УДК 53.088.7
DOI: 10.12737/szf-72202105

Поступила в редакцию 01.03.2021
Принята к публикации 07.04.2021

ПОДХОД К ВОССТАНОВЛЕНИЮ ГЕОМАГНИТНЫХ ДАННЫХ НА БАЗЕ КОНЦЕПЦИИ ЦИФРОВЫХ ДВОЙНИКОВ

GEOMAGNETIC DATA RECOVERY APPROACH BASED ON THE CONCEPT OF DIGITAL TWINS

А.В. Воробьев
Уфимский государственный авиационный технический университет,
Уфа, Россия, geomagnet@list.ru
Геофизический центр РАН,
Москва, Россия, geomagnet@list.ru

В.А. Пилипенко
Геофизический центр РАН,
Москва, Россия, pilipenko_va@mail.ru
Институт физики Земли им. О.Ю. Шмидта РАН,
Москва, Россия, pilipenko_va@mail.ru

A.V. Vorobev
Ufa State Aviation Technical University,
Ufa, Russia, geomagnet@list.ru
Geophysical Center of RAS,
Moscow, Russia, geomagnet@list.ru

V.A. Pilipenko
Geophysical Center of RAS,
Russia, Moscow, pilipenko_va@mail.ru
Schmidt Institute of Physics of the Earth RAS,
Moscow, Russia, pilipenko_va@mail.ru

Аннотация. Ни одна наземная магнитная станция или обсерватория не гарантирует качество получаемой и передаваемой информации. Пропуски данных, выбросы и аномальные значения являются распространенной проблемой, касающейся практически любой сети наземных магнитометров и затрудняющей эффективную обработку и анализ экспериментальных данных. Обеспечение мониторинга надежности и повышение качества работы аппаратно-программных модулей, входящих в состав магнитных станций, возможно за счет разработки их виртуальных моделей, или так называемых цифровых двойников.

В настоящей работе на примере сети высокоширотных магнитометров IMAGE рассматривается один из возможных подходов к созданию моделей такого рода. Обосновано использование цифровых двойников магнитных станций для минимизации ряда проблем и ограничений, связанных с наличием выбросов и пропущенных значений во временных рядах геомагнитных данных, а также для обеспечения возможности ретроспективного прогноза параметров геомагнитного поля со среднеквадратической ошибкой в авроральной зоне до 11.5 нТл. Интеграция цифровых двойников в процессы сбора и регистрации геомагнитных данных реализует возможность автоматической идентификации и замещения отсутствующих и аномальных значений, таким образом повышая за счет эффекта резервирования отказоустойчивость магнитной станции как объекта-источника данных.

На примере цифрового двойника станции «Kilpisjärvi» (Финляндия) показано, что предлагаемый подход реализует восстановление 99.55 % годовой информации, при этом 86.73 % – с ошибкой, не превышающей 12 нТл.

Ключевые слова: цифровые двойники, восстановление временных рядов, статистический анализ, геомагнитные данные, магнитные станции.

Abstract. There is no ground-based magnetic station or observatory that guarantees the quality of information received and transmitted to it. Data gaps, outliers, and anomalies are a common problem affecting virtually any ground-based magnetometer network, creating additional obstacles to efficient processing and analysis of experimental data. It is possible to monitor the reliability and improve the quality of the hardware and software modules included in magnetic stations by developing their virtual models or so-called digital twins.

In this paper, using a network of high-latitude IMAGE magnetometers as an example, we consider one of the possible approaches to creating such models. It has been substantiated that the use of digital twins of magnetic stations can minimize a number of problems and limitations associated with the presence of emissions and missing values in time series of geomagnetic data, and also provides the possibility of retrospective forecasting of geomagnetic field parameters with a mean square error (MSE) in the auroral zone up to 11.5 nT. Integration of digital twins into the processes of collecting and registering geomagnetic data makes the automatic identification and replacement of missing and abnormal values possible, thus increasing, due to the redundancy effect, the fault tolerance of the magnetic station as a data source object.

By the example of the digital twin of the station “Kilpisjärvi” (Finland), it is shown that the proposed approach implements recovery of 99.55 % of annual information, while 86.73 % with M not exceeding 12 nT.

Keywords: digital twins, time series reconstruction, statistical analysis, geomagnetic data, magnetic stations.

ВВЕДЕНИЕ

В настоящее время магнитные обсерватории и вариационные станции являются одним из основных инструментов наблюдения геомагнитного поля (ГМП) и его вариаций. Так, на сегодняшний день существует более 300 наземных магнитных станций, регистрирующих и публикующих информацию о состоянии параметров ГМП в режиме реального (псевдореального) времени. Как правило, магнитные станции объединяются в сети (обычно по территориальному признаку), которые со стороны потребителя представляют собой специализированные веб-сервисы, обеспечивающие доступ к геомагнитным данным и обладающие необходимым для их поиска, предварительного просмотра и загрузки функционалом. По состоянию на начало 2021 г. известно более 20 таких сетей магнитных станций, наиболее крупными из которых являются INTERMAGNET, IMAGE, CARISMA, MACCS, MAGDAS и др.

Распространенной и до сих пор не имеющей окончательного решения проблемой, затрудняющей обработку получаемой геофизической информации, являются выбросы, помехи и пропуски во временных рядах геомагнитных данных. Даже для магнитных обсерваторий сети INTERMAGNET [Love, 2013, Khomutov, 2018], поддерживающей наивысший стандарт качества, пропущенные фрагменты занимают достаточно широкий диапазон и варьируют как во времени, так и от станции к станции. Например, для станции Alma Ata (AAA) в 2015 г. доля пропущенных значений составила 36 % от годовой наработки, для станции Dalat (DLT) — более 12 %, для станции Sodankylä (SOD) — 0.4 % и т. д. [Воробьев, Воробьева, 2018а].

Множественные выбросы и отсутствующие значения, помимо негативного влияния на эффективность самого подхода к наблюдению ГМП, исключают возможность применения к данным такого рода математического аппарата, требующего соблюдения условия непрерывности информационного сигнала (вычисление производной, преобразование Фурье, вейвлет-преобразование и пр.). Кроме этого, отсутствующие значения создают ощутимые проблемы как при моделировании пространственного распределения вариаций ГМП [Воробьев и др., 2020; Reich, Roussanova, 2013], так и связанной с ними экспериментальной информации высокого уровня (индексы геомагнитной активности, карты возмущенности, магнитные кеограммы и др.) [Гвишиани и др., 2019].

До недавнего времени восстановление результатов наблюдений ГМП осуществлялось с помощью линейной интерполяции или кубического сплайна, что в общем допустимо для устранения единичных пропусков, но абсолютно непригодно для импутации больших фрагментов. На сегодняшний день известны более сложные подходы к восстановлению такого рода временных рядов, базирующиеся в основном на аналитической обработке информационного сигнала в окрестности отсутствующих фрагментов, анализе периодических и сезонных составляющих, а также исследовании фурье- и вейвлет-спектров информационного сигнала [Воробьев, Воробьева, 2018б; Гвишиани и др., 2011; Мандрикова, Соловьев, 2012; Kondrashov et al., 2010; Mandrikova, et al., 2018]. Все они, как правило, требуют выполнения достаточно большого количества условий, ограничивающих их эффективное применение, обладают методической погрешностью в пределах 15 %, нуждаются в значительных вычислительных мощностях, непосредственном участии человека и, как следствие, неприменимы к большим объемам данных. Таким образом, обработка и анализ информации, получаемой непосредственно с магнитных станций, сопряжены с рядом затруднений и ограничений, во многом препятствующих эффективному проведению дальнейших исследований.

Перспективным подходом к решению данной проблемы могут быть создание и интеграция в процесс сбора геомагнитных данных проблемно-ориентированных цифровых двойников магнитных станций, позволяющих в некотором приближении имитировать поведение их физических прототипов. Реализация предлагаемой концепции может существенно повысить эффективность контроля качества информации, получаемой отдельными магнитометрами и вывести процессы обработки, анализа и прогнозирования геомагнитных возмущений (ГМВ) на качественно новый уровень.

1. ОЦЕНКА И АНАЛИЗ ПОКАЗАТЕЛЕЙ НАДЕЖНОСТИ НАЗЕМНЫХ МАГНИТНЫХ СТАНЦИЙ

Рассмотрим в качестве примера минутные данные сети магнитометров IMAGE [https://space.fmi.fi/image; Tanskanen, 2009] за 2015 г., т. е. период, соответствующий максимуму активности 24-го солнечного цикла (январь 2009 – май 2020 г.) [https://space.fmi.fi/image/www/index.php?page=user_defined]. В табл. 1 представлены результаты оценки полноты временных рядов 36 станций, где появление пропущенного значения расценивается как отказ технического объекта, т. е. его переход в неработоспособное состояние [ГОСТ 27.002-2015, 2016]. Общее время неработоспособного состояния станции T_F, соответствующее числу пропущенных значений во временном ряду, определится следующим образом

$$T_F = T - T_W, \quad (1)$$

где T — наработка; T_W — число информативных значений (общее время работоспособного состояния) за период времени T.

Среднее время до восстановления рабочего состояния (эквивалент математического ожидания размера отсутствующего фрагмента) и среднее время наработки до отказа системы (эквивалент среднего размера фрагмента без пропусков) можно определить из выражений (2) и (3) соответственно.

$$\langle T2R \rangle = \frac{1}{N_F} \sum_{i=1}^{N_R} T2R_i = \frac{T_F}{N_F}, \quad (2)$$

$$\langle T2F \rangle = \frac{1}{N_W + k} \sum_{i=1}^{N_W + k} T2F_i = \frac{T_W}{N_W + k}, \quad (3)$$

Таблица 1

Оценка показателей надежности магнитных станций сети IMAGE (на примере геомагнитных данных за 2015 г.)

IAGA код	Координаты				T_W		T_F		N_F	<*T2R*> [мин]	<*T2F*> [мин]
	GEO		CGM								
	LAT, [град]	LON, [град]	LAT, [град]	LON, [град]	[мин]	[%]	[мин]	[%]			
NAL	78.92	11.95	76.57	109.96	509551	96.947	16049	3.053	20	802.45	25477.55
LYR	78.20	15.82	75.64	111.03	506314	96.331	19286	3.669	11	1753.27	46028.55
HOR	77.00	15.60	74.52	108.72	466554	88.766	59046	11.234	4	14761.5	116638.5
HOP	76.51	25.01	73.53	114.59	492524	93.707	33076	6.293	49	675.02	10051.51
BJN	74.50	19.20	71.89	107.71	525523	99.985	77	0.015	7	11	75074.71
NOR	71.09	25.79	68.19	109.28	519087	98.761	6513	1.239	144	45.23	3604.77
SOR	70.54	22.22	67.80	106.04	523740	99.646	1860	0.354	43	43.26	12180.0
KEV	69.76	27.01	66.82	109.22	525569	99.994	31	0.006	11	2.82	47779.0
TRO	69.66	18.94	67.07	102.77	524713	99.831	887	0.169	15	59.13	34980.87
MAS	69.46	23.70	66.65	106.36	524144	99.723	1456	0.277	73	19.95	7180.05
AND	69.30	16.03	66.86	100.22	525284	99.94	316	0.06	6	52.67	87547.33
KIL	69.06	20.77	66.37	103.75	523732	99.645	1868	0.355	33	56.61	15870.67
IVA	68.56	27.29	65.60	108.61	486940	92.645	38660	7.355	6	6443.33	81156.67
ABK	68.35	18.82	65.74	101.70	525600	100	0	0	0	–	–
MUO	68.02	23.53	65.19	105.23	492390	93.682	33210	6.318	359	92.51	1371.56
KIR	67.84	20.42	65.14	102.62	525577	99.996	23	0.004	13	1.77	40429.0
SOD	67.37	26.63	64.41	107.33	524905	99.868	695	0.132	12	57.92	43742.08
PEL	66.90	24.08	64.03	104.97	491992	93.606	33608	6.394	8	4201.0	61499.0
JCK	66.40	16.98	63.82	98.94	516366	98.243	9234	1.757	36	256.5	14343.5
DON	66.11	12.50	63.75	95.19	511710	97.357	13890	2.643	19	731.05	26932.11
RAN	65.90	26.41	62.92	106.30	519118	98.767	6482	1.233	130	49.86	3993.22
RVK	64.94	10.98	62.61	93.27	513440	97.686	12160	2.314	61	199.34	8417.05
LYC	64.61	18.75	61.87	99.33	525600	100	0	0	0	–	–
OUJ	64.52	27.23	61.47	106.27	525304	99.944	296	0.056	11	26.91	47754.91
MEK	62.77	30.97	59.57	108.66	511795	97.373	13805	2.627	23	600.22	22251.96
HAN	62.25	26.60	59.12	104.72	520619	99.052	4981	0.948	381	13.07	1366.45
DOB	62.07	9.11	59.64	90.19	524128	99.72	1472	0.28	19	77.47	27585.68
SOL	61.08	4.84	58.82	86.25	512471	97.502	13129	2.498	31	423.52	16531.32
NUR	60.50	24.65	57.32	102.35	525540	99.989	60	0.011	2	30.0	262770.0
UPS	59.90	17.35	56.88	95.95	525600	100	0	0	0	–	–
KAR	59.21	5.24	56.70	85.69	524637	99.817	963	0.183	41	23.49	12796.02
TAR	58.26	26.46	54.88	103.11	525137	99.912	463	0.088	12	38.58	43761.42
BRZ	56.17	24.86	52.66	100.97	523584	99.616	2016	0.384	3	672.0	174528.0
SUW	54.01	23.18	50.21	98.95	487904	92.828	37696	7.172	20	1884.8	24395.2
WNG	53.74	9.07	50.15	86.75	525577	99.996	23	0.004	19	1.21	27661.95
NGK	52.07	12.68	48.03	89.28	525600	100	0	0	0	–	–

Примечание: GEO — географическая система координат; CGM (Corrected GeoMagnetic) — геомагнитная система координат; серым цветом выделены магнитные станции авроралного кластера

где $T2R_i$ и $T2F_i$ — время до i-го восстановления системы после отказа и до i-го отказа системы соответственно; N_F и N_W — число отказов системы и число восстановлений после отказа соответственно; $k=1$ или $k=0$, если в момент начала наблюдения система находилась в работоспособном или неработоспособном состоянии соответственно.

Анализ пропусков временных рядов сети IMAGE показал, что в 50 % случаях магнитных станций математическое ожидание размера пропущенного фрагмента превышает 58.5 мин. Усредненный (по всем станциям) размер пропущенного фрагмента составляет 1066 мин. Математическое ожидание числа отказов с восстановлением по всем станциям превышает 45 в год. При этом 50 % станций испытывают за год более 17 отказов. В крайних случаях общий объем отсутствующих фрагментов одной станции может превышать 11.2 % (более 41 сут) от общего размера годовой выборки, при этом среднее время восстановления может достигать 10 сут и более.

Полученные результаты указывают на то, что применение общеизвестных подходов к восстановлению временных рядов (линейная интерполяция, интерполяция кубическими сплайнами, а также методы, описанные в [Гвишиани и др., 2011; Мандрикова, Соловьев, 2012; Воробьев, Воробьева, 2018б; Kondrashov et al., 2010; Mandrikova et al., 2018]),

для большинства фрагментов отсутствующих значений рассматриваемых здесь источников (главным образом из-за размера отсутствующего фрагмента) может оказаться неэффективным. Кроме того, если речь идет о больших объемах информации (результаты наблюдения параметров ГМП за 1 год и более), применение методов, в алгоритмах работы которых предусмотрено участие человека, также становится весьма затруднительным.

2. КОНЦЕПЦИЯ ЦИФРОВОГО ДВОЙНИКА МАГНИТНОЙ СТАНЦИИ

Обычно под цифровым двойником понимают динамическое виртуальное представление физического объекта (процесса или системы) в течение его жизненного цикла с использованием данных в режиме реального времени для изучения и понимания [Parmar et al., 2020; Zongyan, 2020].

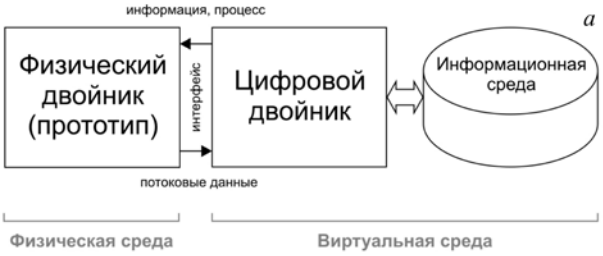
Различают следующие цифровые двойники (ЦД): прототипы (Digital Twin Prototype, DTP), содержащие информацию, необходимую для описания и создания физических версий экземпляров объекта; ЦД-экземпляры (Digital Twin Instance, DTI) описывающие конкретный физический экземпляр объекта, с которым двойник остается связанным на протяжении всего срока эксплуатации, и агрегированные двойники (Digital Twin Aggregate, DTA), представляющие собой информационную систему управления физическими экземплярами семейства объектов, также имеющую доступ ко всем их цифровым двойникам [Grieves, 2014].

На рис. 1, *а* приведена концепция ЦД-экземпляра, в которой, согласно рассматриваемой проблематике, в качестве физического прототипа системы выступает магнитная обсерватория или вариационная станция, а в качестве информационной среды — база геомагнитных данных, алгоритмическое и математическое обеспечение.

На рис. 1, *б* представлена модель интеграции ЦД-экземпляра в процессы сбора и публикации геомагнитных данных. Согласно предложенной схеме, возмущающее воздействие $x(t)$ распространяется на физический прототип магнитной станции (блок 1) и ряд опорных источников данных (блок 2), информация с которых используется в моделях и алгоритмах ЦД (блок 3) и входит в состав его информационной среды (рис. 1, *а*).

В зависимости от числа n доступных на момент времени t опорных источников на основании тестовых выборок выбирается модель ЦД, реализующая синтез значения $y^*(t)$ с минимальной ошибкой относительно $y(t)$ — ожидаемого значения на выходе станции-прототипа (блок 1).

Далее данные, соответствующие состоянию ГМП в момент времени t, с выхода ЦД и ее физического прототипа поступают на устройство сравнения (блок 4), которое путем анализа этих значений, например на основании выражения (4), принимает решение о публикации в качестве результата замера либо данных станции-прототипа (условие выполняется), либо ее ЦД-экземпляра (условие не выполняется). В случае невыполнения условия (4) значение с выхода магнитной станции-прототипа также сохраняется, однако помечается как аномальное. Если сигнал с выхода магнитной станции отсутствует, в качестве результата замера публикуется соответствующее значение с выхода ЦД. Верифицированные значения, хранящиеся в базе геомагнитных данных (блок 6), структурируются в виде векторов ответов и регрессоров и используются для актуализации и корректировки векторов коэффициентов моделей ЦД (блок 7).



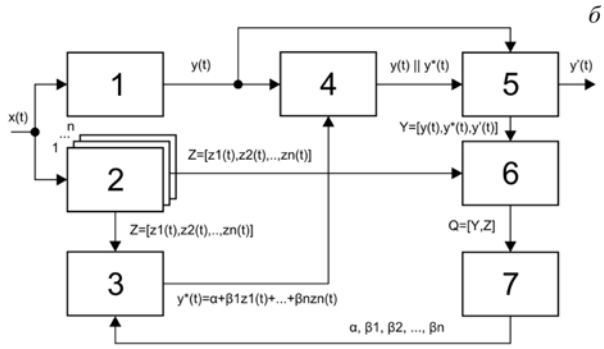


Рис. 1. Общая концепция ЦД (*а*) и модель интеграции ЦД-экземпляра в процессы сбора и публикации геомагнитных данных (*б*): 1 — магнитная станция-прототип; 2 — опорные источники данных (магнитные станции); 3 — математическое и алгоритмическое обеспечение ЦД-экземпляра магнитной станции (1); 4 — устройство сравнения; 5 — выходной буфер; 6 — база геомагнитных данных; 7 — система корректировки весовых коэффициентов

$$\left|y_t - y_t^*\right| < 3\sigma$$

или

$$\left|y_t - y_t^*\right| < 3\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\left(y_i - y_i^*\right) - \bar{y}\right)^2}, \qquad (4)$$

где σ — стандартное отклонение; y_t^* и y_t — значения на выходе цифрового двойника и его физического прототипа соответственно в момент времени t; m — размер тестовой выборки.

3. СИНТЕЗ, МОДИФИКАЦИИ И ВАЛИДАЦИЯ БАЗОВЫХ МОДЕЛЕЙ ЦИФРОВОГО ДВОЙНИКА

Выберем в качестве физического прототипа ЦД магнитометрический модуль, регистрирующий северную составляющую (X-компоненту) вектора ГМП, на станции «Kilpisjärvi» (KIL) и произведем пространственную кластеризацию всего множества магнитных станций с целью идентификации опорных

Таблица 2

Корреляционные связи между X_{KIL} и аналогичным параметром прочих станций

Магнитные станции, входящие в авроральный кластер													
NOR	SOR	KEV	TRO	MAS	AND	IVA	ABK	MUO	KIR	SOD	PEL	JCK	DON
0.872	0.933	0.978	0.985	0.99	0.987	0.975	0.986	0.957	0.958	0.909	0.875	0.845	0.820
Магнитные станции, не входящие в авроральный кластер													
NAL	LYR	HOR	HOP	BJN	RAN	RVK	LYC	OUJ	MEK	HAN	DOB	SOL	NUR
–0.164	-0.129	0.015	0.015	0.427	0.053	0.694	0.642	0.617	0.432	0.384	0.363	0.262	0.274
UPS		KAR		TAR		BRZ		SUW		WNG		NGK	
0.218		0.142		0.176		0.098		–0.045		–0.017		–0.044	

источников данных для последующего моделирования данного параметра.

Оценка пространственной однородности географических объектов посредством индекса Морана по признаку географического соседства по метрике [Демьянов, Савельева, 2010] выявила между рядом станций, располагающихся в диапазоне 66–71° N (табл. 1), наличие положительной пространственной корреляции, что свидетельствует о принадлежности этих станций к единому с KIL пространственному кластеру (далее — авроральный кластер).

Сравнительный анализ корреляционных связей северной (X) составляющей вектора ГМВ станции KIL с аналогичными параметрами прочих станций аврорального кластера (табл. 2), а также ряд дополнительных исследований [Воробьев, Воробьева, 2018в] подтверждают справедливость данного предположения и указывают на возможность использования этих данных в качестве предикатов (признаков) для моделирования параметра X_{KIL}.

Оценка коэффициента детерминации ($R^2=0.999$) показала, что в рамках решаемой задачи подход, основанный на методе множественной линейной регрессии, является наилучшим. Уравнение линейной регрессии, позволяющее восстановить значение искомого параметра $f(x, \beta)$ по известным значениям $x_1, \ldots, x_k$ имеет вид

$$f(x,\beta)=\beta_1 x_1+\beta_2 x_2+ \\ +\ldots+\beta_k x_k=\sum_{j=1}^{k}\beta_j x_j=x^T\hat{\beta}, \quad (5)$$

где $x^T=(x_1, x_2, \ldots, x_k)$ — вектор регрессоров; $\hat{\beta}=(\beta_1,\beta_2,\ldots,\beta_k)^T$ — вектор-столбец коэффициентов; k — число признаков модели.

Принимая во внимание данные табл. 2, запишем выражение (5) следующим образом:

$$X^*_{KIL}=\alpha+\beta_1 X_{NOR}+\beta_2 X_{SOR}+\beta_3 X_{KEV}+ \\ \beta_4 X_{TRO}+\beta_5 X_{MAS}+\beta_6 X_{AND}+\beta_7 X_{IVA}+ \\ +\beta_8 X_{ABK}+\beta_9 X_{MUO}+\beta_{10} X_{KIR}+\beta_{11} X_{SOD}+ \\ +\beta_{12} X_{PEL}+\beta_{13} X_{JCK}+\beta_{14} X_{DON}, \quad (6)$$

где $\alpha=418$ нТл — смещение по оси ординат; $\beta_1, \beta_2, \ldots, \beta_{14}$ — коэффициенты, рассчитанные методом наименьших квадратов:

$\beta_1=-0.0511992$; $\beta_2=-0.0791793$; $\beta_3=0.011932$; $\beta_4=0.5858979$; $\beta_5=-0.2199333$; $\beta_6=-0.203925$; $\beta_7=0.1138129$; $\beta_8=0.6873423$; $\beta_9=0.0020214$; $\beta_{10}=-0.2845333$; $\beta_{11}=0.0170759$; $\beta_{12}=0.0152406$; $\beta_{13}=0.0037965$; $\beta_{14}=-0.0263773$.

Среднеквадратическая ошибка (MSE) модели (6), рассчитанная по тестовой выборке объемом 20 % от исходного (годового) массива данных с применением процедуры кросс-валидации, составила 11.5 нТл, что составляет 0.51 % от размаха значений параметра X_{KIL} за 2015 г. Коэффициент корреляции Пирсона ($r=0.999$) и результаты t-теста Стьюдента (статистический критерий примерно равен нулю, p-значение — порядка 1) указывают на то, что исходные (X_{KIL}) и синтезированные (X^*_{KIL}) данные статистически неразличимы и принадлежат одной и той же выборке. Однако вероятность безотказной работы модели (6) ограничивается вероятностью отказа хотя бы одной из станций, входящих в авроральный кластер (см. табл. 1), и по имеющимся на 2015 г. данным составляет 77.4 %.

Повысить надежность ЦД можно путем модификации модели (6), например за счет применения при оценке ее коэффициентов метода LASSO [She, 2010; Hoerl, 2020], заключающегося во введении ограничения на норму вектора коэффициентов модели $\hat{\beta}$, что приведет к обращению в нуль некоторых ее коэффициентов, т. е. фактически к исключению из выражения (6) одной или нескольких станций. В связи с этим немаловажным положительным эффектом, возникающим в результате использования метода LASSO, является повышение устойчивости и интерпретируемости модели, поскольку в итоге отбираются признаки, оказывающие наибольшее влияние на вектор ответов. Из (7) следует, что при нулевом значении параметра регуляризации λ LASSO-регрессия сводится к обычному методу наименьших квадратов (МНК), а с его увеличением формируемая модель становится все более лаконичной — до тех пор пока не выродится в нуль-модель, дающую на выходе один и тот же результат для всех возможных входов [Токмакова, Стрижов, 2012].

$$\hat{\beta}_{LASSO}=\arg\min_{\beta}\left(\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{k}\beta_j x_{ij}\right)^2+\lambda|\beta|\right), \quad (7)$$

где y — ожидаемый отклик модели; λ — параметр регуляризации.

При $\lambda=1$ можно добиться сокращения выражения (6) на три слагаемых ($\beta_3=\beta_9=\beta_{12}=0$), повысив

тем самым вероятность срабатывания модели до 86.3 %, при этом практически не теряя в точности (MSE~12 нТл) и сохраняя параметры корреляции и статистической однородности оригинальной и синтезируемой выборок на уровне модели (5). Еще более значимо повысить вероятность срабатывания модели, по возможности исключая максимальное число слагаемых из выражения (6), контролируя при этом неизменность параметра корреляции и результатов *t*-теста Стьюдента, а также удерживая MSE в некотором приемлемом диапазоне, например, MSE≤30 нТл.

Однако, как показала практика, осуществление данной операции путем простого увеличения параметра λ малоэффективно и приводит к значительному росту ошибки моделирования при относительно небольшом снижении числа ее слагаемых. Другими словами, дальнейшее применение методов машинной оптимизации (в том числе гребневой регрессии и Elastic-Net [Zou, Hastie, 2005]) нецелесообразно и последующую минимизацию числа признаков следует производить вручную, например путем попарного сравнительного анализа статистик доступных предикатов. С этой целью в соответствии с выражением (8) исключим из временных рядов каждой станции медианное значение, нормализуем гистограмму и, руководствуясь критериями Колмогорова — Смирнова для полученных выборок $|\Delta X|$, подберем функцию, наилучшим образом аппроксимирующую характер распределения ее значений. Эта функция, в свою очередь, помимо однородности генеральных выборок, может указывать на однородность физических механизмов, ответственных за появление возмущений в точках их наблюдения [Vorobev, Vorobeva, 2019].

$$\left|\Delta X_{ij}\right| = \left|X_{ij} - \mathrm{Me}\left(X_j\right)\right|, \qquad (8)$$

где X_{ij} — *i*-е значение за *j*-е сутки *X*-составляющей на данной станции; $\mathrm{Me}(X_j)$ — медианное значение выборки *X* за *j*-е сутки; *i* и *j* соответствуют порядковым номерам минуты в сутках (от 1 до 1440) и дня в году (от 1 до 365) соответственно.

Анализ распределения абсолютных значений возмущенной составляющей *X*-компоненты ГМП на станции KIL ($|\Delta X|_{\mathrm{KIL}}$) показал, что бóльшая часть значений выборки (~95 %) распределена по логнормальному закону (рис. 2, *в*). Однако начиная с 95-го процентиля наблюдается экспоненциальный хвост, указывающий на то, что дисперсия исследуемой величины определяется преимущественно редкими интенсивными (а не частыми небольшими) отклонениями, очевидно, в данном случае имеющими место вследствие суббуревой активности. Дальнейшие исследования показали, что статистически наиболее близкими к $|\Delta X|_{\mathrm{KIL}}$ являются выборки $|\Delta X|_{\mathrm{TRO}}$, $|\Delta X|_{\mathrm{MAS}}$ и $|\Delta X|_{\mathrm{ABK}}$, т. е. абсолютные значения возмущенных составляющих *X*-компоненты ГМП на станциях Tromsø (TRO), Masi (MAS) и Abisko (ABK) соответственно. При этом практически единственным различием является процентиль выборки, соответствующий началу экспоненциального хвоста и обусловленный, по-видимому, широтным расположением конкретной станции (рис. 2, табл. 1).

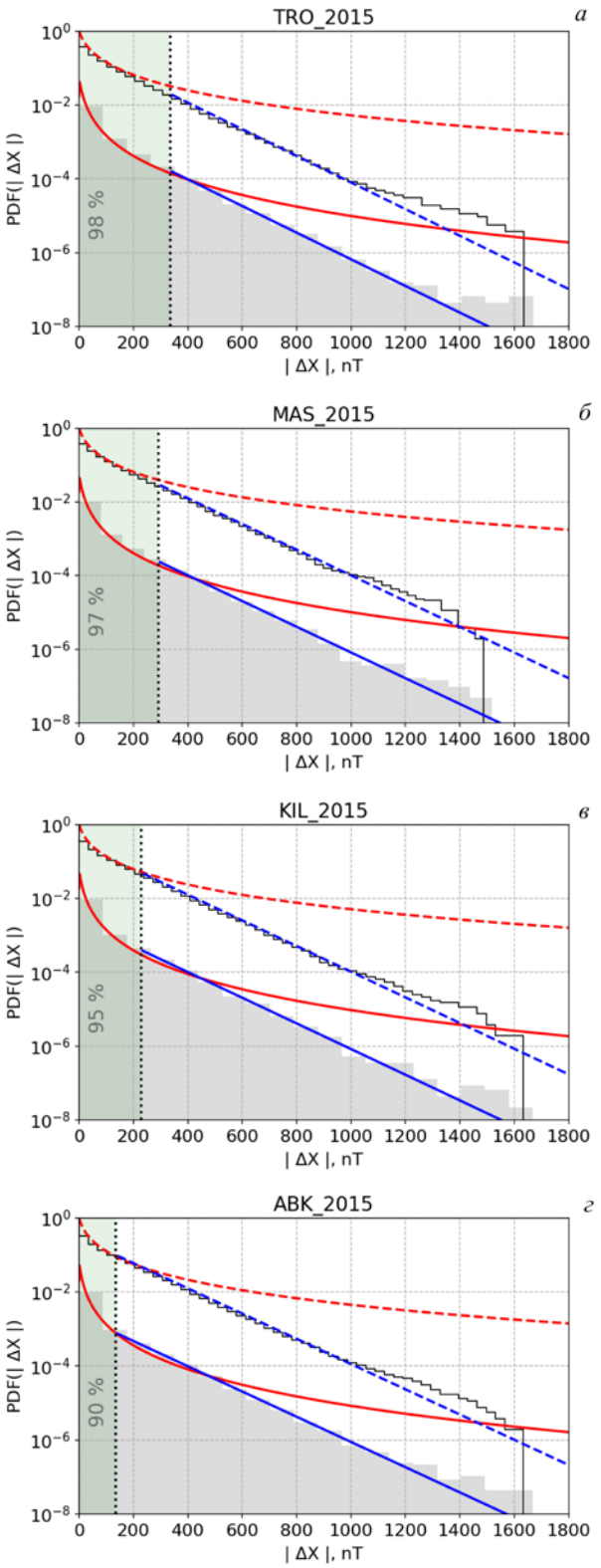


Рис. 2. Статистика ГМВ: красные и синие сплошные (штриховые) линии соответствуют функциям плотности вероятности (выживаемости) логнормального и экспоненциального законов распределения соответственно; черная сплошная линия — эмпирическая функция выживаемости

Кроме этого, анализ уровня корреляции между региональным IL-индексом (интенсивностью западного аврорального электроджета) и *X*-составляющей выделенных четырех станций (рис. 2) выявил соразмерность этих корреляционных связей (в каждом случае коэффициент корреляции Пирсона составляет

~0.7), что опять же свидетельствует о том, что рассматриваемые станции в равной степени подвергаются влиянию одних и тех же внешних факторов. Таким образом, минимальной ошибки при моделировании параметра X_{KIL} на базе минимальных наборов опорных источников данных можно добиться путем включения в эти наборы станций TRO, MAS и ABK. Тогда выражение (6) может быть сведено к следующему:

$$X^*_{KIL} = \alpha + \beta_4 X_{TRO} + \beta_5 X_{MAS} + \beta_8 X_{ABK}, \quad (9)$$

где $\alpha = 250$ нТл; $\beta_4 = 0.2924148$; $\beta_5 = 0.2850315$; $\beta_8 = 0.4408421$.

На рис. 3, *а* показаны магнитограммы исходного и восстановленного на базе регрессионной модели (9) временного ряда, включающего одну из самых мощных магнитных бурь за последние несколько лет наблюдений. Дисперсию результатов моделирования и разность между эмпирическими и синтезированными данными можно оценить из рис. 3, *б*, *в* соответственно. Вероятность срабатывания ЦД на базе модели (9) составляет 99.5 %, а MSE<30 нТл (табл. 3).

Следует отметить, что альтернативным, а в некоторых ситуациях единственным подходом к созданию ЦД могут являться методы, основанные на геопространственной интерполяции. Например, согласно методу обратных расстояний (Inverse Distance Weighting) [Isaaks, Mohan, 1989], интерполируемое значение параметра в заданной точке географического пространства определяется средневзвешенной суммой детерминированных значений, находящихся в ее окрестности. В случае модификации Шепарда [Isaaks, Mohan, 1989] уровень влияния детерминированной точки на искомое значение устанавливается показателем степени p и с удалением от вершины полигона, включающего опорные источники данных, ее влияние на интерполируемое значение ослабевает. Для рассматриваемого случая аналитическая запись IDW-метода имеет вид

$$X^*_{KIL} = \sum_{i=1}^{m} \frac{1}{d_i^p} X_i \Big/ \sum_{i=1}^{m} \frac{1}{d_i^p}, \quad (10)$$

где m — число станций авроpального кластера, X_i — значение X-составляющей i-й станции, d — расстояние между станцией KIL и i-й станцией авроpального кластера, p – весовой коэффициент.

Основным недостатком IDW-метода при интерполяции параметров ГМП является заложенное в него предположение об изотропности поля возмущения, хотя известно, что широтные и долготные масштабы большинства ГМВ существенно различаются. Исследования показали, что применительно к рассматриваемой проблематике среднеквадратическая ошибка модели ЦД, построенной на базе IDW-метода, монотонно увеличивается с уменьшением показателя p, что говорит о том, что искомый параметр определяется главным образом данными станций, наиболее близко расположенных к моделируемому объекту. В результате ошибка моделирования на базе (10) будет несколько выше MSE рассмотренных ранее регрессионных моделей (табл. 3). Несмотря на это, методы геопространственной интерполяции могут быть полезны в ситуациях, когда физический прототип станции отсутствует.

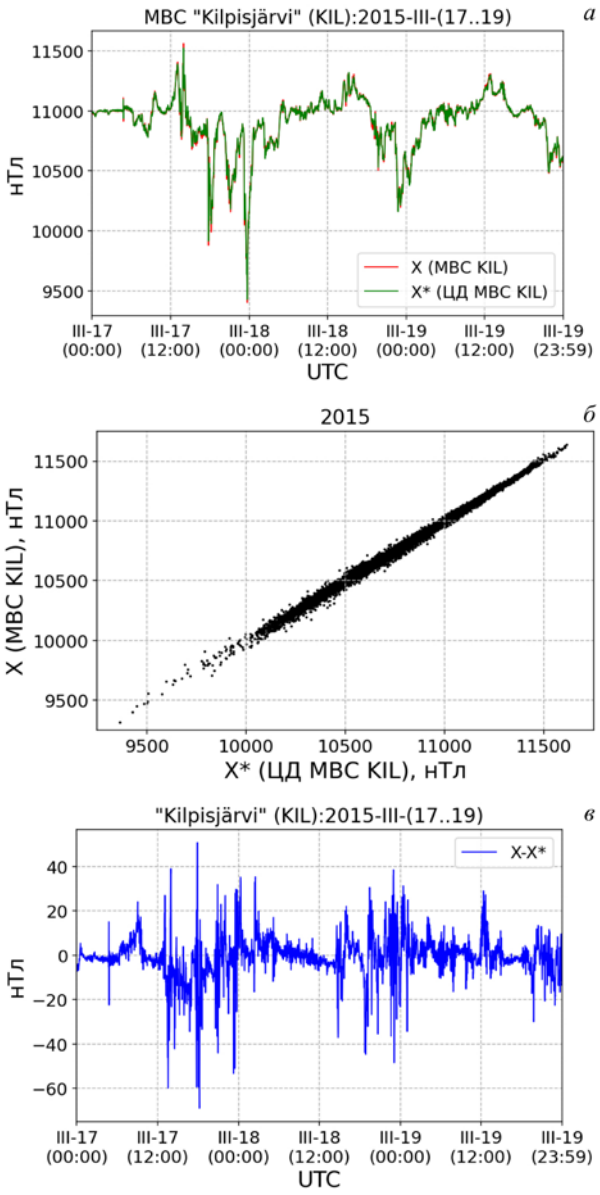


Рис. 3. Верификация цифрового двойника станции «Kilpisjärvi» (KIL)

4. ВЕРИФИКАЦИЯ ЦИФРОВОГО ДВОЙНИКА В ЧАСТОТНОЙ ОБЛАСТИ ИНФОРМАЦИОННОГО СИГНАЛА

Вариации ГМП в диапазоне периодов 2–12 мин хотя и значительно уступают по интенсивности глобальным ГМВ (магнитным бурям и суббурям), но имеют исключительно важное значение.

Возмущения этого частотного диапазона (Pi3-, Ps6-пульсации, Pc5-волны, начала суббурь) приводят к наиболее мощным всплескам геоиндуцированных токов (ГИТ) в линиях электропередач (ЛЭП). Поэтому важным аспектом при функционировании ЦД является идентификация и сохранение информации об этих возмущениях. Выделим посредством фильтра верхних частот Баттерворта в выборках X_{KIL} и X^*_{KIL} диапазон вариаций с времен-

Таблица 3

Параметры валидации моделей цифрового двойника станции KIL

Параметр / Модель	MSE, [нТл]	MSE, [%]	r	T-test Стьюдента		T_W, [мин]	T_F, [мин]	P_W, [%]
				стат. критерий	p-знач.			
Выр. (6) + МНК	11.5	0.51	0.999	~0	~1	406936	118664	77.423
Выр. (6) + LASSO (λ=1)	12.0	0.54	0.999	~0	~1	453819	71781	86.343
Выр. (9) + МНК	29.5	1.27	0.999	~0	~1	523257	2343	99.554
Выр. (10), IDW (p=3)	114.1	4.94	0.995	~0	~1	406936	118664	77.423

Примечание: P_W — ожидаемая вероятность срабатывания модели.

ными масштабами 2–12 мин и сопоставим вейвлет-спектрограммы отфильтрованного информационного сигнала, регистрируемого станцией KIL (рис. 4, *а*), с временными рядами, генерируемыми ее ЦД за аналогичный период времени (рис. 4, *б*).

Таким образом, как следует из рис. 4, а также ряда аналогичных тестов для других фрагментов временного ряда, в области ультранизких частот (периоды 2–12 мин) наблюдаются незначительные (в пределах заявленной в табл. 3 ошибки) отклонения амплитуды, при этом пространственная локализация частотных пакетов остается практически неизменной.

5. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ПЕРСПЕКТИВЫ ИХ ПРИМЕНЕНИЯ

Применение ЦД-экземпляра магнитной станции KIL позволяет восстановить 99.55 % данных за 2015 г., при этом среднеквадратическая ошибка 86.73 % восстановленных значений не превышает 12 нТл. Состояние отказа всей локальной системы сбора и регистрации геомагнитных данных (см. рис 1, *б*) наступает при одновременном отсутствии сигнала на выходе магнитной станции и ее ЦД (блоки 1 и 3 на рис. 1 соответственно). Для станции KIL расчетное значение вероятности наступления такого события составляет менее 0.0016 %, что соответствует восьми пропущенным значениям в год, которые, в свою очередь, можно восстановить методами линейной интерполяции или кубического сплайна.

Таким образом, внедрение ЦД магнитных станций в процессы сбора и регистрации геомагнитных данных за счет эффекта резервирования может (на уровне потребителя) значимо повысить надежность и отказоустойчивость отдельных магнитных станций, а также сократить трудоемкость некоторых процессов предварительной обработки геомагнитных данных, например таких, как поиск и идентификация выбросов во временных рядах.

Однако при реализации данного подхода следует иметь в виду ограничения его эффективного применения, определяемые, в первую очередь, пространственной анизотропией параметров ГМП. Таким образом, MSE ЦД-экземпляра каждой конкретной магнитной станции будет непосредственно зависеть от географического местоположения физического прототипа, а также числа, удаленности и характера взаиморасположения окрестных магнитных станций.

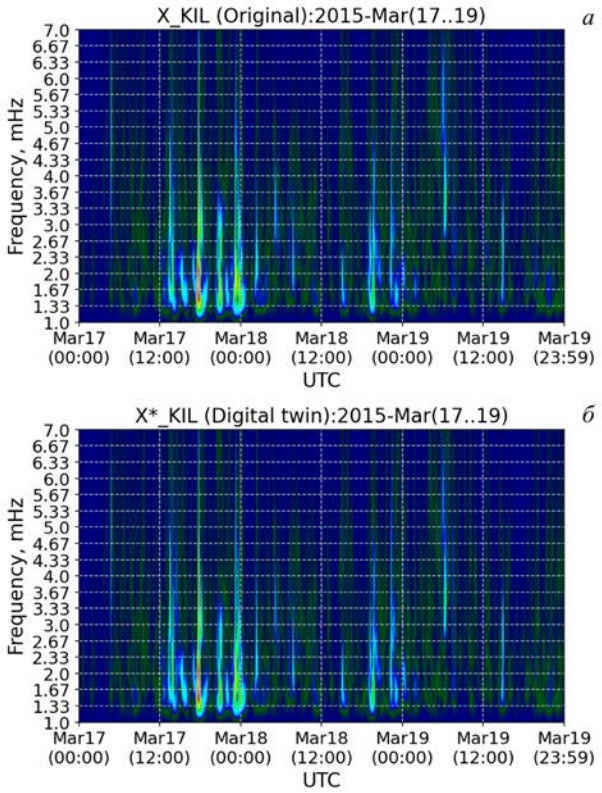


Рис. 4. Верификация ЦД магнитной станции «Kilpisjärvi» (KIL) в частотном диапазоне 1–7 мГц

Логичным направлением развития виртуальных магнитных станций является интеграция в информационную среду ЦД данных спутниковых наблюдений ГМП (например, миссии SWARM, CHAMP и т. п.). Допустимо предположение, что реализация данного подхода, помимо агрегирования дополнительных данных, необходимых при калибровке (настройке моделей) ЦД магнитных станций, способна также ослабить ряд методических ограничений, связанных, например, с отсутствием близлежащих магнитных станций.

Говоря о перспективах применения ЦД магнитных станций, необходимо выделить главным образом следующие задачи:

- восстановление и реконструкция временных рядов геомагнитных данных;
- автоматизированный поиск и идентификация выбросов во временных рядах геомагнитных данных;
- сбор геомагнитных данных в условиях, где использование физических магнитных станций неприемлемо или малоэффективно, например в непо-

средственной близости от объектов, оказывающих сильное зашумляющее действие на магнитные сенсоры и первичные измерительные преобразователи (магистральные трубопроводы, линии электропередач, объекты железнодорожной и нефтегазодобывающей инфраструктуры и т. п.).

- информационное сопровождение процессов наклонно-направленного бурения глубоких скважин в арктической зоне РФ [Гвишиани, Лукьянова, 2015, 2018].

Здесь следует отметить также, что ЦД-экземпляры обладают потенциалом применения в задачах машинного поиска и идентификации локализованных возмущений ГМП, например таких, как MPE (magnetic perturbation events), представляющих собой изолированные всплески интенсивности поля длительностью 5–15 мин в ночные часы [Engebretson et al., 2019], которые могут быть ответственны за интенсивные всплески ГИТ в ЛЭП [Datcu, et al., 2020]. Горизонтальный масштаб такого рода возмущений составляет ~200–300 км, и они регистрируются, как правило, на одной-двух станциях сети. Таким образом, ЦД способны автоматизировать данный процесс путем выделения возмущений, резко отличающихся от модельных значений.

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

На примере магнитной станции KIL в работе показано, что ЦД магнитных станций, построенные на базе LASSO-регрессии, способны обеспечить ретроспективный прогноз и восстановление *X*-составляющей вектора ГМП в авроральной зоне со среднеквадратической ошибкой от 11.5 (в 77.4 % случаев) до 29.5 нТл (в 99.6 % случаев) в зависимости от числа используемых опорных станций.

Сравнительный анализ вейвлет-спектрограмм данных ЦД магнитной станции и ее физического прототипа в диапазоне периодов 2–12 мин (Pi3-, Ps6-пульсации, Pc5-волны, начала суббурь) показал, что в амплитудной области информационного сигнала могут присутствовать незначительные различия, соразмерные с ошибкой моделирования, однако пространственная локализация частотных пакетов остается практически неизменной.

При отсутствии физического прототипа магнитной станции, определяющего вектор ответов обучающей выборки, реализация ЦД возможна путем пространственной интерполяции, например на базе IDW-метода, однако здесь следует ожидать несколько большую по сравнению с регрессионным подходом ошибку моделирования.

Основными факторами, ограничивающими эффективность применения предложенного подхода, являются географическое местоположения конкретного физического прототипа, а также число, удаленность и взаиморасположение окрестных магнитных станций. Минимизировать их влияние возможно путем расширения информационной среды ЦД, например за счет агрегации данных спутниковых наблюдений ГМП.

Мы благодарим институты, поддерживающие сеть магнитометров IMAGE: геофизическую обсерваторию Тромсё Арктического университета Норвегии (Норвегия), Финский метеорологический институт (Финляндия), Институт геофизики Польской академии наук (Польша), Немецкий исследовательский центр наук о Земле (Германия), Геологическую службу Швеции (Швеция), Шведский институт космической физики (Швеция), Геофизическую обсерваторию Соданкюля Университета Оулу (Финляндия) и Полярный геофизический институт (Россия).

Исследование выполнено при поддержке гранта РНФ № 21-77-30010.

СПИСОК ЛИТЕРАТУРЫ

Воробьев А.В., Воробьева Г.Р. Подход к оценке относительной информационной эффективности магнитных обсерваторий сети INTERMAGNET. *Геомагнетизм и аэрономия*. 2018а. Т. 58, № 5. С. 648–652. DOI: 10.1134/S0016793218050158.

Воробьев А.В., Воробьева Г.Р. Индуктивный метод восстановления временных рядов геомагнитных данных. *Труды СПИИРАН*. 2018б. № 2. С. 104–133. DOI: 10.15622/sp.57.5.

Воробьев А.В., Воробьева Г.Р. Корреляционный анализ геомагнитных данных, синхронно регистрируемых магнитными обсерваториями INTERMAGNET. *Геомагнетизм и аэрономия*. 2018в. Т. 58, № 2. С. 187–193. DOI: 10.7868/S0016794018020049.

Воробьев А.В., Пилипенко В.А., Еникеев Т.А., Воробьева Г.Р. Геоинформационная система для анализа динамики экстремальных геомагнитных возмущений по данным наблюдений наземных станций. *Компьютерная оптика*. 2020. Т. 44, № 5. С. 782–790. DOI: 10.18287/2412-6179-CO-707.

Гвишиани А.Д., Лукьянова Р.Ю. Исследование геомагнитного поля и проблемы точности бурения наклонно-направленных скважин в Арктическом регионе. *Горный журнал*. 2015. № 10. С. 94–99. DOI: 10.17580/gzh.2015.10.17.

Гвишиани А.Д., Лукьянова Р.Ю. Оценка влияния геомагнитных возмущений на траекторию наклонно-направленного бурения глубоких скважин в арктическом регионе. *Физика Земли*. 2018. Т. 54, № 4. С. 19–30. DOI: 10.1134/S0002333718040051.

Гвишиани А.Д., Агаян С.М., Богоутдинов Ш.Р., Каган А.И. Гравитационное сглаживание временных рядов. *Труды Института математики и механики УрО РАН*. 2011. Т. 17, № 2. С. 62–70.

Гвишиани А.Д., Лукьянова Р.Ю., Соловьев А.А. *Геомагнетизм: от ядра Земли до Солнца*. М.: РАН, 2019. 186 с.

ГОСТ 27.0022015. Надежность в технике. Термины и определения. М.: Стандартинформ, 2016. 23 с.

Демьянов В.В., Савельева Е.А. *Геостатистика: теория и практика*. М.: Наука, 2010. 327 с.

Мандрикова О.В., Соловьев И.С. Вейвлет-технология обработки и анализа геомагнитных данных. *Цифровая обработка сигналов*. 2012. № 2. С. 24–29.

Токмакова А.А., Стрижов В.В. Оценивание гиперпараметров линейных регрессионных моделей при отборе шумовых и коррелирующих признаков. *Информатика и ее применение*. 2012. Т. 6, № 4. С. 66–75.

Datcu M., Le Moigne J., Loekken S., et al. Special Issue on Big Data From Space. *IEEE Transactions on Big Data*. 2020. Vol. 6, no. 3. P. 427–429. DOI: 10.1109/TBDATA.2020.3015536.

Engebretson M.J., Steinmetz E.S., Posch J.L., et al. Nighttime magnetic perturbation events observed in Arctic Canada: 2. Multiple-instrument observations. *J. Geophys. Res.: Space Phys*. 2019. Vol. 124, iss. 9. P. 7459–7476. DOI: 10.1029/2019JA026797.

Grieves M.W. *Digital Twin: Manufacturing Excellence through Virtual Factory Replication*, Florida Institute of Technology Publ., 2014. 7 p.

Hoerl R.W. Ridge Regression: A Historical Context. *Technometrics*. 2020. Vol. 62, iss. 4. P. 420-425. DOI: 10.1080/00401706.2020.1742207.

Isaaks E.H., Mohan R. *An Introduction to applied geostatistics*. Oxford: Oxford University Press, 1989. 592 p.

Khomutov S.Yu. International project INTERMAGNET and magnetic observatories of Russia: cooperation and progress. *E3S Web of Conferences*. 2018. Vol. 62. P. 02008. DOI: 10.1051/e3sconf/20186202008.

Kondrashov D., Shprits Y., Ghil M. Gap filling of solar wind data by singular spectrum analysis. *Geophys. Res. Lett*. 2010. Vol. 37, iss. 15. L15101. DOI: 10.1029/2010GL044138.

Love J. An International Network of Magnetic Observatories. *EOS, transactions, American geophysical union*. 2013. Vol. 94, no 42. P. 373–384.

Mandrikova O.V., Solovyev I.S., Khomutov S.Y., et al. Multiscale variation model and activity level estimation algorithm of the Earth's magnetic field based on wavelet packets. *Ann. Geophys*. 2018. Vol. 36, iss. 5. P. 1207–1225. DOI: 10.5194/angeo-36-1207-2018.

Parmar R., Leiponen A., Llewellyn D.W.T. Building an organizational digital twin. *Business Horizons*. 2020. Vol. 63, iss. 6. P. 725–736. DOI: 10.1016/j.bushor.2020.08.001.

Reich K., Roussanova E. Visualising geomagnetic data by means of corresponding observations. *International Journal on Geomathematics*. 2013. Vol. 4. P. 1–25. DOI: 10.1007/s13137-012-0043-4.

She Y. Sparse regression with exact clustering. *Electron. J. Statist*. 2010. Vol. 4. P. 1055–1096. DOI: 10.1214/10-EJS578.

Tanskanen E.I. A comprehensive high-throughput analysis of substorms observed by IMAGE magnetometer network: Years 1993–2003 examined. *J. Geophys. Res*. 2009. Vol. 114, iss. A5. P. A05204. DOI: 10.1029/2008JA013682.

Vorobev A., Vorobeva G. Properties and type of latitudinal dependence of statistical distribution of geomagnetic field variations, 2019, In: Kocharyan G., Lyakhov A. (eds). *Trigger Effects Geosystems*. Springer Proc. Earth and Environmental Sciences. Springer Cham. 2019. P. 197–206. DOI: 10.1007/978-3-030-31970-0_22.

Zongyan W. Digital Twin Technology. *Industry 4.0 — Impact on Intelligent: Logistics and Manufacturing. IntechOpen*. 2020. DOI: 10.5772/intechopen.80974.

Zou H., Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Ser. B (Statistical Methodology)*. 2005. Vol. 67, iss. 2. P. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

URL: https://space.fmi.fi/image (дата обращения 1 марта 2021 г.).

URL: https://space.fmi.fi/image/www/index.php?page=user_defined (дата обращения 1 марта 2021 г.).

Статья подготовлена по материалам Шестнадцатой ежегодной конференция «Физика плазмы в Солнечной системе», 8–12 февраля 2021 г., ИКИ РАН.

Как цитировать эту статью:

Воробьев А.В., Пилипенко В.А.. Подход к восстановлению геомагнитных данных на базе концепции цифровых двойников. *Солнечно-земная физика*. 2021. Т. 7, № 2. С. 53–62. DOI: 10.12737/szf-72202105.