Iterative Selection of Essential Input Features under Conditions of their Multicollinearity in Space Weather Time Series Forecasting

Nikolay Shchurov
Physical Department & D.V.Scobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University
Moscow, Russia
shchurov_no2@mail.ru

Igor Isaev
D.V.Scobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University
Moscow, Russia
ORCID 0000-0001-7121-2383

Oleg Barinov
D.V.Scobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University
Moscow, Russia
obar@sinp.msu.ru

Irina Myagkova
D.V.Scobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University
Moscow, Russia
ORCID 0000-0001-6854-5873

Sergei Dolenko
D.V.Scobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University
Moscow, Russia
ORCID 0000-0001-6214-3195

***Abstract*—The paper presents a method for selecting essential input features when predicting the geomagnetic Dst index, based on iterative selection of features with the highest correlation with respect to the target variable and exclusion of features with high cross-correlation. The models were trained on data from October 1997 to 2017. The criterion for the quality of the forecast using selected features was the root-mean squared error of the Dst index forecast based on the selected set of features on independent data (2018-2022).**

Keywords—multivariate time series, prediction, feature selection, artificial neural networks, Earth's magnetosphere, geomagnetic Dst index

I. INTRODUCTION

The study of the states of the Earth's magnetosphere and their prediction are of significant interest from both fundamental and practical points of view.

Clarification of the role of various physical processes in the formation of geomagnetic disturbances remains one of the key tasks in the physics of solar-terrestrial connections to this day. On the other hand, geomagnetic disturbances, or, as they are more often called in the media, magnetic storms, disrupt radio communications and can negatively affect the operation of power lines, railway automation and pipelines caused by geo-induced currents – GIC [1].

The intensity of geomagnetic disturbances is described by the so-called geomagnetic indices, among which one of the most widely used is the Dst (Disturbance storm-time) index - storm variation [2]. The history of the appearance of this index, its sources and characteristic features are described in our previous work devoted to optimizing methods for predicting it using machine learning (ML) methods using time series (TS) containing data on the state of the Earth's magnetosphere, interplanetary magnetic field (IMF) and solar wind (SW) [3].

This study has been performed at the expense of the grant of the Russian Science Foundation, project no. 23-21-00237, https://rscf.ru/en/project/23-21-00237/.

As has been known for quite some time, the main sources of geomagnetic disturbances are coronal mass ejections and high-speed SW streams.

Since the Earth's magnetosphere is a complex dynamic system, the state and behavior of which depends not only on its current state and external influence on it, but also on its history, to predict its state, and therefore geomagnetic disturbances, it is necessary to use sufficient ML methods. In this case, to take into account the history, one can use either recurrent algorithms [4] or delay embedding of a multidimensional TS of input variables (e.g. [5-11]). In the second approach, each data pattern includes, in addition to the current values of the input variables, their previous values with delays from one TS step to a value called the embedding depth.

A significant disadvantage of the TS embedding method is the multiple increase in the number of input features (IF), which increases the requirements for the number of patterns in the training set, and so it can create problems associated with model overtraining. However, based on the features of the method, it follows that the IF obtained as a result of delay embedding of a TS are characterized by multicollinearity, that is, they carry largely similar information. Therefore, to reduce the dimensionality of input data when making forecasts, it is useful to use algorithms for selecting IF that take into account multicollinearity.

In the present study, the selection of significant IF in the problem of predicting the Dst index is carried out by a method based on iterative selection of features with the highest correlation with respect to the target variable, and exclusion of features with high cross-correlation. This method is compared with the traditional selection method - the cross-correlation filter, as well as with limiting the embedding depth of each input parameter by the autocorrelation value without using selection. The comparison criterion is the root mean squared error of the Dst index forecast based on the selected set of features on independent data.

Since the Dst index has a rather long history of observation, ML has been used to predict geomagnetic disturbances by many various scientific groups [4-7], including the authors of this work [8-11].

In their earlier studies, the authors of this paper have shown that the best quality of the Dst index forecast was achieved when building a neural network (NN) model that uses the history of both the Dst index and the parameters of the SW (velocity) and IMF (component Bz) as input data [8]. In the following studies [9-10], each pattern contained hourly average values of the main parameters of the SW and IMF, as well as hourly values of the Dst index itself - with delay embedding of the TS for the depth of 24 hours, which significantly improved the quality of the forecast.

In our previous papers dedicated to reducing the dimensionality of input data [3, 12], the results of data dimensionality reduction based on the ranking of IF by their significance in solving the problem of the geomagnetic Dst index forecasting were considered. To assess the relative significance of features, an iterative approach was used, associated with the search of candidate models by discarding features one by one based on simple linear regression models. An alternative method for selecting IF is described in the paper [13].

In addition to optimizing the operation of NN models, the selection of significant IF can make it possible to draw some conclusions about the relationships between various physical quantities, the values of which are used as IF.

The **purpose** of this study was to investigate the applicability of the method for selecting IF, based on iterative selection of features with the highest correlation with respect to the target variable and on excluding features with high cross-correlation, in predicting the Dst index, as well as to compare the results obtained using the method with various parameters.

II. Input Data

Since it has long been known that the cause of geomagnetic disturbances are changes in the parameters of the IMF and SW (caused by coronal mass ejections and high-speed streams of SW) [14], we used TS of the following physical quantities characterizing the state of near-Earth space and the Earth's magnetosphere as IF when constructing the NN model:

- SW characteristics – SW velocity (v_{SW}) in km/s, proton density (ρ_{SW}) in cm^{-3}, and SW temperature (T_{SW}) in K;
- IMF characteristics – Bx, By and Bz components in the GSM system, and the magnetic field vector module |B| (all in nT);
- The predicted Dst index itself (in nT);
- Time characteristics describing the phase of the Earth's rotation around the Sun and around its axis.

SW and IMF parameters were measured at the Lagrange point L1 in the Sun-Earth system.

Each parameter, except for time characteristics, was described by 25 hourly average values: the current (0^{th}) one, and 24 historical values for the last day (a total of 8 physical values * 25 hours + 4 time characteristics = 204 IF). We used data for the period from November 1997 to the end of 2017 as the training set (from which 20% were randomly allocated to the validation set) and from the beginning of 2018 to the end of 2022 as the independent test set. The target variable was the Dst value one hour ahead.

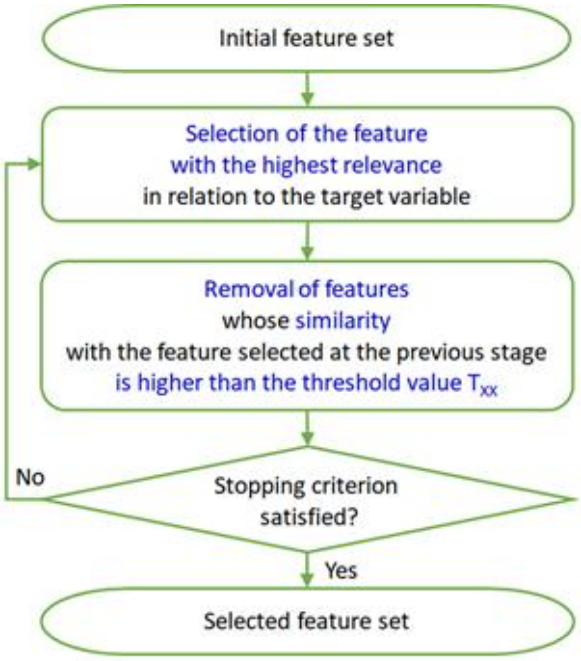

Fig. 1. Iterative algorithm for selection of input features.

III. Feature Selection Algorithm

To rank IFs by importance, the method based on the iterative selection of features with the highest correlation with respect to the target variable and the exclusion of features with high cross-correlation was used. The general scheme of the algorithm is presented in Fig. 1.

The algorithm has two main parameters (thresholds).

T_{XX} threshold is the maximum allowed cross-correlation between any pair of selected features (Fig.1).

T_{XY} threshold is the minimum allowed correlation of any selected feature with the target variable.

The pair of these parameters defines both the number of IF selected by the algorithm and the set of IF selected for each physical quantity.

IV. Neural Network Architecture and Parameters

The following architecture and parameters of the NN were used:

- Architecture: Multilayer perceptron, 1 hidden layer with 32 neurons;
- Optimization algorithm: Stochastic gradient descent (SGD);
- Learning rate 0.01, moment 0.5;
- Mini batch size – 200 patterns;
- Early stopping method – stopping training after 200 epochs without improving the results on the validation set.

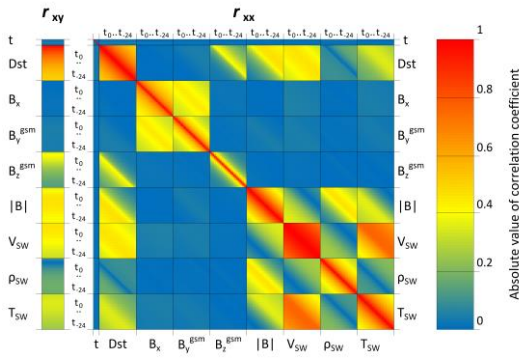

Fig. 2. Heatmap for correlations of input features with target variable and with each other.

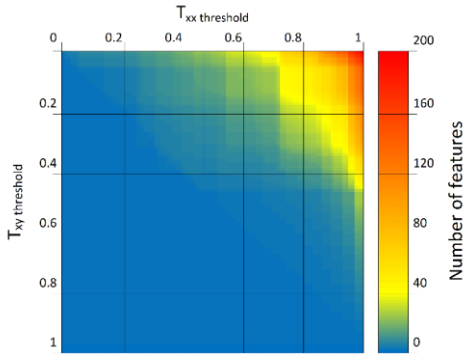

Fig. 3. Heatmap for number of features selected versus values of the thresholds.

Each neural network was trained 5 times, statistics of the application results were averaged.

V. Results

Fig. 2 shows the correlations of input features with target variable and with each other using a heatmap.

As can be seen from the presented diagram, in addition to the Dst index itself, the best correlation with Dst is shown by the IMF module |B| and its Bz component, as well as the velocity and temperature of the SW. Two latest physical variables demonstrate high cross-correlation.

Fig. 3 shows the dependence of the number of selected features on the algorithm parameters described above in Section III (T_{XX} and T_{XY} thresholds).

Table 1 shows the results of NN trained only on the features selected for various combinations of parameters Txx and T_{XY}. Combinations with approximately a 5-fold reduction in the number of input features (down to about 40) were selected. The trivial inertial model is the simplest reference model always used to assess model quality: its prediction is equal to the last known value; models with quality indicators worse than those of the trivial inertial model should be discarded. The threshold combination Txx=1, Txy=0 corresponds to selection of the full set of features (no features discarded). Txx=1 corresponds to a simple correlation filter selecting all features with correlations with target variable exceeding the T_{XY} threshold.

TABLE I. Comparison of statistical indicators of different models on the test dataset (2018-2022).

Parameters of feature selection	Number of features	Test set (2018-2022)		
		R^2	*MAE*	*RMSE*
Trivial inertial model	1	0.9268	0.1229	0.1803
Txx=1, Txy=0	204	**0.9578**	**0.0961**	**0.1351**
Txx=1, Txy=0.45	40	*0.9379*	*0.1120*	*0.1633*
Txx=0.9, Txy=0.3	39	**0.9509**	**0.1037**	**0.1456**
Txx=0.85, Txy=0.2	38	0.9499	0.1053	0.1472
Txx=0.8, Txy=0.15	41	0.9498	0.1050	0.1473
Txx=0.75, Txy=0.15	39	0.9491	0.1057	0.1480

The quality indicators used in Table I are the multiple determination coefficient (R^2), the mean absolute error (MAE) and the root mean squared error (RMSE). R^2 is dimensionless, it compares the tested model with a simple model whose prediction is equal to the average value of the target variable over the whole tested set, and its best possible value is equal to 1. Both types of errors have the same dimension as the target variable (nT).

Figure 4 shows the features selected by the algorithm for the best Txx and Txy pair from Table I - Txx=0.9 and Txy=0.3.

VI. Conclusion

Based on the above results, we can draw the next conclusions:

- As the result of selection of significant input features by the studied algorithm, optimal sets of input features were selected for Dst-index prediction horizon equal to 1 hour. The results are consistent with the physical notions on the groups of the most significant input features.
- The proposed algorithm selects input features that are meaningful from the physical point of view.
- The use of the algorithm makes it possible to reduce the number of input features 5-fold without significantly worsening the results.
- With the same number of input features, the results obtained are better than when using a simple correlation filter not taking into account the multicollinearity of the input features.
- Further optimization of algorithm and neural networks parameters is necessary.

References

[1] A.V. Vorob'ev, V.A. Pilipenko, T.A. Enikeev, G.R. Vorob'eva, "Geoinformation system for analyzing of the dynamics of extreme geomagnetic disturbances from observations of ground stations", Computer Optics, vol. 44, N5, pp. 782-790, 2020. DOI: 10.18287/2412-6179-CO-707.

[2] M. Sugiura, "Hourly values of equatorial Dst for the IGY" Ann. Int. Geophys. Pergamon Press, Oxford. vol. 35. p. 9-45. 1964.

[3] R. Vladimirov, V. Shirokii, O. Barinov, I. Myagkova, S. Dolenko. Investigation of the Importance of Input Features by Linear Regression in Predicting the Geomagnetic Index by Machine Learning. IEEE Proceedings, 2022, VIII International Conference on Information Technology and Nanotechnology (ITNT). DOI: 10.1109/ITNT55410.2022.9848686.

[4] J.-G. Wu and H. Lundstedt, "Prediction of geomagnetic storms from solar wind data using Elman recurrent neural networks," Geophys. Res. Lett., vol. 23, pp. 319 – 322, 1996. DOI:10.1029/96GL00259

[5] H. Lundstedt and P. Wintoft, "Prediction of geomagnetic storms from solar wind data with the use of a neural network," Ann. Geophys., vol. 12, pp. 19 – 24, 1994. DOI: 10.1007/s00585-994-0019-2

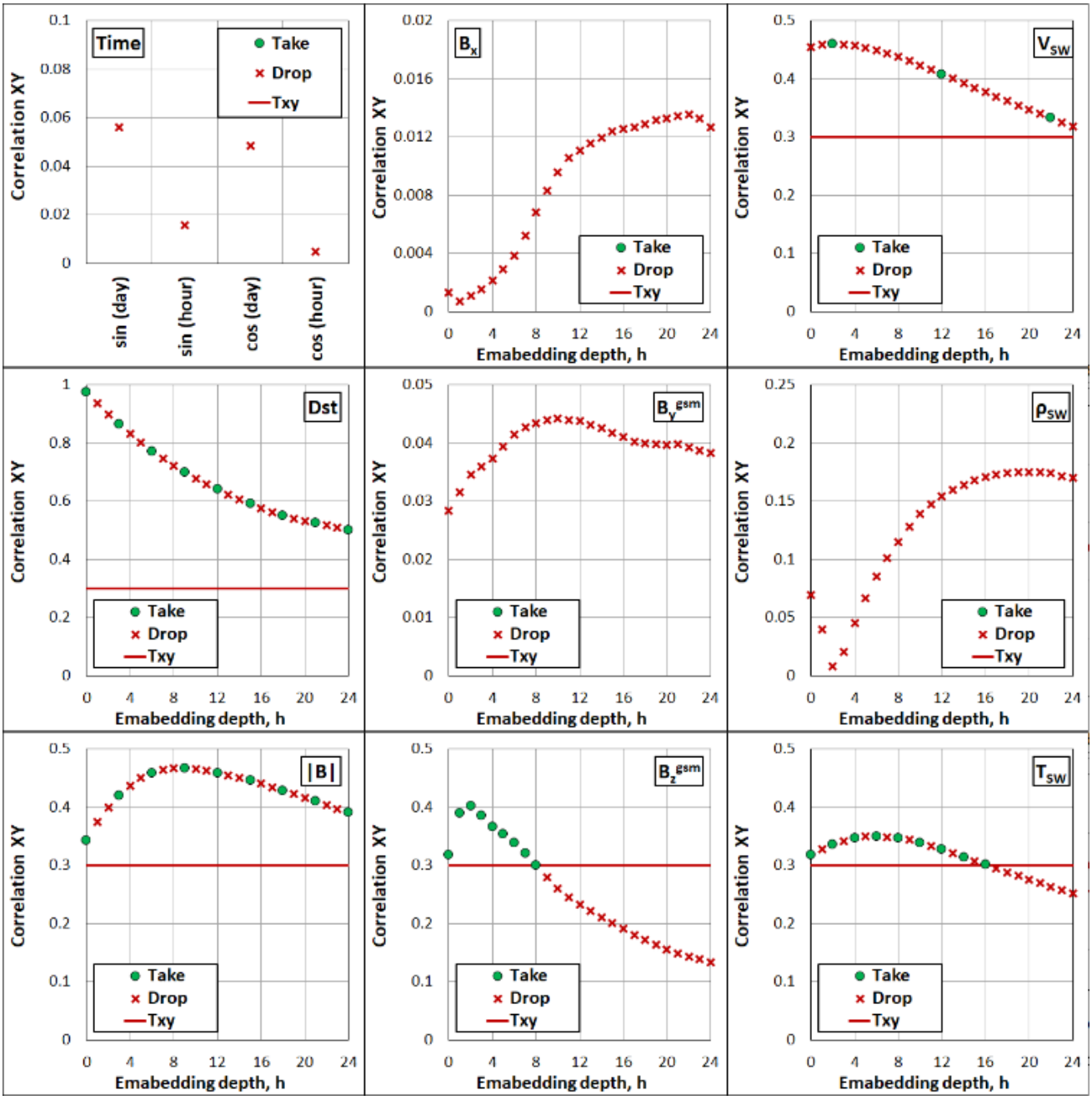

Fig. 4. The results of feature selection for T_{XX}=0.9, T_{XY}=0.3.

[6] G.M. Lindsay, C.T. Russell, J.G. Luhmann, "Predictability of Dst index based upon solar wind conditions monitored inside 1 AU," J. Geophys. Res., vol. 104, number A5, pp. 10335–10344, 1999. DOI: 10.1029/1999JA900010

[7] N.A. Barkhatov et al., "Comparison of efficiency of artificial neural networks for forecasting the geomagnetic activity index Dst," Radiophysics and Quantum Electronics, vol. 43. Number 5, pp. 347-355, 2000. DOI: 10.1007/BF02677150

[8] S.A. Dolenko, Yu.V. Orlov, I.G. Persiantsev, Ju.S. Shugai, "Neural network algorithm for events forecasting and its application to space physics data," Lecture Notes in Computer Science, vol. 3697. pp. 527-532, 2005. DOI: 10.1007/11550907_163

[9] V. Shiroky, I. Myagkova, and S. Dolenko., "Use of ensemble approach and stacked generalization for neural network prediction of geomagnetic dst index," Lecture Notes in Computer Science vol. 9887, pp. 541–541, 2016. DOI: 10.1007/978-3-319-44781-0

[10] I.N. Myagkova, V.R. Shirokii, R.D. Vladimirov, O.G. Barinov, and S.A. Dolenko., "Prediction of the dst geomagnetic index using adaptive methods," Russian Meteorology and Hydrology, vol. 46, issue 3, pp.157–162, 2021. DOI: 10.3103/S1068373921030031

[11] A.O. Efitorov, I.N. Myagkova, V.R. Shirokii, and S.A. Dolenko., "The prediction of dst index based on machine learning methods," Cosmic Research (English translation of Kosimicheskie Issledovaniya), vol. 56, issue 6, pp. 434–441, 2018. DOI: 10.1134/S0010952518060035.

[12] R.D. Vladimirov, V.R. Shirokiy, I.N. Myagkova, O.G. Barinov, and S.A. Dolenko. Comparison of the effectiveness of machine learning methods in studying the importance of input features in the problem of predicting the geomagnetic index Dst. Geomagnetism and Aeronomy, 63(2):190–201, 2023. DOI: 10.1134/S0016793222600795

[13] J.A. Lazzús, P. Vega, P. Rojas, I. Salfate, "Forecasting the Dst index using a swarm-optimized neural network," SpaceWeather, vol. 15, pp. 1068–1089, 2017. DOI: https://doi.org/10.1002/2017SW001608.

[14] S.-I. Akasofu, S. Chapman, "Solar-Terrestrial Physics," Clarendon Press, Oxford. P. 889, 1972.