Annual temperature variation reliably identifies different sites in a large water basin

M G Erunova¹, O E Yakubailik²,4 and M G Sadovsky²,3

¹Federal Research Center Krasnoyarsk Science Center of the SB RAS
Akademgorodok, 50, 660036, Krasnoyarsk, Russia
²Institute of Computational Modelling SB RAS, Akademgorodok, 50/44, 660036, Krasnoyarsk, Russia
³Siberian Federal University, Svobodny pr., 79, Krasnoyarsk, 660041 Russia
⁴E-mail: oleg@icm.krasn.ru

Abstract. Spatial database of morphometric features based on basin approach is developed for Krasnoyarsk region. The water basins are ranged into four levels. The digital model of hydrography network developed by VSEGEI was used, in combination to Whitebox GAT, QGIS and ArcGIS software. The average climate features are calculated due to this digital model, and spatial analysis supported with advanced non-linear statistics is provided.

1. Introduction
Spatial inhomogeneity is essential feature of geography and environment: various sites differ in their characteristics; climate, soil, relief, geological substrate, plantation and fauna are among them. Also, the physical location of a site is of great value; these peculiarities are usually described in terms of zone network. Agriculture and other types of nature consumption are another two important issues affecting the sites. Each of the issues corresponds to the location, and one can identify the least entity to be taken into analysis. Forest stand is that type of entity, for plantation; specific zones with homogeneous climate conditions are the entities determined in terms of climate zone mapping. In similar way one defines the entities in terms of relief. An identification of the least entity meeting all the constraints simultaneously is a hot problem here.

Whether a site falls into a basin of some rive, is the leading factor determining zone attribution of it and play the key role in consumption planning and usage of land. It is very important that one can range those sites into ranked sets, one embedded into another, thus providing a powerful information tool for the analysis and land management, including the site resource determination.

Proper land use is based on morphometry. Relief is the key factor determining the directions and characteristics of matter fluxes, thus affecting the landscape formation; that latter, in turn, might be changed by human activity. The basin based digital modeling stipulates determination of the following items: slope, exposition, curvature of land surface, visibility area determination, etc. [1]. GIS applications are the powerful tool to do it through implementation of digital elevation models (DRM).

DRM are mainly derived from Earth remote sensing data and may differ due to significant variations in the methods of those remote data collection and treatment. GTOPO30 (Global 30 Arc-Second Elevation), SRTM (The Shuttle Radar Topography Mission); SRTM Void Filled; GMTED2010 (Global Multi-resolution Terrain Elevation Data 2010); ACE2 (Altimeter Corrected Elevations 2); ASTER
(Advanced Spaceborne Thermal Emission and Reflection Radiometer) GDEM (Global Digital Elevation Model) are the most widespread digital models. However, all of them are not free from some errors in accuracy [2]. These errors have been eliminated in the latest MERIT Hydrologically Adjusted Elevations model [3].

Here we present some up-to-date results of the modeling of land use based on basin approach in modeling, for Krasnoyarsk region. We used MERIT Hydrologically Adjusted Elevations digital relief model [3]. In particular, we have developed the hierarchy pattern of water basins, for the territory under investigation; determination of morphometric features of relief in river goesystem based on MERIT Hydrologically Adjusted Elevations DRM accompanied with the corresponding spatial database implementation, and advanced statistical analysis of the data.

For each lower level watershed the annual course of the surface temperature has been determined, over ERS data. The data has 8 days gap between the measurements. Each basin was supplied with the mean day temperature averaged over the basin. We checked the distribution of the basins provided by this annual data record, and compared it to other land use features. Speaking in advance, very good correlation has been found between some geographical, environmental and some other issues, and the clustering provided over the annual temperature records.

2. Material and methods

We used the following original data:

- MERIT Hydrologically Adjusted Elevations DRM with spatial resolution ~90 m at the equator derived from the latest elevation data (MERIT DEM) [3];
- Vector layer of the geosystem of river basins developed automatically, due to 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.

We have introduced the four-level system of detail of river network, resulted from the different scales. Fig. 1 shows an example of such kind river networking developed for two rivers located close to the city of Krasnoyars: these are the river of Kan and the river of Mana.

The first level is provided by the most large rivers in the region; here they are the river of Kan and the river of Mana. The second level comprises the greatest tributaries of the rivers of the level 1. Hence, Kan (label 1) has as a tributary the river of Rybnaya labeled as 1.18. The third level separates the second one onto smaller basins; e.g. the river of Balay is the tributary of Rybnaya and is labeled as 1.18.1. The fourth level comprises smaller basins; e.g., the river of Tomna (label 1.18.1.1) of river of Karakchul (label 1.18.1.2).

2.1. Methods of clustering analysis

It should be stressed that the data on annual temperature variation are the time series, indeed. It means that one may not change the order of the records arbitrary. Meanwhile, any time series could be considered as a vector of relevant metric space. The ultimate goal was to identify clusters (if any) in this space, and check whether those clusters comprise some specific basins, or not.

To cluster the data (that are the basins in 49-dimensional space) we used $K$-means [4]. It should be stressed that we have excluded four dates from the analysis, since these dates contain a lot of gaps in the records. Thus, the final data set was the vector set in 45-dimensional space. We tried $K$-means with various $K$, $2 \leq K \leq 8$.

Next, we used elastic map technique [5-7]. This is the method to decrease the data dimension, cluster and visualize. This is basically new method, so we should explain it in few details. The method consists in approximation of the multidimensional data with a manifold of low dimension; we used two-dimensional manifolds. The method consists of four steps.
At the first step the first and the second principal components are determined. These are the one-dimensional spaces generated by two eigenvectors of the covariance matrix of the original data corresponding to the greatest eigenvalue, and the next one. Geometrically, these directions provide the greatest dispersion of data in the original space. As soon as the first and the second principal components are determined, one must develop a plane over them, project each data point on it and choose the minimal square comprising all the projections.

![Cluster pattern of the distribution of 531 basins of the fourth level in the 45-dimensional metric space of surface temperatures.](image)

**Figure 1.** Cluster pattern of the distribution of 531 basins of the fourth level in the 45-dimensional metric space of surface temperatures.

At the second step each data point is connected with its projection with mathematical spring; it means that a spring has infinite elasticity, and that latter remains linear, in dependence on the deformation. At the third step initially rigid square is substituted with elastic membrane. The membrane is stipulated to be homogeneous and uniform: there is no prevalence in the elasticity in a direction, not in a expansibility. Then the system is released to reach the minimal total deformation energy.
Finally, the image of each data point is redefined, on the jammed surface: one must find out the orthogonal projection on this jammed surface. Upon this redefinition, the springs are cut-off and the membrane stretches back to a plane state. Obviously, the location of the newly defined images change their location on this map, thus showing the inner structuredness through the clustering.

3. Results
To begin with, we have not found any stable (or even partly-stable) clustering through $K$-means application, for any $K\leq 8$. Thus, we changed for elastic map technique. This approach yields very distinct and clear cluster structure. Fig. 1 shows this pattern.

We addressed two (dual) questions:
- Whether the basins from the specific area (the level 1 area corresponding to the river Mana) fall into a single cluster, on the elastic map, or not, and
- Whether the basins comprising the core parts of each of five clusters determined in elastic map occupy geographically close areas, in geographic map?

The answers on both questions are positive.

![Figure 2. Clustering mapped into the geographic maps, see explanations in text.](image-url)
We studied cluster structuredness of 531 basins in the 45-dimensional metric space of annual temperature course measured with 8 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 detailed map of $25 \times 25$ size. Background color represents the local density of the points (these are the basins), in the inner coordinates.

To define local density, each point is supplied with bell-shaped function; we used Gaussian one:

$$f_j(r) = \exp\left(-\frac{(r - r_j)^2}{\sigma^2}\right).$$

(1)

Here $r_j$ is the location (inner coordinates) of $j$-th basin, 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 basins, so that the function

$$F(r) = \sum_{j=1}^{531} f_j(r)$$

is obtained. The function (2) is drawn out in Fig.1 as a background. The local density is the key tool to identify clusters in elastic map.

Let now focus on Fig.2. Fig.2 A shows the distribution of the basins marked in the data base with the numbers exceeding some specific one. In fact, these basins belong (at most) to the level 1 basin of the river of Mana. All these basins mainly occupy mainly the cluster 1 shown in Fig.1 located at the upper part of the elastic map. Few basins are located in other clusters, and there is no preference in the attribution to other clusters than 1. Thus, Figs.1 and 2 A answer positively on the question towards the preferable distribution of the basins over the elastic map: yes, the basins belonging to some higher ranked territory preferably occupy some specific cluster in the elastic map, but not disperse themselves randomly over that latter.

The dual question was towards the location of the basins (in terms of geography) belonging to the clusters identified my elastic map. Fig.2 B shows the location of core basins belonging to four clusters shown in Fig1.: these are the clusters 1 to 4. Core basins are those located inside the high local density area of a cluster. In such capacity, he abundance of the core basins is rather small: there are 115 basins belonging to all four cores of these clusters (that makes one fifth of the entire set of basins).

An expansion of a cluster results in greater number of basins, obviously. Fig.2 C shows the location of the basins when taken sufficiently widely. It should be stressed also, that Figs. 2 C and 2 D show all five clusters observed on the elastic map, including that one located in upper right corner (cluster 5). Fig.2 D shows the supposition of the basins from all five clusters, and those observed in the cores of the clusters. There is no direct of evident and simple relation between the core subset of a cluster and its geographical positioning.

Cluster 5 is very specific; it comprises quite few basins. The members of this cluster are shown in deep red in Fig.2. The cluster is highly isolated, in elastic map. Remarkable fact is that this cluster comprises the small (in size) basins located at the watersheds, over grater territories. Thus, Fig.2 ambiguously proves the positive answer on the second question: the clusters gathered into a single cluster on elastic map exhibit a concordant spatial behaviour when traced on geographical map.

4. Discussion

Here we studied the distribution of small (of the fourth level) basins in the 45-dimensional metric space of annual temperature measurements provided by ERS data. The data present an averaged surface temperature of a basin. The key result of this work is that non-linear clustering tool (elastic map
technique) reliably reveals the clusters gathering the small water basins into the groups with similar geographic, climate, land use, etc. properties.

It should be stressed that the clustering is provided by a single variable: temperature. To be exact, the data are not one-dimensional, but multi-dimensional: we tried the clustering over the truncated subsets of temperature data. We selected randomly the subsets comprising 3/4, a half, 1/3 and 1/4 of the original set of temperature records, and tried the elastic map implementation. Monotonous decay of quality of clustering has been observed, as the capacity of the data set goes down with complete loss of that latter, as the number of temperature records used for clustering reached 10 items or less.

A lot of efforts was targeted (and still is) to implementation of some integrative, scalar like index representing, in general, but effectively, the land use potential of a territory. On one hand, this activity seems to be not very faithful: indeed, any scalar may basically represent a vector entity. On the other hand, the inner structure of data manifested in clustering etc. may support this activity. In other words, the efficiency of clustering provided over the temperature data makes a good perspective in implementation of an index derived from a set of various characteristics.

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] Fukunaga K 1990 Introduction to statistical pattern recognition, Academic Press, London. 657 p.
[5] Gorban A, 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
[6] Akinduko A and Gorban A 2014 Multiscale principal component analysis Journal of Physics: Conference Series 490 012081
[7] Mirkes E, Zinovyev A and Gorban A 2013 Lecture Notes in Computer Science 7902 500