ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2022 Управление, вычислительная техника и информатика № 61

Tomsk State University Journal of Control and Computer Science

Научная статья
УДК 004.91
doi: 10.17223/19988605/61/4

Подход к восстановлению временных рядов пространственных данных на основе информационно-корреляционного и геостатистического анализа

Гульнара Равилевна Воробьева[1], Андрей Владимирович Воробьев[2]

[1, 2] Уфимский государственный авиационный технический университет, Уфа, Россия
[1] gulnara.vorobeva@gmail.com
[2] cpu16bit@gmail.com

Аннотация. Обсуждаются вопросы повышения эффективности процессов обработки и анализа пространственных данных в контексте решения задачи восстановления соответствующих временных рядов с точностью, не превышающей заданной допустимой величины. Предложен подход, основанный на резервировании распределенных источников пространственных данных в соответствии с результатами их теоретико-информационного анализа, с одной стороны, и геостатистического анализа – с другой. На примере результатов мониторинга параметров геомагнитного поля показана эффективность применения предложенного подхода.

Ключевые слова: пространственные данные; обработка данных; теоретико-информационный анализ; геостатистический анализ; восстановление временных рядов

***Благодарности*:** Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-07-00011.

***Для цитирования*:** Воробьева Г.Р., Воробьев А.В. Подход к восстановлению временных рядов пространственных данных на основе информационно-корреляционного и геостатистического анализа // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2022. № 61. С. 37–46. doi: 10.17223/19988605/61/4

Original article
doi: 10.17223/19988605/61/4

An approach to the recovery of spatial data time series based on information-correlation and geostatistical analysis

Gulnara R. Vorobeva[1], Andrei V. Vorobev[2]

[1, 2] Ufa State Aviation Technical University, Ufa, Russian Federation
[1] gulnara.vorobeva@gmail.com
[2] cpu16bit@gmail.com

Abstract. The issues of increasing the efficiency of processing and analysis of spatial data in the context of solving the problem of recovering the corresponding time series with an accuracy not exceeding a given allowable value are discussed. An approach based on redundancy of distributed sources of spatial data in accordance with the results of their information-theoretic analysis, on the one hand, and geostatistical analysis, on the other, is proposed. Using the results of monitoring the parameters of the geomagnetic field as an example, the effectiveness of the proposed approach is shown.

Keywords: spatial data; data processing; theoretical information analysis; geostatistical analysis; recovery of time series

***Acknowledgments*:** The research was carried out with the financial support of the Russian Foundation for Basic Research *№ 20-07-00011.*

***For citation*:** Vorobeva, G.R., Vorobev, A.V. (2022) An approach to the recovery of spatial data time series based on information-correlation and geostatistical analysis. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 61. pp. 37–46. doi: 10.17223/19988605/61/4

Одним из магистральных направлений развития технологий обработки пространственных данных является применение методов геостатистического анализа для оценки их пространственно-временной анизотропии в различных прикладных областях науки и техники. Вместе с тем, к примеру, детерминистические модели и методы пространственной интерполяции, крикинга и кокрикинга, моделирования пространственной неопределенности предполагают непрерывность временных рядов данных, синхронно регистрируемых пространственно распределенными источниками данных (геодезическими опорными пунктами) [1. С. 50]. Однако несовершенство используемого для регистрации пространственных данных информационно-измерительного оборудования, сбои в каналах передачи информации, обусловленные причинами различной природы, а также ошибки, вызванные человеческим фактором, приводят к тому, что в настоящее время практически не существует источников данных, гарантирующих их бесперебойную регистрацию.

Ярким примером обозначенной проблемы являются наземные магнитные обсерватории и вариационные станции, которые в режиме реального времени осуществляют регистрацию параметров магнитного поля Земли и его вариаций. При этом проведенные авторами исследования показали, что годовые архивы соответствующих геомагнитных наблюдений содержат от 22,36 до 35,12% пропущенных значений, что крайне негативно сказывается на работе информационных систем, функционирование которых основано на применении указанных геопространственных данных [2. P. 627].

Представляется целесообразным отметить, что в силу ряда причин (преимущественно экономического характера) решить проблему совершенствованием технического обеспечения измерительного (и регистрационного) процесса не представляется возможным, так же как и нивелировать негативное воздействие человеческого фактора. В этой связи необходима и актуальна разработка подхода, позволяющего восстановить временные ряды пространственных данных с точностью, значение которой должно быть максимально приближено к значению, регламентируемому действующими в соответствующей прикладной области деятельности стандартами, спецификациями и рекомендациями. Так, к примеру, в случае геомагнитных данных в качестве такого документа может быть использована спецификация IAGA (International Association of Geomagnetism and Aeronomy – Международная ассоциация геомагнетизма и аэрономии) [3. P. 9], которая определяет допустимую ошибку при регистрации параметров геомагнитного поля и его вариаций не более 1 нТл.

В настоящее время широко известны методы восстановления временных рядов данных, обеспечивающие с некоторой точностью импутацию недостающих единичных значений и целых фрагментов. К примеру, для единичных значений часто применяется сглаживание временного ряда методом скользящей средней, предполагающее замену отсутствующего фрагмента данных усредненным значением соседних по отношению к нему элементов [4. P. 137]. При тех же условиях может быть применена линейная интерполяция, основанная на подборе заданного уравнением прямой полинома первой степени на базе известных значений уровня временного ряда геомагнитных данных [4. P. 137]. Для более длинных фрагментов известны решения, основанные на модели авторегрессии (AR) первого порядка, которая обеспечивает возможность прогнозирования отсутствующих фрагментов временного ряда на основании предшествующих ему значений [4. P. 141]. Аналогично для длительных фрагментов может быть применена интегрированная модель авторегрессии – скользящего среднего (ARIMA), которая обеспечивает лучшую по сравнению с AR метрику качества прогнозирования пропущенного фрагмента за счет гибкой параметризации обработки данных [4. P. 142].

Вместе с тем анализ эффективности перечисленных методов восстановления временных рядов геомагнитных данных (был проанализирован 10-минутный фрагмент) показал невозможность их применения в соответствующих информационных системах ввиду возникновения ошибки, превышающей допустимое спецификациями значение в 1 нТл [5. P. 1062]. К примеру, метод скользящей

средней обеспечивает среднеквадратическую ошибку восстановления данных 1,3 нТл. Метод линейной интерполяции хорошо показал себя при восстановлении единичных пропусков (среднеквадратическая ошибка составила 0,03 нТл), но при увеличении восстанавливаемого фрагмента до 10 значений величина среднеквадратической ошибки превысила 1,43 нТл. Модели AR и ARIMA также показали хорошие результаты на небольших фрагментах, обеспечив восстановление 5-минутного фрагмента с приемлемой точностью в 0,7 нТл, но при увеличении импутируемого фрагмента ошибка возросла до 1,8 нТл.

При этом важно отметить, что известные методы восстановления временных рядов пространственных данных никак не учитывают характерные для процессов особенности статистических и энтропийных характеристик, без понимания которых затруднительна оценка соответствующих параметров в контексте внешних для них факторов. В частности, при восстановлении геомагнитных данных целесообразно учитывать недетерминированную зависимость характера изменения параметров магнитного поля от состояния магнитосферы в соответствующий момент времени. Кроме того, важной характеристикой является пространственная анизотропия параметров геомагнитного поля и его вариаций, что также является одним из критериев, которые необходимо учитывать в процессе восстановления временных рядов.

1. Кластеризация источников пространственных данных на основе их корреляционных и энтропийных характеристик

В основе предлагаемого подхода к восстановлению временных рядов пространственных данных лежит оценка взаимного расположения их источников. В этой связи предлагается объединять в группы (кластеры) источники данных, которые по своим параметрам пространственной корреляции, а также взаимной информации могут считаться условно близкими.

На начальном этапе представим множество источников данных посредством выражений в соответствии с системой аксиом Цермело–Франкеля [6]:

$$A = \langle a_1, a_2, \dots, a_n \rangle, \quad (1)$$

где в интенсиональном описании a_i – группа, объединяющая источники пространственных данных (для заданной предметной области, например, геомагнитные данные) по всему миру.

Группы источников данных на начальном этапе делятся на глобальную, региональные и локальные в зависимости от конкретной прикладной задачи, для них применяются соответствующие данные:

$$A = \langle A^g, A^r, A^l \rangle, A^g = \langle a_1^g, \dots, a_k^g \rangle,\ A^r = \langle a_1^r, \dots, a_m^r \rangle,\ A^l = \langle a_1^l, \dots, a_p^l \rangle, \quad (2)$$

где A^g, A^r, A^l – глобальная, региональная и локальная группы источников данных соответственно.

Группы источников данных более низкого уровня могут представлять собой подмножества одной или нескольких групп более высокого уровня, что в терминах дополнений множеств может быть представлено как

$$A = A^l \cup \left(A^l / A^r\right) \cup (A^r / A^g), \quad (3)$$

где дополнение $\left(A^l / A^r\right)$ задано элементами множества A^r, не входящими во множество A^l, а дополнение (A^r / A^g) задано элементами множества A^g, не входящими во множество A^r. При этом $A^g \supset A^r, A^r \supset A^l$.

На последующем этапе источники данных, относящиеся к одной и той же пространственной группе (кластеру), подвергаются дополнительному агрегированию на основании соответствующих теоретико-информационных характеристик. Для этого анализируется пространственное распределение информационных характеристик соответствующих пространственных данных для оценки степени связности временных рядов данных с различной географической привязкой. Ключевой характеристикой здесь является информационная энтропия (энтропия Шеннона) $H(X)$, определяемая согласно выражению [7. P. 104]:

$$H(X) = \sum_{i=1}^{n} p_i \log_2 p_i, \quad (4)$$

где p_i – вероятность наблюдения значения x_i, X – временной ряд пространственных данных, представленный совокупностью значений $x_1,\ldots, x_n$.

Согласно выражению (4) энтропия является усредненной характеристикой сообщения – математическим ожиданием распределения случайной величины $i_1, i_2,...i_n$, где i_k – итерация наблюдений соответствующего параметра (содержательная / атрибутивная составляющая анализируемых пространственных данных). Являясь в определенной степени мерой рассеяния, информационная энтропия подобна дисперсии, но не зависит от типа распределения и характеризуется универсальностью и аддитивностью. Кроме того, энтропия, в отличие, к примеру, от корреляции, характеризует любую, в том числе нелинейную, связь переменных.

Вообще говоря, информационная энтропия Шеннона, по определению [7. P. 104], количественно характеризует произвольное распределение какого-либо параметра процесса. Если во внешней среде или в самой исследуемой системе происходят какие-либо изменения, то это приводит к изменению распределения ее параметров. В этом смысле информационная энтропия Шеннона может рассматриваться как функция состояния системы, поскольку количественно описывает меру неопределенности значений параметров, характеризующих систему [7. P. 104].

Анализ взаимных энтропийных характеристик ставит своей целью выявление статистически значимых зависимостей произвольного вида между приращениями пар временных рядов. Чем более отдалены друг от друга в пространстве временные ряды, тем менее выражены их взаимные теоретико-информационные характеристики.

При этом основной количественной мерой взаимных энтропийных характеристик является взаимная информация (Mutual Information), которая, согласно [8. P. 31], является мерой взаимной зависимости между двумя переменными. Она квантифицирует количество информации (по мере Хартли [8. P. 31]), полученное от одной величины при наблюдении другой:

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} p(x_i, y_j)\log_2\frac{p(x_i,y_j)}{p(x_i)p(y_j)}, \quad (6)$$

где $p(x_i)$ – вероятность появления значения x_i в точке X, $p(y_j)$ – вероятность появления значения y_j в точке Y, $p(x_i, y_j)$ – вероятность совместного появления значений x_i в точке X и значения y_j в точке Y

Вероятностные составляющие информационной энтропии могут быть интерпретированы таким образом, что чем меньше вероятность появления какого-либо значения исследуемого атрибутивного параметра, тем большую неопределенность снимает регистрирующее его появление сообщения и тем больше информации оно несет. Иными словами, по формуле Бриллюэна $H + I = 1$.

При этом, согласно [7. P. 104; 8. P. 31], для статистически независимых выборок условные энтропии $H(X/Y)$ и $H(Y/X)$ не пересекаются. В противном случае площадь их пересечения характеризует меру взаимной информации между ними. Чем больше величина взаимной информации, тем теснее связь и меньше величина условной энтропии $H(X/Y)$.

По результатам анализа величины взаимной информации внутри пространственного кластера могут выделяться дополнительные группы, представленные источниками данных с наиболее сильной связью. В итоге осуществляется декомпозиция исходного кластера на дочерние пространственные кластеры с соответствующими источниками данных.

На последующем этапе продолжается анализ выделенных пространственных кластеров источников данных на предмет оценки связывающего их корреляционного параметра – коэффициента информационной корреляции, который определяется согласно выражению

$$R(X,Y) = \sqrt{1 - e^{-2I(X,Y)}}, \quad (7)$$

где $I(X, Y)$ – мера взаимной информации для выборок данных, синхронно зарегистрированных источниками X и Y.

Свойства взаимной информации для источников данных X и Y полностью определяют свойства информационного коэффициента корреляции, показывая связь между соответствующим источниками данных. Так, к примеру, для независимых источников данных (независимых как друг от друга, так и от некоторых внешних факторов) $R(X, Y) = 0$.

Здесь представляется целесообразным отметить, что значения и информационной энтропии (на основании которой рассчитывается взаимная информация), и информационной корреляции для физически специфичных величин являются основанием для формулирования суждений о пространственной однородности физических процессов, данные о которых синхронно зарегистрированы источниками *X* и *Y*. Кроме того, перечисленные параметры в совокупности могут быть использованы в качестве индикатора, свидетельствующего о сопоставимом изменении величин *X* и *Y* в результате одинакового воздействия экзогенных факторов.

Полученные значения коэффициента информационной корреляции используются для последующей декомпозиции каждого пространственного кластера. В каждую группу дополнительно выделяются источники данных с наиболее сильной связью с точки зрения попарной информационной корреляции.

Предполагается, что источники данных, представленные в одном пространственном кластере, могут быть использованы для частичного взаимного информационного резервирования, а также применены в составе известных статистических моделей для восстановления недостающих значений соответствующих временных рядов. Однако известно, что проблема пространственной автокорреляции, указывающая на пространственную зависимость источников данных, может внести существенные искажения в процедуру применения стандартных статистических методов, предполагающих зависимость между наблюдениями [9. С. 330]. Так, к примеру, не компенсирующий пространственную зависимость регрессионный анализ может сопровождаться неустойчивыми оценками параметров, что способно привести к недостоверным результатам проводимых тестов значимости [9. С. 339].

Для исключения пространственной автокорреляции необходимо ввести расчет дополнительного геостатистического параметра, который показывает, носят ли случайный характер внешние процессы / факторы, определяющие данные, регистрируемые пространственно распределенными источниками. С этой целью выдвигается нулевая гипотеза, согласно которой пространственные процессы, обусловливающие анализируемые пространственные данные, носят случайный характер. Для подтверждения / опровержения данной нулевой гипотезы для каждой пары пространственных объектов (источников данных) рассчитывается значение глобального индекса Морана.

В этой связи при обработке геопространственных данных необходимо учитывать параметры пространственной неоднородности и пространственной зависимости источников данных [10]:

$$I_G = \frac{\sum_i \sum_j w_{ij}(x_i-\mu)(x_j-\mu)}{\sum_i (x_i-\mu)^2}, \quad (8)$$

где I_G – индекс Морана, x_i, x_j – значения параметра x в пространственных точках i и j, μ – среднее значение параметра x, w – экспертный весовой коэффициент.

В зависимости от соотношения значений I_G и $I(E)$ (где для n точек $I(E) = -1/(n-1)$) возможно определить, являются ли значения в соседних пространственных регионах подобными ($I_G > I(E)$), отличающимися ($I_G < I(E)$) или расположенными случайным образом ($I_G = I(E)$), и оценить зависимость описываемого процесса от экзогенных факторов

Значение индекса Морана может быть использовано для уточнения полученных на предшествующих этапах пространственных кластеров. Данные, синхронно регистрируемые территориально распределенными источниками, традиционно измеряются с помощью интервальных иколичественных шкал. Принимая во внимание их геопространственную привязку, теснота связи между ними может быть определена на основании анализа пространственной неоднородности и пространственной зависимости, что, в свою очередь, предполагает наличие пространственной корреляции (положительной или отрицательной) между пространственными наблюдениями, что и показывает индекс Морана.

Целесообразно заметить, что для корректного расчета коэффициента пространственной автокорреляции, а также перечисленных теоретико-информационных характеристик необходимо, чтобы количество значений, синхронно зарегистрированных в исследуемых распределенных источниках данных *X* и *Y*, было одинаковым, исследуемые временные ряды от *X* и *Y* были распределены нормально и измерены в интервальной шкале или шкале отношений [10]. Установлено, что совокупное

применение теоретико-информационных и геостатистических характеристик источников данных позволяет исключить ложную пространственную автокорреляцию, которая может внести существенные искажения в сформулированные на их основе выводы.

2. Описание предлагаемого подхода

Резюмируя описание приведенных выше методов и моделей теоретико-информационного и геостатистического анализа, представляется целесообразным обобщенное представление предлагаемого подхода к восстановлению временных рядов пространственных данных. Процедуре импутации пропущенных фрагментов временных рядов пространственных данных предшествует составление так называемого доверительного перечня резервных геодезических пунктов. Представленные в составе последнего источники данных являются основой для восстановления временного ряда пространственных данных, относящихся к анализируемому источнику данных. В случае сильной корреляционной связи наиболее близкий по доверительному списку источник данных может использоваться в качестве резервного для анализируемого. Замена пропусков исходного временного ряда нормализованными данными резервного источника данных на основе доверительного списка выполняется посредством сопоставления временных индексов и установления соответствия между ними. Выбранный фрагмент резервного временного ряда копируется под соответствующие временные индексы восстанавливаемого ряда, заменяя в нем обнаруженные пропуски.

Важно отметить, что полученные в ходе применения метода информационного резервирования результаты восстановления данных, как правило, являются смещенными и поэтому 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 [11. P. 56].

Составление доверительного списка источников данных – многоэтапный процесс последовательной пространственной кластеризации опорных геодезических пунктов [12, 13]. На начальном этапе пространственные кластеры формируются в соответствии со спецификой предметной области или прикладной задачи, для которой используются соответствующие временные ряды пространственных данных. Так, распределенные источники данных могут быть изначально кластеризованы, к примеру, по своей ведомственной принадлежности, используемой информационно-измерительной технике, территориальному распределению и пр. При этом допускается наличие как единственного глобального пространственного кластера, который представлен всеми доступными источниками данных, так и множества локальных и региональных пространственных кластеров, каждый из которых семантически отличается от остальных.

На последующем этапе отдельно рассматривается каждый из выделенных ранее пространственных кластеров. Осуществляется декомпозиция пространственного кластера на вложенные кластеры, каждый из которых, в свою очередь, представлен источниками данных, наиболее близкими по своим теоретико-информационным характеристикам. Для этого попарно для всех источников данных в соответствующем пространственном кластере последовательно рассчитываются значения информационной энтропии Шеннона и взаимной информации. Для источников данных с наилучшим показателями по взаимной информации формируются новые включающие их пространственные кластеры. При этом, как и в случае предшествующих этапов кластеризации, возможны ситуации, когда кластер представлен единственным источником данных, для которого не были обнаружены «близкие» к нему иные геодезические пункты.

Далее для выделенных пространственных кластеров осуществляется попарный анализ коэффициента информационной корреляции, в том числе на основании ранее рассчитанного значения по взаимной информации между теми же источниками данных. Результаты такого анализа также позволяют внутри кластеров выделить источники данных с наиболее сильной корреляционной связью, что, в свою очередь, является основой для формирования новых пространственных кластеров при декомпозиции исходного.

На следующем шаге попарно анализируются параметры пространственной неоднородности и пространственной зависимости источников данных внутри кластеров, сформированных на предшествующем этапе, для чего рассчитывается значение индекса Морана по соответствующим временным рядам пространственных данных. При этом выявляются те опорные геодезические пункты из исходного множества, вариации в данных которых определяются одними и теми же внешними факторами. В результате выполняется корректировка полученных ранее пространственных кластеров. Непосредственно при восстановлении временных рядов определяется принадлежность анализируемого источника данных соответствующему пространственному кластеру, на основании чего как непосредственно идентифицируется доверительный список опорных геодезических пунктов с учетом обозначенных энтропийных и геостатистических характеристик, так и определяется тот источник данных, который может быть рассмотрен в качестве резервного.

3. Анализ эффективности предложенного подхода

Оценка эффективности предложенного подхода восстановления временных рядов пространственных данных была выполнена на примере годовых архивов геомагнитных данных, регистрируемых пространственно распределенными магнитными обсерваториями и вариационными станциями. Рассматриваются фрагменты одномерного временного ряда пространственных данных, которые характеризуют минутные наблюдения вариаций горизонтальной составляющей вектора геомагнитного поля (BH, нТл). В качестве анализируемого источника данных была выбрана магнитная обсерватория DOU (Dourbes) глобальной сети магнитных обсерваторий INTERMAGNET (архив свободно распространяется и доступен по адресу https://intermagnet.org) [14. P. 2; 15. P. 110].

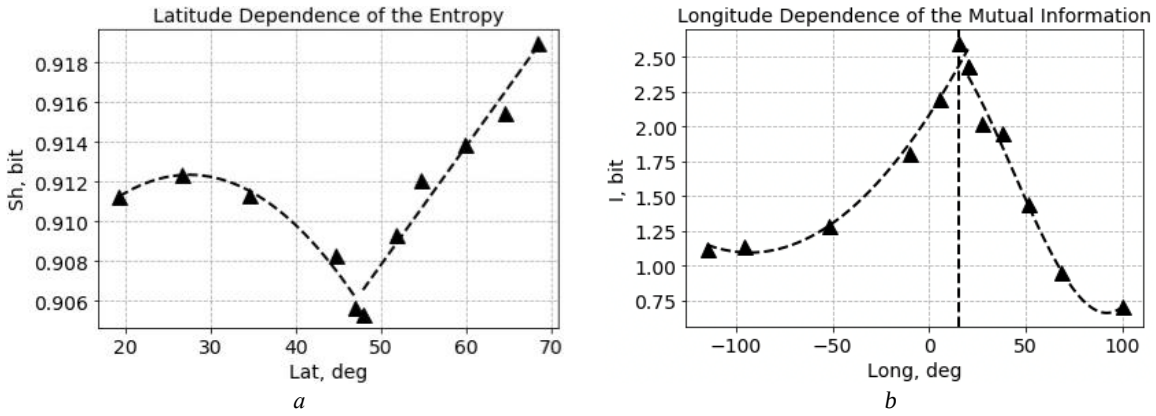
Множество обсерваторий рассматриваемой сети было подвергнуто процедуры многоэтапной пространственной кластеризации. Для этого на начальном этапе определяются взаимные энтропийные характеристики итерационного попарного сравнения основного источника данных с каждым доступным в глобальной группе источником данных, на основании которого анализируются информационная энтропия Шеннона, условная энтропия и основной показатель – взаимная информация. В результате для глобального множества источников данных, сформированного на предыдущем шаге, формируется множество, которое, с одной стороны, включает в себя анализируемый опорный геодезический пункт DOU, а с другой – содержит источники данных, показавшие максимальные или близкие к ним значения энтропийных характеристик (преимущественно рассматриваются значения взаимной информации).

Например, для геомагнитных данных в доверительный список попадают магнитные обсерватории и вариационные станции с близкими к единице показателями взаимной информации и условной энтропии, со значениями информационной энтропии, отличающимися не более чем на 0,01, а также относящиеся к одной пространственной группе и описываемые, как правило, одним статистическим законом распределения.

Примечательно, что расчет информационной энтропии Шеннона для каждого набора геомагнитных данных показал, что наименьшая неопределенность наблюдается в районе средних широт (46,9° N). При этом энтропия возрастает по направлению к низким широтам рис. 1). Аналогичные исследования, проведенные для других значений широт, показали повторяемость полученных зависимостей и подтвердили результаты анализа. Физическое объяснение полученных результатов теоретико-информационного анализа заключается в том, что значения параметров геомагнитного поля и его вариаций на средних широтах зависят от наименьшего количества факторов. При этом по мере приближения к экватору и полюсам число факторов, влияющих на результаты наблюдений, возрастает, что приводит к росту неопределенности данных о состоянии геомагнитного поля и его вариаций, но повышает информативность каждой итерации геомагнитного мониторинга. При этом установлено, что в пределах одного широтного диапазона энтропийные характеристики магнитных обсерваторий почти неизменны. Физический смысл полученных результатов можно интерпретировать таким образом, что на значения параметров геомагнитного поля и его вариаций в пределах одного широтного

диапазона оказывают влияние одни и те же внешние факторы (или их совокупность), что приводит к относительной неизменности информационной энтропии с изменением географической долготы.

Кроме того, при анализе значения условной энтропии было установлено его увеличение по мере пространственного удаления анализируемого источника данных от основного (в данном случае DOU). Аналогичная тенденция была обнаружена и при анализе значений взаимной информации для пар источников геомагнитных данных.



a *b*

Рис. 1. Широтная и долготная зависимости энтропийных характеристик геомагнитных данных
(*a* – информационная энтропия Шеннона, *b* – взаимная информация)
Fig. 1. Latitudinal and longitude dependences of entropy characteristics of geomagnetic data
(*a* – Shannon information entropy, *b* – mutual information)

На последующем этапе для сформированного пространственного кластера с обсерваторией DOU выполняется анализ информационной корреляции. При этом осуществляется итерационная обработка данных посредством попарного сравнения входящих в кластер опорных геодезических пунктов и того источника данных, для которого необходимо выполнить восстановление временного ряда (DOU). В результате получено подмножество в составе исходного пространственного кластера, для которого были достигнуты наилучшие показатели информационной корреляции. Анализ пространственной однородности соответствующих источников данных показал, что сформированный кластер является корректным.

Представленные в окончательно сформированном кластере источники данных образуют доверительный список для DOU. Данные в списке ранжируются по своим энтропийным и корреляционным характеристикам. Проведенные вычислительные эксперименты показали, что данные, зарегистрированные станцией DOUrbes, наилучшим образом коррелируют с наблюдениями обсерватории MAV (за исследуемый год). Это позволяет сделать вывод, что для реконструкции искомого временного ряда геомагнитных данных обсерваторию MAV можно назначить резервной станцией (при условии наличия данных за соответствующий временной интервал).

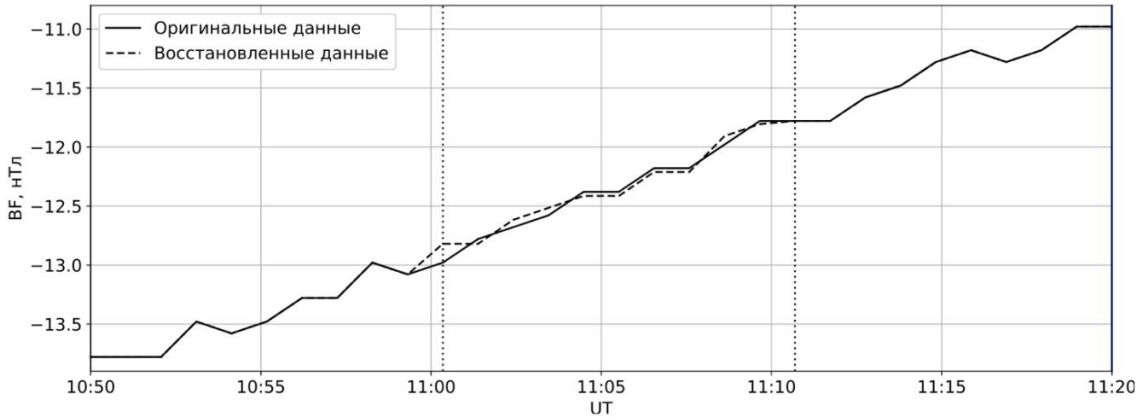


Рис. 2. Результаты восстановления 10-минутного фрагмента временного ряда геомагнитных данных
Fig. 2. Results of recovery of a 10-minute fragment of the time series of geomagnetic data

Дальнейший анализ показал, что применение метода информационного резервирования для восстановления 10-минутного пропуска в экспериментальном временном ряду характеризуется среднеквадратической ошибкой со значением ~ 0,036 нТл, что в целом меньше регламентированной спецификацией / стандартом IAGA-2002 погрешности для измерений параметров геомагнитного поля и его вариаций (~1 нТл) (рис. 2).

Заключение

В настоящей работе предлагается повысить эффективность используемых источниками данных технических систем регистрации с помощью информационного резервирования, которое предполагает использование дополнительных данных, выступающих в качестве вспомогательных. Выбор в пользу метода информационного резервирования обусловлен в наибольшей степени тем, что в большинстве известных случаев, когда отказ в работе технической системы приводит к потере или искажению информации (обрабатываемой локально или передаваемой по каналам связи), повышение надежности достигается преимущественно посредством информационного резервирования. В настоящее время метод информационного резервирования активно используется в системах управления и связи, информационных, измерительных и вычислительных системах сбора и обработки информации и позволяет повысить надежность технической системы регистрации данных при недостаточной надежности носителей информации, невозможности возобновления информации с помощью первичных источников и т.д.

На основании перечисленных характеристик и особенностей метода информационного резервирования как способа повышения надежности технических систем, обеспечивающих данными системы поддержки принятия решений, в работе предложен метод восстановления временных рядов. Метод предполагает определение наиболее вероятных значений посредством формирования доверительного списка источников данных на основании оценки пространственной гетерогенности и зависимости синхронно регистрируемых данных, а также сравнительной оценки соответствующих фрагментов временных рядов геомагнитных данных, зарегистрированных в момент времени, предшествующий восстанавливаемому.

В ходе проведенных вычислительных экспериментов для геомагнитных данных было установлено, что применение предложенного подхода позволяет восстанавливать временные ряды с точностью 0,01–0,5 нТл, что не превышает допустимой величины ошибки геомагнитных измерений.

Список источников

1. Демьянов В.В., Савельева Е.А. Геостатистика: теория и практика. М. : Наука, 2010. 327 с.
2. Vorobev A.V., Vorobeva G.R. Approach to Assessment of the Relative Informational Efficiency of Intermagnet Magnetic Observatories // Geomagnetism and Aeronomy. 2018. V. 58 (5). P. 625–628.
3. St-Louis B.J. Intermagnet Technical Reference Manual. Version 4. 1999. 156 c.
4. Mantas J. Statistical methods // Studies in Health Technology and Informatics. 2002. V. 65. P. 136–147.
5. Vorobeva G.R. Approach to the recovery of geomagnetic data by comparing daily fragments of a time series with equal geomagnetic activity // Computer Optics. 2019. № 43. P. 1053–1063.
6. Swan A. A class of higher inductive types in Zermelo-Fraenkel set theory // Mathematical Logic Quarterly. 2022. № 1. P. 118–127. doi: 10.1002/malq.202100040
7. Ricci L., Perinelli A., Castelluzzo M. Estimating the variance of Shannon entropy // Physical Review. 2021. V. 104 (2-1). Art. 024220. doi: 10.1103/PhysRevE.104.024220.
8. Carrara N., Ernst J. On the Estimation of Mutual Information // Proceedings. 2020. № 33. P. 31. doi: 10.3390/proceedings2019033031.
9. Гудчайлд М.Ф. Пространственный аналитическая перспектива на географических информационных системах // Международный журнал географических информационных систем. 1987. № 1 (4). С. 327–344.
10. Pinheiro P. [et al.] Lacunarity exponent and Moran index: A complementary methodology to analyze AFM images and its application to chitosan films // Physica A: Statistical Mechanics and its Applications. 2021. V. 581. Art. 126192.
11. Liu Q., Li Ch., Wei Y. Condition numbers of multidimensional mixed least squares-total least squares problems // Applied Numerical Mathematics. 2022. № 178. doi: 10.1016/j.apnum.2022.03.014
12. Sugasawa Sh., Murakami D. Spatially clustered regression // Spatial Statistics. 2021. № 44. Art. 100525.
13. Ilmi Nasution B., Kurniawan R., Caraka R. Nature-Inspired Spatial Clustering // The R Journal. 2021. March. P. 1–33. doi: 10.1109/CITSM.2014.7042178
14. Lin J.-W. Real-time Magnetic Observatory Network: A Review // European Journal of Environment and Earth Sciences. 2021. № 2. P. 1–2.

15. Kim J.-H., Chang H.-Y. Geomagnetic field variations observed by INTERMAGNET during 4 total solar eclipses // Journal of Atmospheric and Solar-Terrestrial Physics. 2018. V. 172. P. 107–116.

References

1. Demiyanov, V.V. & Savelieva, E.A. (2010) *Geostatistika: teoriya i praktika* [Geostatistics: Theory and Practice]. Moscow: Nauka.
2. Vorobev, A.V. & Vorobeva, G.R. (2018) Approach to Assessment of the Relative Informational Efficiency of Intermagnet Magnetic Observatories. *Geomagnetism and Aeronomy*. 58(5). pp. 625–628.
3. St-Louis, B.J. (1999). *Intermagnet Technical Reference Manual*. Version 4.
4. Mantas, J. (2002) Statistical methods. *Studies in Health Technology and Informatics*. 65. pp. 136–147. DOI: 10.3233/978-1-60750-909-7-136.
5. Vorobeva, G.R. (2019) Approach to the recovery of geomagnetic data by comparing daily fragments of a time series with equal geomagnetic activity. *Computer Optics*. 43. pp. 1053–1063.
6. Swan, A. (2022) A class of higher inductive types in Zermelo-Fraenkel set theory. *Mathematical Logic Quarterly*. 1. pp. 118–127. doi:10.1002/malq.202100040.
7. Ricci, L., Perinelli, A. & Castelluzzo, M. (2021) Estimating the variance of Shannon entropy. *Physical Review*. 104 (2-1). Art. 024220. DOI: 10.1103/PhysRevE.104.024220
8. Carrara, N. & Ernst, J. (2020) On the Estimation of Mutual Information. *Proceedings.* 33. p. 31. DOI: 10.3390/proceedings2019033031.
9. Gudchild, M.F. (1987) Prostranstvennyy analiticheskaya perspektiva na geograficheskikh informatsionnykh sistemakh [Spatial Analytical Perspective on Geographic Information Systems]. *Mezhdunarodnyy zhurnal geograficheskikh informatsionnykh sistem International – Journal of Geographic Information Systems.* 1(4). pp. 327–344.
10. Pinheiro, P. et al. (2021) Lacunarity exponent and Moran index: A complementary methodology to analyze AFM images and its application to chitosan films. *Physica A: Statistical Mechanics and its Applications*. 581. Art. 126192. DOI: 10.1016/j.physa.2021.126192
11. Liu, Q., Li, Ch. & Wei, Y. (2022) Condition numbers of multidimensional mixed least squares-total least squares problems. *Applied Numerical Mathematics*. 178. DOI: 10.1016/j.apnum.2022.03.014.
12. Sugasawa, Sh. & Murakami, D. (2021) Spatially clustered regression. *Spatial Statistics*. 44. Art. 100525. DOI: 10.1016/j.spasta.2021.100525
13. Ilmi Nasution, B., Kurniawan, R. & Caraka, R. (2021) Nature-Inspired Spatial Clustering. *The R Journal*. March. pp. 1–33. DOI: 10.1109/CITSM.2014.7042178.
14. Lin, J.-W. (2021) Real-time Magnetic Observatory Network: A Review. *European Journal of Environment and Earth Sciences*. 2. pp. 1–2. DOI: 10.24018/ejgeo.2021.2.5.177.
15. Kim, J.-H. & Chang, H.-Y. (2018) Geomagnetic field variations observed by INTERMAGNET during 4 total solar eclipses. *Journal of Atmospheric and Solar-Terrestrial Physics*. 172. pp. 107–116. DOI: 10.1016/j.jastp.2018.03.023.

***Информация об авторах*:**
Воробьева Гульнара Равилевна – доцент, доктор технических наук, профессор кафедры вычислительной математики и кибернетики Уфимского государственного авиационного технического университета. E-mail: gulnara.vorobeva@gmail.com
Воробьев Андрей Владимирович – доцент, кандидат технических наук, доцент кафедры геоинформационных систем Уфимского государственного авиационного технического университета. E-mail: cpu16bit@gmail.com

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:
Vorobeva Gulnara R. (Associate Professor, Doctor of Technical Sciences, Professor, Ufa State Aviation Technical University, Ufa, Russian Federation). E-mail: gulnara.vorobeva@gmail.com
Vorobev Andrei V. (Associate Professor, Candidate of Technical Sciences, Associate Professor, Ufa State Aviation Technical University, Ufa, Russian Federation). E-mail: cpu16bit@gmail.com

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Поступила в редакцию 11.05.2022; принята к публикации 29.11.2022

Received 11.05.2022; accepted for publication 29.11.2022