Computational Simulation Models

TOOLS FOR ANALYSIS OF PROCESSES MEASURED ON SPARSE AND IRREGULAR SPATIAL-TEMPORAL GRID WITH APPLICATION TO DATA OF NATIONAL CENSUSES[3]

Alexander Temruk, Mikhail Alexandrov

Abstract: *In the paper we shortly describe mathematical and program tools for analysis spatial- temporal processes. Speaking mathematical tools we mean filling data on thick and regular grid in space and time and also use of principal component method for process decomposition. Speaking program tools we mean software for end-user, which implements all mentioned operations and provides dynamic visualization for results of modeling. The proposed technology proves to be useful for processing data of National censuses and we demonstrate it on the real example reflecting dynamic of population density in one of the Russian regions. Such a technology was described almost 15 years ago by P. Makagonov, who applied it for analysis of migration flows in the State of Oaxaca in Mexico. In our work we use this experience, but our system is developed on the new platform and it contains new algorithms.*

ACM Classification Keywords: *I.2 Artificial Intelligence*

Keywords: *principal component analysis, density of population, National census, software*

Introduction

1.1 Problem settings

Subject of this article are processes unfolding in time and space. We use the following limitations:

- Process is described by a single parameter. The case of several interrelated parameters is not considered.
- Position of measurement points in time does not change. So we have a grid, but not a chaotic set of points in space and time
- Space is an area of the plane. Thus, we deal with a function defined on the grid 2Dx1D

These limits are illustrated in Fig. 1. It shows the points on the plane, in which the values of function were measured. The numbers indicate the year. These timestamps refer to the years when the censuses were conducted in the USSR and in Russia.

When the grid of observation is dense and regular, one can implement analysis without filling. When this grid is sparse and irregular, we need a stage of filling data. In this paper we consider this case as the general one. To fill the data we offer our own algorithms taking into account the property of smoothness for the function under consideration.

[3] Work done under partial support of the British Petroleum grant (BP-RPANEPA 2013).

We deal with a sparse and irregular grid of observations. Such cases occur when the measurements are either very expensive or they are impossible for several reasons. It may be, for example, analysis of dynamics of parameters reflecting ecology, demography, and economics of large areas. In this paper we consider the dynamics of the population density in a region of Russia for demonstration of the proposed technology.

The term "analysis" involves the two following operations:

- recovery and visualization of spatial dynamics of a parameter under consideration
- decomposition of spatial-temporal processes and their visualization

One should note that both of these operations are possible only when data are given on a regular grid. Naturally that the presence of dense grid in space and time increases the quality of the analysis.

Visualization is to provide a film which can be displayed to an expert. Here each frame reflects distribution of the parameter at plane at successive moments of time. The main problem here is the presentation, which should emphasize changes of the parameter in time. Convenient form of presentation can stimulate intuition of an expert to explain the reasons of the observed dynamics. In this paper, we use our own algorithms to preprocess data before using standard procedures of visualization. Direct utilization of these procedures is ineffective.

To identify processes of different orders, we use principal component analysis (PCA). Despite of its simplicity principal components are rarely used in the analysis of spatial-temporal data. For this reason PCA is not a part of well-known software packages for processing such a data. So, in this paper we briefly describe the operations that allow to calculate these principal components. Dynamics of the first and second principal components (PCs) is usually sufficient to explain the spatial dynamics of the parameter under consideration. The developed system allows to implement its visualization using the method of preprocessing that we use in the film.

The final result of our applied research is the software system in the form of application oriented to ordinary end-user. The software system is developed in Matlab and includes algorithms, which have been described above. It is: filling spatiotemporal data, calculating PCs, dynamics visualization. To demonstrate the developed software system we use two examples. The artificial example shows the result of decomposition on PCs. The real example shows all steps of data processing: filling in time, filling in space, the decomposition on PCs. In this example, we process the data of censuses related to one of the Russian regions

1.2 Related works

Analysis of spatiotemporal data is used in climatology [Torne, 2007], meteorology [Kunitsyn, 2008], physics [Galanin, 2007] and many other applications. Typically, such an analysis only display data without revealing the factors of hidden dynamics.

There are several known software packages for the analysis of spatial data. The most advanced of them are is represented in a the group of software products ArcGIS. [ArcGIS, http: https://www.arcgis.com]. All of them have developed tools for maps presentation in different scales, interactive design of maps, matching maps. However, these packages can't satisfy our needs for the following reasons: algorithms of filling spatial data are rough enough (linear interpolation and triangulation), there is no filling of data in time; there is no analysis of latent factors in dynamics. We do not mention here the time-series analysis. This analysis is available in many packages, but it is not accompanied by spatial analysis.

Our research use results of the work [Makagonov, 2003]. In this work the authors describe the system they developped to study migration flows in Oaxaca, one of the Mexican states. The data they used were data of the National censuses. The difference between Makagonov´s system and our system is: we use our own algorithms for filling data and visualization of results. Besides we use standard MatLab platform for system development and our codes are the open ones.

The paper is structured by the following way. In the section 2 we present algorithms. In the section 3 we describe developed software and results of experiments. Section 4 includes conclusions.

Algorithms

Filling in time, local splines

The function is assumed to be smooth and given on irregular grid in time If we consider the problem of interpolation under these conditions then we can use the cubic spline interpolation [Bahvalov, 1999]. Spline interpolation procedure is available in the package MatLab and it is named *spline()*. To find parameters of cubic interpolation the following conditions are used: (a) equality of polynomials function value $S_k(t_{k-1})= y_{k-1}$; $S_k(t_k)=y_k$; (b) continuity of the first derivative of splines $S'_{k-1}(t_k)=S'_k(t_k)$; (c) continuity of the second derivative of splines $S''_{k-1}(t_k)=S''_k(t_k)$. These conditions lead to a linear system of three diagonal type, which is solved by sweep method. We call this spline as a global spline.

In our system we use the so-called local cubic splines. It does not require equality of the second derivatives at the points of measurement. To calculate these splines the following conditions are used: (a) for all segments $S_k(t_{k-1})=y_{k-1}$, $S_k(t_k)=y_k$; (b) for the first segment $S''_1(t_0)=0$, $S''_1(t_1)=0$; (c) for the second and subsequent segments $S'_k(t_{k-1})= S'_{k-1}(t_{k-1})$, $S''_k(t_k)=0$. In the last condition $S'_{k-1}(t_{k-1})$ is known from the previous calculation of the spline on the previous segment.

The described algorithm is a modification of the algorithm described in [Kostyuk, 1977].

As a result of local spline interpolation we get regular and dense grid in time. It is illustrated on Fig. 2.

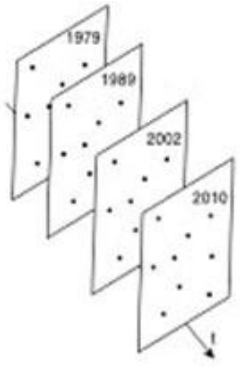


Fig. 1. Original data for analysis

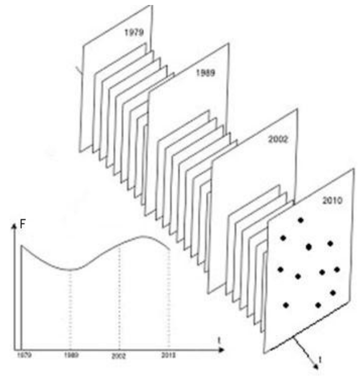


Fig. 2. Filling data in time

Filling in space, method of shades

At each temporary layer we have a grid of irregular points. To fill it to a regular and dense grid the well-known triangulation method is usually offered [Kalitkin, 1978]. This method is simple, but it does not take into account the smoothness of the function. In this paper we use the method of shadows. This method was proposed for filling two-dimensional distribution of geological parameters on a surface. The algorithm of this method is described in [Hakimov, 1986].

1. Function value in a point is determined by linear combination of function values at the points with known values. This contribution decreases with the distance between points and it depends on a characteristic radius *R*.
2. Only *n*-nearest neighbors affect function value in a given point.
3. Each point near a given point creates a shadow, which reduces the influence of points located behind. The shadow is defined by angle *θ*

Value function in some *i*-th point on the plane is given by the formula:

$$\varphi_i = F_0 + \Sigma_j w_j (\Delta F_j) /(1+\alpha r_j), \quad \Sigma_j w_j = 1, \quad j=1,...n$$

Here: φ_i is function value in the *i*-th point, F_0 is the mean value of functions in the known points, ΔF_j is deviation from the mean value in the *j*-th point (where the value of the function is known), w_j is the weight of this point, r_j is the distance from the *j*-th point to a given *i*-th point, *n* is the number of points with the known values of function near *i*-th point, $\alpha=1/R$ is the proportionality factor. Weight w_j is calculated by means of simple formulas that take into account the distances between points and shadow effect. Fig. 3 illustrates this effect: the influence of the points behind the shadow is less. One should note that instead of the function $1/(1+\alpha r_j)$ other functions can be used. For example, $1/(1+(\alpha r_j)^2)$ or $\exp(-\alpha r_j)$. When filling is done we have the regular and dense grid instead of the irregular and sparse one, see Figure 4.

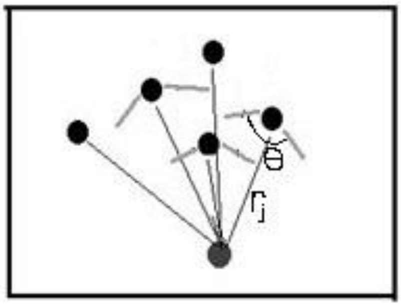


Fig. 3. Simple illustration of the method of shadows

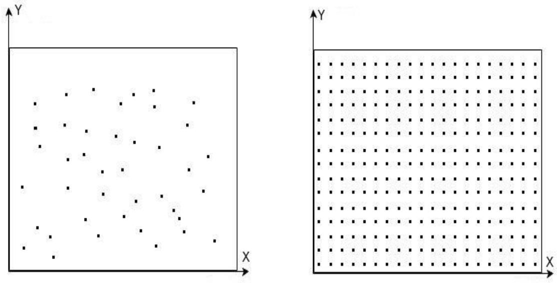


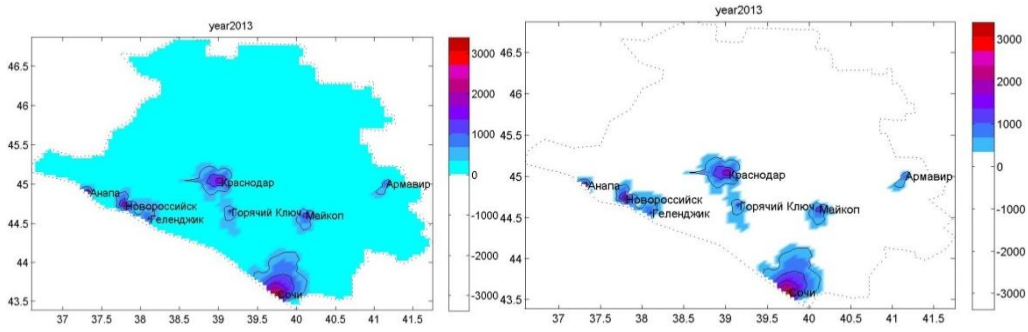
Fig. 4. The original and the resulting grids

Here: *R, n, θ* are parameters of the method.These parameters should be tuned to specific data to be processed. We did such experiments with the data of Primorskiy region of Russia and found the best values of parameters. Namely, *R* is equal to a quarter of the maximum distance between known points, *n* is taken equal to 10% of all points, and *θ* is equal to π/4:

Algorithm of visualization, procedure of 'whitening'

The traditional way for representing 2D functions are isolines. Matlab includes procedures *contour()*, *meshgrid()* to present functions in the form of isolines. However, the direct application of these procedures proves to be non-effective: user doesn't see changes in function dynamics.

We propose preprocessing procedure before visualization, which we call the 'whitening'. Here whitening is the use of white color for those parts of the plane where the function value is less than a certain threshold. To determine the optimal threshold we implemented a series of experiments. The best threshold is the level, which corresponds to the band with the largest square in this area. Figure 5 shows a map of Krasnodarskiy region in Russia with the most major cities. Color reflects the population density in different parts of the region. The numbers on the color scale on the right is the density of the population in relation to the mean density. On Figure: 5a the parts of the map where the density is less than its mean value in the region are whitened. On Figure 5b the parts of the map associated with the band of the largest square are whitened.



(a) (b)

Fig. 5. Map of Krasnodarskiy region, effect of whitening

Calculation of spatial-temporal principal components

The PCA is one of the traditional tools of multivariate data analysis. For the case of objects that are not related to space and time, principal components (PCs) are calculated according to the known algorithm:

- Calculation of correlation matrix of parameters
- Determination of eigenvalues and eigenvectors of the matrix, the eigenvectors define the directions of PCs
- Objects are projected on PC directions

The distribution of objects along the axis of the first PC is considered as an impact of the first factor, the distribution of objects along the axis of the second PC is considered as an impact of the second factor, etc.

Calculating PC for the case of spatial-temporal data is not well known. So, we explain the process of calculating on the simple example. Suppose we have m=20 points on the plane $P_1,P_2,\dots P_{20}$. Assume that at each point we have time series, which contain n=50 values: T_1, T_2, ...T_{50}. All of these values can be represented by the matrix:*U.*

Table 1. Sourse matrix

$U(m,n)$ =

Points	T_1	T_2	...	T_{50}
P_1	1.2	1.35	…	1.10
P_2	0.95	1.14	…	1.56
…	…	…	…	…
P_{20}	3.16	2.68	…	3.05

Then matrix $R(n,n)= U^T(m,n) \times U(m,n)$ is formed (values are not real).

The last matrix is matrix of correlation of time series. Its eigenvalues and eigenvectors define the spatial-temporal PCs.. Each eigenvector contains 50 elements. of n=50. We denote these eigenvectors as: $\gamma_k = (\gamma_{k,1}, \gamma_{k,2}, \dots\dots \gamma_{k,50})$. Here *k* is the number of *k*-th orthonotmal vector. Then we calculate the values of PCs for all points $P_1,P_2,\dots P_{20}$ and all moments of time T_1, T_2, ...T_{50} using the formulas: $C_1= U\,\gamma_1$, $C_2= U\,\gamma_2$, etc. Each C_k is the matrix $C_k\,(m,n)$.

Table 2. Correlation matrix

$R\ (n,n) =$

Points	T_1	T_2	…	T_{50}
T_1	1	0.95	…	0.28
T_2	0.95	1	…	0.36
…	…	…	…	…
T_{50}	0.28	0.36	…	1

Software and experiments

Architecture and functions of the system

Software system was developed on the software platform MatLab. The system is presented in the form of exe-application. It is oriented on end user, so it contains a user-friendly interface. One of the Interface windows is shown on Figure.6.

The system has a modular structure. It contains:

- 3 computing modules: interpolation in time, interpolation in space, calculation of PCs
- 3 input-output modules: reading and writing files, visualization, user interface

Data transfer between modules is implemented through external files. Such an autonomy allows easily to implement modifications of the system.

Typical steps of data processing are the follows:

1. Import tables with the original data;
2. Spline interpolation on a given temporal grid;
3. Spatial interpolation in each temporal layer;
4. Calculation of the first and the second PCs;
5. Visualization of the spatial dynamics for the function under consideration;
6. Visualization of the spatial dynamics for the first and the second PCs.

In those cases when the original data are given on regular and dense grid the steps (2) and (3) are absent.

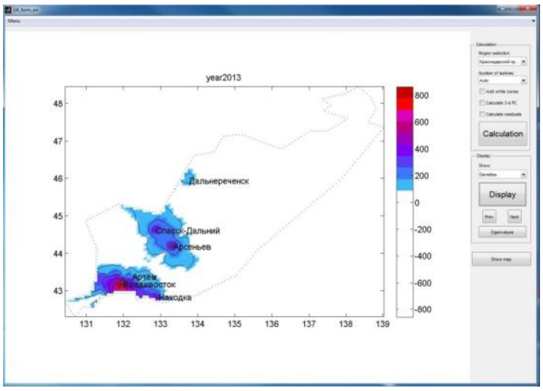


Fig. 6. Interface of the system

Testing the system on artificial example

The problem

In the artificial example the values are generated by the known function on a given spatial-temporal grid. These values are considered as the experimental data. The grid is dense and regular. So, the problem of filling function is absent here. The function itself contains components that can be explicitly linked to the factors of the first and second order. The purpose of the experiment is to identify these factors on the basis of artificial experimental data and their visualization.

Original data

We consider the mathematical function containing 3 “caps”:

$$S_0(x,y,t) = \frac{1+2*t}{1+\alpha R(x,y)} \qquad S_1(x,y,t) = S_2(x,y,t) = \frac{2^t}{1+\alpha R(x,y)}.$$

Here we denote:

$$R(x,y) = \sqrt{(x-x_0)^2 + \left(y-y_0\right)^2}$$

In these formulae: x_o, y_o are the centers of the caps; x, y are coordinates of the grid; α is a parameter specifying variability of the cap. These functions are defined on the plane in the region [0;1]x[0;1] and on time interval [0;5]. Coordinates of centers for the following caps are: $x_o= y_o=0{,}5$ for S_0, $x_o=y_o=0{,}1$ for S_1, $x_o=y_o=0{,}9$ for S_2. Parameter $\alpha =5$ is for the central caps, and $\alpha =10$ is for the caps near corners. Calculation of the function is performed with the steps $\Delta x=\Delta y=0{,}02$ and Δt = 1 in space and in time respectively. Thus we have a spatial grid 50 x 50 for 6- temporal layers.

It is easy to see that there are two developing processes in the example. One is a slow process. It is associated with the cap S_0. Amplitude of the cap grows arithmetically. Another process is the fast one. It is associated with the caps S_1 and S_2. Amplitude of these caps grows exponentially.

Results

Calculations were carried out for the first and second PCs. Values of the first PC in the first and last time points are shown on Fig. 7. The first component shows a growth of the function in the same proportion on the entire region. Values of the second PC in the first and the last time points are shown on Fig. 8. The second PC shows an additional growth of the function caused by the influence of the caps located at corners.

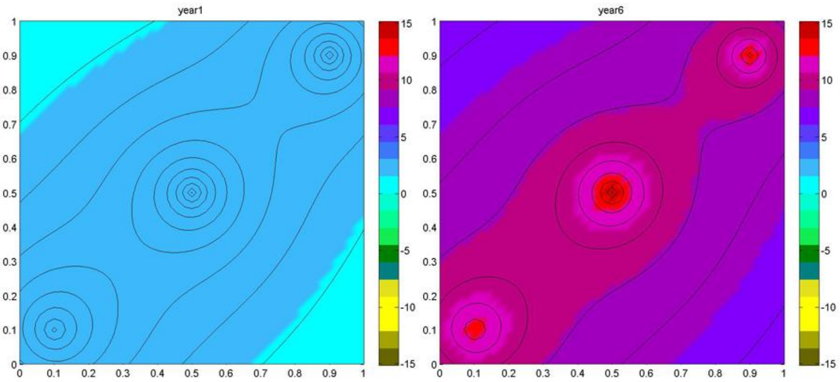


Fig. 7. The first PC in the first and the last moments of time

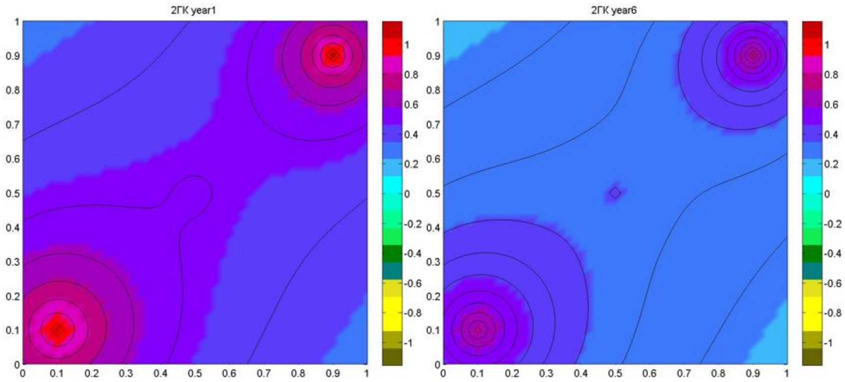


Fig. 8. The second PC in the first and the last moments of time

Therefore, one can see that the software surely distinguishes both processes.

Analysis of census data in Krasnodarskiy region

<u>Geography of the region</u>

Krasnodarskiy region is located in the southern Russia near Ukraine, Abkhazia, Georgia. Sochi, the capital of Olympic Games 2014, is a city in Krasnodarskiy region. The big cities in this region are Krasnodar (capital of the region), Sochi (resort town) and Novorossiysk (port). Smaller cities are Anapa, Gelendzhik and Armavir. The region has very good natural environment, and therefore agriculture is well developed here. As a consequence the percentage of rural population is high in the region. .

The last censuses were held in the USSR in 1979 and 1989 and in Russia in 2002 and 2010. Using this data the prediction for 2013 year was calculated. Thus, we have 5 time layers irregularly distributed along the time axis: 1979, 1989, 2002, 2010 and 2013.

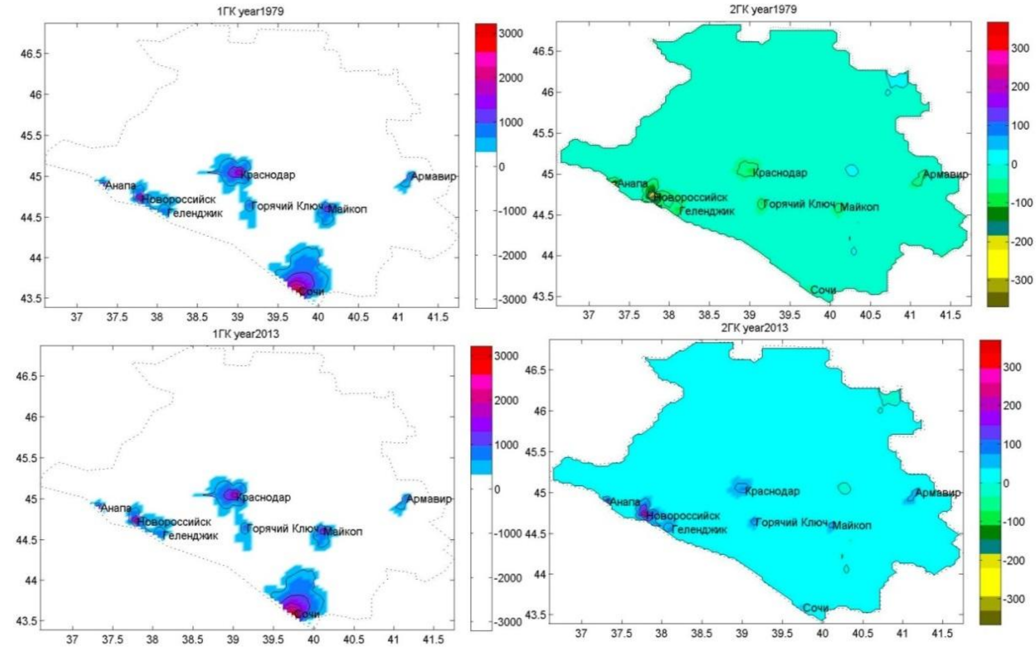
Krasnodarskiy region consists of 38 districts. The data concerning square and population in each district are known. It makes possible to calculate the density of population in these districts.

On the initial stage of processing the data of population density is calculated for each district and for each year 1979,…2013. Thus we have the data on sparse and irregular grid. This step should be considered as preprocessing. It was performed in semi-automatic mode using Excel.

<u>Data processing</u>

Further calculations are performed according to the scheme described in the section 3.1. Namely, the data are imported into the system and then the spatial-temporal interpolation and the calculation of the first two PCs are implemented. The results of calculation are saved in two files. The first file contains a “film”. It is a sequence of maps with population density in the region in the deferent moments of time. The second file contains values of the first and the second PCs. Fig. 9 shows the contents of this file.

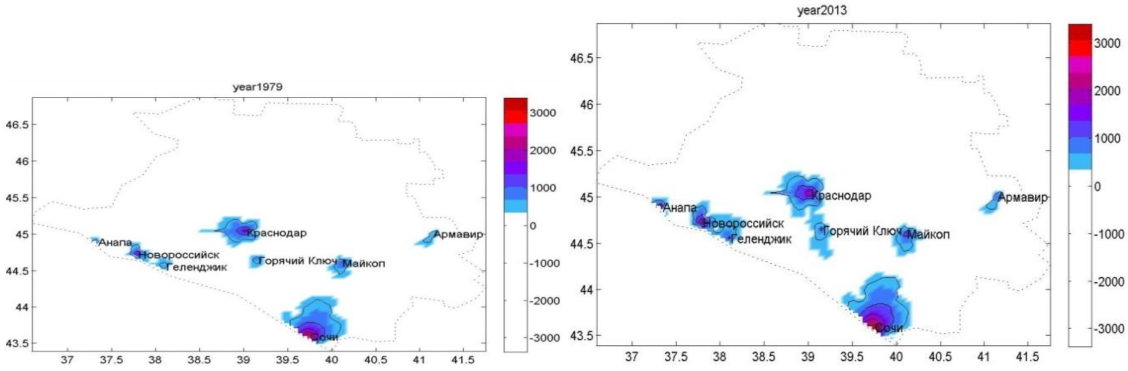
Consider the map on the top left. In the middle of the map there is Krasnodar, the capital of the region. The most southern point is Sochi. Three cities are situated on the west: Anapa, Novorossiysk and Gelendzhik. Krasnodar, Sochi and Novorossiysk are our main points. It is easy to find these cities on all other maps. On the right part of maps there is a color scale. Each color is associated with a certain population density.



(a) (b)

Fig. 9. Principal components, Krasnodarskiy region:
(a) first PC in 1979 and in 2013 years; (b) second PC in 1979 and in 2013 years;

The first PC demonstrates the process of a small growth of population density in the whole region. This growth is 18%. First of all the growth is observed in the cities of Krasnodar, Sochi, Novorossiysk mentioned above. The second PC reflects the second-order process. This is the process of redistribution of population during 44 years and an additional growth of population in Novorossiysk city. The latter can be explained by the accelerated construction of the port with its military infrastructure.



(a) *(b)*

Fig. 10. Density of population in Krasnodarskiy region: (a) in 1979 and (b) in 2013 years

Post-analysis

Effect of PCA can be seen on Fig.10. There is shown the dynamics of population density in Krasnodarskiy region in 1979 and 2013 years. White areas correspond to the threshold of whitening described in the section 2.3. In our case this value is slightly higher than the mean density in the region. The maps say almost nothing about the districts having impact on population redistribution in the region. PCA allows to reveal these hidden processes.

Conclusion

In this paper we propose the technology and software for analysis of spatial dynamics of the data given on sparse and irregular grid. In this work:

- method of shadows for filling data in space is described and realized
- algorithm of local spline interpolation for filling data in time is described and realized
- module for calculation of spatial-temporal PCs is developed; it contains open codes and can be used in other programs
- functionality of developed system is demonstrated on artificial and real data

In future we plan to implement the following modifications in the system:

- to include possibility to manage parameters of methods related to filling data using interface (now these parameters are fixed by default)
- to include module especially oriented on processing data of National censuses

Acknowledgements

The authors are grateful to Dr. Makagonov for his interest to our work and valuable advice.

Bibliography

[ArcGIS, URL] ArcGIS resourse: http://www.arcgis.com/features/

[Bahvalov , 1975] Bahvalov N. *Numerical methods* - Moscow: Nauka, 1975 (rus.)

[Hakimov , 1986] Hakimov B., Garris V. *New approach in interpolation of geological fields, Mathematical methods of investigation in gerology*. N_11, 1986 - pp.6-13 (rus.)

[Galanin, 2007] Galanin M., Guzev M., Nizkaya T. *Numerical solution of thermal plasticity problem with additional parameters* - Preprint, Inst. Appl. Math., the Russian Academy of Science, Moscow 2007 (rus.)

[Kalitkin, 1978] Kalitkin N. *Numerical methods*. Nauka, Moscow, 1978 (rus.)

[Kostyuk, 1977] Kostyuk V. *Overview of graphical output*. / / Automation experiment and computer graphics. - Tomsk: TSU, 1977. - P. 90-102 (rus.)

[Kunitsyn] Kunitsyn V. *Satellite radio probing and tomography of the atmosphere*, (rus.) URL:http://atm563.phys.msu.su/rus/text_direct.htm

[Lomtadze, 1984] Lomtadze V. *Interpolation taking into account field anisotropy and evaluation of result's accuracy, Mathematical methods of investigation in geology*. N_5, 1984. - pp.11-18 (rus.)

[Makagonov , 2003] Makagonov P. Sboychakov, K: *Interaction y diferencia en la utilization de los metodos de analisis de sistemas y metodos estadisticos en las diferentes etapas de la mineria de datos en problemas sociales* - Pachuca, Hidalgo Mexico, 2003, pp.12-15

[Torne, 2007] Torne P., *Tropical tropospheric trends*.

URL:http://www.realclimate.org/index.php/archives/2007/12/tropical-troposphere-trends/

Authors' Information

Temruk Aleksandr – M.Sc. student, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; Moscow Institute of Physics and Technology (State University); Institutskii per 9., Dolgoprudny, Moscow Region, 141700, Russia) e-mail: atem@mail.ru

Major Fields of Scientific Research: mathematical modeling, data mining

Mikhail Alexandrov – Professor, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: MAlexandrov@mail.ru

Major Fields of Scientific Research: data mining, text mining, mathematical modelling