ряда. Так, в работе авторами бы рассмотрен один из методов сглаживания — упрощенный метод скользящей средней. Данный подход реализован таким образом, что ширина окна (фрагмента временного ряда) N фиксирована и равна 3. При этом пропущенное значение ряда рассчитывается как среднее арифметическое предшествующего и последующего замеров:

$$x_i = x_{i-1} + x_{i+1}/2, \quad i = 1, ..., N,$$

где x_i — восстанавливаемое значение; x_{i-1} и x_{i+1} — предшествующее и последующее значения уровня временного ряда соответственно [11].

Поскольку характер изменения регистрируемого информационного сигнала исключает скачкообразные вариации (что в первую очередь обуславливается природой их происхождения), то данный метод требует минимальных затрат машинного времени, что немаловажно для обеспечения оперативности обработки временного ряда геомагнитных данных.

При этом следует оговорить ограничения, накладываемые на число и характер распределения пропущенных значений. Идеализированный вариант использования упрощенного метода скользящей средней предполагает единственное пропущенное значение между двумя известными геомагнитными измерениями. В действительности такая ситуация складывается крайне редко и реальные геомагнитные данные сопровождаются целой серией пропущенных значений, следующих во временном ряду последовательно друг за другом. В этом случае алгоритм предусматривает циклический поиск первого значимого замера (отличного от выброса / пропуска) и его подстановку в выражение расчета среднего арифметического значения. Очевидно, что чем дальше в ряду находится данное значение от восстанавливаемого, тем больше величина среднеквадратического отклонения, возникающего при импутации временного ряда. Поэтому предпочтительно применение метода скользящей средней для восстановления единичных значений с симметричными предшествующим и последующим фрагментами временного ряда. В остальных случаях величина ошибки критически возрастает, что делает применение указанного метода нецелесообразным.

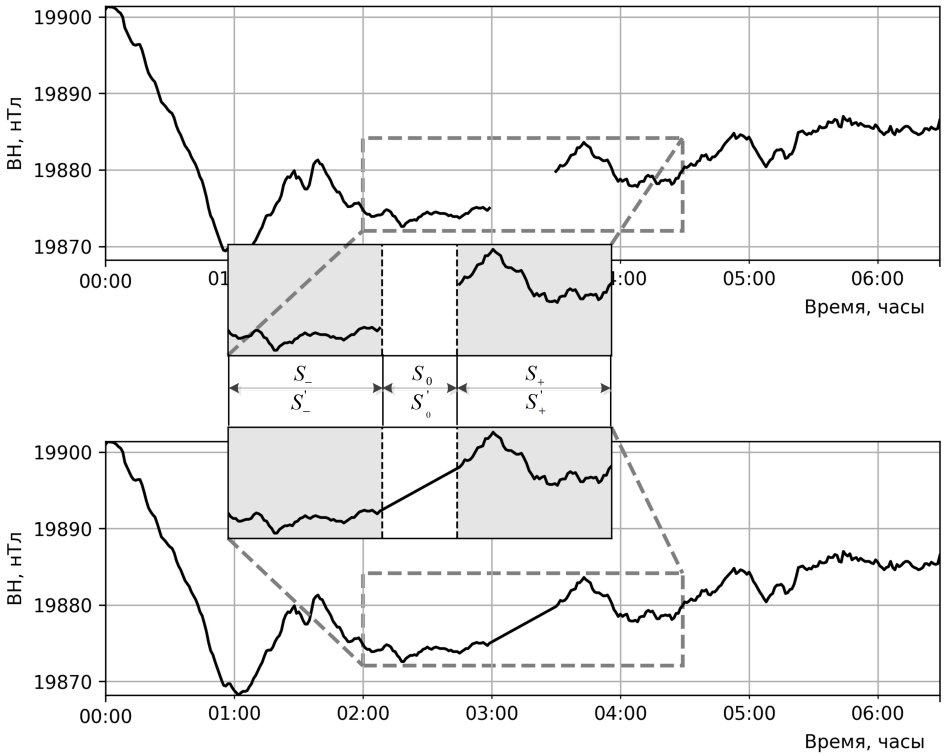

Рис. 2: Схема применения метода линейной интерполяции для восстановления временных рядов геомагнитных данных

Так, к примеру, восстановление пропущенных значений экспериментального временного ряда данным методом (Рис. 3, *б*) сопровождается среднеквадратической ошибкой величиной 1.049 нТл, что также существенно превышает допустимый уровень отклонения значений геомагнитных данных.

Одной из отличительных особенностей временного ряда геомагнитных данных является наличие тренда и сезонной / циклической составляющей, что обусловлено, в первую очередь, природой рассматриваемого геофизического явления. Так, на Рис. 1, *в*, *г* приведены корреллограммы данных экспериментального временного ряда, иллюстрирующие корреляционную

зависимость между его последовательными уровнями, известную как автокорреляция уровней ряда.

Из графика автокорреляции (Рис. 1, *г*) видно, что коэффициент корреляции лагов исследуемого ряда отличен от нуля на протяжении всего временного интервала, что свидетельствует о прогнозируемости временного ряда на основе параметрических прогностических моделей и методов. Эта особенность была использована авторами для исследования эффективности восстановления временных рядов геомагнитных данных методами прогнозирования временных рядов.

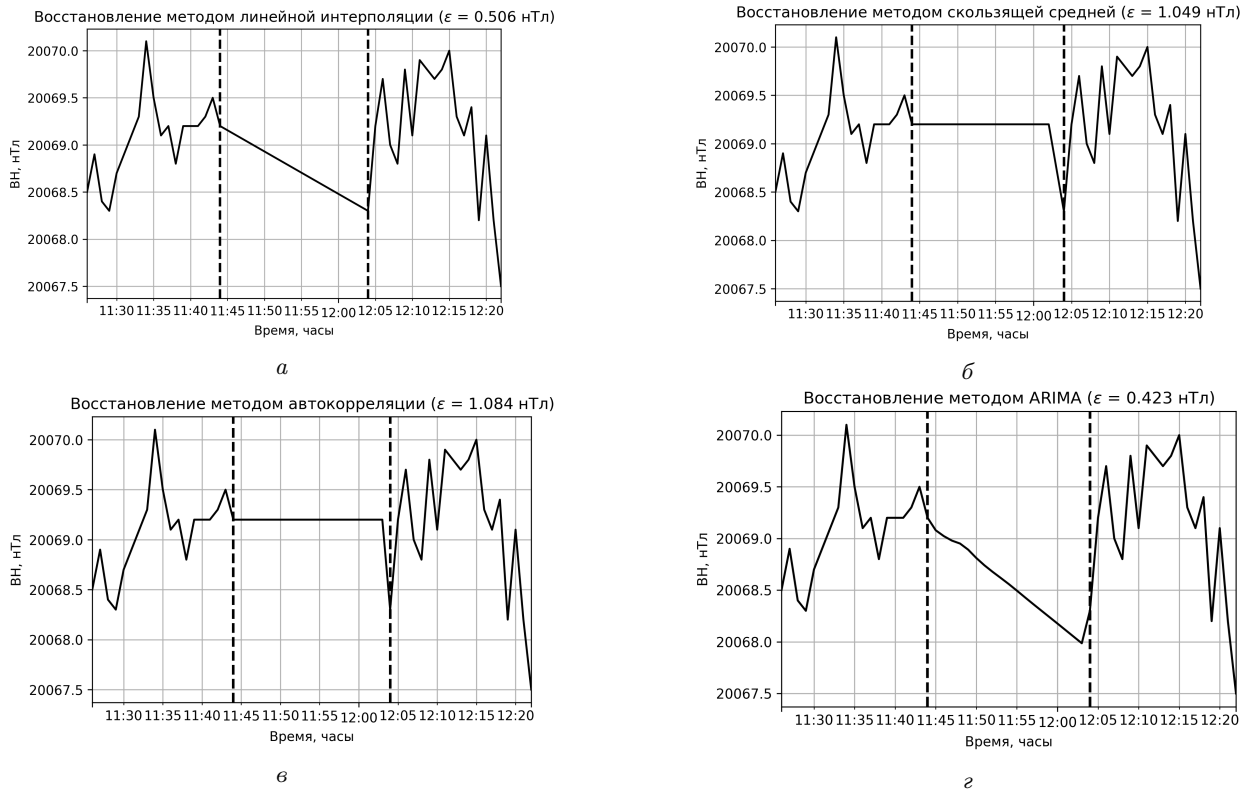

Рис. 3: Результаты восстановления экспериментального временного ряда известными методами импутации пропусков данных

Суть предлагаемого прогностического подхода заключается в том, что известные уровни ряда, предшествующие пропущенному фрагменту, выступают в роли обучающей выборки, на основании которой выполняется прогноз отсутствующего фрагмента искомой длины. Прогностические методы используют в своей основе модель авторегрессии, которую в общем виде можно представить как:

$$Y_i = a_0 + \sum a_i Y_{i-1} + \epsilon_i,$$

где Y_i — целевая переменной (атомарное восстанавливаемое значение); a_0 — коэффициент, описывающий поведение модели при отсутствии внешних факторов, a_i — коэффициенты, описывающие влияние на поведение модели i внешних факторов; Y_{i-1} — прежние значения целевой переменной; ϵ_i — погрешность модели [12], [13].

В рассматриваемом случае экспериментального временного ряда геомагнитных данных имеет место модель авторегрессии первого порядка, что позволяет оценивать изменение целевой переменной в зависимости от единственного фактора — ее собственного значения в про-

шлом периоде авторегрессии. Выбор такой модели авторегрессии обусловлен анализом корреллограммы (Рис. 1, *г*), результат которого свидетельствует о том, что корреляция максимальна между двумя соседними значениями и непрерывно убывает по мере увеличения числа исследуемых лагов.

Та же закономерность прослеживается и при оценке точности прогнозирования пропущенных значений: чем больше восстанавливаемый фрагмент, тем большую погрешность обнаруживает метод и модель авторегрессии. В этой связи рассматриваемый метод авторегрессии используется итерационно для восстановления одного значения в пропущенном фрагменте, которое, в свою очередь, становится частью новой обучающей выборки для восстановления последующего элемента пропущенного фрагмента и т. д.

Как и в предыдущих случаях, восстановление пропущенных значений экспериментального временного ряда данным методом (Рис. 3, *в*) также характеризуется недопустимой среднеквадратической ошибкой величиной 1.084 нТл и существенным искажением формы восстановленного информационного сигнала, что негативно сказывается на результатах анализа полученного временного ряда.

Следующий прогностический метод, исследованный применительно к восстановлению геомагнитных данных, — интегрированная модель авторегрессии — скользящего среднего (ARIMA) [14]. Модель характеризуется тремя параметрами: p — порядок авторегрессии, d — порядок интегрирования, q — порядок скользящего среднего [15]:

$$(\Delta^d X_t) = \sum_{i=1}^{n} \phi_i(\Delta^d X_{t-1}) + \epsilon_t + \sum_{j=1}^{q} \Theta_j(\Delta^d \epsilon_{t-j}), \quad \epsilon_t \approx N(0, {\sigma_t}^2),$$

где $\Delta(\cdot)$, $\Theta(\cdot)$ — полиномы степеней p и q, B — лаговый оператор ($B^j X_t = X_{t-1}, B^j \epsilon_{t-j}, j = 0, \pm 1, ...$), d — порядок взятия последовательной разности ($\Delta X_t = X_{t-1} - X_t = (1 - B)X_t, \Delta^2 X_t = \Delta^2 X_{t+1} - \Delta X_t = (1 - B)^2 X_t, ...$)

Итеративное исследование различных комбинаций перечисленных параметров модели ARIMA было выполнено с помощью «сетчатого поиска» и показало, что лучшее значение информационного критерия Акаике (AIC) [16], [17], [18] достигается при $p = 1$, $d = 0$, $q = 1$. Полученный результат еще раз подтверждает тот факт, что наилучшая корреляция наблюдается между двумя соседними значениями экспериментального временного ряда геомагнитных данных. Восстановление пропущенных значений экспериментального временного ряда данным методом (Рис. 3, *г*) характеризуется наименьшей величиной среднеквадратической ошибки по сравнению с ранее рассмотренными методами импутации, составляя 0.423 нТл. Однако допустимая погрешность измерений по-прежнему превышена.

3. ИНФОРМАЦИОННОЕ РЕЗЕРВИРОВАНИЕ МАГНИТНЫХ ОБСЕРВАТОРИЙ

Круглосуточная работа сети магнитных обсерваторий обеспечивает непрерывный поток оперативной информации о значениях параметров геомагнитного поля и его вариаций, необходимой в различных областях науки, при проведении геологических изысканий, в хозяйственной деятельности ряда предприятий, мониторинге космической погоды и др. [1]. Пропуски в данных являются безвозвратными и могут привести к потере особо важной информации о геофизических явлениях, в том числе предшествующих техногенным катастрофам [19].

В общем случае информационно-измерительная система магнитной обсерватории векторного типа рассматривается как три подобных друг другу ортогонально ориентированных магнитометра. Тогда в соответствии с ГОСТ 27.002-89 надежность как комплексное свойство магнитной обсерватории определяется ее способностью выполнять непрерывную регистрацию

параметров геомагнитного поля и его вариаций в заданных режимах и условиях применения [20]. При этом критерием отказа магнитометрической системы, согласно стандарту IAGA-2002, выступает значение «99999.00» регистрируемого магнитной станцией информационного сигнала, что требует решения, обеспечивающего компенсацию в случае такого отказа потерь регистрируемой магнитной станцией информации о состоянии геомагнитного поля.

Одним из возможных путей повышения надежности информационно-измерительных систем является их резервирование путем введения различных типов избыточности. Известные подходы к резервированию — структурное, временное, информационное, функциональное и др. — нацелены на обеспечение нормального функционирования системы после возникновения отказов в ее элементах. В рассматриваемом случае резервирование должно обеспечить непрерывный мониторинг геомагнитных данных в заданной пространственной точке даже в случае отказа информационно-измерительной системы магнитной станции.

Известно, что в технических объектах, где возникновение отказа приводит к потере или искажению обрабатываемой или передаваемой информации, повышение надежности достигается преимущественно посредством информационного резервирования — метода, предусматривающего использование избыточной информации сверх минимально необходимой для выполнения задач [20]. Информационное резервирование является специфическим видом резервирования, используемым в системах связи, управления, измерительных, информационных, вычислительных системах и других системах сбора и обработки информации, в условиях недостаточной надежности носителей информации, невозможности возобновления информации с помощью первичных источников и пр.

Рассматриваемый метод резервирования предполагает введение дополнительной информации для восстановления основной в случае ее потери или искажения, что может быть обеспечено путем дублирования данных на различных устройствах, коррелированности данных измерений физических полей, использования данных, удовлетворяющих инвариантным соотношениям и пр. [21]. Применительно к магнитным обсерваториям информационное резервирование в соответствии с [21] должно быть описано набором следующих стандартных характеристик:

- кратность резервирования как отношение числа единиц резервной и основной информации: ввиду роста объемов геомагнитных данных и связанных с этим растущих требований к аппаратно-программных средствам их хранения число в знаменателе соответствующей дроби должно быть минимальным;
- область использования резервных ресурсов, определяющая масштаб самого процесса резервирования: параметр показывает, затрагивает ли резервирование каждый из трех магнитометров станции по отдельности (поэлементное резервирование) или задействовано в комплексе (общее резервирование);
- дисциплина резервирования, устанавливающая порядок использования избыточных ресурсов, которые введены в систему для реализации различных способов резервирования: в случае наличия разноуровневого информационного резервирования данный параметр определяет последовательность их ввода для обеспечения непрерывности информационного сигнала;
- дисциплина восстановления ресурсов, основным назначением которой является определение моментов начала и завершения восстановления потерянного или искаженного информационного сигнала, регистрируемого магнитной станцией.

В настоящей работе авторами предлагаются два новых подхода к информационному резервированию применительно к повышению надежности магнитных обсерваторий. Первый из предлагаемых подходов использует результаты корреляционного анализа геомагнитных данных, отнесен к категории информационного пространственного резервирования и получил

название метода резервной станции. Другой подход, получивший название прецедентного резервирования, базируется на индуктивном методе машинного обучения и классифицирован как информационное временное резервирование. Ниже рассматриваются особенности каждого из предложенных подходов в контексте перечисленных характеристик информационного резервирования магнитных станций и приводится сравнительный анализ их эффективности для компенсации пропусков во временных рядах зарегистрированных геомагнитных данных.

4. МЕТОД РЕЗЕРВНОЙ СТАНЦИИ

Геомагнитные данные, синхронно регистрируемые магнитными обсерваториями ИНТЕРМАГНЕТ, измеряются с помощью интервальных и количественных шкал, что позволяет использовать коэффициент корреляции Пирсона [10] для анализа тесноты связи между ними. Математическая мера этой корреляции определяется отношением, попарно сравнивающим анализируемые выборки геомагнитных данных:

$$\sigma_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2} \sqrt{\sum (Y - \overline{Y})^2}},$$

где $\overline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$ и $\overline{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t$ — средние значения выборок X и Y соответственно.

Важно отметить, что для корректного расчета данного коэффициента корреляции необходимо, чтобы количество значений в исследуемых переменных X и Y было одинаковым, исследуемые переменные X и Y были распределены нормально и измерены в интервальной шкале или шкале отношений. Таким образом, учитывая названные ограничения относительно размерности и характера рядов данных, в настоящей работе корреляционный анализ геомагнитных данных, синхронно регистрируемых магнитными обсерваториями INTERMAGNET, производился по наборам качественно гомогенных данных, включающих 16 384 значения, что обеспечило взаимопогашение случайных колебаний.

Результаты анализа степени корреляции геомагнитных данных, регистрируемых станцией DOUrbes, с аналогичными данными, регистрируемыми другими 74 доступными на исследуемый период магнитными станциями (Таблица 1), выявил, что коэффициент корреляции Пирсона занимает интервал от долей процента до практически абсолютной идентичности, что зависит преимущественно от взаимной удаленности двух станций (в большей степени — по широте, в меньшей — по долготе). При этом многие магнитные станции сети ИНТЕРМАГНЕТ расположены таким образом, что происходит преимущественное дублирование регистрируемого информационного сигнала с показателем корреляции Пирсона, превышающим 99 % [19].

Полученные результаты положены в основу метода резервной станции, суть которого состоит в том, что данные, зарегистрированные магнитными обсерваториями с сильной корреляционной связью, считаются взаимозаменяемыми. Это означает, что отсутствующие данные временного ряда одной магнитной станции могут быть заменены уровнями аналогичного отрезка временного ряда коррелирующей с ней (резервной) станции.

На Рис. 4 приведена схема применения метода резервной станции для восстановления временного ряда геомагнитных данных. Верхний график показывает временное распределение уровней восстанавливаемого временного ряда, где S_0 — отсутствующий фрагмент временного ряда, S_- и S_+ — соответственно предшествующий и следующий за ним фрагменты оригинального ряда той же размерности. Нижний график иллюстрирует временной ряд геомагнитных данных, зарегистрированных резервной станцией, где S'_0 — соответствующий восстанавливаемому фрагмент временного ряда, S'_- и S'_+ — соответственно предшествующий и следующий за

Таблица 1: Результат расчета абсолютного значения коэффициента корреляции Пирсона значений полного вектора геомагнитного поля, регистрируемых обсерваторией «Dourbes» (DOU), с аналогичными данными других магнитных станций, %

AAA	ABK	AIA	AMS	ARS	BDV	BEL	BLC	BOU	BOX	BRD
10,5	14,4	46,0	44,7	17,9	89,9	67,2	48,5	23,7	4,0	26,2
BRW	BSL	CBB	CLF	CMO	CYG	CZT	DED	DLT	DMC	DOU
25,5	21,1	36,2	91,1	12,5	19,1	61,9	42,6	22,8	6,6	100,0
DRV	DUR	EBR	FCC	FRD	FRN	FUR	GAN	GCK	GUA	GUI
23,3	66,1	64,3	21,1	3,8	38,0	88,3	9,1	80,3	27,7	14,1
HON	HLP	HRB	HRN	HUA	IPM	IQA	IRT	JAI	KHB	KIV
24,5	49,6	79,8	26,1	7,6	7,9	10,0	7,3	19,5	27,4	68,8
KOU	LYC	MAB	MBO	MEA	MGD	NEW	NVS	ORC	OTT	PAF
20,8	12,7	98,0	42,7	35,2	6,2	9,3	33,7	28,3	31,1	1,6
PET	PHU	PPT	RES	SFS	SHU	SIT	SJG	SOD	SON	SPG
7,0	23,1	32,8	23,8	46,4	8,4	35,8	33,8	18,8	25,0	2,8
SPT	STJ	TAM	TEO	TUC	UPS	VAL	VIC	YKC	—	—
52,2	35,9	18,1	29,0	35,4	8,4	79,8	4,5	31,2	—	—

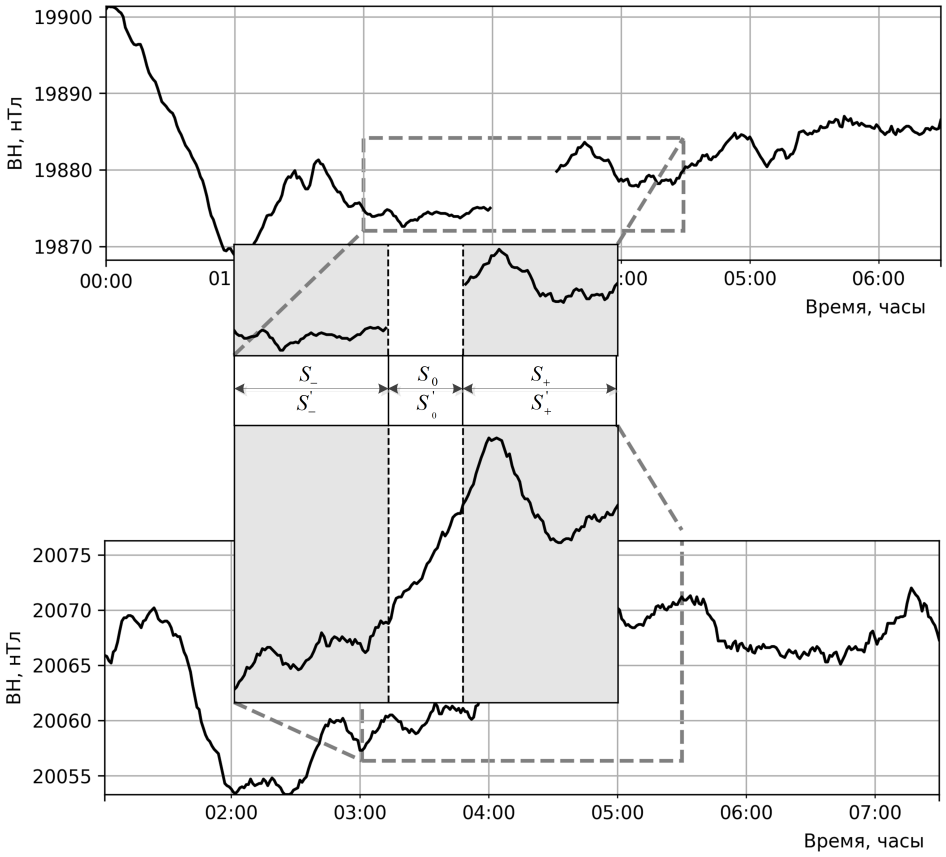

Рис. 4: Схема применения метода резервной станции для восстановления временных рядов геомагнитных данных

ним фрагменты ряда той же размерности. Поскольку в соответствии с предложенным методом фрагменты S_- и S_+ исходного временного ряда и и резервного временного ряда считаются взаимозаменяемыми (при предварительной нормализации данных), то отсутствующий фрагмент S_0 заменяется на фрагмент S'_0 в восстанавливаемом временном ряду геомагнитных данных.

Так, данные, зарегистрированные станцией DOUrbes, наилучшим образом коррелируют с наблюдениями обсерватории MAV (за исследуемый 2017 год). Это позволяет сделать вывод о том, что для восстановления искомого временного ряда геомагнитных данных обсерваторию MAV можно назначить резервной станцией (при условии наличия данных за соответствующий временной интервал).

При этом при наличии во временном ряду последовательно следующих друг за другом пропусков метод резервной станции показывает сопоставимый с другими методами результат. Восстановление дискретных пропусков с помощью данного метода нецелесообразно, поскольку его применение требует длительной и трудоемкой предобработки геомагнитных данных.

Так, в простейшем случае разнонаправленные измерительные устройства могут изменить характер сильной корреляционной зависимости на отрицательный (например, данные магнитных станций DOU и MAV), что требует приведения соответствующих временных рядов к единому виду путем их отражения относительно оси абсцисс и смещения вдоль оси ординат. Замена пропусков исходного временного ряда нормализованными данными резервной станции выполняется посредством сопоставления временных индексов и установления соответствия между ними. Выбранный фрагмент резервного временного ряда копируется под соответствующие временные индексы восстанавливаемого ряда, заменяя в нем обнаруженные пропуски.

5. МЕТОД ПРЕЦЕДЕНТНОГО РЕЗЕРВИРОВАНИЯ

В основе метода прецедентного резервирования лежит концепция индуктивного обучения, заключающаяся в выявлении общих закономерностей по частным эмпирическим данным [6].

Ключевой идеей метода является предположение, что любому фрагменту временного ряда можно с некоторой допустимой степенью точности поставить в соответствие один или несколько фрагментов предшествующих ему значений уровня того же ряда (Рис. 5). В этом случае накопленные магнитной станцией статистические данные выступают в качестве базы прецедентов, где каждый фрагмент временного ряда заданной длины являет собой атомарный прецедент. Тем самым магнитная станция «резервирует» себя собственными ранее выполненными измерениями, которые при определенных ограничениях могут заменить пропущенные сегменты временного ряда.

Обозначим тройку из сегмента отсутствующих значений временного ряда, а также предшествующего и последующего за ним сегментов заданной длины как восстанавливаемую выборку S:

$$S = \{S_-, S_0, S_+\}, S_- = \{s_i\}, i = 1, ..., L,$$
$$S_0 = \{s_j\}, j = L + 1, ..., M, S_+ = \{s_k\}, k = M + 1, ..., N,$$

где S — восстанавливаемая выборка, S_0 — пропущенный фрагмент, S_- — фрагмент, предшествующий пропуску, S_+ — фрагмент, следующий за пропуском.

Предшествующие восстанавливаемой выборке сегменты временного ряда будем называть статистическими и обозначим как S. Тогда паре предшествующего S_- и следующего S_+ за пропуском сегментов временного ряда можно поставить в соответствие пару статистических

сегментов той же длины (S'_- и S'_+):

$$S' = \{s_k\}, k = 1, ..., N;$$
$$S'_- = \{s_i\}, i = 1, ..., |S_-|, S'_- \in S';$$
$$S'_+ = \{s_j\}, j = 1, ..., |S_+|, S'_+ \in S',$$
$$S'_- \to S_-, S'_+ \to S_+,$$

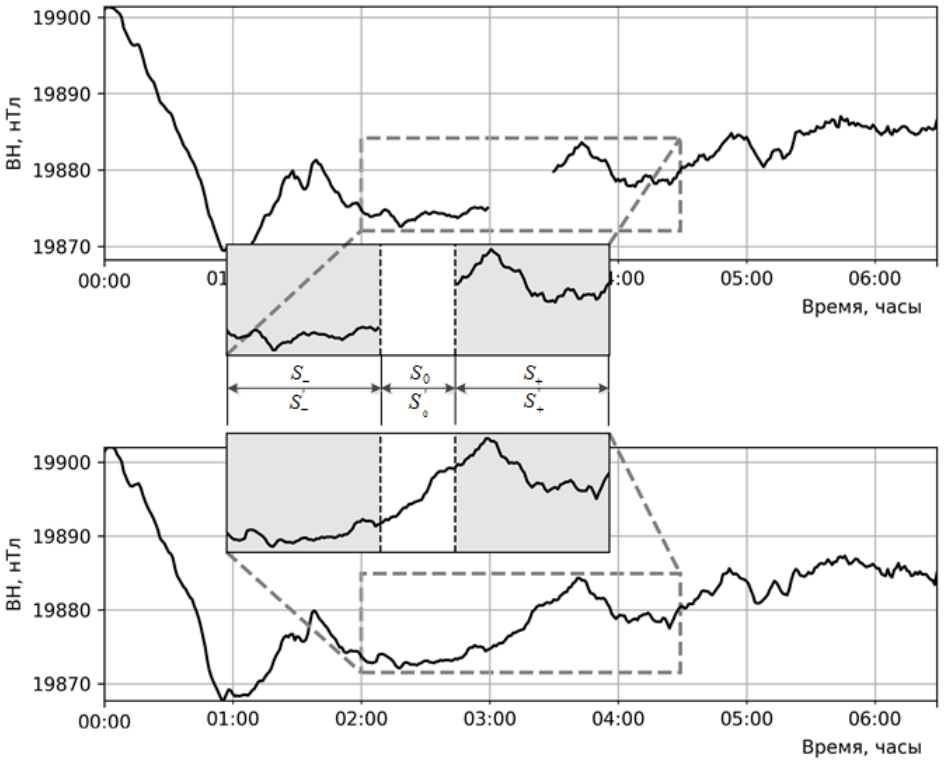

Рис. 5: Схема применения метода прецедентного резервирования для восстановления временных рядов геомагнитных данных

Разделяющие каждую пару сегменты также считаются подобными и взаимозаменяемыми, что позволяет заполнить пропуски соответствующими значениями статистического сегмента временного ряда (с предварительной нормализацией данных):

$$S'_0 = \{s_n\}, n = 1, ..., |S_0|, S'_0 \in S'; S'_0 \to S_0.$$

Мерой соответствия сегментов временного ряда выступает степень их линейной корреляции, что подтверждается выявленной сильной положительной автокорреляцией значений измерений геомагнитных параметров. При этом процедура поиска заменяющего пропуск статистического сегмента выполняется путем последовательного обхода элементов временного ряда в соответствии с жадным алгоритмом перебора значений, где размерность анализируемого фрагмента равна длине восстанавливаемой выборки и обрабатываемые сегменты пересекаются друг с другом.

Обозначим восстанавливаемую выборку с вычтенным из нее сегментом отсутствующих значений как признаковое описание прецедента, а формируемую на каждой итерации перебора

пару статистических сегментов — как обрабатываемую выборку. С учетом введенных обозначений с помощью коэффициента корреляции Пирсона определим степень линейной зависимости между исследуемым сегментом и заданным признаковым описанием:

$$r_{xy} = \frac{\sum\limits_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{m}(x_i - \overline{x})^2}\sqrt{\sum\limits_{i=1}^{m}(y_i - \overline{y})^2}},$$

где x_i — значения обрабатываемой выборки, y_i — значения восстанавливаемой выборки, $\overline{x}$ — среднее арифметическое обрабатываемой выборки; $\overline{y}$ — среднее арифметическое восстанавливаемой выборки.

Абсолютные величины расчетных значений коэффициентов корреляции Пирсона заносятся в предварительно выделенный пул, применяемый для определения наибольшего из значений и, как следствие, соответствующей ему обрабатываемой выборки. Максимальное значение коэффициента корреляции является основанием для подтверждения предположения о соответствии и взаимозаменяемости восстанавливаемой выборки и выделенной тройки статистических сегментов временного ряда.

Для уменьшения уровней шума на завершающем этапе к полученным данным применяется метод медианного сглаживания, главным преимуществом которого является его устойчивость к выбросам. Для заданного медианного интервала временного ряда вычисляется сумма частот значений уровня, рассчитывается половина полученного значения и определяется, какое значение ряда на нее приходится:

$$M = \frac{x_0 + L(0.5\sum f_i - N_{prev}}{f_M)},$$

где — медианное значение, x_0 — начальное значение медианного интервала, L — длина медианного интервала, i — длина временного ряда, $\sum f_i$ — сумма частот временного ряда, f_M — частота медианного интервала, N_{prev} — сумма частот интервалов, предшествующих медианному.

Для предотвращения коллизий начального и конечного значений уровней ряда применительно к временному ряду геомагнитных данных медианный интервал был определен в соответствии с процедурой Тьюки и принят $L = 3$.

В общем виде алгоритм применения метода прецедентного резервирования и его реализация на языке программирования Python приведена на Рис. 6.

Важно отметить, что применение метода прецедентного резервирования в соответствии с принципом бритвы Оккама сопряжено с оценкой минимального объема выборки, достаточного для восстановления данных за период упреждения. Иными словами, требуется определить длину сегментов S_+ и S_- восстанавливаемой выборки при известном числе значений уровня пропущенного сегмента в ней.

Очевидно, что простейшей будет являться модель временного ряда, в которой восстанавливаемая выборка несимметрична и охватывает все актуальные значения уровней, ограничивающие серию пропусков. Программная обработка выборки такой размерности требует значительных затрат вычислительных ресурсов и аппаратного времени, что в условиях восстановления большого числа пропущенных значений временных рядов геомагнитных данных неприемлемо.

Анализ и экспериментальное исследование метода прецедентного резервирования показали, что стратификацию временного ряда следует осуществлять по принципу эквивалентности длины подмножеств: сегменты S_+ и S_- восстанавливаемой выборки выбираются исходя из

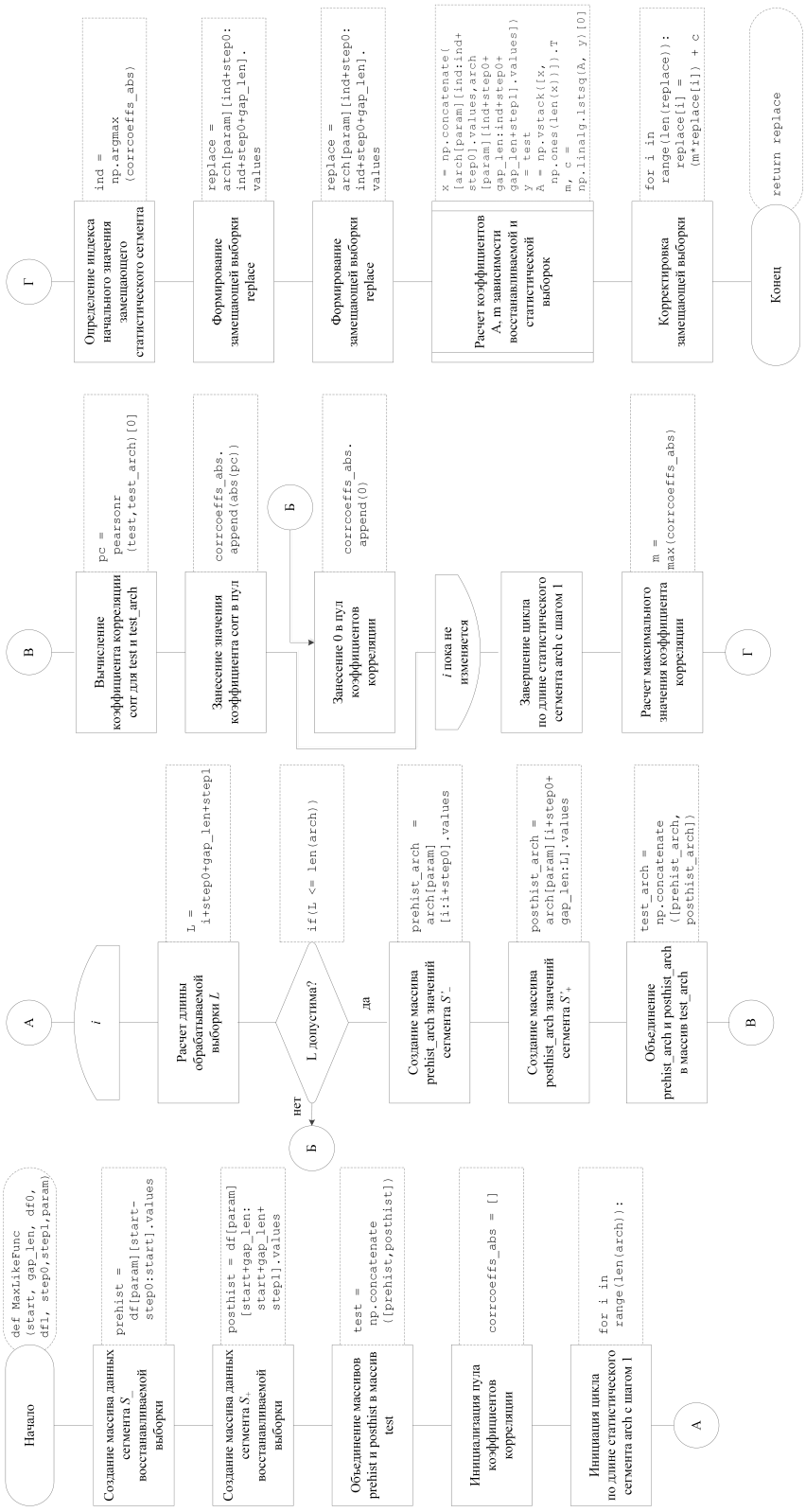

Рис. 6: Алгоритм применения метода прецедентного резервирования для восстановления временных рядов геомагнитных данных и его программная реализация на языке Python

числа значений уровня в периоде упреждения. Так, к примеру, пропуск длиной в 20 значений формирует восстанавливаемую выборку из 60 последовательных значений уровня временного ряда геомагнитных данных со среднеквадратической ошибкой порядка 0.138 нТл.

6. АНАЛИЗ ЭФФЕКТИВНОСТИ МЕТОДОВ РЕЗЕРВИРОВАНИЯ ГЕОМАГНИТНЫХ ДАННЫХ

Для оценки эффективности предложенных методов восстановления данных был использован рассмотренный выше экспериментальный временной ряд геомагнитных данных с искусственно введенной серией пропусков длиной в 20 значений (Рис. 1, *б*).

Анализ показал, что применение метода резервной станции для восстановления пропуска в экспериментальном временном ряду характеризуется среднеквадратической ошибкой величиной 0.836 нТл, что существенно выше допустимой для геомагнитных данных по стандарту IAGA-2002 погрешности измерений (0.1 нТл) и хуже результатов, полученных рассмотренными ранее статистическими и прогностическими методами имптуации (Рис. 7, *а*).

а *б*

Рис. 7: Результаты восстановления экспериментального временного ряда геомагнитных данных методами резервирования

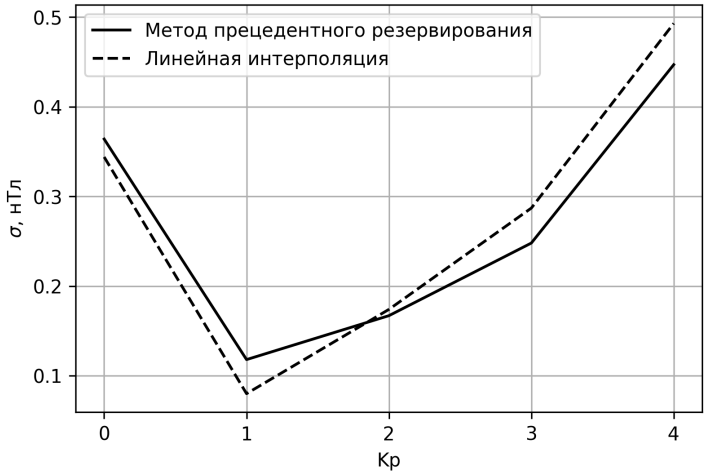

Рис. 8: Сравнительный анализ зависимости величины среднеквадратической ошибки от значения индекса Kp при восстановлении геомагнитных данных методами линейной интерполяции и прецедентного резервирования

Восстановление экспериментального временного ряда геомагнитных данных методом прецедентного резервирования обеспечило среднеквадратическую ошибку импутации величиной 0.138 нТл, что существенно лучше всех рассмотренных ранее методов (в том числе и метода резервной станции) и при грубом приближении укладывается в допустимую стандартом IAGA-2002 погрешность геомагнитных измерений.

Для дополнительной оценки эффективности предложенного авторами метода прецедентного резервирования была проведена серия экспериментов по восстановлению геомагнитных данных, пропущенных при различных значения индекса Kp. Исследования проводились для серий пропусков длиной в 30 значений (минут). Оценка была проведена применительно к двум методам: используемому в настоящее время методу линейной интерполяции и методу прецедентного резервирования. Полученные в ходе эксперимента результаты приведены на Рис. 8.

Как видно из рисунка, в условиях спокойной магнитосферы (при $Kp \leq 2$) метод линейной интерполяции обеспечивает меньшую среднеквадратическую ошибку, чем метод прецедентного резервирования. При приближении Kp к значению 2 величины среднеквадратической ошибки выравниваются, а с достижением индекса значения 2 метод прецедентного резервирования показывает лучший результат по сравнению с методом линейной интерполяции. Следует отметить, что метод прецедентного резервирования обеспечивает приемлемую величину среднеквадратической ошибки только при $Kp \geq 2$ и $Kp \leq 4$. Далее значение отклонения увеличивается в несколько раз и процедура восстановления данных теряет смысл.

7. ЗАКЛЮЧЕНИЕ

Открытый доступ к данным о непрерывных изменениях параметров магнитного поля Земли и объединение наземных высокотехнологичных магнитных станций в единую мировую информационную сеть ИНТЕРМАГНЕТ (INTERMAGNET — International Real-Time Magnetic Observatory Network) объясняют тот факт, что на сегодняшний день именно они являются наиболее распространенным, достоверным и доступным для большинства ученых и специалистов методом наблюдения параметров геомагнитного поля и его вариаций. Однако ввиду несовершенства используемых магнитными станциями аппаратуры и каналов данных исходные временные ряды геомагнитных данных содержат пропуски и «выбросы» за нормальное значение (артефакты), которые являются необратимыми и могут привести к потере важной информации о геофизических явлениях и процессах.

Повышение надежности информационно-измерительных систем магнитных обсерваторий и, в первую очередь, восстановление временных рядов геомагнитных данных является одной из важнейших задач современной геофизики. Вместе с тем известные и широко применяемые в настоящее время способы импутации временных рядов геомагнитных данных не обеспечивают уровня погрешности измерений, регламентированного международным стандартом IAGA-2002. В этой связи в настоящей работе авторами предложены два новых метода восстановления геомагнитных данных, в основу которой положено сочетание элементов теории надежности технических систем и технологий машинного обучения.

Метод резервной станции представляет собой способ пространственного информационного резервирования информационно-измерительных систем магнитных обсерваторий. Суть предложенного метода состоит в том, что сильная корреляционная связь наборов данных, регистрируемых парой наземных магнитных обсерваторий, является основанием для взаимозаменяемости соответствующих временных рядов. Так, пропущенные значения во временном ряду геомагнитных данных, зарегистрированных магнитной обсерваторией, могут быть заменены соответствующим фрагментом данных, зарегистрированных в то же время другой обсерваторией, признанной в соответствии с описанным выше принципом резервной для текущей станции. Метод наиболее эффективен при восстановлении серий пропусков временных рядов

в условиях неспокойной магнитной обстановки, что обусловлено возникающими при этом вариациями параметров геомагнитного поля, которые, в свою очередь, приводят к сложным скачкообразным изменениям уровней временного ряда и разрыву линий тренда, нарушению их цикличности и периодичности. При таких условиях известные методы импутации (в том числе, прогностические, показывающие наилучший по степени точности результат) обеспечивают низкую эффективность восстановления данных.

Другой предложенный в работе метод классифицирован как способ временного информационного резервирования, отличительной особенностью которого является то, что резервирующим устройством выступает сама магнитная станция. Метод, получивший название прецедентного резервирования, базируется на индуктивном методе обучения по прецедентам и отличается тем, что в качестве признаков прецедентов используются данные, предшествующие и последующие за пропуском во временном ряду. Как и в предыдущем случае, метод наиболее эффективен в условиях неспокойной магнитной обстановке и вызванном этом сложном характере регистрируемого информационного сигнала. При этом проведенные эксперименты показали, что метод прецедентного резервирования позволяет в среднем на 79.54 % повысить точность восстановления временного ряда геомагнитных данных в условиях возбужденной магнитосферы по сравнению с известными методами импутации геомагнитных данных. Метод особенно актуален для восстановления серийных пропусков в условиях отсутствия сильной корреляционной связи выбранной магнитной обсерватории с другими станциями в сети.

СПИСОК ЛИТЕРАТУРЫ

1. Мандрикова О. В., Жижикина Е. А. Автоматический способ оценки состояния геомагнитного поля. Компьютерная оптика, 2015, т. 39, № 3, стр. 420–428.

2. INTERMAGNET technical reference manual. Version 4.6 / Ed. by S.-L. Beno?t. Edinburgh : INTERMAGNET, BGS, 2012.

3. Love J .J., Chulliat A. An international network of magnetic observatories. Eos Trans. AGU, 2013, no. 94(42), pp. 373–374.

4. Macmillan S., Olsen N. Observatory data and the Swarm mission. Earth, Planets and Space, 2013, vol. 65, no. 11, pp. 1355–1362.

5. Гвишиани А. Д., Лукьянова Р. Ю. Геоинформатика и наблюдения магнитного поля Земли: российский сегмент. Физика Земли, 2015, № 2, стр. 3–20.

6. Mandea M., Korte M. Geomagnetic Observations and Models. Springer, 2011.

7. Рыбкина А. И. [и др.] Интерполяция данных обсерваторских измерений и визуализация полной напряженности магнитного поля Земли. Вестник Отделения наук о Земле РАН, 2013, т. 5, № 3002, стр. 1—4.

8. Gvishiani A. [et al.] Survey of geomagnetic observations made in the northern sector of Russia and new methods for analysing them. Surveys in Geophysics, 2014, vol. 35(5), pp. 1123–1154.

9. Soloviev A. [et al.] Mathematical tools for geomagnetic data monitoring and the INTERMAGNET Russian segment. Data Science Journal, 2013, vol. 12, pp. WDS114–WDS119.

10. Лившиц М. Случайные процессы — от теории к практике. М.: Лань, 2016.

11. Aue A., Dubart D., H?rmann S. On the prediction of stationary functional time series. Journal of the American Statistical Association, 2015, no. 110(509), pp. 378–392.

12. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003.

13. Vorobev A. V., Vorobeva G. R. Web-oriented 2D/3D-visualization of geomagnetic field and its variations parameters. Scientific visualization, 2017, vol. 9, no.2, pp. 94–101.

14. Gooijer J. Elements of nonlinear time series analysis and forecasting. Cham, Switzerland: Springer, 2017.

15. Jain E., Mallick D. A Study of time series models ARIMA and ETS. International Journal of Modern Education and Computer Science (IJMECS), 2017, vol. 9, no.4, pp. 57–63.

16. Pfaff B. Analysis of integrated and cointegrated time series with R. New York : Springer, 2008.

17. Чучуева И. А. Модель прогнозирования временных рядов по выборке максимального подобия. Автореферат кандидатской диссертации. Московский государственный технический университет им. Н.Э. Баумана, 2012.

18. Box G. [et al.] Time series analysis: forecasting and control. New York: John Wiley & Sons, 2017.

19. Воробьев А.В., Воробьева Г.Р. Метеоинформатика. Геомагнитные вариации и космическая погода: монография. М.: Инновационное машиностроение, 2017.

20. ГОСТ 27.002-89. Надежность в технике. Основные понятия. Термины и определения. М.: Издательство стандартов, 1989.

21. Р 50-54-82-88. Рекомендации. Надежность в технике. Выбор способов и методов резервирования. М: Издательство стандартов, 1988.

Redundancy Methods in Solutions of Geomagnetic Data Time Series Recovery Problem

Vorobev A. V., Vorobeva G. R.

The authors suggest the two approaches to restoration of time series of geomagnetic data, which are based on the principles of information reservation of technical objects and technology of machine training. The methods of the standby station and precedent reservation are respectively spatial and temporal types of information reservation, in which the information redundancy necessary for increasing the reliability of the measuring systems of magnetic stations is provided by a magnetic station having the strongest correlation with the analyzed, in the first case, and statistical data accumulated by the magnetic observatory itself , in the second case. The article presents the results of an experiment to reconstruct a time series of real geomagnetic data using the proposed redundancy methods and shows their effectiveness in a disturbed magnetosphere. Thus, the experiment demonstrated that, for example, the precedent redundancy method allows an average of 79.54% to improve the accuracy of the recovery of the time series of geomagnetic data under conditions of an excited magnetosphere in comparison with known methods of geomagnetic data imputation.

KEYWORDS: time series, geomagnetic data, recovery of time series, reliability of technical systems, information backup, standby station method, precedent redundancy.