УДК 519.246.8 **DOI** 10.15622/sp.57.5

А.В. Воробьев, Г.Р. Воробьева

ИНДУКТИВНЫЙ МЕТОД ВОССТАНОВЛЕНИЯ ВРЕМЕННЫХ РЯДОВ ГЕОМАГНИТНЫХ ДАННЫХ

Воробьев А.В., Воробьева Г.Р. **Индуктивный метод восстановления временных рядов геомагнитных данных.**

Аннотация. В настоящее время интенсивное развитие систем и технологий регистрации параметров магнитного поля Земли способствует экспоненциальному росту объемов геомагнитных данных, основным источником которых выступают постоянные магнитные станции. Несовершенство применяемой аппаратуры и задействованных каналов передачи информации обуславливает наличие пропусков во временных рядах зарегистрированных данных, что вместе с пространственной анизотропией создает серьезное препятствие для обработки геомагнитных данных при решении прикладных задач. Российские и зарубежные научные организации восстанавливают пропущенные геомагнитные данные методом линейной интерполяции, что обеспечивает приемлемые результаты в условиях спокойной магнитосферы, но значительно искажает временные ряды при изменении окружающей магнитной обстановки. В этой связи возникает актуальная научно-техническая задача разработки подхода к восстановлению геомагнитных данных в условиях возбужденной магнитосферы, обеспечивающего оптимальные метрики качества импутации временных рядов.

Авторами предложен метод восстановления временных рядов, основанный на индуктивном методе обучения алгоритмов. Согласно предлагаемому подходу, каждая магнитная станция оперирует собственной базой знаний, формируемой в ходе регистрации параметров геомагнитного поля и его вариаций. Комбинация значений ряда, предшествующих и следующих за пропуском, является признаковым описанием, применяемым для поиска прецедента в базе знаний магнитной станции. Результат содержит искомый фрагмент временного ряда и заменяет пропущенные значения его уровней. Сложность характера информационного сигнала, обусловленная неспокойной магнитной обстановкой, повышает точность поиска по прецедентам, эффективность которого тем выше, чем большей базой знаний располагает магнитная станция.

Проведенный анализ результатов восстановления пропусков временных рядов геомагнитных данных, зарегистрированных в условиях возбужденной магнитосферы, показал, что предложенный индуктивный метод импутации позволяет повысить точность восстановления пропущенных значений в среднем на 79.54 % по сравнению с используемыми в настоящее время методами, что позволит повысить эффективность обработки геомагнитных данных при решении прикладных задач.

Ключевые слова: геомагнитные данные, временные ряды, пропущенные значения, машинное обучение, обучение по прецедентам, импутация временных рядов.

1. Введение. Наблюдаемый в настоящее время рост объемов геомагнитных данных обусловлен интенсивным развитием наземных и спутниковых систем глобального мониторинга, обеспечивающих непрерывную регистрацию параметров магнитного поля Земли в режиме квазиреального времени [1]. Магнитные станции, аэромагнитные, гидромагнитные съемки, спутниковые и подземные скважинные исследования обеспечивают наблюдения, комплексный и своевременный анализ которых является основой для моделирования компонент геомаг-

нитного поля, понимания причин его эволюции и оценки активности, что особенно важно ввиду доказанной опасности геомагнитных вариаций для техногенных объектов и систем (спутников, коротковолновой радиосвязи, высокоточной магнитометрической аппаратуры, систем автоматики высокоширотных железных дорог и др.).

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 [2-4]. Регистрируемые станциями ежеминутные и ежесекундные данные наблюдений агрегированы в наборы односуточных текстовых файлов формата IAGA-2002 и доступны по протоколам FTP и HTTP конечным пользователям и приложениям.

Одной из важнейших задач магнитных станций ИНТЕРМАГНЕТ является обеспечение непрерывности регистрации данных об измеряемых параметрах геомагнитного поля и его вариаций [5]. Однако ввиду несовершенства используемых магнитными станциями аппаратуры и каналов данных исходные временные ряды геомагнитных данных содержат пропуски и «выбросы» за нормальное значение, которые являются необратимыми и могут привести к потере важной информации о геофизических явлениях и процессах [6].

В этой связи совершенствование методов и алгоритмов эффективной обработки больших объемов геомагнитных данных, включая способы восстановления пропущенных значений, входит в число первостепенных проблем современной геофизики.

В настоящее время сетью ИНТЕРМАГНЕТ задача заполнения пропусков геомагнитных данных решается простейшим, но малоэффективным методом — заменой пропусков на зарезервированные значения. Так, стандарт IAGA-2002 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.

Другой известный и широко практикуемый в России и за рубежом подход [7-9] основан на линейной интерполяции временных рядов геомагнитных данных, содержащих пропущенные значения. Простота и вычислительная скорость данного метода являются его безусловным преимуществом, однако его эффективность при этом ограничена восста-

новлением только небольших пропущенных сегментов временного ряда геомагнитных данных при условии спокойной магнитной обстановки.

Ни один из представленных методов не решает выявленную проблему в достаточной мере. В этой связи авторами в данной работе предлагается новый подход к восстановлению временных рядов геомагнитных данных в условиях возбужденной магнитосферы, который позволит дополнить известные подходы к импутации пропусков (здесь и далее под импутацией понимается заполнение пропусков временных рядов) и повысит эффективность существующих методов и средств обработки данных наблюдений параметров геомагнитного поля и его вариаций.

2. Краткий анализ особенностей анализируемых временных рядов. Временные ряды геомагнитных данных, зарегистрированных магнитными станциями ИНТЕРМАГНЕТ, в общем виде обладают сходными характеристиками в контексте стационарности, тренда, регулярных и нерегулярных осцилляций. Краткий анализ этих особенностей в данной работе проводится авторами на примере типичного временного ряда геомагнитных данных.

В качестве экспериментального в работе рассматривается временной ряд, уровни которого представлены результатами поминутных измерений горизонтальной компоненты вектора геомагнитного поля, полученными магнитной станцией DOUrbes (50.1°N, 4.6° E) в 2016 году.

Для нивелирования влияния выбросов геомагнитные данные предлагается анализировать на основе медианы и интерквартильного размаха с визуализацией посредством диаграммы размаха, представленной на рисунке 1.

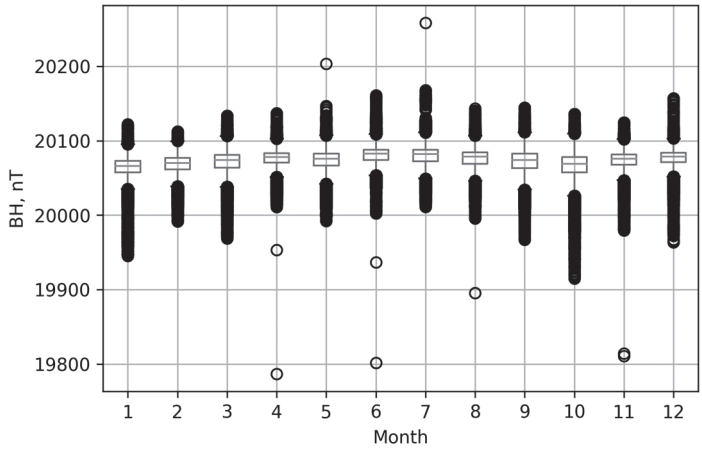


Рис. 1. Результаты анализа временного ряда геомагнитных данных обсерватории DOUrbes за 2016 год

Диаграмма размаха дает представление о медиане, разбросе и асимметрии в распределении значений исследуемого параметра геомагнитного поля для каждого календарного месяца анализируемого года наблюдений. Кроме того, данный способ визуализации является одним из индикаторов наличия выбросов во временном ряду. Так, диаграмма размаха, приведенная на рисунке 1*б*, показывает, что в анализируемом временном ряду на протяжении шести месяцев присутствовали выбросы как за минимальные, так и за максимальные значения. При этом сравнение параллельных графиков показывает, что в течение исследуемого года распределение параметра геомагнитного поля изменялось несущественно, сохраняя значение медианы исследуемого ряда примерно на одном и том же числовом уровне. Аналогичный разброс значений показывает и межквантильный размах: приведенные на диаграмме 25- и 75-процентные квантили относительно стабильны в течение всего рассматриваемого периода, в то время как минимальные и максимальные граничные значения отличаются более существенно (что в целом можно объяснить наличием выбросов в соответствующие календарные месяцы).

Сравнение асимметрии и эксцесса выбросов временного ряда геомагнитных данных с известными моментами нормального распределения посредством теста Харки — Бера [10] показало, что нулевая гипотеза о нормальности распределения выбросов анализируемого ряда отвергается с вероятностью, равной 1.16451551e-08. Это свидетельствует как о нормальном распределении ошибок наблюдения анализируемого геомагнитного параметра поля, так и об однородности всего временного ряда геомагнитных данных.

Кроме того, анализ исследуемого временного ряда геомагнитных данных показал наличие нелинейного тренда, сезонности и цикличности, что определяет аддитивный нестационарный характер его уровней (рисунок 2), при этом:

– нелинейный тренд представляет графическую интерпретацию зависимости уровней временного ряда геомагнитных данных от времени и имеет нейтральный характер флэта, поскольку не обладает выраженным восходящим или нисходящим развитием;

– сезонная компонента временного ряда с характерным для параметров геомагнитного поля 27-дневным периодом характеризует изменение значения горизонтальной составляющей вектора поля в соответствии с указанным циклом;

– циклическая составляющая временного ряда, выделенная в ходе анализа исследуемых уровней, характеризует только длину цикла, поскольку рассматривается в рамках одного календарного года.

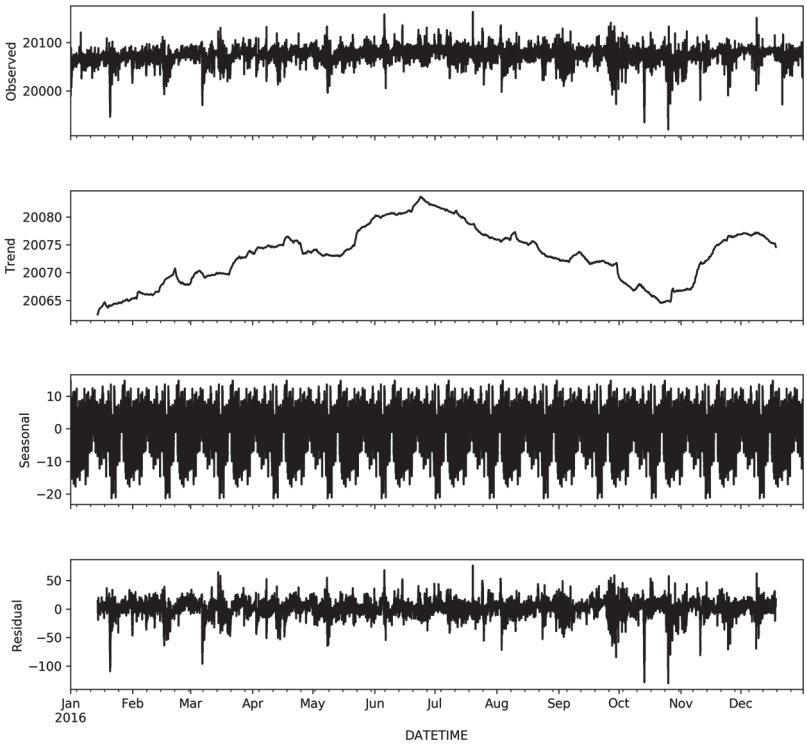


Рис. 2. Декомпозиция временного ряда геомагнитных данных обсерватории DOUrbes за 2016 год

Анализ и визуализация ряда большего временного диапазона подтвердили сделанное таким образом предположение относительно циклической составляющей исследуемого временного ряда.

Данное наблюдение также подтверждается результатами обобщенного теста Дикки — Фуллера (ADF) [11], выявившего наличие во временном ряду единичных корней, что, в свою очередь, свидетельствует о нестационарности уровней ряда.

Применение первой разности приводит временной ряд геомагнитных данных к стационарному виду, что подтверждается результатами того же теста Дикки — Фуллера, отвергающими нулевую гипотезу о наличии единичных корней. Данное наблюдение позволяет описать исследуемые уровни геомагнитных данных как интегрированный временной ряд первого порядка.

3. Классификация пропусков временного ряда геомагнитных данных. Предлагаемая авторами настоящей работы методика

восстановления временных рядов геомагнитных данных базируется на новом подходе к классификации пропущенных значений.

Пусть x_{iT} — уровни временного ряда геомагнитных данных за период T с наблюдениями $i = 1,...,n$. Значение ряда $x \in \{x_{iT} \mid \forall i,T\}$ описывается *измерительной мерой* m_{iT} (здесь и далее для наглядности наличие значения временного ряда наблюдений обозначается как определенное значение уровня ряда):

$$m_{iT} = \begin{cases} 0, \text{ значение } x_{iT} \text{ определено,} \\ 1, \text{ значение } x_{iT} \text{ заменено индикатором,} \\ 2, \text{ значение } x_{iT} \text{ отсуствует.} \end{cases}$$

Тогда временной ряд измерений параметров геомагнитного поля может быть определен как:

$$x_T = \{x_{\text{obs}}, x_{\text{mis}}\},$$

где x_{obs} — наблюдения с определенными значениями параметра $x_{iT} \in \{x_{\text{obs}} \mid \forall i,T\}$, x_{mis} — наблюдения с отсутствующими значениями параметра $x_{iT} \in \{x_{\text{mis}} \mid \forall i,T\}$.

При этом:

$$x_{\text{obs}} \equiv \{(x_{iT} \mid \forall i,T) \wedge (m_{iT} = 0)\};$$
$$x_{\text{mis}} \equiv \{(x_{iT} \mid \forall i,T) \wedge (m_{iT} > 0)\}.$$

Полным считается временной ряд x_T длиной n, для которого справедливо выражение:

$$x_T : \forall x_{iT} \in x_T \; m_{iT} = 0 \; (i = 1,...,n),$$
$$x_T = \{x_{\text{obs}}, x_{\text{mis}}\}, \; x_{\text{mis}} = \varnothing.$$

Если временной ряд x_T длиной n удовлетворяет следующим условиям:

$$x_T : \forall x_{iT} \in x_T \; m_{iT} \geq 0 \; (i = 1,...,n),$$
$$x_T = \{x_{\text{obs}}, x_{\text{mis}}\}, \; x_{\text{obs}} = \varnothing, \; , \; x_{\text{mis}} = \varnothing,$$

то он считается неполным.

Согласно принятой в теории анализа временных рядов классификации пропусков [12], отсутствующие значения геомагнитных данных относятся к типу MCAR (Missing Completely at Random), поскольку вероятность пропуска результата наблюдения одинакова для каждой записи набора зарегистрированных значений. Не предоставляя иной альтернативы (тип MAR или MNAR), данный факт не является основанием для выбора метода заполнения пропусков геомагнитных данных.

В этой связи в данной работе предлагается расширить принятую классификацию пропусков значений уровня временного ряда, введя ряд их дополнительных характеристик.

Предварительно для их определения вводится понятие сегмента временного ряда s, под которым понимается множество последовательных l значений уровня ряда:

$$s = \{s_{jT}\},\ j = 1,...,l;\ s_{jT} \in (x_{iT} \mid \forall i,T),\ i = 1,...,n,\ n > l,$$

где l — длина сегмента временного ряда.

Тогда дискретными будут называться пропущенные сегменты s временного ряда размерности n, длина которых строго равна одному:

$$s = \{s_{jT} : m_{jT} \geq 1\},\ j = 1;\ s_{jT} \in (x_{iT} \mid \forall i,T),\ i = 1,...,n,\ n > l.$$

Серийными будут обозначаться такие пропущенные сегменты s временного ряда размерности n, длина которых превышает один:

$$s = \{s_{jT} : m_{jT} \geq 1\},\ j = 1...,l; s_a \prec s_b,\ s_a,\ s_b \in s,\ a < b;$$
$$s_{jT} \in (x_{iT} \mid \forall i,T),\ i = 1,...,n,\ n > l;\ s \subset \{x_{iT}\},\ i = 1,...,l.$$

Если за пропуском длиной l следует сегмент определенных значений уровня, длина которого равна или превышает значение l, то такой пропуск будет называться интерполируемым:

$$s \subset \{x_{iT} : m_{iT} > 0\},\ i = 1,...,l;\ b \subset \{x_{iT} : m_{iT} = 0\},\ i = 1,...,n-l;\ \#(b) \geq \#(s).$$

В противном случае пропуск значений уровня временного ряда будет считаться экстраполируемым:

$$s \subset \{x_{iT} : m_{iT} > 0\},\ i = 1,...,l;\ b \subset \{x_{iT} : m_{iT} = 0\},\ i = 1,...,n-l;$$
$$\#(b) < \#(s) \text{ or } b = \varnothing.$$

С учетом приведенных характеристик, любой пропуск временного ряда геомагнитных данных может быть отнесен к одному из следующих типов: дискретный интерполируемый; дискретный экстраполируемый; серийный интерполируемый; серийный экстраполируемый.

Анализ экспериментальных данных наблюдений сети магнитных станций ИНТЕРМАГНЕТ показал следующее распределение типов пропусков геомагнитных данных за исследуемый 2016 год:

– дискретные интерполируемые — не обнаружено;

– дискретные экстраполируемые — не обнаружено;

– серийные интерполируемые — 10 пропусков (максимальная длина сегмента — 240 минут);

– серийный экстраполируемый — 27 пропусков (отсутствуют данные наблюдений за 27 непоследовательных суток).

Также для сравнения был проведен анализ временного ряда обсерватории AMS (Martinde Vivies, Amsterdam Island) с худшим по сравнению со станцией DOUrbes показателем относительной информационной эффективности [13]. Типы пропусков геомагнитных данных распределены следующим образом:

– дискретные интерполируемые — 3 пропуска;

– дискретные экстраполируемые — не обнаружено;

– серийные интерполируемые — 4 пропуска (максимальная длина сегмента — 1 108 минут);

– серийный экстраполируемый — 32 пропуска (из них отсутствуют данные 30 суток наблюдений).

Анализ подтверждает неравномерный характер распределения типов пропусков временного ряда геомагнитных данных для любой магнитной станции. Каждое из отсутствующих значений может быть восстановлено, но способ импутации напрямую зависит от характера пропуска. Неверный подбор метода восстановления пропуска может критически исказить весь временной ряд наблюдений и негативно отразиться на результатах его анализа. Поэтому целесообразным представляется подход, согласно которому определяется тип каждого пропуска временного ряда и в соответствии с этим применяется оптимальный метод его восстановления.

4. Анализ эффективности известных методов и моделей восстановления пропусков временного ряда. Основанием для выбора оптимального метода восстановления для каждого типа пропуска во временном ряду геомагнитных данных послужили результаты выполненного авторами анализа эффективности ряда известных моделей и методов импутации и прогнозирования данных, метрикой

качества которых послужило значение среднеквадратической ошибки восстановления ряда.

1) Упрощенный метод скользящей средней — частный случай способа сглаживания временных рядов. Данный подход реализован таким образом, что ширина N сегмента временного ряда фиксирована и равна 3, а пропущенное значение ряда рассчитывается как среднее арифметическое предшествующего и последующего замеров:

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

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

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 и обеспечивает сопоставимую с другими методами метрику качества.

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 пропущенного, тем больше величина среднеквадратического отклонения, возникающего при восстановлении временного ряда. Поэтому предпочтительно применение метода скользящей средней для восстановления дискретных интерполируемых значений с симметричными предшествующим и последующим сегментами временного ряда. В остальных случаях величина ошибки критически возрастает, что снижает целесообразность применения указанного метода.

2) Линейная интерполяция — метод, активно применяемый в настоящее время в геофизике для восстановления геомагнитных данных. Суть метода состоит в том, что крайние точки пропущенного сегмента временного ряда соединяются друг с другом прямой линией,

то есть составляется полином первой степени, поиск коэффициентов которого выполняется в ходе интерполяции:

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1},\ x_1 \le x \le x_2,\ y = ax + b,$$

$$a = \frac{y_2 - y_1}{x_2 - x_1},\ b = y_1 - ax_1,$$

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.

Метод линейной интерполяции обеспечивает наибольшую эффективность при восстановлении дискретных пропусков обоих типов, поскольку интервал, в который попадает искомое значение, в данном случае минимален. Увеличение длины пропущенного сегмента приводит к пропорциональному увеличению значения среднеквадратической ошибки.

3) Метод кубической сплайн-интерполяции известен как более точный по сравнению с описанной выше линейной интерполяцией. Его отличительной особенностью является разбиение интервала интерполяции на отрезки, на каждом из которых функция задается полиномом третьей степени с коэффициентами, обеспечивающими как непрерывность функции, так и ее прохождение через заданные точки:

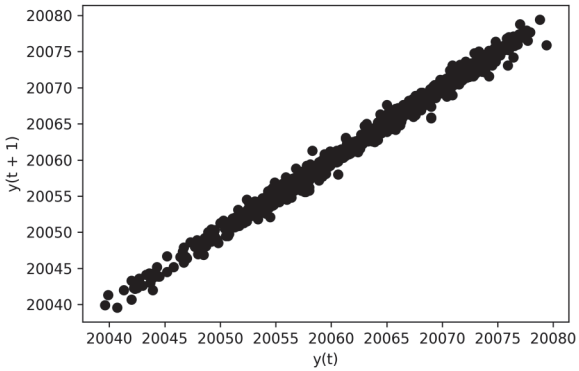
$$F_k(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3;$$
$$F = F_1 \text{ на интервале } [x_0, x_1],$$
$$F = F_2 \text{ на интервале } [x_1, x_2],$$
$$\ldots$$
$$F = F_N \text{ на интервале } [x_{N-1}, x_N],$$

где $F(x)$ — интерполирующая функция, $F_k(x)$ — кубический полином на отрезке $[x_{k-1}, x_k]$; a_k, b_k, c_k, d_k — коэффициенты полинома на отрезке $[x_{k-1}, x_k]$; $[x_0, x_N]$ — интерполируемый отрезок.

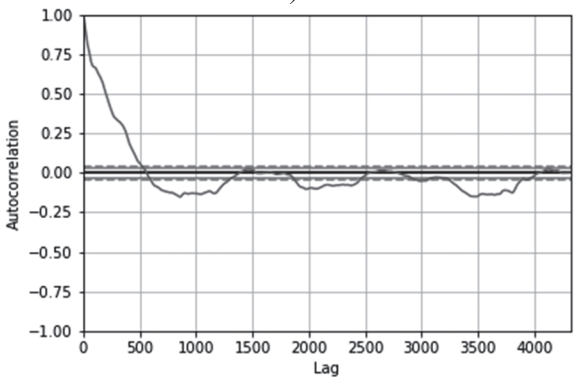
Специфика анализируемого явления позволяет нивелировать склонность кубических сплайнов к осцилляции в окрестностях точки, существенно отличающейся от своих соседей. В этой связи точность рассматриваемого метода повышается по сравнению с результатами, обеспечиваемыми методом кусочно-линейной интерполяции.

4) Модели и методы авторегрессии. Наличие тренда и сезонной / циклической составляющей временного ряда обуславливает корреляционную зависимость между его последовательными значениями, известную как автокорреляция уровней ряда.

Зависимость между измерениями из исследуемого ряда геомагнитных данных и их лагами приведена на рисунке 3*а*. Выбивающиеся из общего массива точек значения являются выбросами уровней ряда и могут быть нивелированы. Оставшиеся измерения визуально кластеризуются по диагонали от левого нижнего к правому верхнему краю диаграммы, что свидетельствует о положительной корреляционной связи между ними.



а)



б)

Рис. 3. Коррелограммы экспериментального временного ряда

Осциллирующая зависимость геомагнитных измерений и их лагов подтверждает детерминированный характер уровней анализируемого временного ряда (рисунок 3*б*). Из графика автокорреляции 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.

Суть предлагаемого подхода заключается в том, что известные уровни ряда, предшествующие пропущенному сегменту, 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:

$$Y_i = a_0 + \sum_{i=1}^{p} a_i Y_{i-1} + \varepsilon_i,$$

где Y_i — целевая переменная (атомарное восстанавливаемое значение); p — порядок модели, 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.

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 прошлом периоде авторегрессии. Выбор такой модели авторегрессии обусловлен анализом коррелограммы (рисунок 3*б*), 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 сектор, тем большую погрешность обнаруживает метод и модель авторегрессии. В этой связи рассматриваемый метод авторегрессии используется итерационно для восстановления одного значения в пропущенном секторе, которое, в свою очередь, становится частью новой обучающей выборки для восстановления последующего элемента пропущенного сегмента и так далее.

Следующий прогностический метод, исследованный применительно к восстановлению геомагнитных данных, — интегрированная модель авторегрессии — скользящего среднего (ARIMA). Модель ха-

рактеризуется тремя параметрами: p — порядок авторегрессии, d — порядок интегрирования, q — порядок скользящего среднего [14-16]:

$$\left(\Delta^d X_t\right) = \sum_{i=1}^{p} \varphi_i \left(\Delta^d X_t\right) + \varepsilon_t + \sum_{j=1}^{q} \theta_j \left(\Delta^d \varepsilon_{t-j}\right),\ \varepsilon_t \sim N(0, \sigma_t^2),$$

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

Итеративное исследование различных комбинаций перечисленных параметров модели ARIMA было выполнено с помощью «сетчатого поиска» и показало, что лучшее значение информационного критерия Акаике (AIC) достигается при $p = 1$, $d = 0$, $q = 1$. Результат подтверждает, что наилучшая корреляция наблюдается между двумя соседними значениями экспериментального ряда данных.

Анализ перечисленных методов восстановления данных был выполнен применительно к решению задачи импутации пропущенных значений экспериментального ряда геомагнитных данных обсерватории DOU за 2016 год (таблица 1).

Таблица 1. Среднеквадратическая ошибка восстановления основных типов пропусков известными методами импутации (нТл)

Тип пропуска \ Метод импутации	Скользящее среднее	Линейная интерполяция	Кубическая сплайн-интерполяция	Модель автокорреляции	ARIMA
Дискретный интерполируемый	0.110	0.153	0.153	0.475	0.482
Дискретный экстраполируемый	0.171	0.157	0.157	0.491	0.522
Серийный интерполируемый (получасовой пропуск при Kp < 3)	0.730	0.679	0.674	0.728	0.953
Серийный экстраполируемый (получасовой пропуск при Kp < 3)	8.402	13.738	12.694	8.508	10.670

Большинство известных методов анализа и восстановления временных рядов реализовано посредством программных средств, ориентированных на конечного пользователя и обладающие интуитивно понятным интерфейсов, и программных библиотек, ориентированных на разработчиков программных сценариев и продуктов [17-23]. Ярким представителем первого типа программных средств является инструментально-программная система «Гусеница» («Caterpillar»), доступная по URL: http://www.gistatgroup.com/gus/. Данная программа предоставляет конечному пользователю возможность загрузки временного ряда, его визуализации и аналитической обработки. Функция восстановления пропусков временного ряда не является здесь основной и реализована только для простейших и краткосрочных отсутствующих сегментов ряда. Основной используемый при этом метод — линейная рекуррентная формула — обеспечивает восстановления временного ряда геомагнитных данных со средним значением среднеквадратической ошибки 2.863 нТл, что существенно хуже методов, проанализированных выше (таблица 1).

Второй подход, ориентированный на разработчиков программных систем, — это прежде всего инструмент реализации известных методов восстановления временных рядов, в частности, перечисленных выше в таблице 1. Так, представленные результаты были получены авторами работы посредством языка программирования Python и его статистических библиотек. Вместе с тем собственными методами восстановления Python и ему подобные языки не располагают и тем более не учитывают специфики временных рядов геомагнитных данных, уровни которых зависят от текущего состояния Kp-индекса и, соответственно, окружающей магнитной обстановки в момент регистрации пропущенного сегмента ряда.

В целом анализ полученных результатов (таблица 1) показал следующее распределение методов восстановления данных по их эффективности. Метод скользящего среднего наиболее эффективен для импутации дискретных интерполируемых пропусков. Если дискретный пропуск является экстраполируемым, то для его восстановления целесообразнее применять метод кубической сплайн-интерполяции, равно как и для импутации серийного интерполируемого пропуска, длительность которого не превышает получасового периода. Серийный экстраполируемый пропуск был восстановлен с помощью метода скользящей средней, хотя значение среднеквадратической ошибки слишком велико и результат импутации недостоверен.

Важно отметить, что дискретные пропуски обоих типов восстанавливаются методами линейной интерполяции и кубической сплайн-интерполяции с равными показателями точности, поэтому могут в равной степени использоваться для восстановления пропущенных значений во временном ряду.

Кроме того, качество восстановления серийных пропусков определяется не только применяемым для импутации методом, но и внешними факторами, в частности, состоянием магнитосферы в соответствующий период времени. Так, указанные методы показывают достаточно хорошие метрики качества восстановления данных при условии спокойной магнитосферы (при значении индекса Kp, не превышающего 2). Однако дополнительно проведенные экспериментальные исследования показали, что среднеквадратическая ошибка восстановления данных увеличивается по мере возрастания значения коэффициента Kp и принимает значения порядка сотни нТл уже при Kp = 4.

Таким образом, результаты проведенного анализа известных методов восстановления временных рядов применительно к геомагнитным данным могут быть интерпретированы следующим образом:

1) дискретные интерполируемые пропуски должны быть восстановлены методом скользящей средней с шириной окна, равной 3. При этом внешние факторы в виде окружающей магнитной обстановки, количественно оцениваемой посредством Kp-индексов, не влияют на результаты импутации.

2) дискретные экстраполируемые пропуски должны быть восстановлены методом кубической сплайн-интерполяции. Как и в предыдущем случае значение Kp-индекса, отражающее состояние магнитосферы в момент регистрации пропуска, не влияет на результат импутации.

3) серийные интерполируемые пропуски в условиях спокойной магнитосферы (при Kp < 3) также должны быть восстановлены методом кубической сплайн-интерполяции, который в ходе проведенных экспериментов демонстрировал стабильно лучший результат;

4) серийные интерполируемые пропуски в условиях спокойной магнитосферы (при Kp < 3) должны быть восстановлены методом скользящей средней с шириной окна, равной 3, также показавшего лучший результат в серии экспериментов по восстановлению временных рядов геомагнитных данных в условиях спокойной магнитосферы.

Изменения уровней временного ряда в условиях неспокойной магнитной обстановки (при Kp > 2) сопровождаются характерными нерегулярными осцилляциями, затрудняющими восстановление пропусков в соответствующие временные интервалы. Сложное изменение характера информационного сигнала проявляется особенно на серийных сегментах ряда. При этом ни один из представленных методов не показал приемлемого результата по восстановлению временных рядов в таких условиях.

В этой связи возникает актуальная задача, заключающаяся в разработке и формализации нового метода восстановления временных рядов геомагнитных данных, обеспечивающего лучшие показатели качества при импутации данных в условиях неспокойной магнитной обстановки.

5. Метод прецедентного резервирования. Для восстановления временных рядов геомагнитных данных в условиях возбужденной магнитосферы в данной работе авторами предлагается метод, получивший название прецедентного резервирования. В его основе лежит концепция индуктивного обучения, заключающаяся в выявлении общих закономерностей по частным эмпирическим данным [17].

Ключевой идеей метода является предположение, что любому сегменту временного ряда можно с некоторой допустимой степенью точности поставить в соответствие один или несколько фрагментов предшествующих ему значений уровня (рисунок 4). В этом случае накопленные магнитной станцией статистические данные выступают в качестве базы прецедентов, где каждый сектор 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.

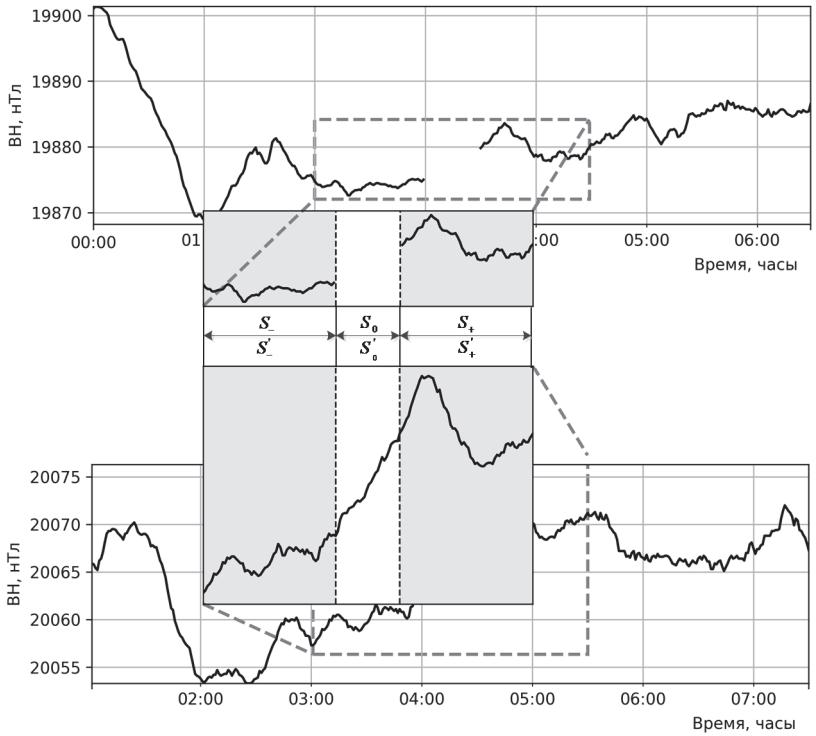


Рис. 4. Схема метода прецедентного резервирования

Пусть S — восстанавливаемая выборка, представленная тройкой из сегмента отсутствующих значений временного ряда, а также предшествующего и последующего за ним сегментов заданной длины:

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

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

Пусть S' — предшествующие восстанавливаемой выборке сегменты временного ряда, именуемые статистическими. 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'_+):

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

Разделяющие каждую пару сегменты 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.

Пусть шаблон поиска релевантных статистических сегментов представлен восстанавливаемой выборкой с вычтенным из нее сегментом отсутствующих значений, а обрабатываемая выборка — формируемой на каждой итерации перебора парой статистических сегментов. Тогда с учетом введенных обозначений с помощью коэффициента корреляции Пирсона [18] можно определить степень линейной зависимости между исследуемым сегментом и шаблоном поиска r_{xy} как:

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

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.

Как и в большинстве задач данного рода, полученные оценки восстановления данных являются смещенными. В этой связи результаты, получаемые посредством указанного метода, должны быть нормализованы таким образом, чтобы результирующие значения были аппроксимированы относительно известных соседних пропуску значений уровней временного ряда. Соответствующая аппроксимация выполняется посредством метода наименьших квадратов, применение которого целесообразно ввиду единичного лага рассматриваемого временного ряда.

Задача аппроксимации в конечном счете сводится к определению значений коэффициентов линейной зависимости, при которых функция двух переменных a и b принимает наименьшее значение. Иными словами, при полученных a и b сумма квадратов отклонений экспериментальных данных от найденной прямой будет наименьшей:

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

Полученные методом наименьших квадратов коэффициенты a и b являются искомыми составляющими аппроксимации полученной выборки, необходимыми для нивелирования полученной в ходе применения метода смещенной оценки.

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:

$$M = x_0 + L\left(\frac{\frac{1}{2}\sum_{i=1}^{K} f_i - N_{\text{prev}}}{f_M}\right),$$

где M — медианное значение, x_0 — начальное значение медианного интервала, L — длина медианного интервала, K — длина временного ряда, $\sum_{i=1}^{K} f_i$ — 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 $L = 3$ [19].

Важно отметить, что применение метода прецедентного резервирования в соответствии с принципом бритвы Оккама [20] 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 числа значений уровня в периоде упреждения. Так, к примеру, пропуск длиной в 30 значений формирует восстанавливаемую выборку из 90 последовательных значений уровня временного ряда геомагнитных данных со среднеквадратической ошибкой порядка 0.4 нТл.

Для оценки эффективности предложенного авторами метода была проведена серия экспериментов по восстановлению геомагнитных данных, пропущенных при различных значениях индекса Kp. Исследования проводились для серийных интерполируемых пропусков длиной в 30 значений (минут). Оценка была проведена применительно к двум методам: используемому в настоящее время методу линейной интерполяции и предложенному методу прецедентного резервирования. Полученные в ходе эксперимента результаты представлены на рисунке 5.

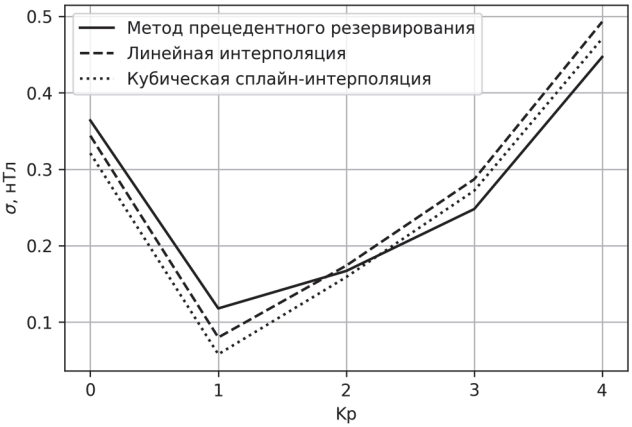


Рис. 5. Сравнительный анализ зависимости величины среднеквадратической ошибки от значения индекса Kp при восстановлении геомагнитных данных методами линейной интерполяции, кубической сплайн-интерполяции и прецедентного резервирования

Как видно из рисунка 5, в условиях спокойной магнитосферы (при $Kp \le 2$) метод линейной интерполяции обеспечивает меньшую среднеквадратическую ошибку, чем метод прецедентного резервирования. При приближении Kp к значению 2 величины среднеквадратической ошибки выравниваются, а с достижением индексом значения 2 метод прецедентного резервирования показывает лучший ре-

зультат 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 ошибки импутации увеличивается в несколько раз и процедура восстановления данных теряет смысл.

6. Методика скользящего регенерирующего окна. Распространенный на сегодняшний день подход к восстановлению пропусков геомагнитных данных предполагает применение линейной интерполяции ко всему временному ряду независимо от типа пропуска и состояния магнитосферы в соответствующий временной период. Вместе с тем результаты проведенного авторами анализа показывают, что не все методы одинаково эффективны и значение среднеквадратической ошибки при их применении зависит от того, к какому типу относится восстанавливаемый пропуск и при каких внешних условиях он был зарегистрирован.

В этой связи в данной работе авторами предложен подход, получивший название скользящего регенерирующего окна. Его суть заключается в том, что временной ряд размерности n с m пропусками разбивается на m пересекающихся окон-сегментов, каждое из которых включает ровно один пропуск и определенные значения уровня, предшествующие и следующие за ним в ряду.

Размер скользящего окна определяется размерностью пропуска, при этом в случае интерполируемого пропуска длины сегментов, предшествующих и следующих за пропуском, равны между собой и составляют величину, равную длине отсутствующего сегмента. Если пропуск экстраполируемый, то длины сегментов, предшествующих и следующих за пропуском, не равны между собой и составляют доступную для каждого конкретного случая величину, меньшую или равную длине отсутствующего сегмента.

В ходе восстановления данных по предложенной методике каждое скользящее окно формирует независимый временной ряд, для восстановления пропуска в котором применяется тот метод, который наилучшим образом подходит для импутации отсутствующего значения соответствующего типа в определенной магнитной обстановке.

Так, к примеру, для восстановления дискретного интерполируемого пропуска подбирается скользящее окно, образующее временной ряд с тремя значениями уровня. Поскольку дискретные значения не зависят от магнитной обстановки, восстановление такого пропуска выполняется методом скользящей средней. Если во временном ряду пропущен сегмент из 30 значений, а последующий и предыдущий за ним сегменты того же размера содержат только определенные значения, то имеет место временной ряд данных, содержащий 90 значений

уровня. В случае, когда такой пропуск был зарегистрирован в условиях спокойной магнитосферы, он восстанавливается методом линейной интерполяции, иначе для импутации отсутствующих значений временного ряда используется метод прецедентного резервирования.

Очевидно, что в результате последовательного обхода временного ряда данных и разделения его на промежуточные временные ряды импутация отсутствующих значений обеспечивает лучшие показатели качества, чем при использовании одного метода, поскольку учитывает особенности каждого пропуска и условия его регистрации. Для подтверждения выдвинутого предположения об эффективности предложенной методики была проведена серия экспериментов, результаты одного из которых приводятся далее.

7. Верификация методики скользящего регенерирующего окна. Экспериментальный временной ряд представлен геомагнитными данными, полученными при измерении горизонтальной компоненты геомагнитного поля BH станцией DOU 1 января 2017 года (рисунок 6).

Оригинальный массив данных (рисунок 6*а*) не содержит пропусков и отражает суточные вариации параметров постоянного магнитного поля при изменении значения Kp-индекса от 2 от 4. Для анализа эффективности предложенной методики восстановления данных во временной ряд были введены 42 дискретных и 3 серийных интерполируемых пропуска. Дискретные пропуски были равномерно распределены по временному ряду, а серийные пропуски подобраны таким образом, чтобы индекс Kp в соответствующие периоды принимал значения 4, 3 и 2 соответственно.

Как показал анализ методов восстановления временных рядов, наименьшую среднеквадратическую ошибку при восстановлении дискретных интерполируемых пропусков геомагнитных данных обеспечивает метод скользящей средней с шириной окна, равной 3.

Выбор метода восстановления серийного пропуска зависит от значения Kp-индекса в период, когда был зафиксирован пропуск (рисунок 5). В рассматриваемом временном ряду определены 30-минутный пропуск при Kp = 4, 20-минутный пропуск при Kp = 3 и 10-минутный пропуск при Kp = 2. Согласно проведенным исследованиям, ожидается, что первые два пропуска могут быть восстановлены с минимальной среднеквадратической ошибкой с помощью метода прецедентного резервирования, а последний — посредством кубической сплайн-интерполяции.

Таким образом, в соответствии с методикой скользящего регенерирующего окна лучшую метрику качества восстановления экспериментального временного ряда обеспечивает последовательное применение следующих методов импутации пропусков:

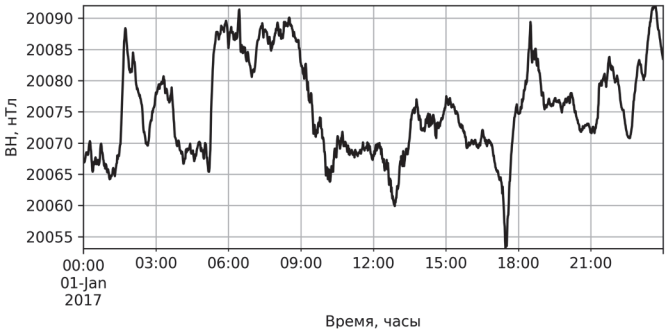
1) скользящая средняя для дискретных пропусков;

2) прецедентное резервирование для 30-минутного пропуска при Kp = 4;

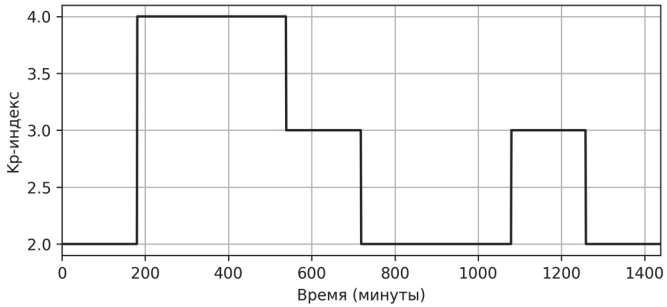
3) прецедентное резервирование для 20-минутного пропуска при Kp = 3;

4) кубическая сплайн-интерполяция для 10-минутного пропуска при Kp = 2.

а)



б)



в)

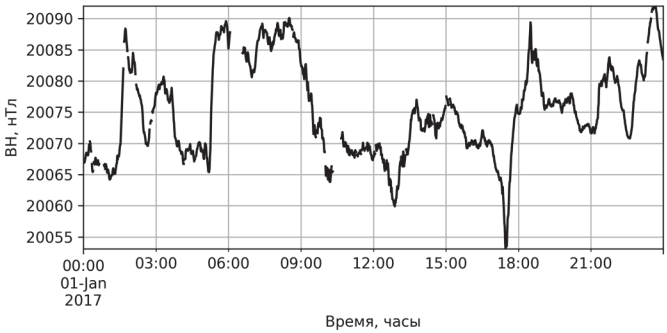


Рис. 6. Экспериментальный ряд геомагнитных данных, зарегистрированных станцией DOU 1 января 2017 года: *а* — оригинальный временной ряд; *б* — изменение Kp-индекса за 1 января 2017 года; *в* — временной ряд с искусственно введенными пропусками данных

Среднеквадратическая ошибка, полученная в результате восстановления экспериментального временного ряда геомагнитных данных посредством предложенных методики и метода, составила 0.219 нТл (рисунок 7). При этом 30-минутный пропуск при Kp = 4 и 20-минутный пропуск при Kp = 3 были восстановлены со значениями среднеквадратической ошибки 0.748 нТл и 0.445 нТл соответственно (рисунок 8).

Для сравнения тот же временной ряд был восстановлен с помощью двух известных подходов: используемого в настоящее время метода восстановления геомагнитных данных — линейной интерполяции, а также метода кубической сплайн-интерполяции, хорошо зарекомендовавшего себя при импутации небольших пропусков в условиях спокойной магнитосферы.

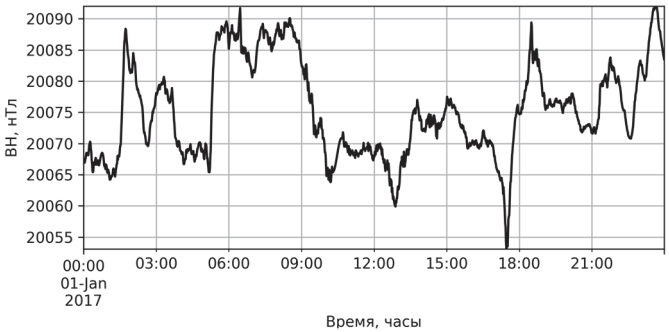


Рис. 7. Результат восстановления экспериментального временного ряда геомагнитных данных в соответствии с методикой скользящего регенерирующего окна

Импутация геомагнитных данных указанными методами показала худшую по сравнению с методикой скользящего регенерирующего окна метрику качества восстановления ряда, при этом общая среднеквадратическая ошибка составила 2.428 нТл и 2.315 нТл. Серийные интерполируемые 30-минутный пропуск при Kp = 4 и 20-минутный пропуск при Kp = 3 были восстановлены со значениями среднеквадратической ошибки соответственно 4.659 нТл и 3.491 нТл методом линейной интерполяции и 4.653 нТл и 3.487 нТл методом кубической сплайн-интерполяции (рисунок 8).

Таким образом, анализ результатов восстановления экспериментального временного ряда показал, что выбор оптимального способа импутации пропусков зависит от их типа и внешних факторов (значения индекса Kp). Так, дискретные пропуски могут быть восстановлены методом скользящей средней независимо от магнитной обстановки (и соответственно значения Kp-индекса). Серийные пропуски различной длительности рекомендуется восстанавливать методом кубической

сплайн-интерполяции (в случае спокойной магнитной обстановки при Kp < 3) или методом прецедентного резервирования (в случае неспокойной магнитной обстановки при Kp > 3). Применение единственного метода импутации без учета указанных факторов приводит к существенному искажению значений восстанавливаемого временного ряда, что независимо от формы анализируемого сигнала приводит к ошибочным выводам касательно локальной магнитной обстановки в отдельно взятой пространственной точке (в рамках магнитной обсерватории).

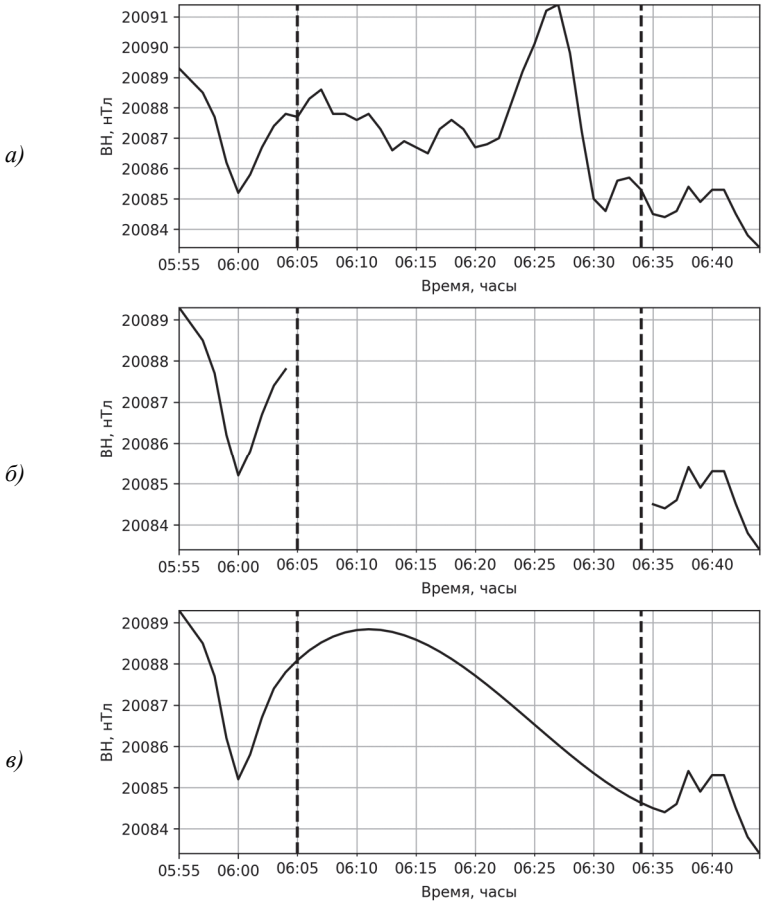


Рис. 8. Восстановление 30-минутного серийного интерполируемого пропуска геомагнитных данных при Kp = 4: *а* — оригинальные данные; *б* — данные с 30-минутным пропуском; *в* — результат восстановления методом кубической сплайн-интерполяции (среднеквадратическая ошибка = 4.653)

г)

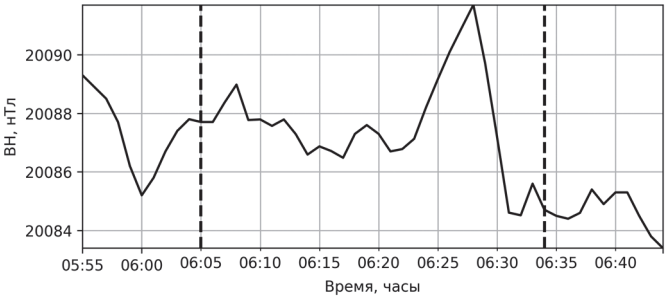


Рис. 8. Восстановление 30-минутного серийного интерполируемого пропуска геомагнитных данных при Kp = 4: *г* — результат восстановления методом прецедентного резервирования (среднеквадратическая ошибка = 0.748)

Основанная на данном принципе методика скользящего регенерирующего окна обеспечила лучшую метрику качества восстановления геомагнитных данных по сравнению с используемым для этих целей в настоящее время методом линейной интерполяции и методом, зарекомендовавшим себя в условиях спокойной магнитосферы — методом кубической сплайн-интерполяции.

8. Заключение. Одной из особенностей временных рядов геомагнитных данных является 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 приводят к сложным скачкообразным изменениям уровней временного ряда и разрыву линий тренда, нарушению их цикличности и периодичности.

Вместе с тем каждая магнитная станция располагает архивами выполненных ей поминутных измерений параметров геомагнитного поля, в совокупности составляющих базу ее прецедентов. Подтверждаемый коррелограммами линейный характер зависимости соседних значений уровня временного ряда позволяет сравнивать сегменты временного ряда по степени их корреляции и на основании этого считать их схожими и взаимозаменяемыми. Очевидно, что чем больше база прецедентов магнитной станции, тем больше корреляция между восстанавливаемым и замещающим сегментами и выше точность импутации значений временного ряда.

Данный подход положен в основу предложенного в настоящей работе метода прецедентного резервирования пропущенных значений временных рядов геомагнитных данных, базирующегося на индуктивном методе обучения по прецедентам и отличающимся тем, что в качестве признаков прецедентов используются данные, предшествую-

щие и последующие за пропуском во временном ряду. 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 данных такого вида.

Кроме того, в настоящей работе авторами предложен и верифицирован подход к восстановлению пропущенных значений временного ряда геомагнитных данных, получивший название методики скользящего регенерирующего окна, суть которого сводится к декомпозиции временного ряда на множество пересекающихся сегментов, количество которых равно числу пропусков в ряду. К каждому выделенному сегменту применяется тот метод восстановления данных, который является наиболее эффективным применительно к выявленному типу пропуска и в условиях заданного значения Kp-индекса геомагнитной активности. Проведенные численные эксперименты показали, что восстановления временного ряда геомагнитных данных посредством данной методики позволяет повысить точность импутации по сравнению с подходом, при котором все пропуски восстанавливаются одним выбранным методом (например, используемым в настоящее время в геофизике методом линейной интерполяции).

По результатам проведенных исследований ожидается, что предлагаемые авторами метод и методика позволят повысить качество обработки и интерпретации геомагнитных данных, регистрируемых постоянными магнитными станциями, за счет более точного восстановления пропущенных при этом значений независимо от состояния магнитосферы в соответствующий период времени.

Литературы

1. *Мандрикова О.В., Жижикина Е.А.* Автоматический способ оценки состояния геомагнитного поля // Компьютерная оптика. 2015. Т. 39. № 3. С. 420–428.
2. INTERMAGNET technical reference manual. Version 4.6 / edited by S.-L. Benoît. Edinburgh: INTERMAGNET. BGS. 2012. 100 p.
3. *Love J.J., Chulliat A.* An international network of magnetic observatories // Eos, Transactions, American Geophysical Union. 2013. vol. 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. С. 3–20.
6. *Mandea M., Korte M.* Geomagnetic Observations and Models // Springer. 2011. 343 p.
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. С. 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. *Damodar N.* Gujarati. Basic Econometrics // The McGraw-Hill Companies. 2004. 1002 p.
11. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс // М.: Дело. 2007. 504 с.
12. *Лукашин Ю. П.* Адаптивные методы краткосрочного прогнозирования временных рядов // М.: Финансы и статистика. 2003. 416 с.
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. *De Gooijer J.G.* Elements of nonlinear time series analysis and forecasting // Springer. 2017. 618 p.
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 // Springer. 2008. 190 p.
17. *Чучуева И.А.* Модель прогнозирования временных рядов по выборке максимального подобия: диссертация канд. тех. наук // М.: Московский государственный технический университет им. Н.Э. Баумана. 2012. 154 с.
18. *Box G. et al*. Time series analysis: forecasting and control // New York: John Wiley & Sons. 2017. 712 p.
19. *Кобзарь А. И.* Прикладная математическая статистика // М.: Физматлит. 2006. 403 с.
20. *Langford J.* Quantitatively tight sample complexity bounds // Carnegie Mellon Thesis. 2002. 130 p.
21. *Moritz S., Sard´a A., Bartz-Beielstein T.* Comparison of different Methods for Univariate Time Series Imputation in R // arXiv preprint arXiv:1510.03924. 2015.
22. *Литтл Р.Дж.А., Рубин Д.Б.* Статистический анализ данных с пропусками // М. 1991. 336 с.
23. *Злоба Е., Яцкив И.* Статистические методы восстановления пропущенных данных // Computer Modelling & New Technologies. 2002. vol. 6. no. 1. pp. 51–61.

Воробьев Андрей Владимирович — к-т техн. наук, доцент, доцент кафедры геоинформационных систем факультета информатики и робототехники, ФГБОУ ВО Уфимский государственный авиационный технический университет (УГАТУ). Область научных интересов: геоинформационные технологии, цифровая обработка сигналов. Число научных публикаций — 138. geomagnet@list.ru, http://www.geomagnet.ru; ул. К. Маркса, 12, Уфа, 450008; р.т.: +7(917)345-2299.

Воробьева Гульнара Равилевна — к-т техн. наук, доцент кафедры вычислительной математики и кибернетики факультета информатики и робототехники, ФГБОУ ВО Уфимский государственный авиационный технический университет (УГАТУ). Область научных интересов: геоинформационные и веб-технологии, системы хранения и обработки информации. Число научных публикаций — 114. gulnara.vorobeva@gmail.com, http://www.geomagnet.ru; ул. К. Маркса, 12, Уфа, 450008; р.т.: +7(917)417-4111.

DOI 10.15622/sp.57.5

A.V. VOROBEV, G.R. VOROBEVA

INDUCTIVE METHOD OF GEOMAGNETIC DATA TIME SERIES RECOVERING

Vorobev A.V., Vorobeva G.R. **Inductive Method of Geomagnetic Data Time Series Recovering.**

Abstract. Today intensive development of systems and technologies for registration of the Earth's magnetic field parameters causes an exponential increase of geomagnetic data quantity, mainly collected by the ground magnetic stations. Imperfection of applied equipment and enabled channels of information transfer leads to the presence of omissions in the registered data time series. Along with spatial anisotropy it creates a serious obstacle to the processing of geomagnetic data. Russian and foreign scientific organizations are used to recover missing geomagnetic data by the linear interpolation. The approach provides admissible results in conditions of a quiet magnetosphere, but significantly distorts time series when changing the surrounding magnetic environment. This fact causes a scientific and technical problem, which is concerned with the development of new approach to recovering geomagnetic data registered in unquiet magnetosphere with acceptable time series imputation quality metrics.

The authors suggest the approach for time series recovering based on inductive method of machine learning. According to the approach each magnetic station operates its own knowledge base, which is formed during the registration of geomagnetic field and its variations parameters. The combination of the values of the series preceding and following the gap is supposed to be a characteristic description, which is used for searching the precedent in the magnetic station knowledge base. The result contains the required fragment of the time series, which replaces the missing values. The complexity of the information signal, caused by an unquiet magnetic environment, increases the accuracy of search by precedents. The greater the knowledge base of the magnetic station, the higher the effectiveness of the search.

Analysis of the results obtained during gap recovering in geomagnetic data time series (registered in conditions of unquiet magnetosphere) demonstrated that the suggested inductive method of imputation allows increasing the accuracy of the missing values recovery by an average of 79.54% compared with the methods currently used. The approach will enhance the efficiency of geomagnetic data processing for solving applied problems.

Keywords: geomagnetic data, time series, missing values, machine learning, learning by precedents, time series imputation.

Vorobev Andrei Vladimirovich — Ph.D., associate professor, associate professor of geoinformation systems department of computer science and robotics faculty, Ufa State Aviation Technical University (USATU). Research interests: geoinformation technologies, digital signal processing. The number of publications — 138. geomagnet@list.ru, http://www.geomagnet.ru; 12, K. Marx St., Ufa, 450008, Russia; office phone: +7(917)345-2299.

Vorobeva Gulnara Ravilevna — Ph.D., associate professor of computational mathematics and cybernetics department of computer science and robotics faculty, Ufa State Aviation Technical University (USATU). Research interests: geoinformation and web technologies, systems of information storing and processing. The number of publications — 114. gulnara.vorobeva@gmail.com, http://www.geomagnet.ru; 12, K. Marx St., Ufa, 450008, Russia; office phone: +7(917)417-4111.

References

1. Mandrikova O.V., Zhizhikina E.A. [Automated approach to estimate geomagnetic field state]. *Komp'juternaja optika – Computer Optics*. 2015. vol. 39. no. 3. pp. 420–428. (In Russ.).

2. INTERMAGNET technical reference manual. Version 4.6. Edited by S.-L. Benoît. Edinburgh: INTERMAGNET. BGS. 2012. 100 p.
3. Love J.J., Chulliat A. An international network of magnetic observatories. *Eos, Transactions, American Geophysical Union*. 2013. vol. 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. Gvishiani A.D., Luk'janova R.Ju. [Geoinformatics and the Earth's magnetic field observations: Russian segment]. *Fizika Zemli – Physics of the Earth*. 2015. vol. 2. pp. 3–20. (In Russ.).
6. Mandea M., Korte M. Geomagnetic Observations and Models. Springer. 2011. 343 p.
7. Rybkina A.I. et al. [Interpolation of observatory measurements data and visualization of the Earth's magnetic field intensity]. *Vestnik Otdelenija nauk o Zemle RAN – Bulletin of the Department of Earth Sciences of the Russian Academy of Sciences*. 2013. vol. 5. no. 3002. pp. 1–4. (In Russ.).
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. Livshic M. *Sluchajnye processy – ot teorii k praktike* [Random processes – from theory to practice]. M.: Lan'. 2016. 320 p.
11. Aue A., Dubart D., Hörmann S. On the prediction of stationary functional time series. *Journal of the American Statistical Association*. 2015. vol. 110(509). pp. 378–392.
12. Lukashin Ju. P. *Adaptivnye metody kratkosrochnogo prognozirovanija vremennyh rjadov* [Adaptive methods of short-term forecasting of time series]. Moscow: Finance and Statistics. 2003. 416 p.
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. De Gooijer J.G. Elements of nonlinear time series analysis and forecasting. Springer. 2017. 618 p.
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. Springer. 2008. 190 p.
17. Chuchueva I.A. *Model' prognozirovanija vremennyh rjadov po vyborke maksimal'nogo podobija. Dissertacija kand. teh. nauk* [Model of forecasting of time series on a sample of the maximum similarity. Ph.D. Thesis]. Moscow. 2012. 154 p. (In Russ.).
18. Box G. et al. Time series analysis: forecasting and control. New York: John Wiley & Sons. 2017. 712 p.
19. Kobzar A.I. *Prikladnaja matematicheskaja statistika* [Applied Mathematical Statistics]. Moscow: Physmatlit. 2006. 403 p. (In Russ.).
20. Langford J. Quantitatively tight sample complexity bounds. Carnegie Mellon Thesis. 2002. 130 p.
21. Moritz S., Sard´a A., Bartz-Beielstein T. Comparison of different Methods for Univariate Time Series Imputation in R. arXiv preprint arXiv:1510.03924. 2015.
22. Little R.J.A., Rubin D.B. *Statistical analysis with missing data*. John Wiley & Sons, 1987. 278 p. (Russ. ed.: Littl R.Dzh.A., Rubin D.B. *Statisticheskij analiz dannyh s propusk*ami. Moscow. 1990. 336 p.).
23. Zloba E., Yackiv I. [Statistical methods for recovering missing data]. *Computer Modelling & New Technologies*. 2002. vol. 6. no. 1. pp. 51–61.