ЖУРНАЛ ФІЗИЧНИХ ДОСЛІДЖЕНЬ
т. 15, № 2 (2011) 2002(8 с.)

JOURNAL OF PHYSICAL STUDIES
v. 15, No. 2 (2011) 2002(8 p.)

REGRESSION MODELLING OF GEOMAGNETIC ACTIVITY

A. S. Parnowski
Space Research Institute NASU & NSAU,
prosp. Akad. Glushkova, 40, build. 4/1, Kyiv–187, 03680 MSP, Ukraine
(Received January 2, 2011)

The regression modelling method is described in detail and applied to the problem of geomagnetic indices forecasting. It provides not only quality forecasts but also gives new information about the underlying physics of the solar wind-magnetosphere interaction.
Key words: space weather, Solar wind, magnetosphere, regression modelling.

PACS number(s): 02.70.Rr, 94.05.Sd, 94.05.sx, 94.30.Lr, 94.30.vf

I. INTRODUCTION

Space weather forecasting is a challenging and non-trivial activity. The most straightforward approach to space weather prediction is to study the whole complex chain of physical processes involved in magnetospheric dynamics and conjugate them in a global model of the evolution of the magnetosphere under the influence of the solar wind. Unfortunately, this is not yet possible due to our poor understanding of the physics of the interaction between the solar wind and the magnetosphere. For this reason, different approaches should be tried. Today the optimal combination of virtues and vices is provided by the methods involving time series analysis and data mining [1]. They provide a moderate prediction time (up to several hours) with the highest available accuracy ($> 80\%$). They are very effective and easy to use but strongly depend on satellite data availability. These are "black box" or "input-output" models, which seek only to reproduce the system's output in response to changes of its inputs. The model terms are usually physically interpretable and thus useful for construction of new phenomenological models. For this reason, this method can not only provide a space weather forecast per se, but also can improve our knowledge of the underlying physics and thus increase the efficiency of other methods.

Multidimensional time series analysis can be performed using the methods of statistics, signal processing, informatics, fuzzy logic, etc. The most widely used variations are artificial neural networks, optimization, and correlation analysis.

Artificial neural networks [2–5] provide short-term predictions up to 4 hours with the correlation coefficient of 0.79 in the paper [4]. Earlier implementations of this approach experienced significant difficulties predicting strong geomagnetic storms with $K_P > 5$, but this approach remains one of the most popular alongside the empirical methods.

Optimization approach [6–12] seems to be more successful being able to provide 8-hour predictions in the paper [8]. However, in the papers based upon the optimization methods the volume of the dataset used usually does not exceed 1 year, which is insufficient to correctly describe the solar cycle.

Correlation analysis [13–17] gives interesting results, but it was used solely for developing and constraining empirical models [16].

However, most of these methods have a common feature: they lead to a regression relation at some point, so it seems natural to skip all the preliminary steps and instantly use the regression analysis without unnecessary multiplication of entities. Regression analysis itself was attempted earlier by Srivastava [18], but it was used to estimate the probability of intense and super-intense storm occurrence depending on the solar and interplanetary parameters. She was able to predict 2 of 4 super-intense and 5 of 5 intense CME-driven storms during the 1996–2002 period using another 46 CME-driven storms for training.

Here we describe a new approach, named "regression modelling", which allows to achieve accurate short-term forecasts of geomagnetic indices, which we will use as quantitative characteristics of space weather. The proposed method is statistical, but has some features of empirical models. It is based upon the regression analysis and mathematical statistics. This approach involves the inductive construction of a regression relation between output and input values. It can provide accurate short-term and, to a certain extent, medium-term forecasts and gives new information about the underlying physics, thus contributing to the solar-terrestrial physics.

Some preliminary descriptions of different aspects of this method can also be found in the articles [19–23].

II. DESCRIPTION OF THE REGRESSION MODELLING METHOD

In this Section we will give a formal description of the regression modelling method in its most general form, providing links to space weather forecasting where necessary.

Consider a discrete dynamical system (in our case, terrestrial magnetosphere) with an unknown number K_{tot} of inputs u_k and one output y (one of geomagnetic indices), which is simply one of inputs u_k. At each step t we know only $K < K_{tot}$ inputs $u_k(t)$, $k = \overline{1, K}$ ($\overline{1, K}$ means all the integer numbers from 1 to K inclusively) and an output $y(t)$. Then at an arbitrary step T we can

write the predicted value of the system's output in the form

$$y(T+\Theta) = y^*(T+\Theta) + \Delta y(T+\Theta), \tag{1}$$

where Θ is the lead time of the forecast (the number of hours the forecasted value is ahead of the last measured value), $y^*(T+\Theta)$ is the estimated predicted value, and $\Delta y(T+\Theta)$ is the uncertainty, which we assume to be random and stochastic. We are also forced to assume that all values are distributed normally to be able to use the methods of mathematical statistics, though this is, of course, not always true. We also assume that the statistical properties of the dynamical system do not change on the time scale Θ. Note that the uncertainty $\Delta y(T+\Theta)$ can be also assumed to be members of some set or to lie in a certain interval [24]. The predicted value $y^*(T+\Theta)$ is expressed through a partial regression relation [25]:

$$y^*(T+\Theta) = C_0 + \sum_{i=1}^{m} C_i x_i(T), \tag{2}$$

where x_i, $i = \overline{1,m}$ are the regressors, which are arbitrary functions of input quantities $u_k(t)$, which are already measured at the time T when the forecast is made, C_i, $i = \overline{0,m}$ are the regression coefficients, C_0 is the coefficient of the constant regressor $x_0 \equiv 1$, and m is the number of variable regressors.

We choose the regressors x_i in the form of products of powers of the input quantities

$$x_i(t) = \prod_{k=1}^{K} u_k^{p_k}(t-l),\ l = \overline{0,L}, \tag{3}$$

where p_k are powers, which can be equal to zero or any natural number, l is the lag, and L is the maximal lag. This is equivalent to using a Kolmogorov-Gabor polynomial [26] as a basis function. In contrast to empirical models we do not add fitting parameters and all the regressors have physical meaning. Note that different sets of regressors should be taken for different values of the lead time Θ.

Of course, $y(T+\Theta)$ is also affected by the inputs $u_k(T+1), \ldots, u_k(T+\Theta)$. However, we don't know the values of these inputs at the moment T and thus cannot use them. This means that by increasing the lead time Θ we sacrifice the ability to take into account the processes with time scales less than Θ.

Now we should determine the coefficients C_i by the generalised least squares method over a large sample of solar wind and geomagnetic data, with equal statistical weights of all points. It is usually advised [27] to use singular value decomposition for this task, but it is a rather slow algorithm and requires multiplying two $n \times m$ matrices, which requires a lot of RAM (each matrix requires approximately 1.67 MB of RAM per regressor for the sample described in Section III). For this reason, we used the simple Gauss–Jordan elimination [27]. The latter also produces a covariance matrix, which is an additional advantage of this algorithm. The expression for C_i has the form [27, 28]:

$$C_i(t) = \sum_j a_{ij}^{-1} b_j, \tag{4}$$

where $(\bullet)^{-1}$ denotes matrix inversion,

$$a_{ij} = \sum_{t=1}^{T} x_i(t) x_j(t), \tag{5}$$

$$b_j = \sum_{t=1}^{T} y(t) x_j(t). \tag{6}$$

Its standard error is given by [27, 28]

$$\Delta C_i = \sqrt{\zeta_{ii}}, \tag{7}$$

where

$$\zeta_{ij} \equiv \mathrm{cov}(x_i x_j) = \sigma^2 a_{ij}^{-1} \tag{8}$$

is the covariance matrix,

$$\sigma = \sqrt{\frac{S}{n-m}} \tag{9}$$

is the residual mean square (RMS) error of the forecast,

$$S = \sum_{t=1}^{n} (y(t) - \overline{y})^2 \tag{10}$$

is the residual sum of squares (RSS), and n is the number of datapoints in the sample. The value $n-m$ here is a number of degrees of freedom.

The statistical significances of the regressors are determined according to Fisher's F-test [25, 29]. This test allows separating significant and insignificant regressors. The insignificant parameters are then rejected and the routine is repeated until the regression contains only significant regressors. This is done in the following way. After processing the data with the least square method, Fisher significance parameter F_i was determined for each regressor by comparing residuals for the full model and the model without the regressor in question [25, 28]:

$$F_i = \frac{S_i - S}{\sigma} = \left(\frac{S_i}{S} - 1\right)(n-m) \tag{11}$$

where S_i is RSS of the model without the i-th regressor. Note that this procedure involves solving m generalised least squares problems whose design matrices have the dimensions $n \times (m-1)$, so it is not too fast. Even when using Gauss–Jordan elimination, we should perform m inversions of $(m-1) \times (m-1)$ matrices, so the runtime growth cubically with m. For this reason, we should not add all the regressors at once, but rather add them gradually. This procedure will be described in Section IV. All the F_i values were compared to the values 2.71, 3.84, 5.02, 6.64, 7.88, 10.83 and 12.10, which correspond to the statistical significance levels of 90%,

95%, 97.5%, 99%, 99.5%, 99.9% and 99.95% respectively. Then, the insignificant regressors are rejected and the routine is repeated until all the regressors are significant. The number of significant regressors thus depends on the selected significance level threshold. All the results given below correspond to a minimal significance level of 90%. Since rejecting any regressor leads to the change of the statistical significances of the others, it is necessary to reject insignificant regressors in several steps, ideally one regressor at a time. In practice, since the F-test has a significant runtime, it is better to reject all insignificant regressors at once, but to gradually increase the significance threshold to the desired value.

After that, new regressors are added. After adding new regressors, all the significances are recalculated, and some of the old regressors can become insignificant.

This routine should be repeated while the addition of new regressors improves some quality characteristics. Such characteristics, depending on the goal, can be the maximum forecast error $\max\limits_{T}|\Delta y(T+\Theta)|$, the RMS error σ or the prediction efficiency (PE)

$$PE = 1 - \left(\frac{\sigma}{\sigma_S}\right)^2, \tag{12}$$

where

$$\sigma_S = \sqrt{\frac{S}{n-1}} \tag{13}$$

is the sample's standard deviation (SD), or the linear correlation (LC) coefficient

$$r(\xi,\eta) = \frac{\sum(\xi-\overline{\xi})(\eta-\overline{\eta})}{\sqrt{\sum(\xi-\overline{\xi})^2\sum(\eta-\overline{\eta})^2}} \tag{14}$$

between $y^*(T+\Theta)$ and $y(T+\Theta)$. Note that these values for the developed model should be compared to the same values for the persistence model

$$y_0^*(T+\Theta) = y(T), \tag{15}$$

which is, obviously, the simplest possible model, which states that the output value will not change since the last measurement, so it just shifts the time series one unit of time to the future.

There is one more thing worth noting about the evaluation of the models. It is possible that the developed model will be too sample-specific and fail to work on different samples, no matter how good its quality characteristics are. To avoid such a situation, the sample should be divided into 2 subsamples. The first subsample, which is commonly called the training sample, is intended for the determination of the model structure and parameters. The second subsample, which is called the test or the validation sample, is for the evaluation of the model.

The regressors x_i are generally nonlinear, so from the control theory's point of view, this method is able to describe discrete dynamical systems with strong nonlinearity. This is an essential feature of the regression modelling method. To obtain a forecast of the sought geomagnetic index, one has to sum up the regression relation over a given sample.

Now we have only one question left: the accuracy of the obtained models. It is important to understand that there are several different sources of errors. First of all, there is an error caused by the incompleteness of our model, which, assuming that Δy is distributed normally, is equal to

$$\Delta y_M^2 = \sum_{i=1}^{m}\sum_{j=1}^{m}\zeta_{ij}x_i x_j. \tag{16}$$

Then, there are errors caused by uncertainties in the determination of the input quantities. We can take them into account by calculating the partial derivatives of the output with respect to the inputs:

$$\Delta y_u^2 = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\frac{\partial y}{\partial u_{k_1}}\frac{\partial y}{\partial u_{k_2}}\mathrm{cov}(\mathrm{u}_{\mathrm{k}_1},\mathrm{u}_{\mathrm{k}_2}). \tag{17}$$

They divide into 3 types: measurement errors, which are negligible, errors caused by the filtration of the input data, which are provided in the OMNI 2 database, and intrinsic temporal irregularities of the input parameters. Finally, there are spatial variations Δy_S, which are due to the fact that we measure the interplanetary magnetic field (IMF) and the solar wind plasma parameters locally and assume that they are uniformly distributed in space. To estimate this type of errors we need to perform simultaneous multipoint measurements of input parameters, which became possible since the launch of STEREO mission in 2006, but lies beyond the scope of this method.

This method has two implementations: static and adaptive. The static version involves calculation of the regressors and the coefficients on the training sample with forecasting made on the validation sample. The adaptive version involves calculation of the regressors on the training sample and calculation of the coefficients simultaneously with the prediction, so that the coefficients are recalculated at each step as new data become available.

III. DESCRIPTION OF DATA USED

We used the OMNI 2 database [30], which contains IMF, solar wind and geomagnetic data, averaged over 1-hour intervals (at the time of publication it contained 54 parameters in total, starting from 1 January 1963). This database covers a vast number of spacecraft. In recent years the data come from spacecraft located in the first Lagrange (L1) point, also called a libration point, which is situated along the Earth–Sun axis approximately 0.01 AU (1.5 millions of kilometres) from the Earth. For typical interplanetary conditions ($V = 470\,\mathrm{km\,s^{-1}}$) a spacecraft located there provides real-time data with a 40-minute lead time.

The data before 1976 are scarce and of poor quality and their inclusion in the dataset negatively impacts

its characteristics. Also, the final $D_{\rm ST}$ index is available only up to 2003, and we should reserve a validation sample to test our models. For this reason we used a training sample that ranges from 1 January 1976 to 31 December 2000, thus forming a continuous 25-year time series with a total of $n = 219168$ datapoints. For the $D_{\rm ST}$ index the mean is $\overline{D_{\rm ST}} = -18.3\,{\rm nT}$, the median is $-23\,{\rm nT}$, the mode is $8\,{\rm nT}$, and the standard deviation is $\sigma_{D_{\rm ST}} = 24.6\,{\rm nT}$. The distribution of the $D_{\rm ST}$ index visually represents a normal one, but the Pearson's χ^2 test [25] disproves this null-hypothesis at the 99.99% confidence level ($\chi^2 = 416125.8$). This is due to flatter wings of the distribution, which are caused by the periodicities of the ACF. For the a_P index the mean is $\overline{a_P} = 14.9\,{\rm nT}$, the median is $5\,{\rm nT}$, the mode is $27\,{\rm nT}$, and the standard deviation is $\sigma_{a_P} = 20.0\,{\rm nT}$. In Section VI we will also use a sample ranging from 1 January 2001 to 31 December 2003 (the latest value of the final $D_{\rm ST}$ index) and its subsamples to test the developed models.

Unfortunately, during intense storms the instruments aboard the spacecraft are often turned off to prevent permanent damage to them and some or all of the input values are unavailable. By rejecting filled values from the time series, we obtain the sample, which can be divided into different subsamples for specific purposes. Of course, the resulting sample will vary according to the exact dependences of the regressors on the input quantities. For example, if our model contains a regressor, which depends on the ion density with a lag of 5 hours, then we have to reject each datapoint whose 5th predecessor contained a filled value of the ion density. Also, before the February 26, 2009 update of the OMNI 2 database we were forced to insert some missing data from other databases, but now all the available data are included for all years.

IV. SELECTION OF THE REGRESSORS AND THE MEMORY OF GEOMAGNETIC INDICES

Now, all that remains is to choose some initial set of regressors. It seems natural to start from the previous values of the output value itself. This will also give us a possibility to investigate temporal variations of the geomagnetic indices. For this purpose, we should construct an autoregression model [23]

$$y^*_{\rm AR}(T+\Theta) = C_0 + \sum_{l=0}^{L} C_l y(T-l) \qquad (18)$$

or, in other words,

$$x_l(t) = y(t-l),\ l = \overline{0, L}. \qquad (19)$$

This model alone is not sufficient to correctly forecast space weather, but it sets a basis for the construction of models that are able to do so.

Let us determine the maximum reasonable value of L. For this purpose, we plot the autocorrelation function (ACF) at $\Theta = 1$ for the $D_{\rm ST}$ (Figure 1) and the a_P index (Figure 2). A brief glance at the ACF is enough to tell that neither of the geomagnetic indices can be treated as a Markov process. In fact, both the $D_{\rm ST}$ and the a_P indices are periodically correlated.

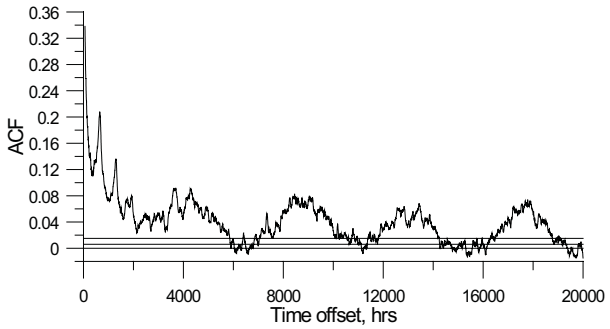

Fig. 1. Autocorrelation function of the $D_{\rm ST}$ index.

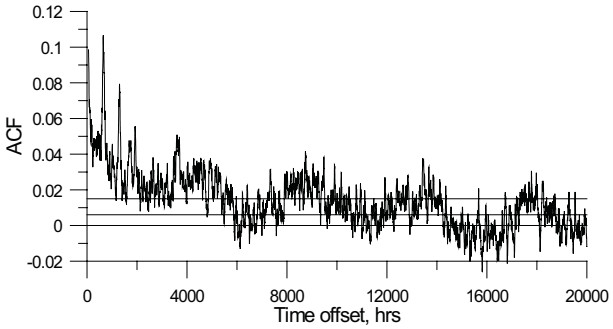

Fig. 2. Autocorrelation function of the a_P index.

One can see that in both cases the ACF tends to a sinusoid with a period close to half a year. Besides the half-a-year periodicity one can also notice the 27-day periodicity, caused by Carrington rotation of the Sun. The former is caused by seasonal variations, which yields a question: if there were no temporal variations, what would ACF tend to at large offsets? If the distribution of $D_{\rm ST}$ and a_P was normal, the answer would be zero. However, the distribution slightly deviates from the normal one, so ACF can tend to some non-zero quantity.

To determine this quantity we need to remove temporal variations. For this purpose we need to calculate the ACF of a random sample with the same statistical characteristics as the actual sample. The easiest way to get such a sample is to process the actual sample with a permutation method, which is widely used in astronomy e.g. for calculation of 2-point correlation functions. This method involves random shuffling of the sample. Using this method many times (10000 times in our case) and calculating the correlation coefficient each time, we get the distribution of the correlation coefficient by Monte Carlo method.

The distribution of the coefficient for this sample appeared to be very close to a normal distribution. For the $D_{\rm ST}$ index the mean was equal to $0.008\,{\rm nT}$ and the variance — to $5.1 \cdot 10^{-6}\,{\rm nT}^2$. For the a_P index the mean was equal to $0.0052\,{\rm nT}$ and the variance — to $8.4 \cdot 10^{-6}\,{\rm nT}^2$. The maximum recorded value in 10000 trials was equal

to 0.015 nT for both indices. The top and the mean values are depicted on Figures 1 and 2 by the horizontal lines. As one can see, in reality the correlation coefficient exceeds this value at most times due to temporal variations. The ACF of the $D_{\rm ST}$ index crosses the top line for the first time at about 6000 hours, though the difference between the ACF and the sine with a half-year period crosses it at about 2000 hours, which is about 3.5 27-day Carrington periods. The ACF of the a_P index crosses the top line at about 1200 hours, though the difference between the ACF and the sine with 27-day period crosses it at about 1000 hours. We will take the latter value as a rough estimation of L. This hints that rather old values of the geomagnetic indices can be quite significant.

Let us return to equation (18). Applying the F-test we can determine which previous $D_{\rm ST}$ and a_P values are statistically significant. We did not search statistically significant values for $L \geqslant 1000$, but it is possible that there are even older statistically significant values. A similar situation was reported by Johnson and Wing [16] regarding K_P: "the significance is often quite large for extended periods of time (10–20 days)". In fact, after adding the regressors corresponding to satellite data, we still found statistically significant values of $D_{\rm ST}$ as far as 850 hours ago (over 35 days). The statistical significance of these oldest values can be over 99.9%.

After determining which previous values of the geomagnetic indices are statistically significant, we added all the spacecraft-measured parameters available in the OMNI 2 database with lags up to 24 hours for hourly values and without lag for daily values. Naturally, common sense also counts: for example it would be silly to add IMF components in GSE and GSM coordinates at the same time. If some regressors x_i have large statistical significance (we used a threshold $F_i \geqslant 100$), we also added all their possible cross-products and powers $\prod\limits_i x_i^{p_i}$. For practical purposes we used non-negative integer values of p_i, limited by a total power of 4: $\sum\limits_i p_i \leqslant 4$. The total number of regressors in final models varies roughly from 50 to 250. Since it is quite large, we will not give here any lists of regressors or coefficients.

Of course, this method does not guarantee that all the significant regressors will enter the regression, since the regressors are not orthogonal and thus more than one expansion is possible.

V. TEMPORAL VARIATIONS OF GEOMAGNETIC INDICES

On Figure 1 one can see a clear seasonal dependence of the $D_{\rm ST}$ index. Indeed, if we select two subsamples, corresponding to the summer and the winter in northern hemisphere, bounded by vernal and autumnal equinoxes, and verify the hypothesis that the difference between the corresponding average $D_{\rm ST}$ values is statistically significant using a one-sided Student's test [25], we obtain $t_\infty = 54.7$, which is well over 99.95% significant. Values of t_∞ corresponding to 99% and 99.95% confidence levels are equal to 2.3 and 3.3 respectively. For the diurnal asymmetry the Student's test gives $t_\infty = 3.3$, which corresponds to a significance level of 99.95%. Note that formally the Student's test is applicable only to normally distributed values and the distribution of $D_{\rm ST}$ has flatter wings than the normal one. However, taking into account the obtained large values of t_∞, we can be sure in qualitative conclusions made.

This dependence was described in many articles, for example in [31–34], but the reason behind it is still disputed. Most authors believe these asymmetries are caused by either of two cusps turning to the sunlit side due to annual rotation of the Earth with respect to the Sun. However, O'Brien and McPherron [34] state that this mechanism would give only 17% of the observed asymmetry. Takalo and Mursula [33] connected the diurnal variations of $D_{\rm ST}$ with an inhomogeneous distribution of $D_{\rm ST}$ network stations with respect to the longitude.

In our opinion, this behaviour is most likely caused by an asymmetry of the $D_{\rm ST}$ stations with respect to the geomagnetic equator. In fact, only the Hermanus station is located in the southern hemisphere (dipole latitude $-33.3°$), while the other 3 stations are located in the northern hemisphere (Kakioka $+26.0°$, Honolulu $+21.1°$, San Juan $+28.0°$). Thus, when the subsolar point is in the northern hemisphere, there are 3 stations nearby, but when it is in the southern hemisphere, there is only one station.

The official definition of the $D_{\rm ST}$ index is [35]

$$D_{\rm ST}(t) = \frac{\langle \Delta H(t) - S_q(t) \rangle}{\langle \cos\theta \rangle}, \qquad (20)$$

where $\Delta H(t)$ is the difference between the observed and the baseline values of the H-component of the geomagnetic field,

$$S_q(t,s) = \sum_m \sum_n A_{mn} \cos(mt + \alpha_m) \cos(ns + \beta_n), \quad (21)$$

is the solar quiet daily variation, s is the current month, θ is the geomagnetic latitude, and $\langle \bullet \rangle$ is an average over 4 contributing stations. Since it does not depend on the sign of θ, any sources in the northern hemisphere will affect the $D_{\rm ST}$ index 3 times stronger than their southern counterparts. Note that temporal variations of the $D_{\rm ST}$ index are provided not only by actual temporal variations of the H-component, but also by the term (21), which explicitly contains them.

Usually, the $D_{\rm ST}$ index is associated with the ring current, which is highly asymmetric during the geomagnetic storm [36–38]. Moreover, the $\Delta H(t)$ term includes all sources of the magnetic field, such as ionospheric currents, power lines, industrial facilities, railroads and so on [39]. This fact explains, among others, the 7-day periodicity, which is of purely anthropogenic origin. Also, initially the $D_{\rm ST}$ index is derived from very noisy data, and it is possible to introduce additional errors during its processing.

This means, in particular, that if some input parameter has high statistical significance, it is not necessary

geoeffective, but if it is statistically insignificant, it is most certainly not geoeffective.

Taking this known geoeffective factor as an example we demonstrate how easily one can take it into account using regression approach. To do so one should simply add the synthetic inputs

$$u_{K+1}(t) = \sin((\mathrm{DOY(t)} - 80)\pi/182.625) \tag{22}$$

and

$$u_{K+2}(t) = \cos((\mathrm{DOY(t)} - 80)\pi/182.625) \tag{23}$$

Here DOY is the day of the year, 80 is the number of days between the beginning of the year and the vernal equinox, and 182.625 is the number of days in half a year. The first of these terms is significant and describes summer-winter asymmetry, and the second one (which appears statistically insignificant) describes an absent spring-autumn asymmetry. Likewise, for the diurnal asymmetry the corresponding synthetic inputs will be

$$u_{K+3}(t) = \sin((\mathrm{UT(t)} - 2)\pi/12) \tag{24}$$

and

$$u_{K+4}(t) = \cos((\mathrm{UT(t)} - 2)\pi/12) \tag{25}$$

Here 2 is the time difference between UT and the local time at the northern magnetic pole, and 12 is the number of hours in half a day. Both these terms are significant.

The coefficient of the regressor, equal to $u_{K+3}(t)$, is less than the actual difference between the summer and the winter mean D_{ST} values by an order of magnitude. This can be explained in the following way: there are other regressors, which depend on the parameters with statistically significant summer-winter asymmetry, e.g. previous D_{ST} values. They provide the lion share of the summer-winter asymmetry of the D_{ST} index. A good example of such a parameter is the international sunspot number R, which has a 27-day periodicity due to Carrington's rotation of the Sun. Nevertheless, there is a small difference which can not be expressed with these terms. Including it into regression, we obtain these statistically significant regressors. To further illustrate this point let us consider as an example a value

$$X(t) = \mathrm{const} + A\sin\omega t \tag{26}$$

In the regression it will look like

$$\begin{aligned} X(t+\Delta t) &= X(t) + A(\sin\omega(t+\Delta t) - \sin\omega t) \\ &= X(t) + A((\cos\omega\Delta t - 1)\sin\omega t + \cos\omega t\sin\omega\Delta t). \end{aligned} \tag{27}$$

The first term in brackets is of order $(\omega\Delta t)^2$, and the second one is of order $(\omega\Delta t)$ in the natural assumption that $\omega\Delta t \ll 1$. So, it will seem that the coefficient is $A(\omega\Delta t)$ rather than A. Note that this is just an example and has nothing to do with the actual regressors.

However, the distribution of mean D_{ST} values vs. the day of the year is much more complicated. Among its features there is a strong asymmetry between the summer and the winter on the one hand and the spring and the autumn on the other. To take it into account we introduced additional terms into our regression, which are the powers of $u_{K+1}(t)$ and their products with the powers of $u_{K+2}(t)$. The sum of regressors with the corresponding coefficients is very similar to the actual distribution. Note that the coefficients were obtained independently from the distribution.

We did the same thing with the diurnal asymmetry. The cross-product $u_{K+1}(t) \cdot u_{K+3}(t)$ is also significant and should be included in the regression. After this we obtained a joint distribution of seasonal and diurnal variations of D_{ST} index, which contains 18 regressors. Increasing the number of the regressors describing temporal variations of the geomagnetic activity we can improve the accuracy of this distribution. In particular, one could add 11-year Schwabe's and 22-year Hale's solar cycles, higher powers of $u_{K+1}, \ldots, u_{K+4}(t)$ etc.

Thus, we demonstrated how easily one can take into account known effects in this method's framework. Note that these regressors do not depend on satellite parameters and can be used alongside the previous values of geomagnetic indices. In this case they improve the forecasting skills of the model and make some of the autoregression terms insignificant.

More details on the temporal variations of the D_{ST} index and the identification of new geoeffective parameters can be found in the article [21].

VI. FORECAST RESULTS

Now let us return to the main goal of this article and discuss forecasting skills of the developed models.

First of all, we determined the statistical characteristics of the developed models with 3 hours lead time over the main sample and listed them in Table 1. It is divided into 3 parts for D_{ST}, a_P and K_P indices. For D_{ST} and a_P indices we provide RMS errors σ, prediction efficiencies and linear correlation coefficients for the developed (r) and the persistence (r_0) models, and for K_P index we provide the percentages of points with deviations lying within $\pm\frac{1}{3}$ and ±1 bins. All these data are provided for 4 models: persistence, autoregression, linear and nonlinear models.

However, some other approaches work well inside the training sample, where they "saw" the correct answer, but perform poorly during the out-of-sample validation. Let us verify that our approach holds no such vices using the validation sample ranging from 1 January 2001 to 31 December 2003. These results are also provided in Table 1. One can see that our models worked well over the training sample with an exception of the nonlinear model for the D_{ST} index which appeared to be too sample-specific.

Index	$D_{\rm ST}$				a_P				K_P	
Model	σ, nT	PE, %	r, %	r_0, %	σ, nT	PE, %	r, %	r_0, %	$\pm\frac{1}{3}$, %	±1, %
Training sample (Jan 1, 1976 — Dec 31, 2000)										
Persistence	10.38	82.2	91.1	91.1	13.65	53.5	76.8	76.8	46.2	81.8
Autoregr.	9.92	83.7	91.5	91.1	12.74	59.6	77.2	76.8	80.1	93.9
Linear	7.89	85.6	92.5	89.2	8.06	68.5	82.9	71.6	80.0	94.3
Nonlinear	7.38	87.8	93.7	89.3	7.69	71.3	84.6	71.6	81.6	95.6
Validation sample (Jan 1, 2001 — Dec 31, 2003)										
Persistence	12.00	82.3	91.2	91.2	15.72	53.8	76.9	76.9	46.9	82.6
Autoregr.	11.59	83.9	91.6	91.2	15.07	58.8	76.9	76.8	80.3	94.0
Linear	9.26	87.7	93.7	90.6	10.35	67.2	82.4	76.6	81.1	94.8
Nonlinear	12.11	79.8	89.6	90.9	10.80	64.3	80.9	76.6	83.2	96.1

Table 1. Statistical characteristics of the developed models for short-term forecasting (3 hours ahead).

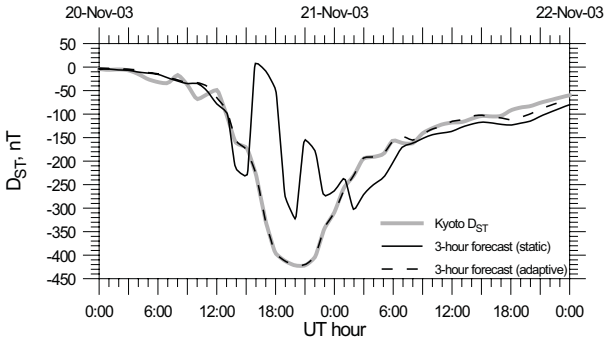

Fig. 3. Forecasted and measured values of the $D_{\rm ST}$ index for the Halloween storm of 2003.

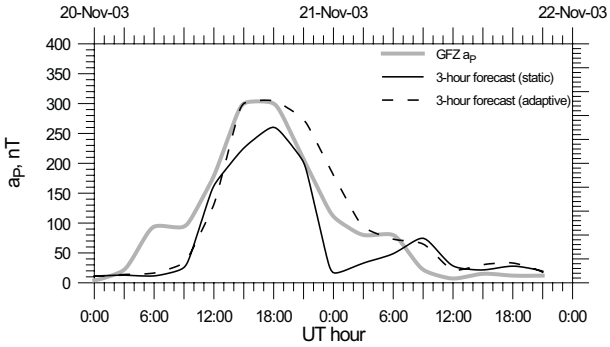

Fig. 4. Forecasted and measured values of the a_P index for the Halloween storm of 2003.

Since the mean square error cannot tell if the deviations are random or systematic, let us plot the forecasted indices over a few subsamples corresponding to strong geomagnetic storms. Figures 3 and 4 show the forecasted and the measured values of the indices for a series of geomagnetic storms in October-November 2003. As you can see, the regression modelling method appeared to be more than adequate to forecast space weather indices.

VII. CONCLUSION

In this article we have described the regression modelling method of space weather forecasting. It provides precise and reliable short-term forecasts of geomagnetic indices at least 3 hours ahead. It is possible to increase the lead time by sacrificing the ability to describe fine temporal structure of the output value, i.e. using output values with lower temporal resolution. Thus it seems reasonable to try forecasting daily values like A_P or C_9 a few days ahead. In addition, the regression modelling method is not limited to forecasting geomagnetic indices, so we can apply it to forecast solar activity as well. It can also tell which quantities are the most geoeffective and in what way they are related to geomagnetic indices, thus contributing to the understanding of the underlying physics. Last but not least, the regression modelling method is suitable for issuing real-time space weather forecasts since it takes about one minute on an average PC to calculate the coefficients and just a few seconds to issue the forecast.

VIII. ACKNOWLEDGEMENTS

The OMNI data were obtained from the GSFC/SPDF OMNIWeb interface at http://omniweb.gsfc.nasa.gov/

This research was partially supported by the Fundamental Research Programme "GEO-UA", Contract No. 12-7/10 and the Research Grant for Young Scientists, Contract No. 10-7/10 of the National Academy of Sciences of Ukraine.

[1] A. S. Parnowski *et al.*, Space Sci. Technol. **16**(2), 90 (2010).
[2] S. Kugblenu *et al.*, Earth Planets Space **51**, 307 (1999).
[3] S. Watanabe *et al.*, Earth Planets Space **54**, 1263 (2002).
[4] S. Wing *et al.*, J. Geophys. Res. **110**, A04203 (2005). doi:10.1029/2004JA010500
[5] G. Pallocchia *et al.*, Mem. Soc. Astron. It. Suppl. **9**, 120 (2006).
[6] X.-Y. Zhou, F.-S. Wei, Earth Planets Space **50**, 839 (1998).
[7] M. A. Balikhin *et al.*, Geophys. Res. Lett. **28**, 1123 (2001).
[8] R. F. Harrison, P. M. Drezet, in *Les Woolliscroft memorial Conference "Multipoint measurements versus theory"*, 141–146 (2001).
[9] O. K. Cheremnykh, V. A. Yatsenko, Space Sci. Technol. **14**(1), 77 (2008).
[10] O. K. Cheremnykh *et al.*, Ukr. J. Phys. **53**(5), 502 (2008).
[11] O. V. Semeniv *et al.*, J. Autom. Inform. Sci. **40**(4), 115 (2008).
[12] V. A. Yatsenko *et al.*, J. Autom. Inform. Sci. **41**(12), 58 (2009).
[13] G. K. Rangarajan, L. M. Barreto, Earth Planets Space *51*, 363 (1999).
[14] S. Y. Oh, Y. Yi, J. Korean Astron. Soc. **37**, 151 (2004).
[15] H. L. Wei *et al.*, Nonlin. Proc. Geophys. **11**, 303 (2004).
[16] J. R. Johnson, S. Wing, U.S. DoE Report PPPL-3919rev, http://www.pppl.gov/pub_report/2004/PPPL-3919rev.pdf (2004).
[17] J. R. Johnson, S. Wing, J. Geophys. Res. **110**, A04211 (2005); doi:10.1029/2004JA010638
[18] M. Srivastava, Annales Geophysicae **23**, 2969 (2005).
[19] A. S. Parnowski, Space Sci. Technol. **14**(3), 48 (2008a).
[20] A. S. Parnowski, J. Phys. Studies **13**(4), 4003 (2008b).
[21] A. S. Parnowski, Astrophys. Space Sci. **323**(2), 169 (2009a).
[22] A. S. Parnowski, J. Autom. Inform. Sci. **41**(3), 128 (2009b).
[23] A. S. Parnowski, Earth Planets Space **61**, 621 (2009c).
[24] V. M. Kuntzevich, M. M. Lychak, *Guaranteed estimates, adaptation and robustness in control systems* (Springer-Verlag, Berlin–Heidelberg–New York–London–Paris–Tokyo–Hong Kong–Barcelona–Budapest, 1992).
[25] D. J. Hudson, *Statistics Lectures on Elementary Statistics and Probability* (Geneva, CERN, 1964).
[26] H. R. Madala, A. G. Ivahnenko, *Inductive Learning Algorithms for Complex Systems Modeling* (CRC Press, Boca Raton–Ann Arbor–London–Tokyo, 1994).
[27] W. H. Press *et al.*, *Numerical Recipes in FORTRAN. The Art of Scientific Computing*, 2nd ed. (Cambridge Univ. Press, Cambridge–New York–Melbourne, 1992).
[28] G. A. F. Seber, *Linear Regression Analysis* (Wiley, New York–London–Sydney–Toronto, 1977).
[29] R. A. Fisher, *Statistical methods for research workers* (London, Oliver and Boyd, 1954).
[30] J. H. King, N. E. Papitashvili, J. Geophys. Res. **110**, A02104 (2005); doi:10.1029/2004JA010649
[31] E. W. Cliver *et al.*, J. Geophys. Res. **105**(A2), 2413 (2000).
[32] W. Lyatsky *et al.*, Geophys. Res. Lett. **28**, 2353 (2001).
[33] J. Takalo, K. Mursula, J. Geophys. Res. **106**(A6), 10905 (2001); doi:10.1029/2000JA000231
[34] T. P. O'Brien, R. L. McPherron, J. Geophys. Res. **107**(A11), 1341 (2002); doi:10.1029/2002JA009435
[35] M. Sugiura *et al.*, IAGA Bulletin, 40 (1991); http://wdc.kugi.kyoto-u.ac.jp/dstdir/dst2/onDstindex.html
[36] E. E. Antonova *et al.*, Annales Geophysicae **27**, 4069 (2009a).
[37] E. E. Antonova *et al.*, Adv. Space Res. **43**, 628 (2009b).
[38] A. Y. Ukhorskiy *et al.*, J. Geophys. Res. **111**, A11S03 (2006); doi:10.1029/2006JA011690.
[39] W. H. Campbell, J. Atm. Terr. Phys. **58**, 1171 (1996).

РЕҐРЕСІЙНЕ МОДЕЛЮВАННЯ ГЕОМАГНІТНОЇ АКТИВНОСТІ

А. С. Парновський
Інститут космічних досліджень НАН України та НКА України,
просп. Акад. Глушкова, 40, корп. 4/1, Київ–187, 03680 МСП, Україна

У статті докладно описаний метод реґресійного моделювання. Його застосовано до задачі прогнозування геомагнітних індексів. Він забезпечує не тільки якісне прогнозування, а й нові відомості щодо механізмів взаємодії сонячного вітру з магнітосферою.