Solar-Terrestrial Physics. 2021. Vol. 7. Iss. 2. P. 48–56. DOI: 10.12737/stp-72202105. © 2021
Vorobev A.V., Pilipenko V.A. Published by INFRA-M Academic Publishing House. Original Russian version: Vorobev A.V., Pilipenko V.A., published in Solnechno-zemnaya fizika. 2021. Vol. 7. Iss. 2. P. 53–62. DOI: 10.12737/szf-72202105. © 2021 INFRA-M Academic Publishing House (Nauchno-Izdatelskii Tsentr INFRA-M)

GEOMAGNETIC DATA RECOVERY APPROACH BASED ON THE CONCEPT OF DIGITAL TWINS

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

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 MSE not exceeding 12 nT.

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

INTRODUCTION

Nowadays, magnetic observatories and variation stations are one of the main instruments for observing the geomagnetic field (GMF) and its variations. Today, there are over 300 ground-based magnetic stations capable of recording and publishing information on GMF parameters in real (pseudoreal) time mode. Magnetic stations are generally integrated into networks (usually according to geographic location), which, for users, are specialized web services that provide access to geomagnetic data and have necessary software and hardware modules for its search, preview, and download. As at the beginning of 2021, there are over 20 networks of magnetic stations, the largest of which are INTERMAGNET, IMAGE, CARISMA, MACCS, MAGDAS, etc.

A widespread and still unsolved problem that hinders geophysical data processing is outliers, noise, and gaps in time series of geomagnetic data. Even for INTERMAGNET magnetic observatories [Love, 2013, Khomutov, 2018] maintaining the highest quality standards, missing fragments occupy a fairly wide range and vary both in time and from station to station. For example, for the station Alma Ata (AAA) in 2015, the percentage of missing values was 36 % of annual information; for Dalat (DLT), over 12 %; for Sodankylä (SOD), 0.4 %, etc. [Vorobev, Vorobeva, 2018a].

Multiple outliers and missing values, besides the negative impact on the effectiveness of the approach to monitoring GMF, preclude the application of the mathematical apparatus to such data, which requires the continuity condition of information signal (derivation, Fourier transform, wavelet transform, etc.) be satisfied. Furthermore, missing values create serious problems in both modeling spatial distribution of GMF variations [Vorobev et al., 2020; Reich, Roussanova, 2013] and their related high-level experimental information (geomagnetic activity indices, perturbation maps, magnetic keograms, etc.) [Gvishiani et al., 2019].

Until recently, GMF observational results have been reconstructed using linear or spline interpolation, which is generally suitable for elimination of single gaps, but is entirely unsuitable for imputation of large fragments. More complex approaches to reconstructing such time series are currently known which are mainly based on the analytical processing of information signal in the vicinity of missing fragments, on the analysis of periodic and seasonal components, as well as on the study of the Fourier and wavelet spectra of information signal [Vorobev Vorobeva, 2018b; Gvishiani et al., 2011; Mandrikova, Solovyev, 2012; Kondrashov et al., 2010; Mandrikova, et al., 2018]. They all, as a rule, require the fulfillment of a fairly large number of conditions limiting their effective use, have a methodological error up to 15 %, need significant computational capability, direct human involvement, and, consequently, are inapplicable to large volumes of data. Thus, processing and analysis of the information collected directly from the magnetic stations involve a number of difficulties and limitations strongly impeding further research.

A promising approach to solving this problem may be creation and integration of problem-oriented digital twins of magnetic stations, which allow, in an approximation, to simulate the behavior of their physical prototypes, into acquisition of geomagnetic data. Implementing the proposed concept may significantly improve the efficiency of quality control of the output information from individual magnetometers and bring processing, analysis, and prediction of geomagnetic perturbations (GMP) to the next level.

1. ASSESSMENT AND ANALYSIS OF RELIABILITY INDEX OF GROUND-BASED MAGNETIC STATIONS

Consider minute data from the magnetometer network IMAGE [https://space.fmi.fi/image; Tanskanen, 2009] for 2015 as an example, i.e. the period corresponding to the maximum of solar cycle 24 (January 2009 – May 2020) [https://space.fmi.fi/image/www/index.php?page=user_defined]. Table 1 lists estimates of the completeness of time series from 36 stations, where the appearance of a missing value is regarded as a failure of a technical object, i.e. transition to disabled state [GOST 27.002-2015, 2016]. The total time of disabled state T_F corresponding to the number of missing values in a time series, is found as follows:

$$T_\mathrm{F} = T - T_\mathrm{W}, \qquad (1)$$

where T is the operation time; T_W is the number of infor-

Table 1

Estimated reliability indices of IMAGE magnetic stations (by the example of geomagnetic data for 2015)

IAGA code	Coordinates				T_W		T_F		N_F	<*T2R*> [min]	<*T2F*> [min]
	GEO		CGM								
	LAT, [deg]	LON, [deg]	LAT, [deg]	LON, [deg]	[min]	[%]	[min]	[%]			
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	–	–

Note: GEO — geographic coordinate system; CGM (Corrected GeoMagnetic) — geomagnetic coordinate system; gray color indicates magnetic stations of the auroral cluster mative values (the total time of operable state) over the time period T.

The mean time to return to operation (equivalent to the expected value of missing fragment size) and the mean operation time to failure of the system (equivalent to the mean fragment size without gaps) can be determined from expressions (2) and (3) respectively.

$$\langle T2R\rangle = \frac{1}{N_\mathrm{F}}\sum_{i=1}^{N_\mathrm{R}} T2R_i = \frac{T_\mathrm{F}}{N_\mathrm{F}}, \tag{2}$$

$$\langle T2F\rangle = \frac{1}{N_\mathrm{W}+k}\sum_{i=1}^{N_\mathrm{W}+k} T2F_i = \frac{T_\mathrm{W}}{N_\mathrm{W}+k}, \tag{3}$$

where $T2R_i$ and $T2F_i$ are the times to the ith recovery of the system after a failure and before the ith failure respectively; N_F and N_W are the number of failures of the system and the number of recoveries after the failure respectively; k=1 or k=0 if at the beginning of observation the system was serviceable or unserviceable respectively.

Analysis of gaps in IMAGE time series has shown that in 50 % of magnetic stations the expected value of missing fragment size exceeds 58.5 min. The missing fragment size averaged over all stations is 1066 min. The expected value of number of failures with recovery for all the stations exceeds 45 per year. At the same time, 50 % of the stations experience more than 17 failures per year. In extreme cases, the total amount of missing fragments in one station may exceed 11.2 % (over 41 days) of the total annual data, with a mean recovery time to 10 days or more.

The results indicate that the use of well-known approaches to reconstruct time series (linear interpolation, spline interpolation, and the methods described in [Gvishiani et al., 2011; Mandrikova, Solovyev, 2012; Vorobev, Vorobeva, 2018b; Kondrashov et al., 2010; Mandrikova et al., 2018]), for most fragments of missing values of the sources we examine (mainly due to the missing fragment size) may appear to be ineffective. In addition, in the context of large amount of information (observation of GMF parameters for 1 year and more), the application of the methods that require human participation also becomes very difficult.

2. CONCEPT OF DIGITAL TWIN OF MAGNETIC STATION

By a digital twin is usually meant a dynamic virtual representation of a physical object (process or system) during its life cycle with the use of real-time data to study and delve into [Parmar et al., 2020; Zongyan, 2020].

There are the following digital twins (DT): digital twin prototypes (DTP) containing information required for description and creation of physical versions of object instances; digital twin instances (DTI) describing a specific physical instance of an object with which the twin remains connected during the whole period of operation, and digital twin aggregates (DTA) representing an information system for monitoring physical instances of a family of objects, which also has access to all their digital twins [Grieves, 2014].

Figure 1, *a* presents the DTI concept in which, in terms of the problem addressed, a physical prototype of the system is a magnetic observatory or a variation station, and the information environment is a geomagnetic database, algorithmic and mathematical support.

Figure 1, *b* shows a model of integration of DTI into collection and publication of geomagnetic data. According to the proposed scheme, the perturbation effect $x(t)$ applies to a physical prototype of magnetic station (block 1) and a number of reference data sources (block 2) whose information is used in DT models and algorithms (block 3) and is included in its information environment (Figure 1, *a*).

Depending on the number n of reference sources available at the time t, from test data we choose a DT model able to synthesize $y^*(t)$ with a minimum error with respect to $y(t)$ — an expected value at the output of the prototype station (block 1).

Then, the data corresponding to GMF conditions at t from the output of DT and its physical prototype arrives at the compare facility (block 4) which, by analyzing these values, for example, based on Expression (4), takes a decision on publication, as a measurement result, of either data from the prototype station (the condition is satisfied) or its DTI (the condition is not satisfied).

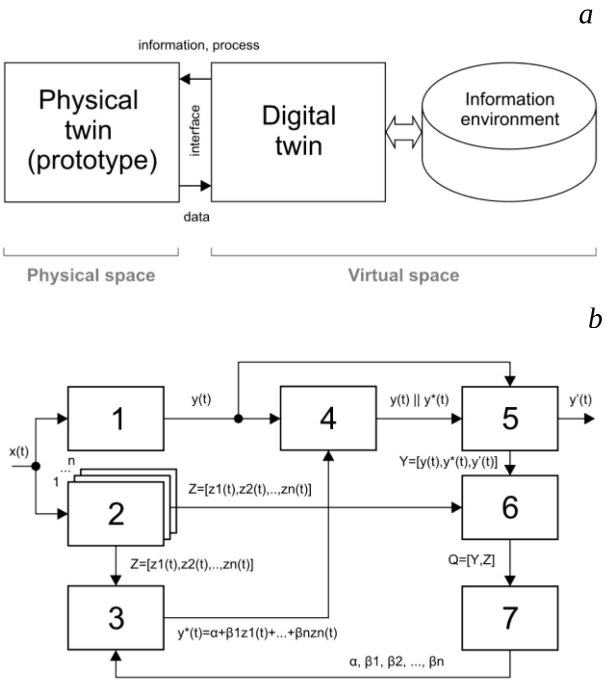

Figure 1. General concept of DT (*a*) and model of DTI integration into collection and publication of geomagnetic data (*b*): 1 — prototype magnetic station; 2 — reference data sources (magnetic stations); 3 — mathematical and algorithmic support of DTI (1); 4 — compare facility; 5 — output buffer; 6 — geomagnetic database; 7 — system for correcting weight coefficients

If condition (4) is not satisfied, the value from the output of the prototype magnetic station is also saved, but flagged as abnormal. If there is no signal from the output of a magnetic station, as a measurement result a corresponding value from the DT output is published. The verified values stored in the geomagnetic database (block 6) are structured as response and regressor vectors and are used to update and adjust vectors of coefficients of DT models (block 7).

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

or

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

where σ is the standard deviation; y_t^* and y_t are values at the output of a digital twin and its physical prototype respectively, at time t; m is the size of test data.

3. SYNTHESIS, MODIFICATION, AND VALIDATION OF FUNDAMENTAL DIGITAL TWIN MODELS

Take as a physical prototype of DT a magnetometric module recording the northern component (X component) of GMF vector at the station Kilpisjärvi (KIL) and perform a spatial clustering of all magnetic stations to identify reference data sources for further modeling of this parameter.

Estimating spatial homogeneity of geographic objects by the Moran index on the basis of geographical proximity in metric [Demyanov, Savelyeva, 2010] has revealed a positive spatial correlation between some stations located between 66 and 71° N (see Table 1), which suggests that these stations belong to the same cluster as KIL (hereinafter, the auroral cluster).

Comparative analysis of correlations between the northern (X) component of GMD vector of KIL and analogous parameters of other stations of the auroral cluster (Table 2), as well as a number of additional studies [Vorobev, Vorobeva, 2018c] confirm the validity of this assumption and indicate the possibility of using the data as predicates for modeling the parameter X_{KIL}.

Estimated determination coefficient (R^2=0.999) has shown that in terms of the problem to be solved the approach based on the method of multiple linear regression is the best. The linear regression equation allowing us to restore the desired parameter $f(x,\beta)$ from known values of x_1, ..., x_k has the form

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

Table 2

Correlations between X_{KIL} and analogous parameter of other stations

Magnetic station included in the auroral cluster													
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
Magnetic stations outside the auroral cluster													
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	

where $x^T = (x_1, x_2, \ldots, x_k)$ is the regressor vector; $\hat{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_k\right)^T$ is the column-vector of coefficients; k is the number of indicators of the model.

Taking into account the data from Table 2, write Equation (5) as follows:

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

(6)figure 2

where α=418 nT is the displacement along the Y-axis; β_1, β_2, ..., β_{14} are the coefficients calculated by the method of least squares:

β_1=–0.0511992; β_2=–0.0791793; β_3=0.011932;
β_4=0.5858979; β_5=–0.2199333; β_6=–0.203925;
β_7=0.1138129; β_8=0.6873423; β_9=0.0020214;
β_{10}=–0.2845333; β_{11}=0.0170759; β_{12}=0.0152406;
β_{13}=0.0037965; β_{14}=–0.0263773.

The mean square error (MSE) of model (6), calculated from the test data of volume 20 % of the initial (annual) data array under the cross-validation procedure was 11.5 nT, which is 0.51 % of the range of values of the parameter X_{KIL} for 2015. The Pearson correlation coefficient (r=0.999) and the results of Student's t-test (statistic is approximately equal to zero; the p value is of the order of 1) indicate that the initial (X_{KIL}) and synthesized (X^*_{KIL}) data is statistically indistinguishable and belongs to the same sample. The probability of reliable operation of model (6) is, however, limited by the probability of failure of at least one of the stations included in the auroral cluster (see Table 1) and, according to the data available for 2015, is 77.4 %.

The DT reliability may be improved by modifying model (6), for example, through the use of the LASSO method in estimating its coefficients [She, 2010; Hoerl, 2020], which involves introducing restriction on the vector norm of model coefficients $\hat{\beta}$. This makes some of its coefficients vanish, i.e. leads to the exclusion of one or more stations from Equation (6). In this regard, an important positive effect arising from the use of the

LASSO method is improvement of stability and interpretability of the model because eventually we can select features that have the greatest impact on the response vector. From (7) it follows that at zero value of the regularization parameter λ the LASSO regression reduces to the ordinary method of least squares (MLS), and as it increases the model developed becomes ever more concise until it degenerates into a zero model giving at the output the same result for all possible inputs [Tokmakova, Strizhov, 2012].

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

where y is the expected response of the model; λ is the regularization parameter.

When λ=1, we can reduce Equation (6) by three components ($\beta_3=\beta_9=\beta_{12}=0$), thereby increasing the probability of model operation to 86.3 % with virtually no loss in accuracy (MSE~12 nT) and with parameters of correlation and statistical homogeneity of the original and synthesized data kept at the level of model (5). Even more significant is to increase the probability of operation of the model, excluding, where possible, the maximum number of terms from (6), monitoring the constancy of the correlation parameter and Student's t-test results, as well as holding MSE in an acceptable range, e.g., MSE≤30 nT.

Nonetheless, as evidenced by practice, the implementation of this operation by simply increasing the parameter λ is inefficient and leads to a significant increase in modeling error at a relatively small reduction in the number of its terms. In other words, further use of the computer-aided optimization methods (including Ridge Regression and Elastic-Net [Zou, Hastie, 2005]) is impractical, the number of indicators should be further minimized manually, for example, through pairwise comparative analysis of statistics of available predicates. For this purpose, according to Expression (8), exclude the median from time series of each station, normalize the histogram, and, on the basis of Kolmogorov–Smirnov tests for $|\Delta X|$, select a function that best approximates distribution of its values. This function, in turn, in addition to the homogeneity of statistical population may indicate the uniformity of physical mechanisms responsible for the occurrence of perturbations at points of their observation [Vorobev, Vorobeva, 2019].

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

where X_{ij} is the ith value per the jth day of the X component at this station; Me(X_j) *is* the median of X per the jth day; i and j correspond to serial numbers of minute in the day (from 1 to 1440) and day in the year (from 1 to 365) respectively.

Analysis of distribution of absolute values of the perturbed GMF X component at the KIL station ($|\Delta X|_{\text{KIL}}$) has shown that most values of the sample (~95 %) are distributed according to the lognormal law (Figure 2, *c*). From the 95th percentile there is, however, an exponential tail indicating that the variance of the value under study is mainly determined by rare intense (but not frequent small) deviations occurring obviously due to substorm activity in this case. Follow-up studies have shown that $|\Delta X|_{\text{TRO}}$, $|\Delta X|_{\text{MAS}}$, and $|\Delta X|_{\text{ABK}}$, i.e. absolute values of perturbed components of GMF X at the stations Tromsø (TRO), Masi (MAS), and Abisko (ABK) respectively, are statistically the closest to $|\Delta X|_{\text{KIL}}$.

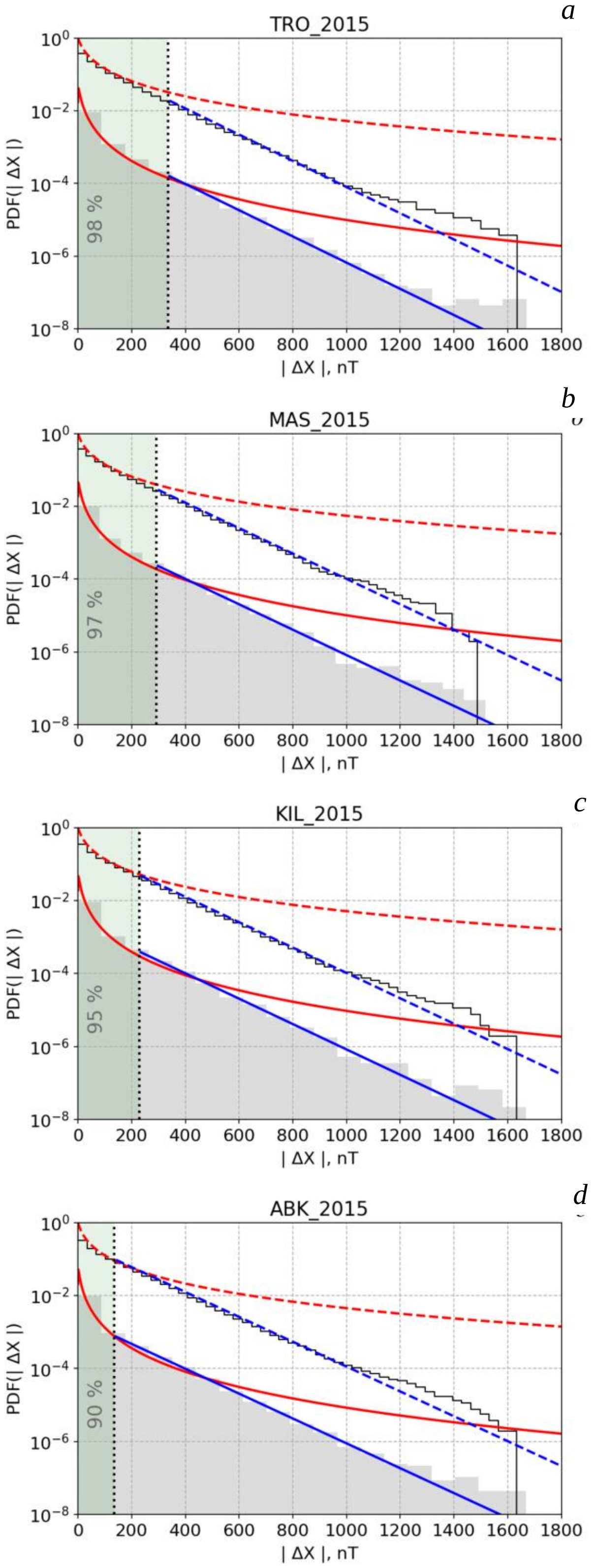

Figure 2. GMD statistics: red and blue solid (dashed) lines show density functions of the probability (survival) of lognormal and exponential distribution laws respectively; black solid line indicates the empirical survival function

In this case, virtually the only difference is the sample percentile corresponding to the beginning of the exponential tail and likely resulting from latitudinal position of a particular station (Figure 2, Table 1).

Besides, analysis of the level of correlation between the regional IL index (intensity of the westward auroral electrojet) and the X component in the four stations considered (Figure 2) has revealed the proportionality of these correlations (in each case the Pearson correlation coefficient is ~0.7), which again suggests that these stations are equally affected by the same external factors. Thus, the error in modeling the parameter X_{KIL} on the basis of minimum sets of reference data sources can be minimized by including the TRO, MAS, and ABK stations in these sets. Then, Expression (6) can be reduced to the following:

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

where α=250 nT; β_4=0.2924148; β_5=0.2850315; β_8=0.4408421.

Figure 3, *a* presents magnetograms of time series initial and reconstructed from regression model (9), which covers one of the most powerful magnetic storms observed over the past few years. The variance of simulation results and the difference between empirical and synthesized data can be estimated from Figure 3, *b*, *c* respectively. The probability of operation of DT based on model (9) is 99.5 %, and MSE<30 nT (Table 3).

Note that an alternative and in some cases the only approach to creating DT may be methods based on geospatial interpolation. For example, according to the inverse distance weighting method [Isaaks, Mohan, 1989], the interpolated parameter at a given point of geographic space is defined by the sum of the weighted mean values in its vicinity. In the case of Shepard modification [Isaaks, Mohan, 1989], the level of influence of the determinate point on the desired value is specified by the power p and with distance away from the polygon vertex containing reference data sources its influence on the interpolated value decreases. For the case of interest, the analytical form of the IDW method is as follows

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

where m is the number of stations in the auroral cluster, X_i is the X component in the ith station, d is the distance between KIL and the ith station of the auroral cluster, p is the weighting factor.

The main drawback of the IDW method in interpolating GMF parameters is its inherent assumption about perturbation field isotropism, although it is known that latitude and longitude scales of most GMDs differ significantly. Studies have shown that in terms of the problem addressed the mean square error in the DT model relying on the IDW method monotonically increases with decreasing p, which indicates that the desired parameter is mainly determined by data from the stations closest to the simulated object. As a result, the error in modeling on the basis of (10) is slightly higher than MSE of the previously considered regression models (Table 3). That said, geospatial interpolation methods may be useful in situations when there is no physical prototype of a station.

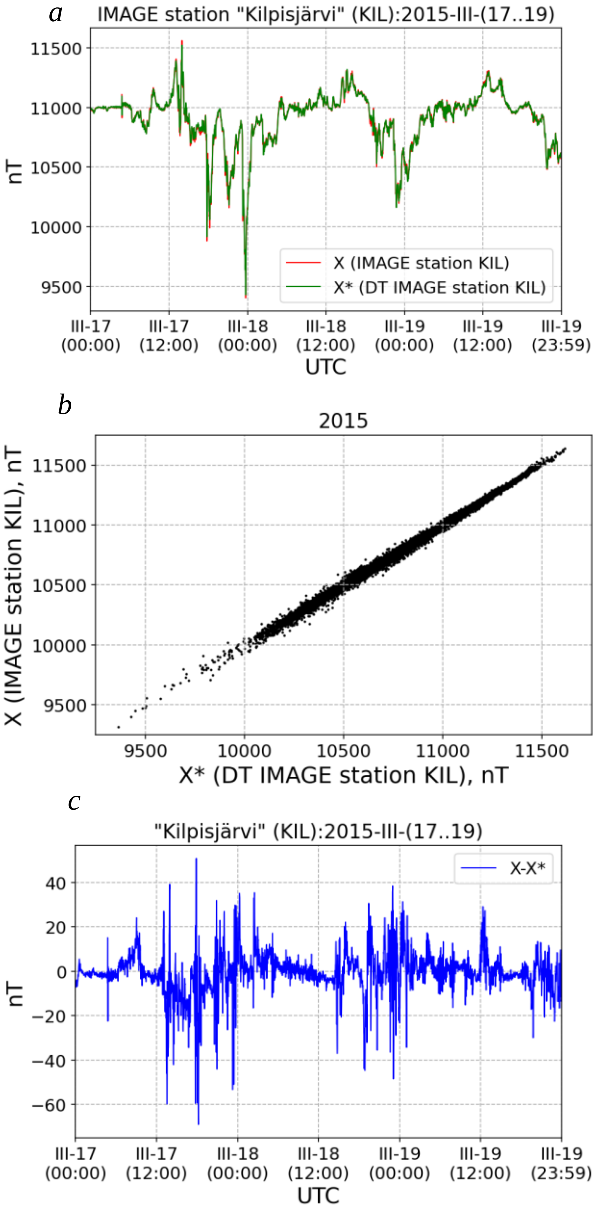

Figure 3. Verification of the digital twin of the station Kilpisjärvi (KIL)

4. DIGITAL TWIN VERIFICATION IN THE FREQUENCY DOMAIN OF INFORMATION SIGNAL

GMF variations in the range 2–12 min, despite their being less intensive than global GMDs (magnetic storms and substorms), are extremely important. Perturbations in this frequency range (Pi3, Ps6 pulsations, Pc5 waves, substorm onsets) generate the most powerful bursts of geomagnetically induced currents (GIC) in electric power transmission lines. Therefore, an important aspect in DT operation is to identify and store information about these perturbations. Identify the 2–12 min variation range, using the Butterworth high pass filter in X_{KIL} and X^*_{KIL}, and compare wavelet spectrograms of the filtered information signal recorded by the KIL station (Figure 4, *a*) with the time series generated by its DT during the same period (Figure 4, *b*).

Table 3

Validation parameters of KIL digital twin models

Parameter / Model	MSE, [nT]	MSE, [%]	r	Student's T -test		T_{W}, [min]	T_{F}, [min]	P_{W}, [%]
				statistic	p value			
Exp. (6) + MLS	11.5	0.51	0.999	~0	~1	406936	118664	77.423
Exp. (6) + LASSO (λ=1)	12.0	0.54	0.999	~0	~1	453819	71781	86.343
Exp. (9) + MLS	29.5	1.27	0.999	~0	~1	523257	2343	99.554
Exp. (10), IDW (p=3)	114.1	4.94	0.995	~0	~1	406936	118664	77.423

Note: P_W is the expected probability of model operation.

Thus, as follows from Figure 4 and from a number of similar tests for other time series fragments, in the ultra low frequency range (2–12 min periods), there are minor (within the error presented in Table 3) amplitude deviations, with spatial localization of frequency packages remaining virtually unchanged.

5. DISCUSSION OF RESULTS AND PROSPECTS OF THEIR APPLICATION

Using KIL DTI allows us to recover 99.55 % of data for 2015, with the mean square error in 86.73 % of recovered values not exceeding 12 nT. The entire local system of collecting and recording geomagnetic data (see Figure 1, *b*) fails when there is no signal at the output of the magnetic station and its DT (blocks 1 and 3 in Figure 1 respectively). For the KIL station, the estimated probability of occurrence of such an event is less than 0.0016 %, which corresponds to eight missing values per year, which in turn can be restored by linear or spline interpolation methods.

Thus, the integration of magnetic station DTs into process of geomagnetic data collection and registration due to the redundancy effect can (at the consumer level) significantly improve the reliability and fault tolerance of some magnetic stations, as well as reduce complexity of certain processes of geomagnetic data preprocessing such as search and identification of outliers in time series.

However, in the implementation of this approach we should take into account limitations of its effective use, defined primarily by spatial anisotropy of GMF parameters. Thus, DTI MSE of each specific magnetic station directly depends on the geographical location of its physical prototype as well as on the number, distance, and relative position of nearby magnetic stations.

A logical direction of development of virtual magnetic stations is integration of satellite GMD observations into the information environment of DT (e.g., SWARM mission, CHAMP, etc.). We may assume that the implementation of this approach, in addition to aggregating supplementary information required for calibration (model setup) of magnetic station DTs, can also ease some methodological limitations associated, e.g., with the absence of nearby magnetic stations.

As for promise of application of magnetic station DTs, we should highlight the following tasks:

- restoration and reconstruction of time series of geomagnetic data;

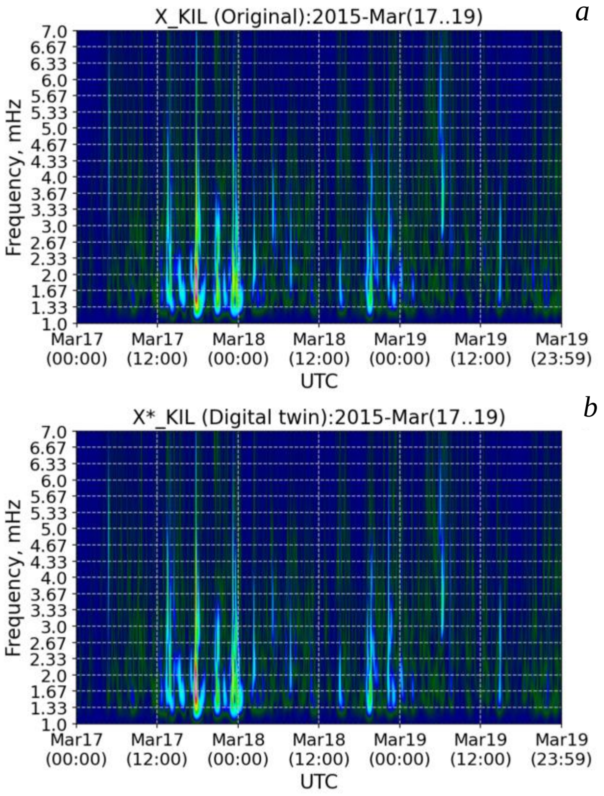

Figure 4. Verification of the magnetic station Kilpisjärvi (KIL) DT in the frequency range 1–7 MHz

- automated search and identification of outliers in time series of geomagnetic data;
- acquisition of geomagnetic data under conditions where the use of physical magnetic stations is unacceptable or ineffective, for example, in the immediate vicinity of the objects exerting a strong noise effect on magnetic sensors and primary detectors (pipelines, electric power transmission lines, railway and oil-and-gas infrastructure, etc.);
- information support of directional deep-well drilling in the arctic zone of the Russian Federation [Gvishiani, Lukyanova, 2015, 2018].

It should also be noted here that DTIs have the potential for being used in problems of machine search and identification of localized GMF perturbations such as MPE (magnetic perturbation events) representing isolated bursts of field intensity lasting for 5–15 min at night [Engebretson et al., 2019], which may be responsible for the intense bursts of GIC in electric power transmission lines [Datcu, et al., 2020]. The horizontal scale of perturbations of this type is ~200–300 km, and they are usually detected by one or two stations of the network. Thus, DTs can automatize this process by identifying perturbations sharply differing from model values.

CONCLUSIONS

By the example of the magnetic station KIL, we have shown that magnetic station DTs, constructed based on the LASSO regression, can provide retrospective forecast and reconstruction of the GMF vector X component in the auroral zone with a mean square error from 11.5 (in 77.4 % of cases) to 29.5 nT (in 99.6 % of cases), depending on the number of reference stations in use.

Comparative analysis of wavelet spectrograms of data on DT of the magnetic station and its physical prototype in the time range 2–12 min (Pi3, Ps6 pulsation, Pc5 wave, substorm onsets) has revealed that there are minor differences, proportional to modeling error, in the amplitude range of information signal, but the spatial localization of frequency packages remain virtually unchanged.

In the absence of the physical prototype of a magnetic station, which defines training data response vector, DT may be implemented through spatial interpolation, e.g., by the IDW method; in this case, however, we should expect a somewhat larger modeling error as compared to the regression approach.

The main factors limiting the effectiveness of the proposed approach are the geographical location of a specific physical prototype, the number, distance, and relative position of nearby magnetic stations. Their effect may be minimized by expanding the information environment of DT, for example, through aggregation of satellite GMF observations.

We thank the institutes who maintain the IMAGE Magnetometer Array: Tromsø Geophysical Observatory of UiT the Arctic University of Norway (Norway), Finnish Meteorological Institute (Finland), Institute of Geophysics Polish Academy of Sciences (Poland), GFZ German Research Centre for Geosciences (Germany), Geological Survey of Sweden (Sweden), Swedish Institute of Space Physics (Sweden), Sodankylä Geophysical Observatory of the University of Oulu (Finland), and Polar Geophysical Institute (Russia).

The work was financially supported by RSF (Grant No. 21-77-30010).

REFERENCES

Datcu M., Le Moigne J., Loekken S., Soille P., Xia G.-S. Special Issue on Big Data From Space. *IEEE Transactions on Big Data*, 2020, vol. 6, no. 3, pp. 427-429. DOI: 10.1109/TBDATA. 2020.3015536.

Demyanov V.V., Savelyeva E.A. *Geostatistics: theory and practice*. Moscow, Nauka Pabl., 2010, 327 p. (In Russian).

Engebretson M.J., Steinmetz E.S., Posch J.L., Pilipenko V.A., Moldwin M.B., Connors M.G. Nighttime magnetic perturbation events observed in Arctic Canada: 2. Multiple-instrument observations. *J. Geophys. Res.: Space Phys*. 2019, no. 124, pp. 7459–7476. DOI: 10.1029/2019JA026797.

GOST 27.0022015. Reliability in technology. Terms and Definitions. Moscow.: Standartinform, 2016.23 p.

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

Gvishiani A.D., Agayan S.M., Bogoutdinov Sh.R., Kagan A.I. Gravitational smoothing of time series. *Trudy Instituta matematiki i mekhaniki UrO RAN* [Proceedings of the Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences]. 2011, vol. 17, no. 2, pp. 62–70. (In Russian).

Gvishiani A.D., Lukyanova R.Yu. Study of the geomagnetic field and the problem of the accuracy of drilling directional wells in the Arctic region. *Gorny Zhurnal* [Mining Journal]. 2015, no. 10, pp. 94–99. DOI: 10.17580/gzh.2015.10.17. (In Russian).

Gvishiani A.D., Lukyanova R.Yu. Assessment of the impact of geomagnetic disturbances on the trajectory of directional drilling of deep wells in the Arctic region. *Fizika Zemli* [Physics of the Earth]. 2018, no. 4, pp. 19–30. DOI: 10.1134/S0002333718040051. (In Russian).

Gvishiani A.D., Lukyanova R.Yu., Soloviev A.A. *Geomagnetism: from the Core of the Earth to the Sun*. Moscow, RAS Pabl., 2019. 186 p. (In Russian).

Hoerl R.W. Ridge Regression: A Historical Context. *Technometrics*. 2020, vol. 62, iss. 4, pp. 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, pp. 373–384.

Mandrikova O. V., Soloviev I. S. Wavelet technology for processing and analyzing geomagnetic data. *Tsifrovaya obrabotka signalov* [Digital Signal Processing]. 2012, no. 2, pp. 24–29. (In Russian).

Mandrikova O.V., Solovyev I.S., Khomutov S.Y., Geppener V.V., Klionskiy D.M., Bogachev M.I. 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. pp. 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, no. 6, pp. 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, pp. 1–25. DOI: 10.1007/s13137-012-0043-4.

She Y. Sparse regression with exact clustering. *Electron. J. Statist*. 2010, vol. 4, pp. 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.

Tokmakova A.A., Strizhov V.V. Estimation of hyperparameters of linear regression models in the selection of noise and correlated features. *Informatika i yeye primeneniye* [Informatics and its application]. 2012, vol. 6, no. 4, pp. 66–75. (In Russian).

Vorobev A.V., Vorobeva G.R. Approach to Assessment of the Relative Informational Efficiency of Intermagnet Magnetic Observatories. *Geomagnetism and Aeronomy*. 2018a, vol. 58, no. 5, pp. 625–628. DOI: 10.1134/S0016793218050158.

Vorobev A.V., Vorobeva G.R. Inductive method for reconstructing time series of geomagnetic data. *Proc. SPIIRAS* [Trudy SPIIRAN]. 2018b, no. 2, pp. 104–133. DOI: 10.15622/sp.57.5. (In Russian).

Vorobev A.V., Vorobeva G.R. Correlation analysis of geomagnetic data synchronously recorded by INTERMAGNET magnetic laboratories. *Geomagnetism and Aeronomy*. 2018c, vol. 58, no. 2, pp. 178–184. DOI: 10.1134/S0016793218020196.

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 in Geosystems. Springer Proceedings in Earth and Environmental Sciences*. Springer Cham. 2019. P. 197–206. DOI: 10.1007/978-3-030-31970-0_22.

Vorobev A.V., Pilipenko V.A., Enikeev T.A., Vorobeva G.R. Geographic information system for analyzing the dynamics of extreme geomagnetic disturbances based on observations of ground stations. *Komp'yuternaya optika* [Computer Optics]. 2020, vol. 44, no. 5, pp. 782–790. DOI: 10.18287/2412-6179-CO-707. (In Russian).

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: Series B (Statistical Methodology).* 2005, vol. 67, iss. 2. pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

URL: https://space.fmi.fi/image (accessed 1 March 2021).

URL: https://space.fmi.fi/image/www/index.php?page=user_defined (accessed 1 March 2021).

This paper is based on material presented at the 16th Annual Conference on Plasma Physics in the Solar System, February 8–12, 2021, IKI RAS.

How to cite this article

Vorobev A.V., Pilipenko V.A. Geomagnetic data recovery approach based on the concept of digital twins. *Solar-Terrestrial Physics*. 2021. Vol. 7. Iss. 2. P. 48–56. DOI: 10.12737/stp-72202105.