An interplay of annual temperature variation and NDVI figures in clustering of small watersheds

M G Erunova¹ and M G Sadovsky²,³

¹Federal Research Center Krasnoyarsk Science Center of the SB RAS, Akademgorodok, 50, Krasnoyarsk, 660036, Russia,
²Institute of Computational Modelling SB RAS, Akademgorodok, 50/44, Krasnoyarsk, 660036, Russia,
³Siberian Federal University; Svobodny pr., 79, Krasnoyarsk, 660041 Russia,

E-mail: oleg@icm.krasn.ru

Abstract. Previously, highly informative clustering of small watersheds provided over the annual temperature course has been reported. Here we report on the interplay between the pattern and the distribution of NDVI figures. Very high concordance between the cluster structure and the indices values is found.

1. Introduction

Previously, a reasonable progress in substantially small water basins clustering has been reported. The remarkable fact was that the clustering was developed over the annual temperature trend determined with 7 days gap, through Earth remote sensing (ERS) data. Very high concordance of the cluster composition, and other features has been found; geographic location and environmental features of small water basins are among the issues. It looks like the annual temperature variation to be the key factor to separate the basins into the groups of rather homogeneous entities, when the separation is provided over the temperature course, only, regardless any other properties of the basins.

It is a common fact that relief is expected to be the key factor in classification of land sites into the groups with relatively homogeneous geographical and environmental properties. Surprisingly, the geography here exhibits a strong correlation to an annual temperature course; this fact indeed seems quite intriguing since it allows conversion of multidimensional complicated data analysis into the clustering problem of the basins determined into significantly simpler metric space. Practically, nonetheless, there are some more indices used in land use management and related activities to analyze and make decisions, and NDVI is the most popular among them. Indeed, this index is “responsible” for vegetation description over a site, and in such capacity may hardly be reduced to simple geographic variables observed for a site.

Here we present some preliminary results of the modeling of land classification based on basin approach in modeling, for Krasnoyarsk region. We used MERIT Hydrologically Adjusted Elevations digital relief model [1-3]. A hierarchy classification of water basins for the territory under investigation has been developed; that latter is a tree-like structure, where the nodes are basins of rivers, and edges correspond to the tributary structure. Formally, such classification is expected to be complete. Practically, it is based on determination of morphometric features of relief in river ecosystem based on MERIT Hydrologically Adjusted Elevations DRM accompanied with the corresponding spatial database.
implementation, and advanced statistical analysis of the data. We identified four levels in the tree pattern so that the annual course of the surface temperature has been determined, over ERS data, for each watershed of the lower (the fourth one) level. The temperature records have 7 days gaps between the measurements. Each basin was supplied with the mean day temperature averaged over the basin. Additionally, each watershed has been supplied with NDVI figures. The key idea of this paper is to compare two patterns provided by clustering: the former is provided by annual temperature course, and the latter is provided by annual NDVI course observed over the same watersheds.

This paper aims to investigate the relations between the structuredness revealed through the clustering of small watersheds provided by the data on annual temperature course, and NDVI index observed over the same land sites. Speaking in advance, good concordance between the pattern revealed through temperature data, and the indices of vegetation was found.

2. Material and methods
Following original data were used:

- MERIT Hydrologically Adjusted Elevations DRM with spatial resolution ~90 m at the equator derived from the latest elevation data (MERIT DEM) [1–3].
- Automatically generated vector layer of the geosystem of river basins using the software indicated above.
- Hydrography network transferred from topography maps of the scales 1:10 00 000, 1:2 500 000 and 1:100 000 in vector formats.

Four-level hierarchical classification of watersheds was developed. The lower level watersheds were then clustered with elastic map technique, in two versions of the variables: (1) annual temperature course and (2) annual NDVI course. Hence, each watershed makes a point in either 49-dimensional metric space (for annual temperature course records), or in 52-dimensional metric space (for annual NDVI course records).

To reveal the cluster structure, we used elastic map technique [4-6]. This method decreases data dimension, clusters and visualizes them. The method consists in approximation of the multidimensional data with a manifold of low dimension; we used two-dimensional manifolds (a square and a sphere). It takes four steps.

The first and the second principal components for raw data are determined, at the first step. They yield the two-dimensional space generated by two eigenvectors of the covariance matrix of the original data corresponding to the greatest eigenvalue. Geometrically, these directions correspond to the greatest dispersion of data in the original space. As soon as the components are determined, one must develop a plane over them and project each data point on it; the minimal square comprising all the projections must be determined. This is the manifold to be used for data approximation.

Connect each data point to its projection on the square with mathematical spring, at the second step. Springs have an infinite elasticity following linear expansion rule, regardless a deformation. Then change the rigid square for elastic membrane, at the third step. The membrane is stipulated to be homogeneous and uniform: elasticity remains the same in any direction, as well, as an expansibility. Then the system is released to reach the minimal total deformation energy.

At the fourth step, the image of each data point is redefined on the jammed surface: one must find the orthogonal projection on this jammed surface. That latter is the point on the jammed surface with the closest location to the membrane. Upon this redefinition, the springs are cut-off and the membrane is released to reach back a plane form. Obviously, the newly redefined images change the location on this map, in comparison to the original ones thus showing the inner structuredness through the clustering.

There are various ways to define cluster pattern, on the membrane. We used the local density approach to do it. To do it each point is supplied with bell-shaped function; we used Gaussian one:
\[
f_j(r) = \exp\left\{ -\frac{(r - r_j)^2}{\sigma^2} \right\}. \tag{1}\]

Figure 1. Cluster pattern of the distribution of 573 watersheds (triangles) in the 45-dimensional metric space of annual course of surface temperature is shown in grey scale. Coloring represents NDVI values distribution over the pattern; scheme details see in text.

Here \( r_j \) is the location (in inner coordinates) of \( j \)-th point, and \( \sigma \) is the smoothness parameter; it resembles to some extent the standard deviation for normal distribution, while it is not. In fact, \( \sigma \) determines a contrast of the contouring on elastic map. Next, one must sum up the functions (1) over the entire set of points, so that the function

\[
F(r) = \sum_{j=1}^{N} f_j(r) \tag{2}
\]
of the local density is obtained. It is the key tool to identify clusters in elastic map.

3. Results
Previously [7], the cluster pattern observed through the annual temperature course data recorded for considerably small watersheds, for two middle-range revers has been reported. Figure 1 shows this pattern, in grey scale of local density (2) of the points (these are the watersheds). Seven clusters are apparent in this figure. Besides, this figure shows the distribution of average (over a year) NDVI values, for both greater rivers (the river of Kan and the river of Mana). Coloring scheme is the following: NDVI annual average values range from 0.16 to 0.60. Thus, the range was divided into six equal intervals, and the watersheds with NDVI value falling into a specific interval were colored as follows:

- the watersheds with NDVI ranged from 0.16 to 0.23 are colored in light blue;
- the watersheds with NDVI ranged from 0.23 to 0.31 are colored in light green;
- the watersheds with NDVI ranged from 0.31 to 0.38 are colored in light lime green;
- the watersheds with NDVI ranged from 0.38 to 0.45 are colored in light yellow;
- the watersheds with NDVI ranged from 0.45 to 0.53 are colored in light orange, and finally
- the watersheds with NDVI ranged from 0.53 to 0.60 are colored in light red.

Clear, unambiguous and distinct prevalence of the occurrence of the watersheds with specific NDVI values is evident, from figure 1.

The distribution of watersheds with various NDVI values is shown in figure 2. This figure shows two geographic maps with the watersheds taken into consideration. Figure 2 A shows the coarse grain pattern of NDVI figures distribution, and figure 2 B shows the fine grain pattern of the distribution. The ranges of NDVI variation re shown in insets. Numbers 1 and 2 indicate the basins of the river Kan, and the river Mana, respectively.

![Figure 2. NDVI distribution over the geographic maps, A is coarse grain, and B is fine grain patterns.](image)

We studied cluster structuredness of 573 basins in the 45-dimensional metric space of annual temperature course measured with 7 days gaps. Four days have been excluded from the analysis since there is a lot of lacunae in the data records, for these days. These are 25, 201, 297 and 345 days, that corresponds to January 25, 2019, July 20, 2019, October 24, 2019 and December 11, 2019 dates. We
used soft map of 16 × 16 size with the by default parameters values, for freely distributed VidaExpert software [7].

4. Discussion

Here we studied the patterns of two distributions of small watersheds provided in the 45-dimensional metric space of annual temperature measurements provided by ERS data. More exactly, the distribution of the watersheds has been developed over the truncated annual temperature record (some dates contain too many gaps in the records, see details in [7]); later the data on an averaged NDVI values were mapped on the distribution. The key result of this work is that there is a high concordance between the clusters observed due to annual temperature course, and over a year averaged NDVI values.

The observed concordance allows to estimate some lower stable indices with the temperature data records. Such estimation makes sense in several cases, for example when the data gaps should be healed, in a database. We expected the average annual NDVI figures to be rather good characteristics of a site (that is a watershed, in our case); surprisingly, we have found that it is not. Indeed, the maximal annual NDVI values observed over the watersheds exhibit better concordance to the clustering pattern developed over the annual temperature course. Consider figure 2 B, the right upper corner. There is a distinct site with low average NDVI figure (colored in dark blue); meanwhile, this is the agricultural site with the best soil and ecosystem productivity. Remarkable fact is that this site is pretty close in terms of NDVI figures to that one located at the bottom right part of the map. That latter comprises rather mountain area with extreme climate and decreased ecosystem productivity. There is no miracle here, nonetheless: agricultural sites are covered with plantation significantly less time, in comparison to natural forests etc. This very simple and obvious fact coming from the nature of an agriculture itself makes the average NDVI figure less informative in comparison to the maximal one.

Let now focus on figure 1 again. Previously we have found a sounding specificity in the composition of the clusters provided over the annual temperature course, of the watersheds comprising the basins of two greater rivers: Kan and Mana. In such capacity, further studies are expected to reveal the specificity of NDVI distribution shown in figure 1. In other words, a combination of the pattern provided by annual temperature course, for these two basins, and NDVI distribution is of great interest. Similar studies with other indices (such as LAI, etc.) are of great interest, as well.

References

[1] Erunova M and Yakubailik O 2020 Zoning of the territory on the basis of morphometric analysis of basin geosystems IOP Conf. Ser.: Earth Environ. Sci. 421 062039
[2] Yamazaki D, Ikeshima D, Tawatari R, Yamaguchi T, O’Loughlin F, Neal J, Sampson C, Kanae S and Bates P 2017 A high accuracy map of global terrain elevations Geophysical Research Letters 44 5844
[3] Yamazaki D, Ikeshima D, Sosa J, Bates P, Allen G and Pavelsky T 2019 MERIT Hydro: A high resolution global hydrography map based on latest topography datasets Water Resources Research 55 5053
[4] Gorban A and Zinovyev A 2015 Fast and user-friendly non-linear principal manifold learning by method of elastic maps IEEE International Conference on Data Science and Advanced (Campus des Cordeliers Paris) pp 1-9
[5] Akinduko A and Gorban A 2014 Multiscale principal component analysis Journal of Physics: Conference Series 490 012081
[6] Mirkes E, Zinovyev A and Gorban A 2013 Lecture Notes in Computer Science 7902 500
[7] Gorban A and Zinovyev A 2015 Fast and user-friendly non-linear principal manifold learning by method of elastic maps IEEE International Conference on Data Science and Advanced Analytics (DSAA) IEEE pp 1-9