Use of principal component analysis for identification of temporal and spatial patterns in the dynamics of ionospheric equatorial anomaly

Yu.S. Maslennikova, V.V. Bochkarev, D.S. Voloskov

Kazan Federal University, Kremlevskaya str.18, Kazan 420018, Russia

E-mail: JSMaslennikova@kpfu.ru, vbochkarev@mail.ru

Abstract. In this paper we describe results of the principal components analysis of the dynamics of Total Electronic Content (TEC) data with the use of global maps presented by the Jet Propulsion Laboratory (NASA, USA) for the period of 2007–2011. We show that the result of the decomposition in principal components essentially depends on the method used for pre-processing the data, their representation (the used coordinate system), and the data centering technique (e.g., daily and seasonal components extracting). The use of momentarily co-moving frame of reference and other special techniques provide opportunity for the detailed analysis of the ionospheric equatorial anomaly. The covariance matrix of decomposition was calculated using Spearman's rank correlation coefficient that allows reducing statistical relationship between principal components.

1. Introduction

Principal component analysis (PCA), also called *empirical orthogonal functions* method (EOF) has been used for wide variety of geophysical data analysis tasks for many years. It should be noted, that for correct analysis of investigated data dynamics one should take into account data representation features and presented trends. For example in paper [1] the authors modelled TEC dynamics using EOF analysis, representing data in local time domain. As a result, the first mode with the highest energy contained equatorial anomaly information. The second mode reflected seasonal variations of ionosphere. In other words, the first high-energy modes, obtained using such approach, contain information only about well-known ionosphere dynamics features. An approach, proposed in paper [2] involves transition to local time domain and coordinate system, taking into account the angle of magnetic axis declination. As a result first mode contains daily TEC variations and the second mode contains seasonal variations. All these examples show, that while using PCA for geophysical data analysis, one should pay particular attention to factoring in the way, data is presented, prevailing periodicities and statistical properties of the data. This paper is dedicated to the development of the methods allowing to take these factors into account. Also for decomposition we use covariance matrix based on Spearman's rank correlation coefficient. It allows reducing statistical relationship between components of decomposition.

2. Total Electronic Content

Total electron content (or TEC) is an important descriptive quantity for the ionosphere of the Earth. TEC is the total number of electrons integrated between two points, along a tube of one squared meter cross section, i.e., the electron columnar number density. For ionosphere dynamics analysis global

TEC maps, provided by JPL (Jet Propulsion Laboratory, NASA, USA), were used for period from 2004 to 2011 years. TEC maps were presented in geographical coordinate system with spatial resolution of 2.5° for latitude and 5° for longitude and time resolution of 2 hours. TEC daily dynamics for 13.11.2006 is presented on the Fig.1. It shows that the maximum of TEC field is located in the subsolar point and migrates, following after motion of the Sun.
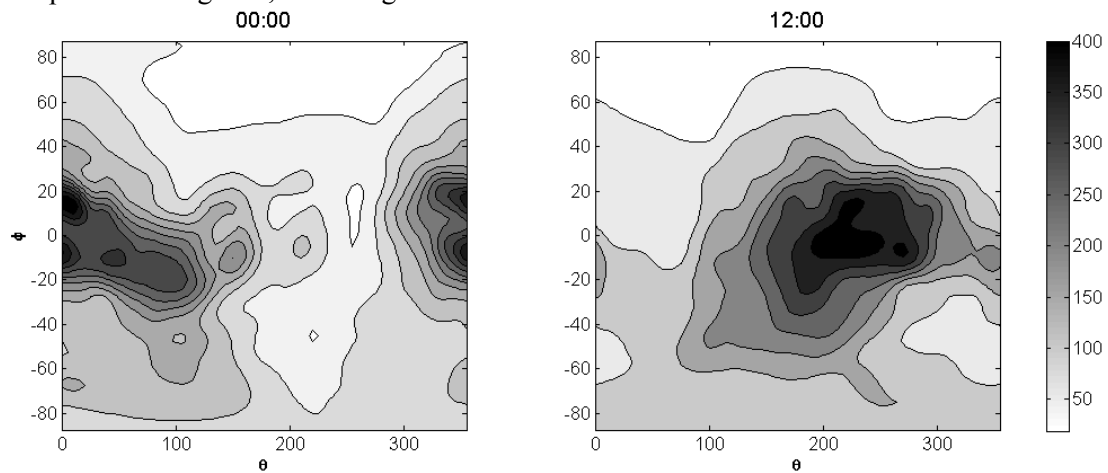
Figure 1. Global TEC maps at 00:00 and 12:00 of 13th November, 2006.

3. Principal component analysis

We can represent a matrix of TEC observations as a matrix $X_{i,j}$, where the index i indicates a point on the Earth's surface and index j corresponds to a time moment. Principal component analysis allows representing the matrix of TEC data in the following form: $X = \sum_k \sigma_k U_k V_k^H$. The orthonormalized vector U_k defines the spatial distribution of TEC variations that corresponds to the kth principal component. The orthonormalized vector V_k defines temporal evolution of the corresponding component. For centralized data, value σ_k^2 corresponds to the variance of the kth principal component.

For the matrix $X_{i,j}$ of TEC data we can estimate the covariance matrix $\mathbf{K_x}$ that defines a character of the time-spatial dependence of TEC global maps. The variance of the 1st component can be defined as follows: $\sigma_1 = \varphi_1^* \mathbf{K_x} \varphi_1$. The search of the best transform φ_1 is reduced to the maximization of the function σ_1 subject to the normalizing condition: $\left(\varphi_1^*, \varphi_1 \right) = \sum_{j=1}^{p} \varphi_{i,j}^2 = 1$. This optimization problem can be solved with the help of the Lagrange function. Remaining components can be found by analogy. The calculation procedure is reduced to the maximization of the variance subject to conditions of normalizing conditions of components. The process should be performed until we find all eigenvalues and the corresponding eigenvectors of the covariance matrix $\mathbf{K_x}$, that represent variances and coefficients of linear combinations of principal components. More detailed information about principal component analysis is presented in [3].

It was shown in the paper [4] that the result of the principal component analysis essentially depends on the data pre-processing technique, used coordinate system and the data centering technique. Following the recommendations from [4], in current work we use momentarily co-moving frame of reference to reduce diurnal periodicity in TEC data and take into account transfer matrix of the spherical coordinate system.

4. Reducing statistical relationship between principal components

In principal component analysis U_k components can be found as eigenvectors of spatial covariance matrix. Using of covariance matrix is equivalent to estimating statistical connection between values in

different spatial points using Pearson's correlation coefficient. However, Pearson's coefficient can be considered as correct estimation of statistical connection only when distribution of fluctuations values is close enough to normal distribution. If fluctuation distribution is different from normal one, components, obtained from principal component analysis could be not statistically independent

In order to assess distribution of $V_k(t)$ amplitude fluctuations, we found discrete wavelet-transform coefficients of these vector (Daubechies-5 wavelet was used). Distribution functions of wavelet coefficients were fitted using t-distribution function. v parameters of approximating t-distribution were following: *2.12±0.13* for the 1st mode, and- *$v=12.62±3.21$* for the 15th mode. Large value of t-distribution's v parameter for high modes shows that their distribution functions are close enough to normal ones whereas for the first modes deviation from the normal distribution is significant.

In such situation we can use Spearman's rank correlation coefficient to estimate statistical relationship. This coefficient is independent from the absolute values of the data and the distribution shape, taking into account the position of the element in ranked set. We used covariance matrix analogue, built the following way: $B_{ij} = \sigma_i \sigma_j \rho^S_{i,j}$, where σ_i –standard deviation of TEC fluctuations in i^{th} point and $\rho^S_{i,j}$ –Spearman's correlation coefficient between values in i^{th} and j^{th} points. Having built the B matrix we can find U_k components like it's eigenvectors.

The most restrictive quantitative characteristic of statistical interconnection is the mutual information measure I [5]. Having large enough datasets we found estimations of 2-dimensional distribution densities of $V_k(t)$ components amplitudes, after that we calculated the mutual information measure I. For more convenient interpretation one can apply the following functional transform to I:

$$R = \sqrt{1 - 2^{-I}} \ .$$

Thus we obtain the value between 0 and 1 with properties which are similar to ones of correlation coefficient. If values have normal distribution, equivalent correlation coefficient R is equal to absolute value of Pearson's coefficient.

Fig. 2 represents mutual distribution functions of 1st and 2nd modes as well as 14th and 15th ones, obtained using the kernel density estimation method. From the Fig.2 one can see that 1st and 2nd components are statistically dependent, what means that decomposition is unsuccessful. Equivalent correlation coefficient between 1st and 2nd modes is $R_{12}=0.714$, between 1st and 3th modes is $R_{13}=0.539$, between 2nd and 3th ones – $R_{23}=0.416$, between 14th and 15th $R_{14,15}=0.121$.

For the same dataset decomposition $X = \sum_k \sigma_k U_k V_k^H$ was performed where covariance matrix was calculated using Spearman's correlation coefficient. In Fig.3 the comparison between results of classic PCA and proposed approach for the first 3 components is presented. Estimations of statistical connection between first modes of decomposition were also performed. The correlation between components decreased by more than 40% ($R'_{12}=0.39$, $R'_{13}=0.29$). Proposed above approach provides opportunity for analyzing of temporal patterns of TEC data. Fig. 4 shows dynamics of the first 2 components. After median filtering we can see 28-30 days periodic fluctuations.
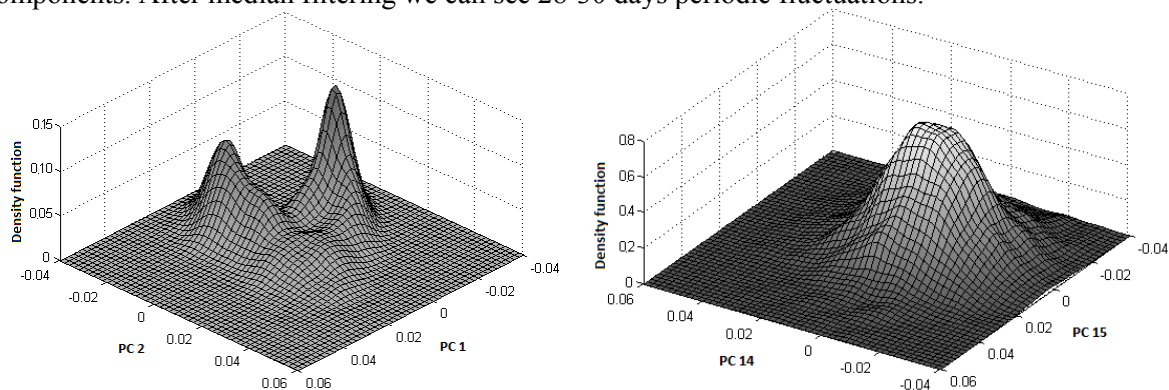
Figure 2. Mutual distribution functions for 1st and 2nd components (left side); for 14th and 15th components (right side).
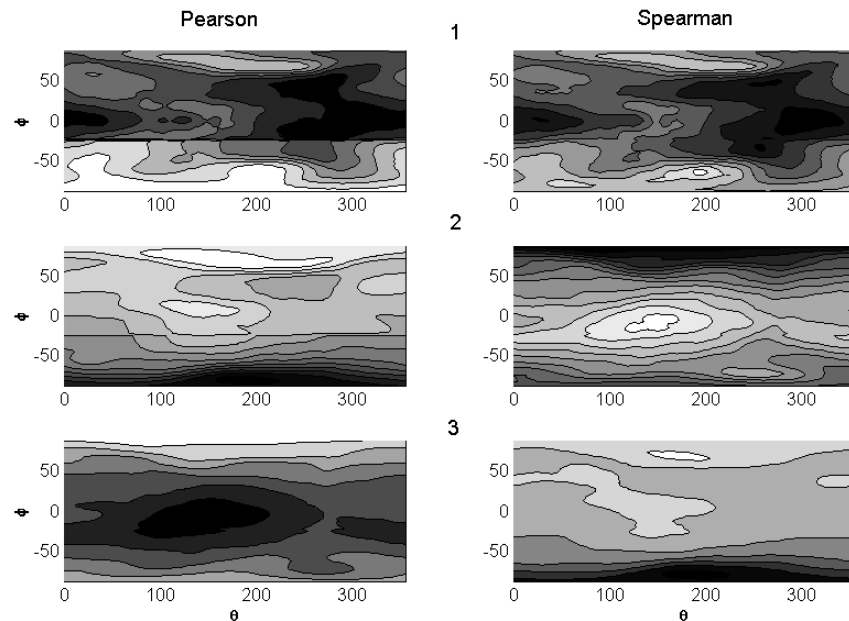
Figure 3. Spatial patterns of TEC data using classic PCA method (left side) and proposed approach using Spearman's correlation coefficient for the first 3 components (right side).
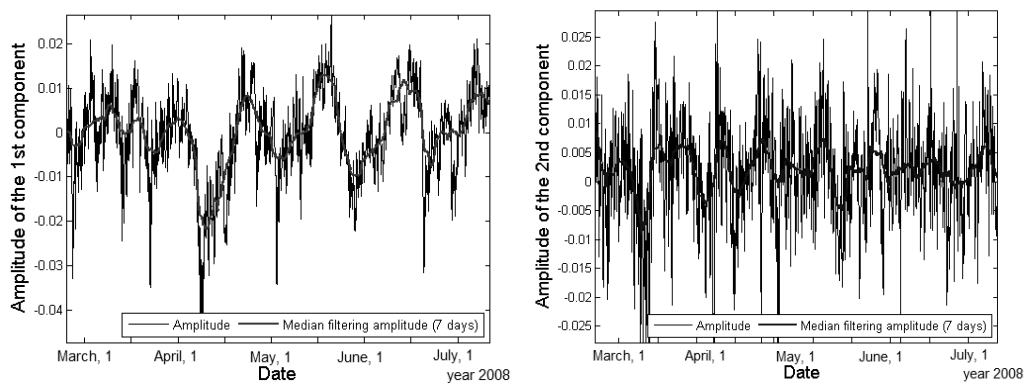
Figure 4. Temporal patterns of TEC data for the first 2 components.

References
[1]   Wan, W., Ding, F., Ren, Z., Zhang, M., Liu, L., and Ning, B., Modeling the global ionospheric total electron content with empirical orthogonal function analysis, Science China Technological Sciences, 2012, vol. 55, pp. 1161–1168.
[2]   Zhang D, Ridley A J, Xiao Z, Hao Y A global model: Empirical orthogonal function analysis of total electron content 1999–2009 data 2012 *J. Geophys. Res.* p. 117
[3]   Jolliffe I T Principal component analysis 2002 *Series: Springer Series in Statistics, 2nd ed., NY*
[4]   Maslennikova Yu S, Bochkarev V V Principal Component Analysis of Global Maps of the Total Electronic Content 2014 *Geomagnetism and Aeronomy* **54(2)** 216–223
[5]   Borda Monica Fundamentals in Information Theory and Coding 2011 *Springer*