International Journal of Mechanical Engineering and Technology (IJMET)
Volume 10, Issue 02, February 2019, pp. 1085-1109. Article ID: IJMET_10_02_115
Available online at http://www.iaeme.com/ijmet/issues.asp?JType=IJMET&VType=10&IType=2
ISSN Print: 0976-6340 and ISSN Online: 0976-6359

CLUSTER ANALYSIS FOR STRUCTURING SPATIAL DATA

I.I. Pivovarova
Department of Informatics and Computer Technology, Saint Petersburg Mining University, 2, 21 Line, Vasilievsky Island, St. Petersburg, 199026, Russia

ABSTRACT

The article discusses cluster analysis algorithms using various sets of source spatial data and software capabilities for processing them in accordance with the task. The work result is the selection of the most accurate cluster procedure and the construction of a map of the research area with selected homogeneous hydrological areas.

The possibility of using cluster analysis methods to detect structures inherent in complex data sets has been substantiated. The estimation of spatial interpolation of runoff characteristics for the regionalization of hydrological data has been performed.

The quality and character of hydrometeorological data is of great importance in hydrology; therefore, all characteristics must comply not only with quantitative criteria, but also be subject to accurate spatial analysis.

Keywords: cluster analysis, regionalization, spatial data, modeling, hydrological forecast.

Cite this Article: I.I. Pivovarova, Cluster Analysis for Structuring Spatial Data, *International Journal of Mechanical Engineering and Technology*, 10(2), 2019, pp. 1102-1109.
http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=10&IType=2

1. INTRODUCTION

The quality and character of hydrometeorological data are of great importance in hydrology, therefore all characteristics must comply not only with quantitative criteria, but also with objective analysis.

Regionalization is a method used to compensate for the lack of data in a catchment [1]. In this process, model parameters for an ungauged catchment are estimated with the use of information from hydrologically similar catchments [2]. The similarity between the gauged and ungauged catchment can be established based on proximity (i.e. selecting a close by catchment assumed to exhibit ahydrologic response similar to the ungauged catchment).

Estimating design flow of ungauged basins is very crucial in the planning and management of hydraulic and water resources engineering. Regionalization for identifying

homogeneous hydrologic regions is a well-accepted technique in this area. Regionalization is defined as determination of hydrologically similar units, and is one of the most challenging tasks in surface hydrology [3].

One of the problems that is usually expressed in an attempt to understand the limiting factors and underlying mechanics of regionalization methods (and why they sometimes either work well or fail miserably) is the quality of climate and hydrometric data [4-5]. Prediction of runoff in unexplored basins, synthesis between processes, places and scales, as well as difficulties in obtaining consistent and representative reservoir descriptors [6]. Are physically similar catchment areas hydrologically similar?

Uncertainty in measurements is inherently reflected in the effectiveness of regionalization methods. Meteorological observations are rare, biased, and often riddled with missing data. The same is true for measured hydrological time series. In addition, meteorological data often needs to be homogenized at a catchment scale and concentrated to model applications, which adds another level of uncertainty. Descriptors, such as land use, may change over time, while soil and bedrock properties are usually unknown or approximate in sparsely populated areas. The overall uncertainty of the entire system makes it difficult to parse its individual elements [7].

2. MATERIALS AND METHODS

There are a number of criteria for verifying spatial data for homogeneity. These criteria allow us to determine whether two samples (data on two different objects) are related to one general population or not [8]. If the samples belong to the same population, then the difference between the samples within the limits of random variations of the quantities and there are no fundamental differences between the objects. In this case, parametric criteria require that the distribution of the sample is subject to a specific distribution law. Thus, the classical criteria of Student and Fisher require that the law of distribution of samples be sufficiently close to the normal law [9]. Parametric criteria allow us to directly estimate the level of the main parameters of the general populations, the difference in the means and the difference in variances. The criteria can identify trends in data changes, evaluate the interaction of two or more factors. Recently, the Cramer and Welch criteria [10-11] have also been used to estimate the homogeneity of data. An additional advantage of these criteria is the optional equality of the variances of the compared samples. Parametric criteria are considered to be more powerful than nonparametric ones, provided that the characteristics are measured in an interval scale and are normally distributed.

Nonparametric criteria do not have the above limitations. The term "nonparametric method" means that it is not necessary to assume that the distribution functions of the results of observations belong to any particular parametric group while it used. Non-parametric criteria do not impose conditions for the recognition of the distribution law. However, criteria of this type do not allow a direct assessment of the level of such important parameters as the average or variance. Using nonparametric criteria impossible to estimate the interaction of two or more conditions or factors affecting the change in characteristics. Many nonparametric methods have been developed - Smirnov's criteria [12], such as the omega-square (Leman-Rosenblatt) [13-14], Wilcoxon (Mann-Whitney) [15-16], Van der Waerden [17], Savage etc.

A separate group of evaluation criteria is cluster analysis. Cluster analysis is one of the static processing methods collecting data that contains information about the selection of objects and then order the objects into relatively homogeneous groups.

In fact, cluster analysis is not so much a common statistical method, as a “set” of various algorithms for “distributing objects among clusters” [18-19]. There is an opinion that, unlike many other statistical procedures, cluster analysis methods are used in most cases when there

are no a priori hypotheses about classes, but you are still in the descriptive stage of the study. Cluster analysis identifies the “most likely significant solution” [20].

Cluster analysis accompanies a wide variety of methods for detecting structures inherent in complex data sets. The basis of the data is most often a sample of objects, each of which is described by a set of separate variables. The task is to combine all the elements into such clusters so that the elements within one cluster would have a high degree of “natural proximity” between them, and the clusters themselves would be “quite different” from the other .

Essentially, cluster analysis suggests that internal dataset is little known. This analysis presumes that structure is remaining unknown. All that is at our disposal is a collection of data. The purpose of the analysis in this case is to detect some "categorical" structure that would be consistent with the observations and would allow to identify homogeneous areas in the research area [21].

In general, the main stages of cluster analysis are the following: selection of comparable objects:

- selection of a set of features that will be used for comparison and description of objects;
- calculating the measure of similarity between objects (or the measure of the difference of objects) in accordance with the selected metric;
- grouping objects into clusters using one or another merge procedure;
- validation of the applicability of the obtained cluster solution.

In the work, the cluster analysis of variables was carried out by two methods: hierarchical and non-hierarchical, namely, the method of constructing a dendrogram and the method of k-means.

3. HIERARCHICAL CLUSTERING

Hierarchical clustering is a set of data ordering algorithms that are visualized using graphs.

Algorithms for ordering data of the specified type assume that a certain set of objects is characterized by a certain degree of connectivity. The presence of nested groups (clusters of different order) is assumed. Algorithms, in turn, are divided into agglomerative (unifying) and divisive (dividing). By the number of signs, monothetical and polythetic classification methods are sometimes distinguished. Like most visual ways of representing dependencies, graphs quickly lose their visibility as the number of objects increases. There are a number of specialized programs for graph construction. [22].

A dendrogram is usually understood as a tree - a graph without cycles, constructed from a matrix of proximity measures. Dendrogram allows to depict the mutual connections between objects from a given set. Creating a dendrogram requires a matrix of similarity (or differences), which determines the level of similarity between pairs of objects. Most commonly used are agglomerative methods.

Next, it should be chosen a method for constructing a dendrogram, which determines the method for recalculating the matrix of similarity (difference) after combining (or dividing) the next two objects into a cluster.

In the works on cluster analysis, a rather impressive series of methods for constructing dendrograms are described:

- Single linkage method. Also known as the “nearest neighbor method”
- Complete linkage method. Also known as the “far neighbor method”

- Pair-group method using arithmetic averages
- Unweighted
- Weighted
- Pair-group method using the centroid average
- Unweighted
- Weighted (median)
- Ward's method

By their approach, all hierarchical clustering algorithms are divided into two types: top-down and bottom-up. The first begin with one large cluster of all the elements, and then divide it into smaller and smaller clusters. The latter, on the contrary, begin with individual elements gradually merging them into larger and larger clusters. There is no any fundamental difference between these two approaches, much more significantly the way in which the algorithm uses to calculate distances between the clusters [23].

In our research, data were processed for 37 hydrological posts evenly distributed throughout the Oka River basin. A cluster analysis was performed on a set of 6 signs (latitude, longitude, flow coefficient, flow modulus, coefficient of variation, frequency of occurrence of dangerous hydrological phenomena).

A preliminary assessment of the cluster procedure was carried out using a dendrogram, which displayed the distances of the measure of similarity between individual values at observation points and groups of the same characteristics. The square of the Euclidean distance is taken as the main measure of similarity. The data were previously normalized. In clustering, the analysis method, the type of formula for the distance, and the number of clusters in the reference algorithm were determined. Next methods were tested: Average Linkage (Between Groups), Average Linkage (Within Groups), Single Linkage, Complete Linkage, Centroid Linkage, Median Linkage, Ward Linkage. Three methods (Single Linkage, Complete Linkage, Ward Linkage) (Fig. 1). The result is the division of the territory into 3 districts (Fig. 2). What, in fact, corresponds to the nature of the formation of the hydrological regime in the study area and emphasizes local features of the characteristics of the flow.

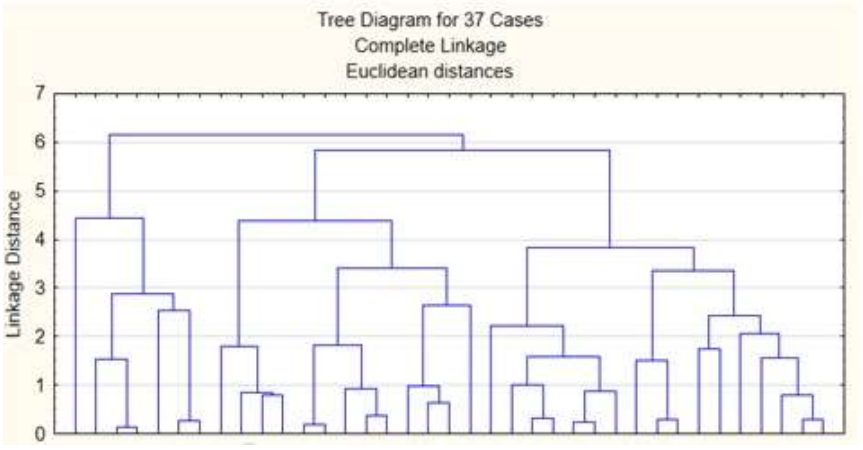

Figure 1 Hierarchical clustering

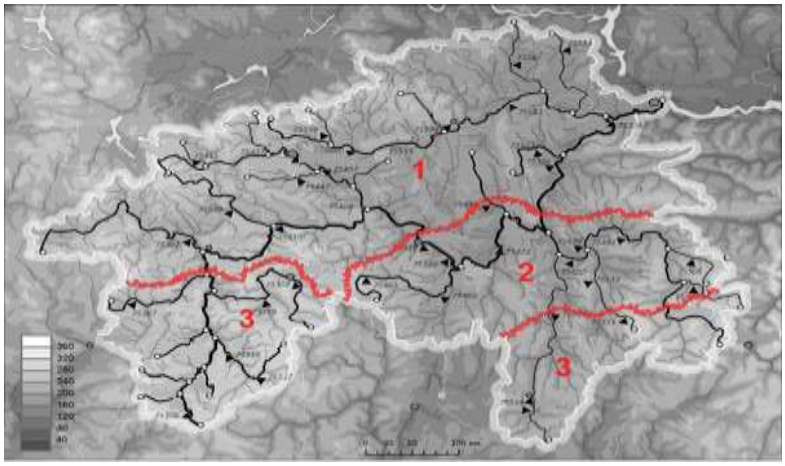

Figure 2 Homogeneous hydrological zones

4. ALGORITHM OF A QUADRATIC ERROR - K- MEANS

K- means or dispersion- analysis turnover is a very well-known method for determining the membership of cluster elements by minimizing the difference between cluster elements and maximizing the distances between them [24].

However, the user statistical software SPSS Statistics, which carried out data processing, he himself must at random determine how many clusters he will end up with. In our case, from the previous clustering experience, the hierarchical method was clear that there should be three clusters.

In general, the calculation algorithm is an iterative procedure in which the following steps are performed:

1. A certain number of k clusters should be selected.
2. K-entries should be randomly selected from initial data set to serve as the initial cluster centers.
3. For each record of the initial sample, the nearest to it center of the cluster should be determined. Those records that are "drawn" by a certain center form initial clusters..
4. Centroids that are clusters of gravity should be calculated. Each centroid is a vector whose elements are the average values of attributes calculated over all records of the cluster. Then Centre cluster shifts into its centroid.

In our case, the algorithm found a set of stable clusters in several tens of iterations. As a result, three districts were highlighted. (Fig. 3, 4)

Figure 3 K- means analysis

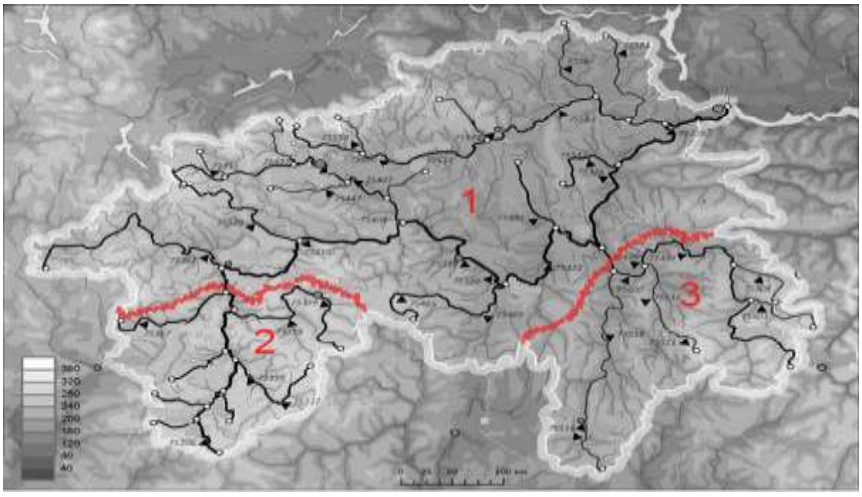

Figure 4 Homogeneous hydrological zones

5. RESULTS AND DISCUSSION

Thus, the result of the work is the selection of the most accurate cluster procedure for the processing of hydrological data and the construction of a map with selected homogeneous hydrological areas.

Cluster analysis of variables was carried out by two methods: hierarchical and non-hierarchical, namely, the dendrogram construction method and the k-means method. According to the results, it can be concluded that the hierarchical methods of combining, although accurate, are laborious: at each step, it is necessary to build a distance matrix for all current clusters. The estimated time increases in proportion to the third degree of the number of observations, which, if there is a large amount of data, can lead to errors in calculations even in such powerful software environments as SPSS Statistics. The advantage of the k-means algorithm is the speed and ease of implementation. It should also be noticed that there are disadvantages, namely: selection of the initial uncertainty cluster centers, and that the number of clusters have to be set initially. However, in our case, all the listed disadvantages are easily compensated by the presence of a priori information obtained at the previous stage of work.

6. CONCLUSION

The assessment of the degree of hydrological homogeneity of river catchments is necessary for developing common approaches to modeling and forecasting hydrological processes. Since it is not a secret that each new research team, when solving problems of hydrological modeling on a specific river basin, in order to improve the accuracy of forecasts, is trying to develop and implement its own model or set of models. As a result, we have many models for the same water object, often with ill-defined or limited access to data for calibration[25-26].

Flow forecasting for poorly studied river basins has been at the forefront of Hydrological Sciences for decades [27]. During this time, the community observed minor improvements, but even today there is an important gap in our ability to predict runoff in the absence of a reliable amount of hydrological data. Therefore, the most promising and widely used methods remain approaches to the regionalization of the parameters of the hydrological model in the conditions of homogeneous hydrological basins.

REFERENCES

[1] Blöschl, G. & Sivapalan, M. 1995. Scale issues in hydrological modelling—a review. Hydrol. Processes 9, pp. 251–290.

[2] Sivakumar, B. and Berndtsson, R. 2010. Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting. World Scientific Publishing Company, Singapore, pp. 53-105.

[3] Jonas MIttet , 2017, Regionalisation technique for urban ungauged catchments. Norwegian University of Science and Technology, p.33.

[4] Blöschl, G., et al., 2013. Runoff prediction in ungauged basins. Synthesis across processes, places and scales. Cambridge: Cambridge University Press.

[5] Sellami, H., et al., 2014. Uncertainty analysis in model parameters regionalization: a case study involving the SWAT model in Mediterranean catchments (Southern France). Hydrology and Earth System Sciences, 18, pp. 2393–2413.

[6] Oudin, L., et al., 2010. Are seemingly physically similar catchments truly hydrologically similar? Water Resources Research, 46, W11558.

[7] Richard Arsenault & François Brissette (2016) Analysis of continuous streamflow regionalization methods within a virtual setting, Hydrological Sciences Journal, 61(15), pp. 2680-2693.

[8] Parks D., Tyson G., Hugenholtz P, Beiko R., 2014. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics, vol.30, p. 3123.

[9] Fisher R. A. 1928.On a distribution yielding the error functions of several well known statistics, "Proc. Intern. Math. Congr. Toronto", vol. 2, p. 805.

[10] Cramer H. 1946. Mathematical methods of statistic, University of Stockholm, p. 648.

[11] Welch B.L. 1938. The significance of the difference between two means when the population variances are unequal. Biometrika. vol.29, pp. 350–36212.

[12] Bolshev LN, Smirnov N.V. 1983. Tables of mathematical statistics / 3rd ed. - Moscow: Nauka publ. p. 474.

[13] Lehmann E.L. 1951. Consistency and unbiasedness of certain nonparametric tests / Annals of Mathematical Statistics, vol.22, no 1, pp. 165-179.

[14] Rosenblatt M. 1952. Limit theorems associated with variants of the von Mises statistic. Annals of Mathematical Statistics, vol.23, no 4, pp. 617-623.

[15] Mann H. B., Whitney D. R. 1983. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, vol. 18, pp. 50-60, 1947.

[16] Hollender M, Wolf D. Methods of nonparametric statistics. Moscow, Finance and Statistics Publ., p. 518.

[17] Bartel van der Waerden. 1957. Mathematische statistik. Springer. Berlin. Gottingen. Heidelberg, p. 436.

[18] Akhmetshin, E. M., Kolpak, E. P., Sulimova, E. A., Kireev, V. S., Samarina, E. A., & Solodilova, N. Z. 2017. Clustering as a criterion for the success of modern industrial enterprises. International Journal of Applied Business and Economic Research, 15(23), pp. 221-231.

[19] Akhmetshin, E. M., Barmuta, K. A., Yakovenko, Z. M., Zadorozhnaya, L. I., Mironov, D. S., & Klochko, E. N. 2017. Advantages of cluster approach in managing the economy of the Russian federation. International Journal of Applied Business and Economic Research, 15(23), pp. 355-364.

[20] Bochkareva, T. N., Drozdov, V. A., Akhmetshin, E. M., Prikhodko, A. N., Gorbenko, A. V., & Zakieva, R. R. 2018. Improving information and technical support of HR management system in the educational establishment. Paper presented at the Proceedings

of the 31st International Business Information Management Association Conference, IBIMA 2018: Innovation Management and Education Excellence through Vision 2020, pp. 3582-3589.

[21] Voronkova, O. Y., Zadimidcenko, A. M., Goloshchapova, L. V., Polyakova, A. G., Kamolov, S. G., & Akhmetshin, E. M. 2018. Economic and mathematical modeling of regional industrial processes. European Research Studies Journal, 21(4), 268-279.

[22] Pashkevich M.A., Petrova T.A. 2017. Assessment of Widespread air Pollution in the Megacity Using Geographic Information Systems. Zapiski Gornogo instituta. vol. 228, pp. 738-742.

[23] Durant B., Odell P. 1977. Cluster analysis. Per. from English E.3.Demidenko ed. and with foreword. A.Ya. Boyarsky. - M .: Statistics, p.128.

[24] Pivovarova, I. 2014. Evaluation of spatial uniformity of hydrological characteristics. Journal of Engineering and Applied Sciences, 9(7), pp. 268-272.

[25] Popov Y. M., Avdeev, S. M., Hamitova S.M., Tesalovsky A.A., Kostin, A.E., Lukashevich V. M., Lukashevich M. V., Kozlov A.V., Koposova N. N., Uromova I. P. Monitoring of green spaces' condition using GIS-technologies. 2018. International Journal of Pharmaceutical Research, 10(4). pp. 730-733

[26] Kozlov A.V., Uromova I.P., Koposova N.N., Novik I.R., Vershinina I.V., Avdeev Y.M., Hamitova S.M., Naliukhin A.N., Kostin A.E., Mokretsov Y.V. 2018. Optimization of the Productivity of Agricultural Crops at Application of Natural Minerals as Ameliorants and Mineral Fertilizers on Sod-Podzolic Soils. Journal of Pharmaceutical Sciences and Research. 10(3), pp. 667-680

[27] Hamitova S. M., Avdeev Y. M., Babich N. A., Pestovskiy A. S., Snetilova V. S., Kozlov, A. V., Uromova I. P , Koposova N. N., Pimanova N. A., Novik I. R. 2018. Toxicity assessment of urban soil of Vologda oblast. International Journal of Pharmaceutical Research. 10(4). pp. 651-654.

[28] K.Padmapriya and Dr. S.Sridhar, Authenticated Indexing for the Query Dependent K-Nearest Neighbours In Spatial Database, *International Journal of Computer Engineering and Technology (IJCET)*, Volume 4, Issue 6, November - December (2013), pp. 70-77.

[29] Mr. Manugula, S. S. Dr. Veeranna, B. and Dr. Patel, S. GeoSpatial Data Foundation For Dam Sites. International Journal of Civil Engineering and Technology, 6(7), 2015, pp. 61-68.