ЧИСЛЕННЫЕ МЕТОДЫ И АНАЛИЗ ДАННЫХ

Подход к восстановлению геомагнитных данных путём сопоставления суточных фрагментов временного ряда с равной геомагнитной активностью

Г.Р. Воробьёва [1]

[1] *ФГБОУ ВО Уфимский государственный авиационный технический университет, Уфа, Россия*

Аннотация

Мониторинг параметров геомагнитного поля и его вариаций осуществляется главным образом с помощью наземных магнитных обсерваторий и вариационных станций. Однако несовершенство применяемой аппаратуры и задействованных каналов связи обусловливает наличие пропусков во временных рядах геомагнитных данных, что, наряду с пространственной анизотропией источников данных, создаёт значимые препятствия на пути их автоматизированной обработки. При этом известные методы импутации пропусков во временных рядах обеспечивают среднеквадратическую ошибку восстановления, существенно превышающую уровень, принятый для такого рода геофизических наблюдений. В настоящей работе предложен метод восстановления геомагнитных данных, основанный на статистических методах обработки временных рядов и принципах машинного обучения с использованием размеченных данных. Его отличительной особенностью является то, что признаковым описанием фрагмента временного ряда выступает пара предшествующего и следующего за ним фрагментов того же ряда, в совокупности образующих обучающую выборку для поиска недостающего фрагмента по набору его признаков с последующим линейным масштабированием для восстановления исходного тренда информационного сигнала. Приводятся оценки параметров временных рядов геомагнитных данных, при которых возможно применение предложенного метода для восстановления как суточных вариаций, так и фрагментов длительностью в несколько минут.

Ключевые слова: восстановление временных рядов, обработка временных рядов, геомагнитные данные, машинное обучение, статистический анализ.

Цитирование: **Воробьёва, Г.Р.** Подход к восстановлению геомагнитных данных путём сопоставления суточных фрагментов временного ряда с равной геомагнитной активностью / А.В. Воробьёв, Г.Р. Воробьёва // Компьютерная оптика. – 2019. – Т. 43, № 6. – С. 1053-1063. – DOI: 10.18287/2412-6179-2019-43-6-1053-1063.

Введение

Интенсивное развитие систем и технологий регистрации параметров магнитного поля Земли способствует экспоненциальному росту объёмов геомагнитных данных, основным источником которых выступают магнитные обсерватории и вариационные станции. Однако несовершенство применяемой аппаратуры и задействованных каналов передачи информации обуславливает наличие пропусков во временных рядах зарегистрированных данных, что, наряду с пространственной анизотропией, создаёт серьезное препятствие для обработки геомагнитных данных при решении прикладных задач [1–3]. При этом допустимый порог в 10 % пропусков во временных рядах, принятый Международной ассоциацией по геомагнетизму и аэрономии (*International Association of Geomagnetism and Aeronomy*, IAGA) [4], к настоящему моменту уже превышен.

Распространённым решением задачи восстановления временных рядов геомагнитных данных является замена отсутствующих данных зарезервированным значением. Так, к примеру, спецификация формата IAGA-2002 представления геомагнитных данных предусматривает замену пропущенных данных значениями «88888.88» или «99999.99» [5], что решает проблему обеспечения целостности временного ряда, но приводит к аномалиям и выбросам, отражаясь на достоверности результатов обработки данных.

В некоторых случаях задача заполнения пропусков во временных рядах геомагнитных данных решается методом линейной интерполяции [6]. Однако, демонстрируя достаточно высокую точность на единичных пропусках, метод линейной интерполяции обеспечивает значительную величину среднеквадратической ошибки при восстановлении фрагментов большей длины.

Аппаратное решение задачи поддержания целостности временного ряда геомагнитных данных основано на введении информационного резервирования, предполагающего регистрацию параметров геомагнитного поля одновременно двумя и более устройствами [7]. Однако такой подход связан с дополнительными финансовыми затратами на закупку и сопровождение оборудования, что не всегда осуществимо.

Таким образом, ни один из известных методов не решает выявленную проблему неполноты геомагнитных данных в достаточной мере. Усложняет ситуа-

цию и тот факт, что одной из особенностей результатов наблюдений за параметрами геомагнитного поля 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. В этой связи возникает актуальная задача разработки и реализации метода восстановления временных рядов геомагнитных данных в пределах заданного значения погрешности и для любых наблюдений параметров магнитного поля Земли и его вариаций (в том числе в условиях возбуждённой магнитосферы).

На сегодняшний день активное развитие алгоритмов и моделей машинного обучения в области обработки больших данных нашло свое отражение и в восстановлении временных рядов (например, [8 – 12]). В настоящей работе элементы подхода, основанного на методах машинного обучения, применены к восстановлению пропусков в результатах измерений параметров геомагнитного поля и его вариаций.

1. Анализ методов восстановления геомагнитных данных

Распространённой метрикой качества методов восстановления временных рядов является значение среднеквадратической ошибки, вычисляемое как частное от суммы квадратов разницы восстановленных и фактических значений и длины временного ряда. Величина такой ошибки не должна превышать некоторого принятого значения, что для геомагнитных данных, согласно рекомендациям международной ассоциации IAGA, составляет не больше 1 нТл [5].

Статистические методы восстановления временных рядов базируются на аналитической обработке известных и формировании новых значений массива данных и отличаются используемыми ими математическими моделями, связывающими непоследовательные фрагменты уровней ряда.

Простейшая модель восстановления данных основана на сглаживании временного ряда методом скользящей средней [13] и предполагает замену пропущенного элемента данных средним значением соседних ему элементов. Исследования показали, что метод обеспечивает хорошую метрику качества в случае единичных пропусков, но с увеличением числа пропущенных значений наблюдается экспоненциальный рост среднеквадратической ошибки. Так, к примеру, восстановление 10-минутного фрагмента временного ряда геомагнитных данных методом скользящей средней обеспечивает среднеквадратическую ошибку величиной 1,3 нТл, что выше допустимой погрешности измерений.

Метод линейной интерполяции основан на подборе коэффициентов полинома первой степени, который задан уравнением прямой, соединяющей известные значения уровней ряда [14]. Линейная интерполяция широко используется для восстановления геомагнитных данных, обеспечивая приемлемую величину среднеквадратической ошибки на небольших фрагментах временного ряда и стремясь к нулю на единичных пропусках. Исследования показали, что восстановление 10-минутного фрагмента временного ряда методом линейной интерполяции обеспечивает величину ошибки в среднем до 0,5 нТл.

Обнаруженная в ходе анализа геомагнитных данных высокая степень автокорреляции свидетельствует о тесной линейной осциллирующей связи соседних уровней временного ряда, что позволяет применять к ним прогностические модели и методы. Тогда задача восстановления временного ряда сводится к прогнозированию недостающего фрагмента на основании предшествующих ему. Так, для восстановления фрагмента временного ряда может быть использована модель авторегрессии (AR) первого порядка, которая позволяет оценивать изменение целевой переменной в зависимости от единственного фактора – её собственного значения в прошлом периоде авторегрессии [15]. Исследования показали, что в AR хорошая метрика качества обеспечивается при небольших пропусках (не более 5 значений) и достигает порядка 0,3 нТл. При восстановлении фрагментов большего размера значение среднеквадратической ошибки экспоненциально возрастает и, к примеру, для 10-минутного фрагмента составляет 0,72 нТл.

Интегрированная модель авторегрессии – скользящего среднего (ARIMA) обеспечивает лучшую по сравнению с AR метрику качества прогнозирования пропущенного фрагмента за счёт гибкой параметризации обработки данных [15]. Исследования показали, что наилучший результат модель показывает при первом порядке авторегрессии с нулевым порядком интегрирования и единичным порядком скользящего среднего. Тогда ошибка восстановления небольших фрагментов (порядка 5 – 7 значений) невелика и составляет около 0,2 нТл, а для более значительных пропусков увеличивается, достигая, к примеру, при обработке 10-минутного фрагмента величины 0,63 нТл.

В работе [16] предложен индуктивный метод восстановления геомагнитных данных, основанный на принципах машинного обучения с использованием размеченных данных. Отличительной особенностью метода является то, что признаковым описанием фрагмента временного ряда выступает пара предшествующего и следующего за ним фрагментов того же ряда, в совокупности образующих обучающую выборку для поиска недостающего фрагмента по набору его признаков с последующим линейным масштабированием для восстановления исходного тренда информационного сигнала. Исследования показали, что метод обеспечивает лучшую по сравнению с другими методами восстановления метрику качества на фрагментах временного ряда длительностью порядка 30 минут. Так, к примеру, 10-минутный фрагмент восстанавливается в среднем с ошибкой 0,07 нТл.

Однако серьёзным недостатком метода является низкая вычислительная скорость программного исполнения, обусловленная большим объёмом анализируемых данных, поскольку обучающая выборка формируется из годовых результатов наблюдений магнитной обсерватории. При этом эмпирически доказано [16], что для небольших фрагментов временных рядов геомагнитных данных может быть подобран хотя бы один схожий фрагмент из архивных данных, зарегистрированных обсерваторией в год, предшествующий году, к которому относятся восстанавливаемые сутки.

Так, к примеру, при таком подходе восстановление 10-минутного фрагмента временного ряда требует в среднем более 2 минут процессорного времени, что существенно превышает общепринятый в теории юзабилити порог отклика приложения в 10 с. При этом размер обучающей выборки составляет в среднем 514 650 объектов (в каждом из 365 суточных файлов может быть выделено 1 410 объектов).

Поскольку индуктивный метод обеспечивает лучшую метрику качества, то целесообразна его оптимизация для оперативного получения наилучшего результата. В работе предлагается повышение вычислительной скорости метода восстановления путём сокращения объёма обучающей выборки за счёт использования временных рядов данных, зарегистрированных магнитной обсерваторией в сутки с магнитной активностью (значения Kp-индекса), идентичной наблюдаемой в исследуемые сутки. При таком подходе восстановление, к примеру, 10-минутного фрагмента временного ряда требует в среднем около 5,3 с процессорного времени при объёме обучающей выборки в 1410 объектов. Установлено, что размер обучающей выборки при этом не влияет на величину среднеквадратической ошибки восстановления временного ряда геомагнитных данных.

Анализ показал, что ежегодно наблюдается от 10 до 200 пар суток с одинаковым изменением значений Kp-индекса, что позволяет сформировать из соответствующих данных обучающую выборку. В случае, если магнитная активность исследуемых суток уникальна, то для минимизации количества экспериментальных данных целесообразно выделить обучающие пары из результатов, полученных магнитной обсерваторией в ходе наблюдений за параметрами магнитного поля Земли в течение года, предшествующего восстанавливаемому.

При этом специфика исследования суточного хода параметров геомагнитного поля позволяет пренебречь минутными вариациями и исключить из восстанавливаемого информационного сигнала низкочастотные составляющие. Анализ показал близость зарегистрированных в сутки с одинаковой геомагнитной активностью временных рядов геомагнитных данных, к которым применён ФНЧ Баттерворта. Выявленный факт может быть использован для восстановления соответствующих временных рядов.

2. Индуктивный метод восстановления геомагнитных данных

Метод восстановления фрагмента суточного временного ряда геомагнитных данных основан на определении массива наиболее вероятных значений путём максимизации значения коэффициента корреляции между массивом, образованным предшествующими и последующими за пропущенным фрагментом значениями, и массивами, построенными по аналогичному принципу из множества известных значений (рис. 1). В качестве последних выступают архивные данные, накопленные магнитной обсерваторией за период, предшествующий восстанавливаемому.

Основная идея метода заключается в предположении, что если пара непоследовательных фрагментов временного ряда, разделённых отсутствующим фрагментом, оказывается близкой к паре фрагментов, разделённых известным фрагментом, то промежуточные между ними (а следовательно, и восстанавливаемые) значения в соответствии с теоремой Такенса [17, 18] будут отличаться статистически незначительно.

Исходными данными выступают временные ряды данных, зарегистрированных магнитной обсерваторией в восстанавливаемый (восстанавливаемые геомагнитные данные на рис. 1) и предшествующие ему периоды (архивные геомагнитные данные на рис. 1). При этом для упрощения обработки длительность периода эмпирически выбрана равной одному году (проведённые ранее исследования [16] показали, что увеличение периода приводит только к существенным вычислительным затратам, но не влияет на эффективность метода восстановления). Кроме того, каждые сутки, соответствующие восстанавливаемому и предшествующим ему периодам, характеризуются параметрами геомагнитной активности, представленной наборами усреднённых значений Kp-индекса, также относящихся к исходным данным предлагаемого метода.

На основании исходных данных на начальном этапе наблюдаемые периоды сравниваются на предмет одинаковой геомагнитной активности в соответствующие сутки (этап 1 на рис. 1). В результате отбирается временной ряд геомагнитных данных, зарегистрированных в сутки прошедшего года, когда геомагнитная активность была идентична значениям Kp-индекса в восстанавливаемые сутки (массивы D2 и D1 для восстанавливаемых и идентичных им по геомагнитной активности суток соответственно, этап 1 на рис. 1).

Далее (этап 2 на рис. 1) для приведения временных рядов к общему виду проводится их центрирование, направленное на удаление из каждого элемента ряда ненулевого среднего значения, которое может затруднить интерпретацию результатов [14]. Результатом выполнения этапа являются выровненные временные ряды геомагнитных данных, один из которых содержит пропущенные значения.

На следующем шаге (этап 3 на рис. 1) для повышения вычислительной скорости метода восстановления и исключения неоднозначности результатов сравнения данных из полученных временных рядов исключается низкочастотная составляющая применением фильтра низких частот (ФНЧ) Баттерворта, об-

ладающего (по сравнению с другими ФНЧ) гладкой амплитудно-частотной характеристикой как в полосе пропускания, так и в полосе задержки [20, 21].

Обработка временных рядов, полученных в результате фильтрации и центрирования исходных данных, выполняется в зависимости от длительности восстанавливаемого фрагмента ряда геомагнитных данных. Так, в случае восстановления временного ряда суточных значений наиболее вероятными значениями для замены выступает фильтрованный ряд геомагнитных данных, зарегистрированных той же магнитной обсерваторией в день, когда магнитная активность была такой же, как в восстанавливаемые сутки. Полученные данные аппроксимируются относительно известных значений восстанавливаемого временного ряда (этап 4 на рис. 1), в результате чего формируется искомый фрагмент геомагнитных данных (этап 5 на рис. 1).

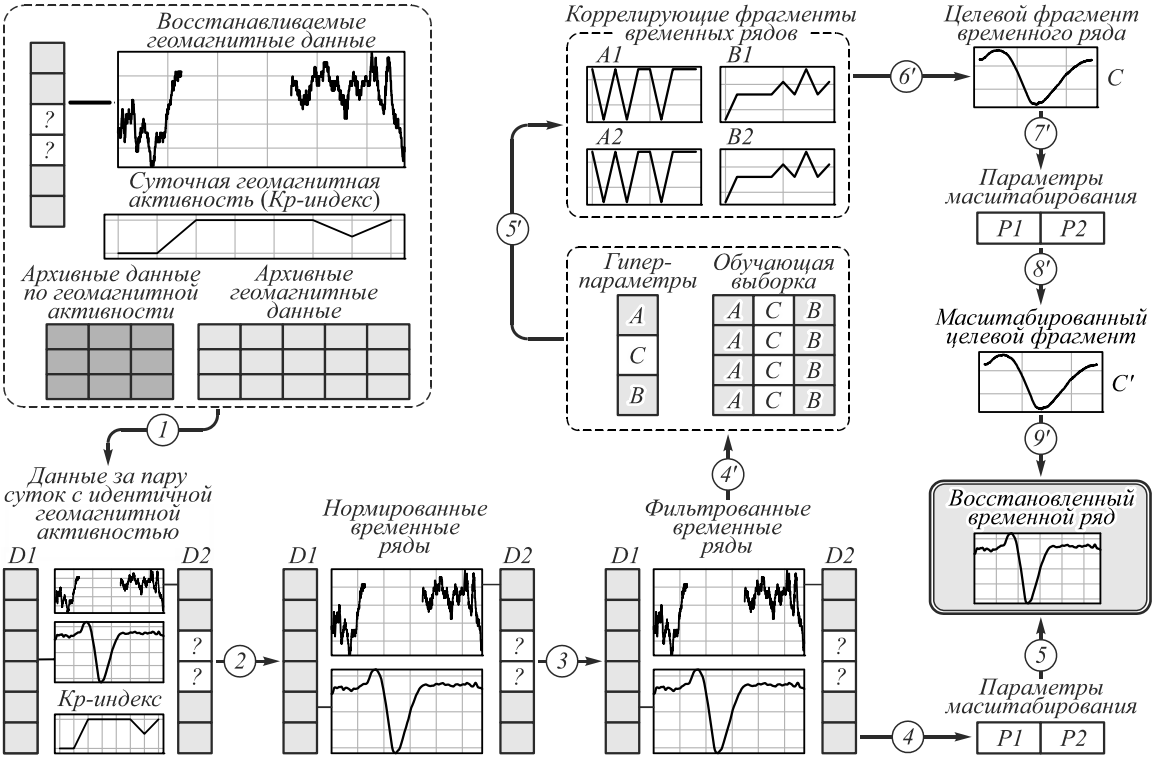

Рис. 1. Схема метода восстановления временных рядов геомагнитных данных

В случае импутации фрагментов меньшей длительности восстанавливаемый временной ряд анализируется на предмет извлечения гиперпараметров (этап 4' на рис. 1), заданных размерами пропущенного (*B* на рис. 1), предшествующего и следующего за ним фрагментов (*A* и *C* на рис. 1 соответственно). На основании полученных гиперпараметров из фрагментов, предшествующих и следующих за пропущенным в восстанавливаемом временном ряду формируется шаблон поиска (этап 4' на рис. 1), а также составляется обучающая выборка из архивных геомагнитных данных, где каждый объект представлен парой непоследовательных фрагментов с заданным друг от друга отступом, а ответ определён разделяющим их фрагментом (этап 5' на рис. 1). При этом проведённые ранее исследования [16] показали, что предлагаемый метод обеспечивает наилучший результат восстановления при равной длине всех указанных фрагментов временного ряда геомагнитных данных.

На следующем шаге (этап 5' на рис. 1) каждая пара объектов обучающей выборки сравнивается с шаблоном поиска, сформированным непоследовательными фрагментами восстанавливаемого временного ряда, разделенных отсутствующими значениями. Наиболее близкая пара объектов принимается как целевой фрагмент временного ряда (этап 6' на рис. 1), который аппроксимируется относительно шаблона поиска (этапы 7' и 8' на рис. 1). Полученный набор значений является искомым фрагментом временного ряда (этап 9' на рис. 1).

3. Формализация метода восстановления геомагнитных данных

Восстановление фрагмента суточного временного ряда

Пусть задан временной ряд $y(t)$ суточных наблюдений $y(t_1)$, $y(t_2)$, ..., $y(t_n)$ параметра геомагнитного поля, полученных в последовательные моменты времени:

$$y(t) = \{(y_1(t), ... y_k(t), ..., y_l(t), ..., y_m(t)\}, \quad (1)$$

где $\{y_1(t), ..., y_{k-1}(t)\}$, $\{y_{l+1}(t), ..., y_m(t)\}$ – наблюдаемые значения уровней временного ряда; $\{y_k(t), ..., y_l(t)\}$ – пропущенные значения уровней временного ряда.

Тогда шаблон поиска T состоит из трёх последовательных фрагментов временного ряда, один из которых (T_2) представлен набором пропущенных значений, а два других – фрагментами, предшествующим (T_1) и следующим (T_3) за первым и равными ему по размерности:

$$T=\{y_{2k-l}(t),...,y_k(t),...,y_l(t),...,y_{2l-k}(t)\}=\{T_1,T_2,T_3\};$$
$$T_1=\{y_{2k-l}(t),...,y_k(t)\},$$
$$T_2=\{y_k(t),...,y_l(t)\},\quad (2)$$
$$T_3=\{y_l(t),...,y_{2l-k}(t)\}.$$

При этом пространство объектов X обучающей выборки задано множеством пар непоследовательных фрагментов временного ряда, разделённых набором значений, число которых равно размерности пропущенного фрагмента исследуемого временного ряда:

$$X=(a_i,b_i)_{i=1}^{l},\quad (3)$$

где l – размер обучающей выборки, a, b – фрагменты временного ряда (независимые переменные выборки).

Для простоты пара переменных каждого элемента из пространства объектов интегрирована в один, что достигается структурным сдвигом соответствующих фрагментов временного ряда и формированием новых массивов значений уровня:

$$X=\{x_i\}: x_i=a_i\to b_i\ R,\ i=\overline{1,l},\quad (4)$$

где x_i – экземпляр пространства объектов (независимая переменная обучающей выборки), a_i, b_i – исходные фрагменты временного ряда, R – квантор сдвига фрагмента временного ряда вправо.

Пространство объектов Y обучающей выборки задано множеством фрагментов временного ряда (целевых переменных), образованных значениями уровней, число которых равно количеству пропущенных значений в исследуемом временном ряду:

$$Y=\{y_i\},i=\overline{1,l}.$$

Между множествами X и Y установлено взаимно однозначное соответствие, такое что все элементы этих двух множеств разбиты на пары вида (x,y), где $x\in X$, $y\in Y$, причём каждый элемент из X и каждый элемент из Y участвует ровно в одной паре. В результате модель, связывающая пространства объектов X и ответов Y, описывается биективной функцией отображения двух множеств

$$y=f(x):\ X\to Y,\ \ Y\to X.$$

Установлено, что если две функции отображения элементов множеств равны между собой, то их аргументы тоже равны, т.е. функция отображения $X\to Y$ проявляет свойство инъективности:

$$f(x_1)=f(x_2)\Rightarrow x_1=x_2\ \ \forall x_1\in X,\forall x_2\in X.$$

При этом каждый элемент из области X является образом ровно одного элемента из области Y, т.е. функция проявляет и свойство сюръективности:

$$f(x)=y\ \ \forall y\in Y,\ \forall x\in X.$$

Сравнение объектов обучающей выборки с шаблоном поиска предполагает нормализацию последнего путем устранения фрагмента T_2, структурного сдвига фрагментов T_1 и T_3 временного ряда и формирования нового массива значений уровня T_0:

$$T_0=T_1\cup T_3=\{y(t_{2k-l}),...y(t_k),y(t_l),...,y(t_{2l-k})\}.$$

Мера близости (сходства) между фрагментом T_0, составленным из предшествующих и последующих за отсутствующим фрагментом значений, и фрагментами из обучающей выборки ($x\in X$) вычисляется на основе коэффициента корреляции Пирсона [22]. Наиболее близким к шаблону поиска принимается тот фрагмент x_t пространства объектов X, которому соответствует максимальное значение коэффициента корреляции:

$$T_0=\{t^i\}_1^m;\ x_t=\{x_t^i\}_1^m\in X:$$
$$\forall x=\{x^i\}_1^m\in X,\ \ x\neq x_t,$$
$$\frac{\sum_{i=1}^{m}\left(x_t^i-\overline{x_t}\right)\left(t_i-\overline{t}\right)}{\sqrt{\sum_{i=1}^{m}\left(x_t^i-\overline{x_t}\right)^2\sum_{i=1}^{m}\left(t_i-\overline{t}\right)^2}}>\frac{\sum_{i=1}^{m}\left(x^i-\overline{x}\right)\left(t_i-\overline{t}\right)}{\sqrt{\sum_{i=1}^{m}\left(x^i-\overline{x}\right)^2\sum_{i=1}^{m}\left(t_i-\overline{t}\right)^2}},$$

где $\overline{x}$ и $\overline{t}$ – средние значения соответствующих массивов со значениями уровня.

Временной ряд геомагнитных данных является нестационарным ввиду стохастического характера изменения значений регистрируемых параметров во времени. Исследования показали, что в результате в пространстве объектов может быть выделено более одного элемента, близкого к шаблону поиска при условии небольшой размерности последнего (до 10 значений уровня).

Для устранения неоднозначности результатов сравнения фрагментов временного ряда целесообразно применение к ним общего центрирования и фильтра низких частот (ФНЧ) Баттерворта. Полученные в результате фильтрации массивы T_f шаблона поиска из (2) и X_f пространства объектов (4) сравниваются на предмет поиска наиболее близких друг к другу значений:

$$T_f=\{t_f^i\}_1^m;\ x_{f_t}=\{x_{f_t}^i\}_1^m\in X:$$
$$\forall x_f=\{x_f^i\}_1^m\in X, x_f\neq x_{f_t},$$
$$\frac{\sum_{i=1}^{m}\left(x_{f_t}^i-\overline{x_{f_t}}\right)\left(t_f^i-\overline{t_f}\right)}{\sqrt{\sum_{i=1}^{m}\left(x_{f_t}^i-\overline{x_{f_t}}\right)^2\sum_{i=1}^{m}\left(t_f^i-\overline{t_f}\right)^2}}>$$
$$>\frac{\sum_{i=1}^{m}\left(x_f^i-\overline{x_f}\right)\left(t_f^i-\overline{t_f}\right)}{\sqrt{\sum_{i=1}^{m}\left(x_f^i-\overline{x_f}\right)^2\sum_{i=1}^{m}\left(t_f^i-\overline{t_f}\right)^2}},$$

где $\overline{x_f}$ и $\overline{t_f}$ – средние значения соответствующих массивов с фильтрованными значениями уровня.

Исследования показали, что в фильтрованном пространстве объектов, как правило, выделяется единственный элемент, близкий шаблону поиска. Позиционно соответствующий ему фрагмент оригинального временного ряда x_t принимается наиболее близким к искомому замещающему:

$$d:\ x_{f_t} \equiv x_f^d \in X_f,\ X_f = \left\{x_f^i\right\}_{i=1}^{l},$$

$$x_t = x_d \in X,\ \ X\ = \left\{x_i\right\}_{i=1}^{l},$$

где d – индекс фрагмента во временном ряду, максимально близкого к искомому.

Полученные результаты восстановления данных являются смещёнными и должны быть аппроксимированы относительно известных соседних пропущенному фрагменту значений уровней временного ряда. Нормализация данных выполняется посредством метода наименьших квадратов и предполагает вычисление значений коэффициентов линейной зависимости двух массивов данных:

$$F(a,b) = \sum_{i=1}^{n}\left(y_i - \left(ax_i + b\right)\right)^2,\ n = \overline{0,d},$$

где a, b – коэффициенты линейной аппроксимации, $x_i \in X$ – временной ряд с зарегистрированными значениями уровня, $y_i \in Y$ – временной ряд с пропущенными значениями уровня, d – длительность пропуска.

Искомый замещающий фрагмент примет вид:

$$z = \left\{ax^k + b\right\}_{k=0}^{d},\ x^k \in x_d,$$

где z – замещающий фрагмент, x_d – фрагмент, наиболее близкий к искомому, d – длина пропущенного фрагмента исследуемого временного ряда.

Гиперпараметрами метода восстановления выступают размеры предшествующего A и последующего C за восстанавливаемыми данными фрагментов временного ряда (и размер самого восстанавливаемого фрагмента B), а также частота среза ФНЧ Баттерворта. Их значения были установлены эмпирически на основании данных, зарегистрированных магнитной обсерваторией Grocka (GCK) за 2008 год. Результаты проведенных исследований [16] показали, что наименьшую ошибку восстановления обеспечивает формирование обучающей выборки из пар фрагментов временного ряда, длина которых равна числу восстанавливаемых значений. Аналогичным образом, для минимизации ошибки метода частота среза ФНЧ не должна превышать величины восстанавливаемого фрагмента.

Восстановление суточного временного ряда

Пусть задан временной ряд $y(t)$ суточных наблюдений $y(t_1), y(t_2), ..., y(t_n)$ параметра геомагнитного поля с неопределёнными (NaN) значениями, зарегистрированными в последовательные моменты времени (минуты) за сутки:

$$y(t) = \{(y(t_1),...,y(t_m)\},\ \forall y(t_m) = \mathrm{NaN},$$

где $\{y(t_1),...,y(t_m)\}$ – значения уровней временного ряда.

Параметры геомагнитной активности заданы значениями Kp-индекса за наблюдаемые сутки:

$$K = \left\{k_i\right\}_{i=1}^{n},$$

где k_i – усреднённые значения Kp-индекса за каждые 3 часа суток, т.е. $n = 24/3 = 8$.

Поскольку порядок следования элементов в множестве K имеет значение, целесообразно заменить представленное множество на упорядоченный набор элементов K, заданный координатами k_i.

$$K = \left(k_1, k_2, ..., k_n\right).$$

Для минимизации объёма экспериментальных данных в работе использованы результаты наблюдений за параметрами геомагнитного поля, полученные магнитной обсерваторией за год, предшествующий исследуемому. Каждые сутки года могут быть описаны упорядоченным набором координат, характеризующих наблюдаемую геомагнитную активность:

$$K^i = \left(k_1^i, k_2^i, ..., k_n^i\right)_{i=1}^{l},$$

где l – среднее количество суток в году ($l = 365$).

Сутки, в течение которых наблюдалась та же геомагнитная активность, что и в восстанавливаемые, могут быть определены как

$$j: K^j = \left(k_k^j\right)_{k=1}^{n} \in \left\{K^i\right\}_{i=1}^{l},$$

$$\left(k_k^j\right)_{k=1}^{n} = \left(k_k\right)_{k=1}^{n},\ K = \left(k_k\right)_{k=1}^{n}.$$

Тогда исходными данными для формирования обучающей выборки метода выступают результаты наблюдений, полученные магнитной обсерваторией в j-е сутки:

$$y^j(t) = \{(y^j(t_1),...,y^j(t_m)\},$$

где $\{y^j(t_1),...,y(t_m)\}$ – минутные значения уровней временного ряда $y_f^j(t)$, m – количество минутных значений в сутки ($m = 1440$).

В результате применения к полученным данным фильтра Баттерворта формируется временной ряд

$$y_f^j(t) = \{(y_f^j(t_1),...,y_f^j(t_m)\},$$

где $y_f^j(t_1),...,y_f^j(t_m)$ – фильтрованные минутные значения уровней временного ряда.

Для аппроксимации полученных данных применительно к восстанавливаемым временным рядам выполняется их нормализация с помощью метода наименьших квадратов. При этом коэффициенты линейной зависимости вычисляются для фильтрованного массива значений и набора данных, зарегистрированных магнитной обсерваторией за сутки, предшествующие восстанавливаемым:

$$F(a,b) = \sum_{i=1}^{m}\left(y_f(t_i) - \left(ay_f^j(t_i) + b\right)\right)^2,$$

где a, b – коэффициенты линейной аппроксимации, $y_f^j(t)$ – временной ряд с фильтрованными значениями уровня, зарегистрированными магнитной обсерваторией в j-е сутки года, предшествующего исследуемому; $y_f(t)$ – временной ряд с фильтрованными значениями, зарегистрированными магнитной обсерваторией в сутки, предшествующие восстанавливаемым.

Искомый замещающий временной ряд примет вид:

$$y(t)=\left\{ay_f^j(t_i)+b\right\}_{k=0}^m,$$

где $y(t)$ – замещающий временной ряд, $y_f^j(t)$ – временной ряд, наиболее близкий к искомому.

Для определения гиперпараметров были проанализированы результаты восстановления суточных временных рядов геомагнитных данных для магнитной обсерватории Grocka (GCK) за 2008 год. Результаты проведённых исследований показали, что наименьшую ошибку восстановления обеспечивает частота среза ФНЧ, не превышающая 10 минут.

4. Экспериментальные исследования

Для минимизации влияния на результат сторонних факторов (распределение источников геомагнитных данных по долготе, сезонные вариации, повышенная геомагнитная активность, асимметрия магнитосферы) исследования были проведены на геомагнитных данных, полученных в среднеширотной обсерватории Grocka (GCK) в условиях спокойной магнитосферы.

Экспериментальные данные были представлены 1 440 значениями полного вектора геомагнитного поля, зарегистрированными магнитной обсерваторией GCK 11 апреля 2008 г. в условиях спокойной магнитной обстановки (среднесуточное значение индекса Kp составило ~1). Для устранения незначимых для поставленной задачи шумов к исходным данным были применён низкочастотный фильтр с частотой среза $f_c=10$.

Исследования показали (рис. 2), что применение метода обеспечивает восстановление временного ряда геомагнитных данных со значением среднеквадратической ошибки в 0,098 нТл, что существенно меньше предельно допустимого значения погрешности геомагнитных измерений. Установлена сильная корреляционная связь между оригинальными и восстановленными геомагнитными данными (значение коэффициента корреляции Пирсона $r_{xy}=0{,}97$). Форма восстановленного сигнала незначительно отличается от исходной, что также свидетельствует об эффективности метода восстановления временного ряда.

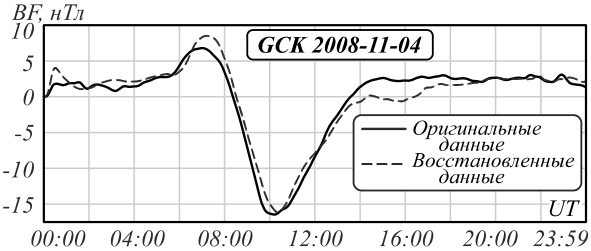

Рис. 2. Результаты восстановления суточных геомагнитных вариаций

При этом предложенный метод обеспечивает существенно меньшее значение среднеквадратической ошибки восстановления экспериментального временного ряда по сравнению с вышеописанными методами реконструкции временных рядов. Так, среднеквадратическая ошибка восстановления представленного временного ряда методом скользящей средней составила 5,41 нТл, линейной интерполяции – 3,19 нТл, модели авторегрессии – 4,85 нТл, модели ARIMA – 4,47 нТл. При этом значение коэффициента корреляции Пирсона для указанных методов составило 0,21; 0,35; 0,23 и 0,24 соответственно.

Исследование возможностей метода для восстановления небольших (длительностью менее суток) фрагментов временного ряда геомагнитных данных было основано на экспериментальных данных, также полученных среднеширотной магнитной обсерваторией GCK 11 апреля 2008 г. Пропущенный фрагмент ряда представлен 10 последовательными значениями полного вектора геомагнитного поля (что соответствует 10-минутному интервалу наблюдений), зарегистрированными в период 11.00–11.10 указанных суток.

Согласно результатам проведённых исследований восстановление фрагмента временного ряда предлагаемым методом обеспечивает значение среднеквадратической ошибки, равное 0,006 нТл, что значительно меньше предельно допустимой стандартами погрешности измерений. Достигнута сильная корреляционная связь между исходными и восстановленными геомагнитными данными, что численно подтверждается значением коэффициента Пирсона, составившим $r_{xy}=0{,}94$. Форма восстановленного сигнала также незначительно отличается от исходной, что свидетельствует о целесообразности применения метода (рис. 3).

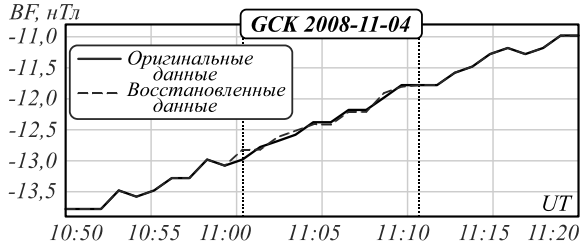

Рис. 3. Результаты восстановления 10-минутного фрагмента временного ряда геомагнитных данных

Проведённый сравнительный анализ также показал, что предложенный метод обеспечивает существенно меньшее значение среднеквадратической ошибки восстановления 10-минутного фрагмента экспериментального временного ряда по сравнению с вышеописанными методами реконструкции временных рядов. Так, среднеквадратическая ошибка восстановления представленного временного ряда методом скользящей средней составила 1,27 нТл, линейной интерполяции – 0,43 нТл, модели авторегрессии – 0,71 нТл, модели ARIMA – 0,64 нТл. При этом значение коэффициента корреляции Пирсона для указанных методов составило 0,48; 0,79; 0,56 и 0,68 соответственно.

Для экспериментального анализа эффективности метода восстановления геомагнитных данных, зарегистрированных высокоширотными магнитными обсерваториями, были использованы результаты наблюдений полного вектора магнитного поля Земли, полученные в обсерватории Abisko (ABK) в те же магнитоспокойные сутки (11 апреля 2008 г.).

Исследования показали, что при восстановлении суточного временного ряда (1 440 фильтрованных

значений) среднеквадратическая ошибка составляет 29,7 нТл, а при импутации 10-минутного фрагмента – 0,225 нТл, что в обоих случаях выше предельно допустимой величины погрешности измерений. Установлено, что независимо от длительности восстанавливаемого фрагмента наблюдается средняя корреляционная связь между оригинальными и восстановленными геомагнитными данными, при этом значение коэффициента Пирсона r_{xy} не превышает 0,7. Формы восстановленных информационных сигналов также существенно отличаются от оригинальных. Такие наблюдения показывают низкую эффективность и нецелесообразность применения метода восстановления для геомагнитных данных, зарегистрированных высокоширотными магнитными обсерваториями.

5. Границы применимости метода восстановления

Сложный характер пространственно-временного распределения параметров геомагнитного поля и его вариаций, а также ряд случайных факторов обуславливают необходимость определения условий, при которых восстановление геомагнитных данных предлагаемым методом является целесообразным. Основным критерием при этом выступает значение ошибки восстановления, ограниченное стандартизованной величиной предельно допустимой погрешности геомагнитных измерений (1 нТл).

В периоды повышенной геомагнитной активности наблюдаются сложные скачкообразные изменения уровней временного ряда геомагнитных данных, что становится значимым препятствием для применения метода восстановления временных рядов. Исследования точности восстановления геомагнитных данных, регистрируемых среднеширотной обсерваторией в различные периоды геомагнитной активности, показали, что значение ошибки восстановления данных тем больше, чем выше величина индекса Kp. Применение метода показало наилучшие результаты в условиях спокойной магнитосферы в периоды, когда среднесуточное значение планетарного индекса геомагнитной активности не превышает 1. При увеличении значения индекса ошибка восстановления существенно возросла и превысила предельно допустимую погрешность геомагнитных измерений (рис. 4).

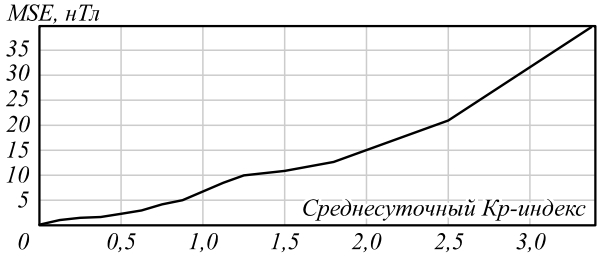

Рис. 4. Зависимость погрешности восстановления данных от геомагнитной активности

Пространственная неоднородность геомагнитного поля проявляется преимущественно в широтной зависимости значений его параметров. Анализ эффективности метода восстановления геомагнитных данных был проведён на основе полученных в условиях спокойной магнитосферы результатов наблюдений магнитных обсерваторий сети INTERMAGNET, расположенных вдоль меридиана 19,1 ± 1,75 ° в.д. (табл. 1).

Исследования показали, что ошибка восстановления геомагнитных данных имеет выраженную широтную зависимость, такую, что её значение минимально для области средних широт и монотонно возрастает в направлении как полюсов, так и экватора (рис. 5). Установлено, что при этом для среднеширотных станций погрешность восстановления фрагментов временного ряда геомагнитных данных тем выше, чем большей длительностью обладает отсутствующий сегмент. Результаты исследований показали, что с наименьшей ошибкой (0,098 нТл) могут быть восстановлены десятиминутные фрагменты (пропуски меньшей длительности целесообразно восполнять методом линейной интерполяции [3]).

Табл. 1. Магнитные обсерватории сети INTERMAGNET, расположенные вдоль меридиана 19,1 ±1,75 ° в.д.

Магнитная обсерватория		**Географические координаты**	
Название	**Код IAGA**	**Широта, с.ш.[град.]**	**Долгота, в.д.[град.]**
Abisko	ABK	68,358	18,823
Lycksele	LYC	64,612	18,748
Uppsala	UPS	59,903	17,353
Hel	HLP	54,608	18,817
Belsk	BEL	51,836	20,789
Tihany	THY	46,900	17,893
Grocka	GCK	44,630	20,770

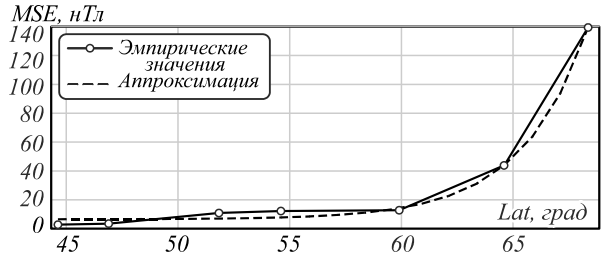

Рис. 5. Зависимость погрешности восстановления данных от географической широты магнитной обсерватории

При этом пропуски длительностью до 30 минут могут быть восстановлены с большей погрешностью, значение которой, однако, не превышает предельно допустимой (рис. 6). В остальных случаях значение погрешности восстановления временного ряда велико, что может привести к недостоверности результатов обработки и анализа полученных таким образом геомагнитных данных. Следует отметить, что исключение составляют данные, представленные 1 440 фильтрованными значениями параметров геомагнитного поля. Такие временные ряды в условиях спокойной геомагнитной обстановки полностью восстанавливаются предлагаемым методом, обеспечивающим значение ошибки в пределах допустимой погрешности геомагнитных измерений.

Таким образом, эмпирическая оценка возможности применения метода восстановления геомагнитных данных показала, что она ограничена следующими условиями:

– спокойная геомагнитная обстановка: значение планетарного индекса геомагнитной активности не должно превышать 1;
– источники геомагнитных данных должны быть размещены на средних географических широтах, поскольку при движении к полюсам и экватору ошибка восстановления превышает предельно допустимую погрешность геомагнитных измерений;
– длительность восстанавливаемого сигнала не должна превышать 10 значений уровня для оригинальных временных рядов и 1440 – для фильтрованных геомагнитных данных.

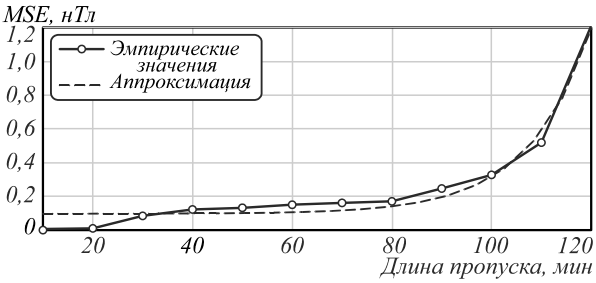

Рис. 6. Зависимость погрешности восстановления данных от длительности пропущенного фрагмента временного ряда

Заключение

Одной из проблем современных систем обработки и анализа геомагнитных данных является восстановление пропущенных фрагментов их временных рядов. Известные методы и средства реконструкции временных рядов недостаточно эффективны и сопровождаются появлением среднеквадратической ошибки, величина которой превышает максимально допустимую погрешность геомагнитных измерений.

В работе предложен подход к восстановлению временных рядов геомагнитных данных, основанный на принципах машинного обучения размеченных данных и методах статического анализа временных рядов. Его отличительной особенностью является то, что наиболее вероятные значения временного ряда определяются из уже известных результатов наблюдения магнитной обсерватории, фрагменты временных рядов которых наиболее близки к искомым. Высокая вычислительная скорость метода при этом обеспечивается сокращением размера обучающей выборки за счёт анализа геомагнитной активности в соответствующие сутки и в предшествующие им периоды.

Результаты эмпирической оценки возможности применения метода восстановления геомагнитных данных свидетельствуют, что они ограничены следующими параметрами: длительность восстанавливаемого сигнала, географическое расположение его источника и текущее состояние магнитосферы.

Благодарности

Результаты, представленные в статье, основаны на данных, собранных магнитными обсерваториями. Авторы благодарят национальные институты, их поддерживающие, за обеспечение высоких стандартов магнитных измерений (*www.intermagnet.org*).

Литература

1. **Love, J.J.** Missing data and the accuracy of magnetic-observatory hour means / J.J. Love // Annals of Geophysics. – 2001. – Vol. 27. – P. 3601-3610.
2. **Barkhatov, N.A.** The method of artificial neuron networks as a procedure for reconstructing gaps in records of individual magnetic observatories from the data at other stations / N.A. Barkhatov, A.E. Levitin, S.Y. Sakharov // Geomagnetism and Aeronomy. – 2002. – Vol. 42. – P. 187-190.
3. **Vorobev, A.V.** Approach to assessment of the relative informational efficiency of intermagnet magnetic observatories / A.V. Vorobev, G.R. Vorobeva // Geomagnetism and Aeronomy. – 2018. – Vol. 58, Issue 5. – P. 625-628.
4. Geomagnetic observations and models / ed. by M. Mandea, M. Korte. – Netherlands: Springer, 2011. – P. 149-181.
5. INTERMAGNET technical reference manual. Version 4.6 / ed. by S.-L. Benoît. – Edinburgh: INTERMAGNET, BGS, 2012. – 100 p.
6. **Soloviev, A.** Mathematical tools for geomagnetic data monitoring and the INTERMAGNET Russian segment / A. Soloviev [et al.] // Data Science Journal. – 2013. – Vol. 12. – P. WDS114-WDS119.
7. **Хомутов, С.Ю.** Обработка магнитных данных в обсерваториях (описание специализированного программного пакета): монография / С.Ю. Хомутов. – Паратунка, Камчатский край: ИКИР ДВО РАН, 2017. – 114 с.
8. **Richman, M.** Missing data imputation through machine learning algorithms / M. Richman [et al.] // Artificial Intelligence Methods in the Environmental Sciences. – 2009. – P. 153-169.
9. **Bertsimas, D.** From predictive methods to missing data imputation: An optimization approach / D. Bertsimas // Journal of Machine Learning Research. – 2018. – Vol. 18. – P. 1-39.
10. **Lopes, N.** Handling missing values via a neural selective input model / N. Lopes // Neural Network World. – 2012. – No. 4/12. – P. 357-370.
11. **Abidin, N.Z.** Performance analysis of machine learning algorithms for missing value imputation / N.Z. Abidin [et al.] // International Journal of Advanced Computer Science and Applications (IJACSA). – 2018. – Vol. 9(6). – P. 442-447.
12. **Liu, Yu.** An overview and evaluation of recent machine learning imputation methods using Cardiac imaging data / Yu. Liu, V. Gopalakrishnan // Data. – 2017. – Vol. 2, Issue 1. – 8.
13. **Кулаичев, А.П.** Методы и средства комплексного анализа данных / А.П. Кулаичев. – М.: «ФОРУМ: ИНФРА», 2006. – 512 с.
14. **Волков, Е.А.** Численные методы / Е.А. Волков. – М.: Наука, 1987.
15. **Бокс, Дж.** Анализ временных рядов (прогноз и управление) / Дж. Бокс, Г. Дженкинс. – Вып. 1. – М.: Мир, 1974. – 408 с.
16. **Воробьёв, А.В.** Индуктивный метод восстановления временных рядов геомагнитных данных / А.В. Воробьёв, Г.Р. Воробьёва // Труды СПИИРАН. – 2018. – № 57. – С. 104-133.
17. **Takens, F.** Detecting strange attractors in turbulence / F. Takens. – In: Dynamical systems and turbulence, Warwick 1980 / ed. by D. Rand, L.-S. Young. – Berlin, Heidelberg: Springer, 1981. – P. 366-381.
18. **Юмагулов, М.Г.** Введение в теорию динамических систем / М.Г. Юмагулов. – СПб.: Лань, 2015. – 272 с.
19. **Лукашин, Ю.П.** Адаптивные методы краткосрочного прогнозирования временных рядов / Ю.П. Лукашин. – М.: Финансы и статистика, 2003. – 416 с.
20. **Гадзиковский, В.И.** Цифровая обработка сигналов / В.И. Гадзиковский. – М.: СОЛОН-Пресс, 2015. – 766 с.

21. **Аллен, Б.Д.** Think DSP. Цифровая обработка сигналов на Python / Б.Д. Аллен. – М.: ДМК Пресс, 2017. – 160 с.

22. **Харченко, М.А.** Корреляционный анализ / М.А. Харченко. – Воронеж: Изд-во ВГУ, 2008. – 31 с.

Сведения об авторе

Воробьёва Гульнара Равилевна, 1983 года рождения, в 2005 году окончила Уфимский государственный авиационный технический университет по специальности «Автоматизированные системы обработки информации и управления», кандидат технических наук, доцент, работает доцентом факультета информатики и робототехники ФГБОУ ВО Уфимский государственный авиационный технический университет. Область научных интересов: геоинформационные и веб-технологии, системы хранения и обработки информации.
E-mail: *gulnara.vorobeva@gmail.com* .

ГРНТИ: 83.77.31
Поступила в редакцию 21 марта 2019 г. Окончательный вариант – 21 мая 2019 г.

Approach to the recovery of geomagnetic data by comparing daily fragments of a time series with equal geomagnetic activity

G.R. Vorobeva [1]
[1] *Ufa State Aviation Technical University, Ufa, Russia*

Abstract

Monitoring of geomagnetic field parameters and its variations is mainly carried out using ground-based magnetic observatories and variational stations. However, the imperfection of equipment used and the communication channels involved causes the presence of gaps in the time series of geomagnetic data, which, along with the spatial anisotropy of data sources, creates significant obstacles to their automated processing. In addition, the well-known methods for imputation of time series gaps provide the root-mean-square recovery error significantly exceeding the level acceptable for geophysical observations. Thus, the paper proposes a method for recovering geomagnetic data based on statistical methods for processing time series and machine learning principles using marked data and characterized by the fact that a pair of the time series fragments preceding and succeeding a missing fragment provide an indicative description of the time series fragment of interest, which together form a training sample to search for the missing fragment by a set of its attributes, followed by linear scaling to restore the original trend of an information signal. Analytical estimates of parameters of geomagnetic data time series are given, under which it is possible to apply the proposed method to recover both daily variations and several-minutes-long fragments.

Keywords: time series recovery, time series processing, geomagnetic data, machine learning, statistical analysis.

Citation: Vorobeva GR. 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. DOI: 10.18287/2412-6179-2019-43-6-1053-1063.

Acknowledgements: The results presented in the article are based on data collected by magnetic observatories. The authors thank the national institutions that support them for ensuring high standards of magnetic measurements (*www.intermagnet.org*).

References

[1] Love JJ. Missing data and accuracy of magnetic-observatory hour means. Ann Geophys 2001; 27: 3601-3610.

[2] Barkhatov NA, Levitin AE, Sakharov SY. The method of artificial neuron networks as a procedure for reconstructing gaps in records of individual magnetic observatories from the data at other stations. Geomag Aeron 2002; 42: 187-190.

[3] Vorobev AV, Vorobeva GR. Approach to assessment of the relative informational efficiency of intermagnet magnetic observatories. Geomagn Aeron 2018; 58(5): 625-628.

[4] Mandea M, Korte M, eds. Geomagnetic observations and models. Netherlands: Springer; 2011: 149-181.

[5] INTERMAGNET technical reference manual. Edinburgh: INTERMAGNET, BGS; 2012.

[6] Soloviev A. Mathematical tools for geomagnetic data monitoring and the INTERMAGNET Russian segment. Data Sci J 2013; 12: WDS114-WDS119.

[7] Khomutov SYu. Processing of magnetic data in observatories (description of a specialized software package) [In Russian]. Paratunka: IKIR FEB RAS Publisher; 2017.

[8] Richman M, et al. Missing data imputation through machine learning algorithms // Artificial Intelligence Methods in the Environmental Sciences 2009: 153-169.

[9] Bertsimas D, et al. From predictive methods to missing data imputation: An optimization approach. Journal of Machine Learning Research 2018; 18: 1-39.

[10] Lopes N. Handling missing values via a neural selective input model. Neural Network World 2012; 4/12: 357-370.

[11] Abidin NZ, et al. Performance analysis of machine learning algorithms for missing value imputation. International Journal of Advanced Computer Science and Applications (IJACSA) 2018; 9(6): 442-447.

[12] Liu Yu, Gopalakrishnan V. An overview and evaluation of recent machine learning imputation methods using Cardiac imaging data. Data 2017; 2(1): 8.

[13] Kulaichev AP. Methods and tools for integrated data analysis [In Russian]. Moscow: "FORUM-INFRA" Publisher; 2006.
[14] Volkov EA Numerical methods [In Russian]. Moscow: "Nauka" Publisher; 1987.
[15] Box J, Jenkins G. Time series analysis (Forecast and management) [In Russian]. Vol 1. Moscow: :Mir" Publisher; 1974.
[16] Vorobev AV, Vorobeva GR. Inductive method for recovering geomagnetic data time series [In Russian]. SPIIRAS Proceedings 2018; 57: 104-133.
[17] Takens F. Detecting strange attractors in turbulence. In Book: Rand D, Young L-S. Dynamical systems and turbulence, Warwick 1980. Berlin, Heidelberg: Springer, 1981: 366-381.
[18] Yumagulov MG. Introduction to the theory of dynamical systems [In Russian]. Saint-Peterburg: "Lan" Publisher; 2015.
[19] Lukashin YuP. Adaptive methods for short-term time series forecasting [In Russian]. Moscow: "Financy i statistika" Publisher; 2003.
[20] Gadzikovsky VI. Digital signal processing [In Russian]. Moscow: "SOLON-Press" Publisher; 2015.
[21] Allen BD. Think DSP: Digital signal processing with Python [In Russian]. Moscow: "DMK Press" Publisher; 2017.
[22] Kharchenko MA. Correlation analysis [In Russian]. Voronezh: "VGU" Publisher; 2008.

Authors' information

Gulnara Ravilevna Vorobeva (b. 1983) graduated from Ufa State Aviation Technical University in 2005, majoring in Automated Systems of Data Processing and Control, PhD. Currently she works as the associate professor of Computer Science and Robotics department in Ufa State Aviation Technical University. Research interests are geoinformation and web technologies, systems of information storing and processing. E-mail: *gulnara.vorobeva@gmail.com* .

Received March 21, 2019. The final version – May 21, 2019.