Regional Spatial Analysis Combining Fuzzy Clustering and Non-parametric Correlation

Bülent Tütmez and Uzay Kaymak

Abstract. In this study, regional analysis based on a limited number of data, which is an important real problem in some disciplines such as geosciences and environmental science, was considered for evaluating spatial data. A combination of fuzzy clustering and non-parametrical statistical analysis is made. In this direction, the partitioning performance of a fuzzy clustering on different types of spatial systems was examined. In this way, a regional projection approach has been constructed. The results show that the combination produces reliable results and also presents possibilities for future works.

Keywords: Fuzzy clustering, rank correlation, spatial data.

1 Introduction

In spatial analysis, each observation is associated with a location and there is at least an implied connection between the location and the observation. Geostatistical (probabilistic) and soft computing methods can be applied for assessing spatial distributions in a site [1]. When observations are made in space, the data can exhibit complex correlation structures. The correlation can be two-dimensional if the data are taken only over a spatial surface [11].

Bülent Tütmez
School of Engineering, İnönü University,
44280 Malatya, Turkey
e-mail: bulent.tutmez@inonu.edu.tr

Uzay Kaymak
School of Industrial Engineering, Eindhoven University of Technology,
5600 MB, Eindhoven, The Netherlands
e-mail: u.kaymak@ieee.org

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 219–227.
springerlink.com

It is obvious that the spatial patterns of individual sampling locations in any study area have different patterns and observations depend on the relative positions of observed locations within the site. The classical geostatistical tools such as variogram, although suitable for irregularly-spaced data, have practical difficulties. One of the main drawbacks is that it is insufficient to analyze the regional heterogeneous behavior of a spatial parameter [5]. In general, spatial systems have heterogeneous properties rather than homogeneous structures. Heterogeneity means that the properties observed at different locations do not have the same value, and that different zones are observed in the site.

One of the practical problems encountered in spatial systems such as in geosciences, ecology and geography is the limited number of data. Often, correlations are estimated from a small number of observations. The correlation coefficient is particularly important in cases with sparse data such as pollution and offshore petroleum data [10]. In these cases, because the measure is expensive and time consuming, it may be necessary to work with limited number of data. Hence, a regional analysis with limited data becomes an important task in spatial systems.

The main objective of a cluster analysis is to partition a given data set of data or objects into clusters [9]. Because most of the clustering algorithms employ the distances between the observations, for a spatial system, the clusters provided by clustering can be considered as distinguished regions [12]. Analyzing a spatial system based on structural properties is a difficult task and applicability of clustering for this purpose should be examined. In this study, the performance of the Fuzzy c-means Algorithm (FCM), which is the well-known clustering algorithm, in conditioned spatial systems is investigated. The partitioning capacity of the algorithm with limited number of data is appraised using Rank Correlation Method (RCM) that is also a well-known non-parametric method.

The rest of the paper is structured as follows. Sect. 2 describes the basics of weighted fuzzy arithmetic and the hybrid fuzzy least-squares regression. Confidence interval-based approach for coefficients and predictions is presented in Sect. 3. Finally, Sect. 4 gives the conclusions.

2 Methodology

Fuzzy clustering and non-parametric correlation analysis are well-known methods. The algorithm proposed in this study aims a combination to appraise a spatial system based on an areal analysis. In this section, a brief review and the basis of the combination is presented.

2.1 Fuzzy Clustering

The main purpose of clustering is to recognize natural groupings of data from a large data set to produce a concise representation of a system's behavior. The FCM is a well-known data clustering method in which a data set is grouped into clusters (regions) with every data point in the data set belonging to every cluster to a certain degree. As a suitable algorithm, the FCM was also proposed to make spatial evaluations [2].

Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of N data objects represented by p-dimensional feature vectors $\mathbf{x}_k = [x_{1k}, \ldots, x_{pk}]^T \in \mathbb{R}^p$. A set of N feature vectors is then represented as $p \times N$ data matrix $\mathbf{X}$. A fuzzy clustering algorithm partitions the data $\mathbf{X}$ into M fuzzy clusters, forming a fuzzy partition in $\mathbf{X}$. A fuzzy partition can be conveniently represented as a matrix $\mathbf{U}$, whose elements $u_{ik} \in [0, 1]$ represent the membership degree of $\mathbf{x}_k$ in cluster i. Hence, the i-th row of $\mathbf{U}$ contains values of the i-th membership function in the fuzzy partition.

Objective function based fuzzy clustering algorithms minimize an objective function of the type:

$$J(\mathbf{X}; \mathbf{U}, \mathbf{V}) = \sum_{i=1}^{M} \sum_{k=1}^{N} (u_{ik})^m \; d^2(\mathbf{x}_k, \mathbf{v}_i), \tag{1}$$

where $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M]$, $\mathbf{v}_i \in \mathbb{R}^p$ is M-tuple centers which have to be computed, and $m \in (1, \infty)$ is a weighting exponent which defines the fuzziness of the clusters. The conventional FCM uses Euclidean distance. The optimization is constrained, amongst others, by the constraint

$$\sum_{i=1}^{M} u_{ik} = 1, \qquad \forall k. \tag{2}$$

2.2 Non-parametric Rank Correlation

Nonparametric statistics can be an effective tool when data is observed on a discrete scale of values or when the assumptions required by parametric statistics can not be satisfied. This time we cannot rely on the central limit theorem which is a concept to justify use of parametric tests and we must turn to a category of alternative procedures named nonparametric techniques. The nonparametric tests use information of a lower rank, such as nominal or ordinal observations. No assumptions about the form of the parent population are required [6].

Spearman's rank correlation is one of the statistical tools to calculate non-parametric correlations between pairs of samples. If we make two sets of ordinal observations on a number of objects, we can designate one of the sets as x and the other as y. We then rank each observation and call the two sets of ranks $R(x_i)$ and $R(y_i)$. Spearman's coefficient measures the similarity between these two ranks [4],

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\left[R(x_i) - R(y_i)\right]^2}{n(n^2 - 1)}. \tag{3}$$

The term inside the brackets of the numerator is simply the difference between the rank of property x and the rank of property y as observed on the i-th object. The following assumptions can be given for conducting the implementation.

- The correlation between the variables should be linear.
- The two variables have been reduced to an ordinal scale of observation.
- If a test of significance is applied, the sample has been selected randomly from the population.

The rank correlation r_s, is analogous to simple correlation r in that it varies from +1.0 (perfect correspondence between the ranks) to −1.0 (perfect inverse relationship between the ranks). A rank correlation of $r_s = 0$ shows that the two sets of ranks are independent. Note that the rank correlation analysis is insufficient, if the number of observations is bigger than 60 [8].

2.3 Regional Appraisal with Memberships

Generally, in natural world spatial systems have heterogeneous property and different zones are observed in a site. Due to these available separate regions, from a clustering algorithm a better partition is expected for heterogeneous sites rather than homogeneous sites. From this point, it could be anticipated that the correlations provided between the clusters should be bigger for a heterogeneous system than a homogeneous system.

In some circumstances, a relatively small sample, whose size cannot be increased and whose underlying population may be distinctly non-normal, has to be studied. When the sample size is small, the uncertainty about the value of the true correlation can be very large, particularly when the estimated correlation is low [10]. Considering this condition, to measure the correlations between the clusters, membership values and their ranks could be used on the ground of a non-parametric correlation analysis. The algorithm of the analysis can be presented by a flowchart as in Fig. 1.

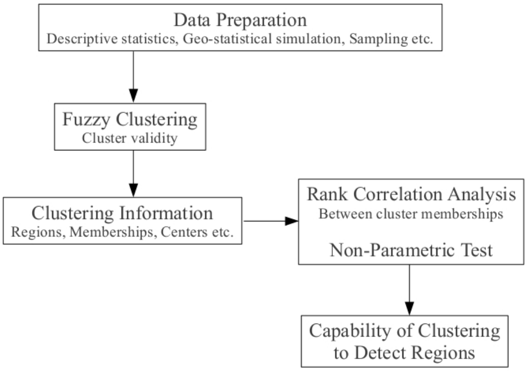
Data Preparation
Descriptive statistics, Geo-statistical simulation, Sampling etc.
Fuzzy Clustering
Cluster validity
Clustering Information
Regions, Memberships, Centers etc.
Rank Correlation Analysis
Between cluster memberships
Non-Parametric Test
Capability of Clustering
to Detect Regions

Fig. 1 Flow chart of the analysis

3 Simulation Studies

3.1 Data Set

Experimental studies have been carried out using two simulated data sets. In the applications, the effectiveness and partitioning capacity of the FCM algorithm on different types of spatial systems has been investigated. The spatial real data set (108 observations) used in [13] was handled. This data set comprised of Elasticity Modulus (EM) values of rock samples collected from an Andesite quarry in Ankara.

To perform the simulation studies, the real set was conditioned by a geostatistical simulation technique which is lower-upper (LU) decomposition technique [7]. For the first case study two simulated sets, one of which has homogeneous and other has heterogeneous properties, were provided based on conditional simulation. In the heterogeneous site, the EM values generate different zones and the spatial variability of the site can be modeled by a function such as Spherical or Gaussian type functions.

Each simulation is conducted on a 21 × 21 regular grid, yielding a total of 441 values. After that, a similar procedure was followed for the second case study. This time, a simulation was carried out on a 20 × 20 regular grid and a total of 400 values.

3.2 Simulation Study 1

Two sample sets (49 records) were randomly drawn from the simulated data sets, each of them including 441 observations. To illustrate the different spatial characteristics, a semi-variance analysis, which is a well-known

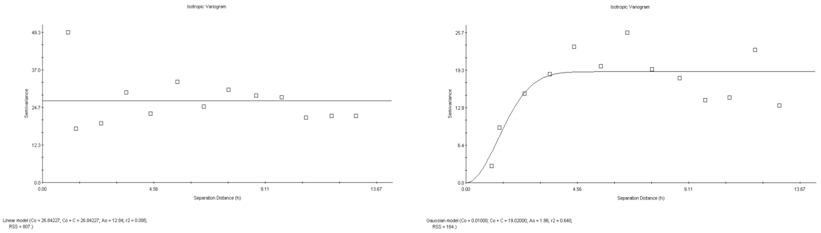

Fig. 2 Spatial behaviors for homogeneous (left) and heterogeneous (right) data sets (simulation study 1)

geostatistical analysis, has been performed. Figure 2 shows the variogram models provided for the sampled distinguished sets. As can be seen in the Fig. 3, although the homogeneous site can not show a spatial relationship (pure nugget model), the heterogeneous site has a Gaussian character. To specify the regions in the sites, fuzzy clustering applications have been performed both for homogeneous and heterogeneous data sets. These different data sets used by the clustering algorithm have the same coordinates and different EM values. Therefore, data matrix **X** contains three dimensions (spatial positions and EM). As a result of the clustering validity studies [3], the optimal number of clusters was defined as four for both sites.

Statistically, if the coefficient of skewness S_f is zero, then the distribution is symmetrical and must be zero for the normal distribution. Similarly, if the Kurtosis is zero, then the distribution of data is approximately normal [14]. Based on these criteria, the memberships have been appraised and use of a nonparametric rank correlation analysis method is decided. Table 1 summarizes the non parametrical (cross) correlation coefficients with the average values for both homogeneous and heterogeneous sets. The values under $N(0,1)$ describe the approximated values of the coefficients required from the large number of data.

Table 1 Rank correlation coefficients among the clusters

Cross Correlation	Homogeneous	Homogeneous $N(0,1)$	Heterogeneous	Heterogeneous $N(0,1)$
r_{12}	−0.380	−2.630	0.101	0.696
r_{13}	−0.023	−0.158	−0.487	−3.370
r_{14}	0.028	0.192	0.168	1.162
r_{23}	0.042	0.288	−0.089	−0.619
r_{24}	0.132	0.913	−0.482	−3.338
r_{34}	−0.377	−2.610	0.010	0.072
Average Correlation	−0.096	−0.668	−0.130	−0.900

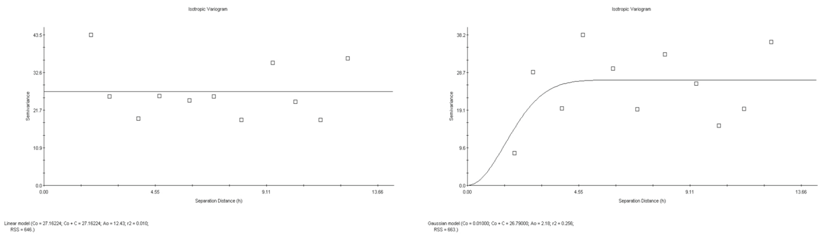

Fig. 3 Spatial behaviors for homogeneous (top) and heterogeneous (bottom) data sets (simulation study 2)

3.3 Simulation Study 2

For the second application, a similar procedure to the one followed in the first application is performed. Firstly, two data sets (each of 25 records) were randomly sampled from the simulated data sets, including 400 observations each one. In order to measure the spatial variability of the observations variogram models have been obtained. Figure 3 illustrates the models. In the homogeneous site, no meaningful spatial dependence is recorded. On the other hand, the heterogeneous site shows a spatial model that is Gaussian.

Based on clustering validity, the optimal number of clusters has been determined as four for both data structures. By using the memberships provided from the clustering application, the nonparametric rank correlation analysis method is applied. Table 2 indicates the cross correlation coefficients with the average values for both data sets.

Table 2 Absolute rank correlation coefficients among the clusters

Cross Correlation	Homogeneous	Heterogeneous
r_{12}	0.439	0.132
r_{13}	0.070	0.125
r_{14}	0.013	0.476
r_{23}	0.237	0.402
r_{24}	0.350	0.066
r_{34}	0.385	0.335
Average Absolute Correlation	0.249	0.256

3.4 Results and Discussion

Because limited number of data may not be increased and the underlying population may be distinctly non-normal in spatial environmental systems, the applications were conducted in the proposed manner. First application showed that the clustering algorithm has a capability to separate the regions. Both the average correlation coefficients are negative and the value obtained for the heterogeneous site is bigger than the homogeneous site. This point indicates the expected result that more clear partition should be carried out for a heterogeneous site.

To test the study, a null hypothesis can be established that the clusters are independent (i.e. $\rho = 0$). The alternative hypothesis is $\rho \neq 0$, so the test is two-tailed, with either very large positive or very large negative correlations leading to rejection. Our analysis shows that the null hypothesis is not rejected both for the homogeneous and the heterogeneous case, indicating independence of clusters.

Second case study was performed by relatively small data sets. Both the average correlation coefficients address the inverse correlations and the clustering algorithm has a capability to determine the regions. In this application, to overcome a possible compensation that may be resulted from pairs close to $+1$ and -1, the study has been carried out using the absolute values. The null hypothesis is that cluster memberships are independent, or that $\rho = 0$. The alternative hypothesis is $\rho \neq 0$, the test is one-tailed. Again, it is found that the null hypothesis is not rejected. Depending on the limited number of data, a crisp difference between two data sets has not been recorded.

4 Conclusions

The partitioning performance of a fuzzy clustering algorithm on different type spatial systems is examined. To appraise the conditioned spatial systems via limited number of data, fuzzy clustering and non-parametric rank correlation method is integrated. By this way, a regional projection method has been constructed. In conclusion, the combination of fuzzy clustering and non-parametric correlation analysis has produced some reliable results and provide possibilities for future studies in depth.

References

1. Bardossy, G., Fodor, J.: Evaluation of Uncertainties and Risks in Geology. Springer, Berlin (2004)
2. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: the fuzzy c-means clustering algorithm. Computers & Geosciences 10(2-3), 191–203 (1984)
3. Bezdek, J.C., Pal, N.R.: Some new indexes of cluster validity. IEEE Transactions on Systems, Man and Cybernetics, Part B 28(3), 301–315 (1998)
4. Conover, W.J.: Practical Nonparametric Statistics. Wiley, New York (1999)
5. Şen, Z.: Spatial Modelling Principles in Earth Sciences. Springer, New York (2009)
6. Davis, J.: Statistics and Data Analysis in Geology. Wiley, New York (2002)

7. Deutsch, C.V., Journel, A.G.: GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, New York (1998)
8. Dudzic, S.: Companion to Advanced Mathematics and Statistics. Hodder Education, London (2007)
9. Höppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis: methods for classification, data analysis and image recognition. Wiley, New York (1999)
10. Niven, E.B., Deutsch, C.V.: Calculating a robust correlation coefficient and quantifying its uncertainty. Computers & Geosciences 40, 1–9 (2012)
11. Piegorsch, W.W., Bailer, A.J.: Analyzing Environmental Data. Wiley, Chichester (2005)
12. Tutmez, B.: Spatial dependence-based fuzzy regression clustering. Applied Soft Computing 12(1), 1–13 (2012)
13. Tutmez, B., Tercan, A.E.: Spatial estimation of some mechanical properties of rocks by fuzzy modelling. Computers and Geotechnics 34, 10–18 (2006)
14. Wellmer, F.W.: Statistical Evaluations in Exploration for Mineral Deposits. Springer, Heidelberg (1998)