Journal of Uncertain Systems
Vol.1, No.3, pp.207-221, 2007
Online at: www.jus.org.uk

# Assessment of Uncertainty in Geological Sites Based on Data Clustering and Conditional Probabilities

Bulent Tutmez[1*], A. Erhan Tercan[2]

[1]*Department of Mining Engineering, Inonu University, 44280 Malatya, Turkey.*
[2]*Department of Mining Engineering, Hacettepe University, 06532 Ankara, Turkey.*

Received 24 May 2007; Accepted 20 June 2007

**Abstract**

Geological investigations are carried out under particularly high uncertainties. Assessment of these uncertainties is very important for project planning. This paper presents a methodology to evaluate uncertainty associated with geological structures such as ore deposits and aquifers. In order to assess the uncertainty of the geological systems observed, fuzzy clustering and spatial measures are used. Then, the heterogeneous zones are evaluated using conditional probabilities. The posterior probabilities obtained from testing data provide useful information for assessing the uncertainty. 

**Keywords:** heterogeneity, fuzzy clustering, spatial measure, conditional probability

## 1. Introduction

Geoscientists are always interested in understanding and evaluating the behaviours of geological structures. Assessment of these structures is directly related to available information and uncertainties. Uncertainty represents partial ignorance or the lack of perfect knowledge on the part of the analyst [1]. According to Bardossy and Fodor [4], two main sources of uncertainties can be distinguished in geosciences:

- Uncertainties due to natural variability
- Uncertainties due to human imperfections and incompetency.

This paper focuses on the first type of geological uncertainty and we assume that variability represents diversity or heterogeneity in a geological population that is irreducible by additional measurements. In the heterogeneous structures, the geological properties observed at different locations do not have the similar values and different zones are observed [15].

The treatment of uncertainty in analysis is going through a paradigm shift from a probabilistic framework to a generalized framework that includes both probabilistic and nonprobablistic methods. The well known probability theory and related statistics are the most common traditional tools to handle uncertainties. In recent years, geostatistical methods have been broadly applied to quantify the geological uncertainties. The geostatistical approach allows the quantitative assessment of spatial uncertainty using geostatistical simulation procedures [13]. Although integrating geostatistics with fuzzy set theory is a novel direction its applications in uncertainty evaluation are very limited. Most geological investigations deal with random (stochastic) events and this is the reason why the probability theory and statistics are at present the basic

* Corresponding author: btutmez@inonu.edu.tr.

tools to handle uncertainties in geosciences [5]. On the other hand, another important type of uncertainty, referred to as fuzziness, is uniquely connected with fuzzy sets. It is thus applicable to fuzzy set theory. Due to its soft property, in recent years, fuzzy set approach has been widely used in uncertainty analyses [3].

In the present study, heterogeneities of geological structures are analyzed using both fuzzy and probabilistic tools. In the first stage, structure identification of the site considered is carried out using the combination of two different tools: fuzzy clustering and geostatistical techniques. Then, the structure defined in the previous step is tested based on conditional posterior probabilities derived from testing observations.

The rest of the paper is organized as follows: Section 2 describes the proposed methodology. In this section, fuzzy clustering, spatial variation and conditional probability approaches are introduced, respectively. Section 3 presents two applications for both simulated and real data sets. Section 4 gives a brief evaluation and the conclusions.

## 2. Methodology

The methodology considers uncertainty in geological sites combining fuzzy clustering with spatial measures and conditional probabilities. The method first uses fuzzy clustering for determining the heterogeneous zones in geological site considered. After that, search (control) domains are established for each cluster based on point semimadogram (PSM) measures which determine the zone of influence (search radius) around each cluster center. In the final stage, the uncertainty evaluation for each zone is carried out using the domains.

### 2.1 Clustering

Data clustering is employed to organize observed data into meaningful structures. In general, cluster analysis refers to a broad spectrum of methods which try to subdivide a data set X into c subsets (clusters) which are pair-wise disjoint, all nonempty, and reproduce X [6]. Cluster analysis encompasses a number of different classification algorithms. Recently, fuzzy algorithms have been widely used as a method for data clustering. Fuzzy clustering partitions a data set into fuzzy clusters such that each data point can belong to multiple clusters.

Fuzzy c-means (FCM) clustering algorithm [6] is a well-known clustering technique that generalizes the classical (hard) c-means algorithm and can be employed when the number of clusters is not known. The FCM algorithm has been used in grade estimation studies for evaluation of uncertainty in geosciences [17].

FCM partitions a collection of $n$ vector $x_i$, $i=1,...,n$ into $c$ fuzzy groups, and finds a cluster center in each group such that a cost function of dissimilarity measure is minimized [10]. The fuzzy partition matrix is of probabilistic property as follows:

$$\sum_{i=1}^{c} \mu_{ij} = 1, \qquad \forall j = 1,...,n \tag{1}$$

The objective function for FCM is given by Eq. 2.

$$J_m(U,c) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^m \; d_{ij}^2 \tag{2}$$

where $\mu_{ij}$ takes values between 0 and 1, $c_i$ is the cluster center of fuzzy group $i$, $d_{ij} = \left\| c_i - x_j \right\|$ is the Euclidean distance between $i$th cluster center and jth data point and $m>1$ is a weighting exponent. Cluster centers and membership matrix are calculated as follows:

$$c_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m x_j}{\sum_{j=1}^{n} \mu_{ij}^m} \tag{3}$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \tag{4}$$

The FCM algorithm determines the cluster centers $\boldsymbol{c}_i$ and the membership matrix $\boldsymbol{U}$ using the following steps [11]:

1. Initialize the cluster centers $\mathbf{c}_i$ *(i = 1,2,…,c).*
2. Determine the membership matrix $\boldsymbol{U}$ by (4).
3. Compute the objective function according to (2). Stop if it is below a certain threshold level or its improvement over the previous iteration is below a certain tolerance.
4. Update the cluster centers according to (3).
5. Go to step 2.

Determining the optimal number of clusters is an important step in fuzzy clustering. In this study, the cluster validity index proposed by Tutmez et al. [18] is used. This method is based on reproducing the variability in the value of cluster centers with minimum number of clusters.

### 2.2 A Stochastic Approach to Spatial Variation: Point Semimadogram

Fuzzy clustering focuses on the dissimilarity between data values based on metric distances. Therefore, the clustering algorithm may be used to identify heterogeneous areas [18]. The fact that spatial variation appears to be random suggests a way forward [19]. The madogram is a function of underlying stochastic process. In geosciences, two sampled values *g(x)* and *g(x+h)* at two points *x* and *x+h* separated by the vector *h* are spatially correlated. As the distance between these values increases, one would expect that the spatial correlation decreases and vice versa. This type of behaviour is modelled by variogram functions in geostatistics, which is commonly used in evaluation of uncertainties in geology. The variogram function is defined as the variance of the differences between two attribute values.

$$2\gamma(h) = Var\left[g(x) - g(x+h)\right] \tag{5}$$

where *g(x)* and *g(x+h)* are random measurements defined at locations *x* and *x+h*, *Var* is the variance operator and *2γ(h)* is the variogram at distance *h*. In practice, the variogram function (5) is estimated as follows:

$$2\gamma(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} \left[g(x_i) - g(x_i + h)\right]^2 \tag{6}$$

*N(h)* is the number of pairs used in calculating the variogram at distance *h*.

The variogram characterizes the structures of the spatial distribution of the attribute considered. Starting from the origin, *γ(0)=0*, the variogram increases in general with increasing distance. The variogram may become stable beyond some distance $h = a$ called the range. Beyond this distance *a*, the mean square deviation between two grades *g(x)* and *g(x+h)* no longer depends on distance *h* between them and the two values are not correlated (Fig.1). The range gives an exact sense to the conventional concept of the zone of influence of a sample [8]. However, the classical variogram is not suitable for describing the local variability. Therefore point semivariogram (PSV) has been proposed [15].

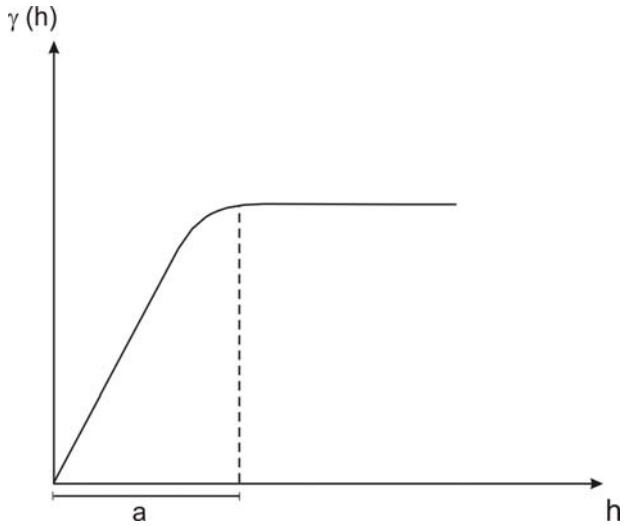


**Fig.1.** A typical variogram

Madograms are particularly useful for establishing the range parameter [7]. In a recent work, point semimadogram (PSM) was suggested as an alternative measure for evaluating the local spatial behavior of data [18]. By this measure, the zone of influence around each observation can be determined. This function uses the absolute difference instead of squaring the difference between $g_m$ and $g_{m+h}$. If data set employed includes the outlier values and the number of data is limited, the PSM is more resistant to outlier values [18],

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left| g_m - g_{(m+h_i)} \right|. \tag{7}$$

Point cumulative semimadogram (PCSM), which is an extended form of PSM, is computed by the cumulative sum of PSMs. The PCSM leads to a non-decreasing function with distance. Because the PCSM gives the regional effect of all the other data locations within the study area on the location concerned, it can be used for determining the search (control) domains around each cluster center. According to Eq. (8), a location $x$ is defined to belong to domain $\Omega$ if the Euclidean distance between cluster center $c_i$ and $x_j$ is not greater than the range $a$ of the location considered.

$$x_j \in \Omega \qquad \text{if } d(c_i, x_j) \le a \quad j = 1,2,...,N \quad i = 1,2,...,c \tag{8}$$

where $N$ is the number of data. If a location belongs to more than one cluster domain, the closest cluster center is considered.

### 2.3 Conditional Probability

For a given site clustered by fuzzy clustering, the following question can be asked: "What is the chance of an observed data belonging to cluster1 on the second sampling given either a cluster 2 or a cluster 3 on the first sampling"-. This is a problem of conditional probability, the probability of an event given that specified events have occurred in the past. Conditional probability can be formalized by determining the probability that event $A$ occurs given that arbitrary event $E$, where

$$\Pr(A \setminus \mathrm{E}) \equiv \frac{\Pr(\mathrm{A} \cap \mathrm{E})}{\Pr(\mathrm{E})}. \tag{9}$$

Rearranging this definition gives

$$\Pr(A \cap E) = \Pr(A \setminus \mathrm{E}) \Pr(\mathrm{E}). \tag{10}$$

Eq. (10) is known as the multiplication rule of conditional probability. Further, since $\Pr(E \cap A) = \Pr(A \cap E)$, Equation (10) implies that

$$\Pr(A \setminus \mathrm{E}) \Pr(\mathrm{E}) = \Pr(\mathrm{E} \setminus \mathrm{A}) \Pr(\mathrm{A}). \tag{11}$$

If $A_1, A_2, ..., A_n$ denotes the prior information, or a partition of a universal set *X*, and $E \subset X$ represents the arbitrary event as shown in Figure 2 [2], the theorem of total probability can be stated as follows

$$\Pr(E) = \Pr(A_1)\Pr(E \setminus A_1) + \Pr(A_2)\Pr(E \setminus A_2) + ... + \Pr(A_n)\Pr(E \setminus A_n) \quad (12)$$

where $\Pr(A_i)$ = the probability of the event $A_i$ and $E \setminus A$ = the occurrence of *E* given $A_i$, where *i = 1,2,…,n.* This theorem has considerable importance in computing the probability of the event E, especially in practical cases where the probability cannot be computed directly, but the probabilities of the partitioning events and the conditional probabilities can be computed [3].

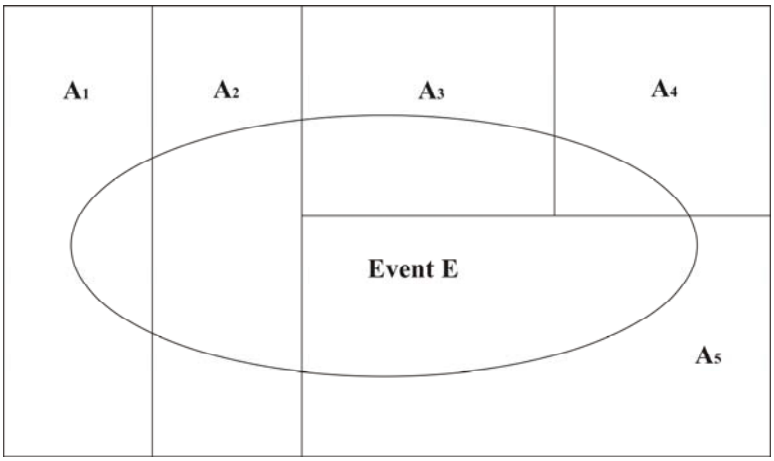


**Fig.2.** Bayes' theorem.

The theorem presented by Bayes is based on the same conditions of partitioning and events as the theorem of total probability and is very useful in computing the posterior (reverse) probability of the type $\Pr(A_i \setminus E)$, for *i = 1,2,…,n.* The posterior probability can be computed as follows:

$$\Pr(A_i \setminus E) = \frac{\Pr(A_i)\Pr(E \setminus A_i)}{\Pr(A_1)\Pr(E \setminus A_1) + \Pr(A_2)\Pr(E \setminus A_2) + ... + \Pr(A_n)\Pr(E \setminus A_n)} \quad (13)$$

According to Bayes approach, from prior probabilities $\Pr(A_i)$ and new evidence expressed in terms of conditional probabilities $\Pr(E \setminus A_i)$, posterior probabilities $\Pr(A_i \setminus E)$ are computed. When further evidence becomes available, the posterior probabilities are employed as prior probabilities and the procedure of probability updating, is repeated [12]. In this paper, posterior probabilities are employed to evaluate the heterogeneity of the geological structure considering clustering information and spatial measures.

## 3 Case Studies

The methodology presented in this paper is illustrated using two data sets. In the first case, the method is used for evaluating a simulated data obtained from an Andesite (rock) quarry. In the second case study, a real data set is used and this time, the proposed method is used for the purpose of evaluating aquifer porosity.

### 3.1 Case Study 1

The data set was obtained via conditional simulation using the lower-upper (LU) simulation technique [7]. Simulation is carried out on a 21x21 regular grid with 7 unit grid spacing, yielding a total of 441 values. A standardized set (49 records) is randomly drawn from this data set and are shown in Fig.3.

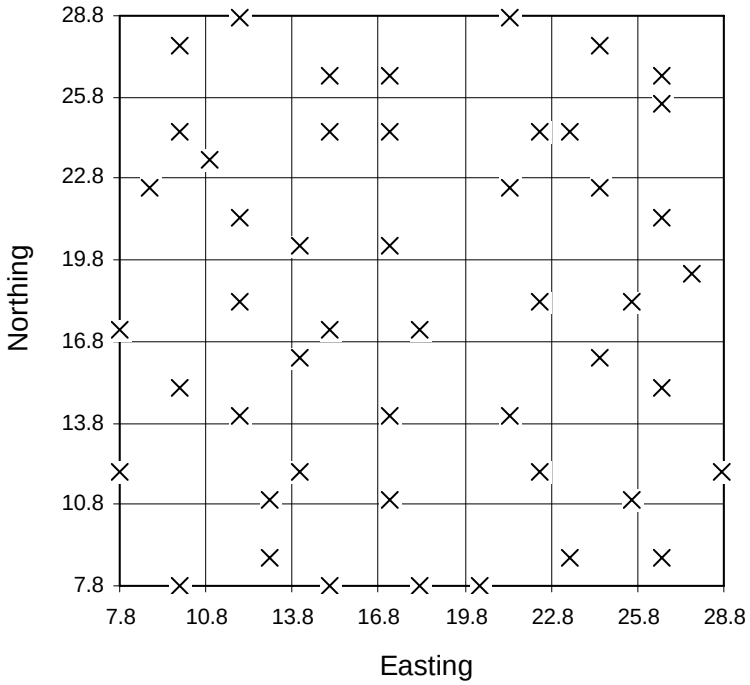


**Fig. 3.** Simulated data set.

## Clustering

In the first stage, data clustering was carried out. The most important component of this stage is determining number of clusters. If the number of clusters is unknown, various methods might be employed to find a suitable number of clusters [14]. In this study, a novel cluster validity approach, which was performed in different problems [16,18] for appraising the geological structures, has been used. It is based on reproducing the variability of the sample data in the value of cluster centers with minimum number of clusters. For the simulated data set, the optimum number of clusters is determined to be five.

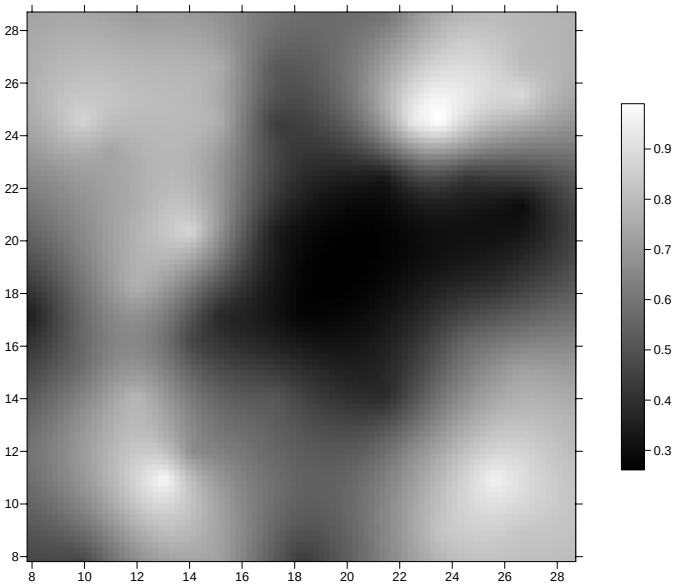


**Fig. 4.** The map of membership values for the simulated data.

In addition to cluster centers, the FCM algorithm produces a partition (membership) matrix, whose *ij*th element $\mu_{ij} \in [0,1]$ is the membership degree of data $x_j$ in cluster *i*. The one-dimensional fuzzy sets $\mu_{A_{ik}}$ can be projected onto the space of the input variables $x_k$. where, the *i*th row of *U* contains a pointwise definition [9] of a multidimensional fuzzy set. For this procedure, the expresion $\mu_{A_{ik}}(x_{kj}) = \text{proj}_k(\mu_{ij})$ has been employed. As a result of this application, the maximum memberships for each observation (location) are computed. A map of membership values is shown in Fig. 4.

**Spatial measure**

When the spatial distribution of an attribute is heterogeneous, individual PCSMs for the sample values show different behaviour. Calculated PCSMs are plotted on vertical axis versus the corresponding distances on horizontal axis. As an example, the PCSM plot of cluster1 and its domain is shown in Figures 5 and 6, respectively.

As seen in Fig. 4, the quarry has five different zones. The number of observations in each zone denotes the probabilities belonging to clusters. These probabilities can be computed as follows:

$P(C_1)$ = 15/49 = 0.306 $P(C_2)$ = 9/49 = 0.184 $P(C_3)$ = 7/49 = 0.143
$P(C_4)$ = 9/49 =0.184 $P(C_5)$ = 9/49 = 0.184.

where, $C_{1\text{-}5}$ are the clusters and *P(C)* is the probability related to the cluster considered.

Similarly, the prior probabilities can be obtained from the membership matrix using a basic categorization. For this purpose, five levels (intervals) have been designed and the prior probabilities of each level have been determined as shown in Table 1, where *N* denotes the number of data values within the interval defined and *P(p)* denotes the prior probabilities. For example, the prior probability 0.510 is obtained from the division 25/49.

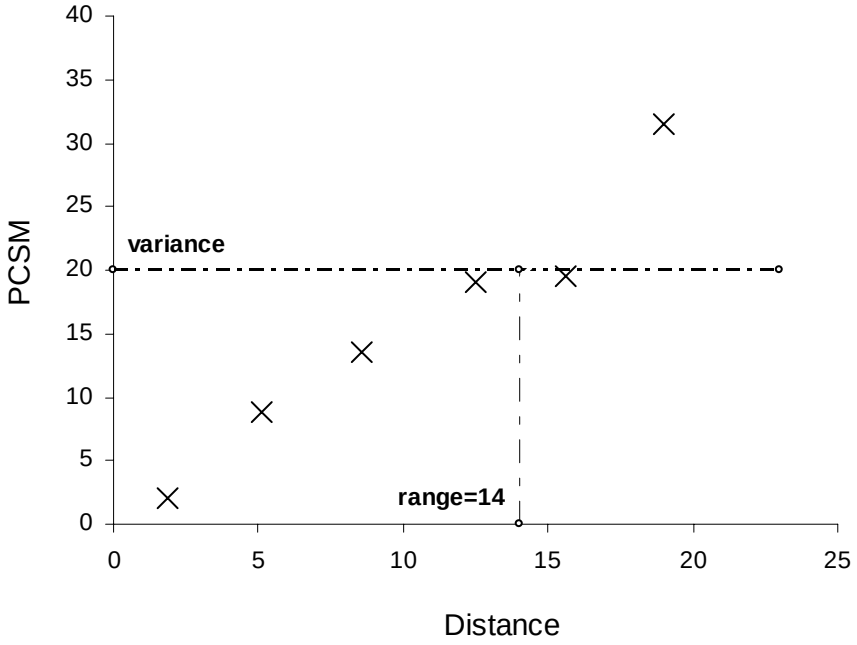


**Fig. 5.** Experimental PCSM plot for cluster 1.

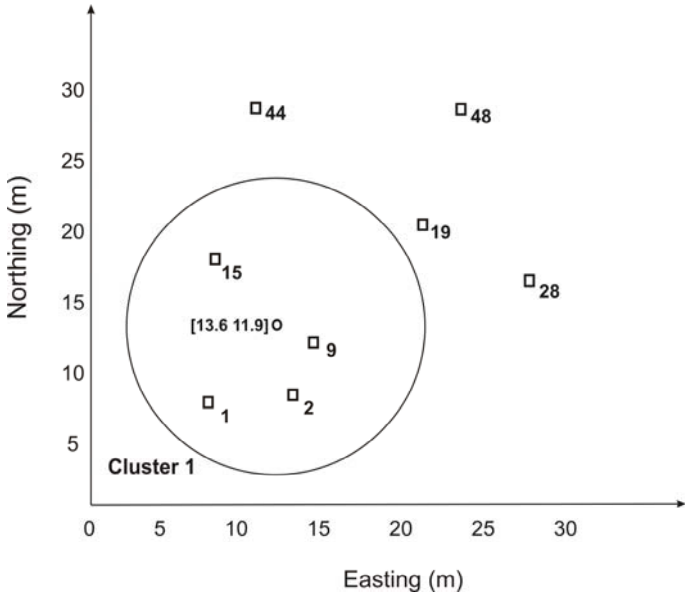


**Fig. 6.** Control domain for cluster 1.

According to prior discrete probabilities, the following mean probabilities for each cluster can be computed by Eq. (14):

$$\bar{P}_X = \sum_{i=1}^{5} p_i P(p_i) \tag{14}$$

$C_1 \rightarrow P_X$ = 0.05(0.510) + 0.2(0.204) + 0.4(0.102) + 0.6(0.102) + 0.85(0.082) = 0.238
$C_2 \rightarrow P_X$ = 0.05(0.531) + 0.2(0.224) + 0.4(0.082) + 0.6(0.061) + 0.85(0.102) = 0.228
$C_3 \rightarrow P_X$ = 0.05(0.531) + 0.2(0.326) + 0.4(0.082) + 0.6(0) + 0.85(0.061) = 0.177
$C_4 \rightarrow P_X$ = 0.05(0.592) + 0.2(0.184) + 0.4(0.102) + 0.6(0.020) + 0.85(0.102) = 0.206
$C_5 \rightarrow P_X$ = 0.05(0.510) + 0.2(0.286) + 0.4(0.041) + 0.6(0.041) + 0.85(0.122) = 0.227

**Table 1.** Levels and prior probabilities.

| Levels | Cluster 1 | | Cluster 2 | | Cluster 3 | | Cluster 4 | | Cluster 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N[a] | P(p)[b] | N | P(p) | N | P(p) | N | P(p) | N | P(p) |
| μ < 0.1 | 25 | 0.510 | 26 | 0.531 | 26 | 0.531 | 29 | 0.592 | 25 | 0.510 |
| 0.1 ≤ μ ≤ 0.3 | 10 | 0.204 | 11 | 0.224 | 16 | 0.326 | 9 | 0.184 | 14 | 0.286 |
| 0.3 ≤ μ ≤ 0.5 | 5 | 0.102 | 4 | 0.082 | 4 | 0.082 | 5 | 0.102 | 2 | 0.041 |
| 0.5 ≤ μ ≤ 0.7 | 5 | 0.102 | 3 | 0.061 | 0 | 0 | 1 | 0.020 | 2 | 0.041 |
| μ > 0.7 | 4 | 0.082 | 5 | 0.102 | 3 | 0.061 | 5 | 0.102 | 6 | 0.122 |

**[a]: Number of data [b]: Prior probability**

Because the events computed above are not independent, the joint probabilities can be determined as follows:

$$C_1 \to P(X \cap C_1) = P(X \setminus \mathrm{C}_1)P(C_1) = 0.238 * 0.306 = 0.0728$$
$$C_2 \to P(X \cap C_2) = P(X \setminus \mathrm{C}_2)P(C_2) = 0.228 * 0.184 = 0.0420$$
$$C_3 \to P(X \cap C_3) = P(X \setminus \mathrm{C}_3)P(C_3) = 0.177 * 0.143 = 0.0253$$
$$C_4 \to P(X \cap C_4) = P(X \setminus \mathrm{C}_4)P(C_4) = 0.206 * 0.184 = 0.0379$$
$$C_5 \to P(X \cap C_5) = P(X \setminus \mathrm{C}_5)P(C_5) = 0.227 * 0.184 = 0.0418.$$

As can be seen from the outcomes of calculations outlined above, cluster 1 has the biggest contribution capacity. Therefore, the further analyses may be presented based on the parameters of this cluster.

Assume that a testing data was sampled from the quarry investigated and this observation was found in cluster1 *($x_1$)*. The prior distribution in Table 1 needs to be re-evaluated to reflect the new information. This revised distribution is named the *posterior distribution* $P_p^{'}(p)$ and can be calculated as follows:

$$P_p^{'}(0.05) = P(p_{x1} \setminus \mathrm{x}_1)\frac{P(x_1 \setminus \mathrm{p}_{x1})P(p_{x1})}{P_x} = \frac{0.510(0.05)}{0.238} = 0.107$$

The other posterior probabilities can be determined as follows:

$$P_p^{'}(0.2) = \frac{0.204(0.2)}{0.238} = 0.171$$
$$P_p^{'}(0.4) = \frac{0.102(0.4)}{0.238} = 0.171$$
$$P_p^{'}(0.6) = \frac{0.102(0.6)}{0.238} = 0.257$$
$$P_p^{'}(0.85) = \frac{0.082(0.85)}{0.238} = 0.293$$

The resulting probabilities computed above add up to 1. The average probability of 0.238 can be accepted as a normalizing factor for calculating these probabilities. The posterior mean probability $\bar{p}(X)$ is determined using new posterior distribution as follows:

$$\bar{p}(X) = 0.05(0.107) + 0.2(0.171) + 0.4(0.171) + 0.6(0.257) + 0.85(0.293)$$
$$= 0.511.$$

The posterior probability (0.511) is bigger than the prior mean probability (0.238). This outcome resulted from the sampled testing location. For this observation, the mean complement probability $\bar{p}(CX)$ is 1-0.511 = 0.489. Now, assume that a testing data was sampled from the quarry and this observation was found out of cluster1. This time, the revised posterior distribution is determined based on complement posterior probability $\bar{p}(CX)$. Where, the relationship $\bar{p}(X) + \bar{p}(CX) = 1.0$ should be noted. The revised distribution is the posterior distribution $P_p^{'}(p)$ can be computed as follows:

$$P_p^{'}(0.05) = \frac{0.107(1-0.05)}{1-0.511} = 0.208$$
$$P_p^{'}(0.2) = \frac{0.171(1-0.2)}{1-0.511} = 0.280$$

$$P_p^{'}(0.4) = \frac{0.171(1-0.4)}{1-0.511} = 0.210$$
$$P_p^{'}(0.6) = \frac{0.257(1-0.6)}{1-0.511} = 0.210$$
$$P_p^{'}(0.85) = \frac{0.293(1-0.85)}{1-0.511} = 0.090$$

The mean complement probability $\bar{p}(CX)$ is computed as follows:

$$\bar{p}(CX) = 0.95(0.208) + 0.8(0.280) + 0.6(0.210) + 0.4(0.210) + 0.15(0.090)$$
$$= 0.646.$$

From this, the posterior mean probability is 1-0.646 = 0.354. It can be stated that the posterior mean probability decreases as testing observation is beyond the control domain.

### 3.2 Case Study 2

For the real application, the coastal area between Mersin and Tarsus cities is investigated. This area is located in Southern part of Turkey and it contains agricultural, industrial and settlement areas. Groundwater is used as a main source of water in this region. Since the groundwater is widely used for water supplies, efficient groundwater management is important for this area [16]. Porosity distribution of the area has been analyzed using 27 well logs.

The FCM clustering was performed for determining number of clusters and partition (membership) matrix. For this application, the appropriate number of clusters has been defined as four. Figure 7 shows spatial data set and cluster centers determined.

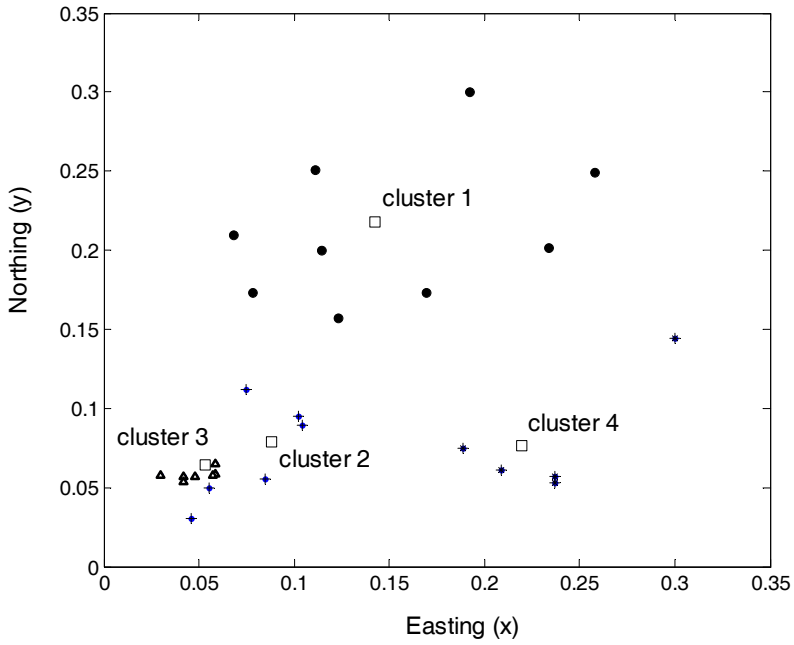


**Fig. 7.** Real set and cluster centers.

By using information derived from data clustering, spatial measures have been performed. As a result of these measures, ranges for search domains have been determined (Fig. 8) using PCSM graphs.

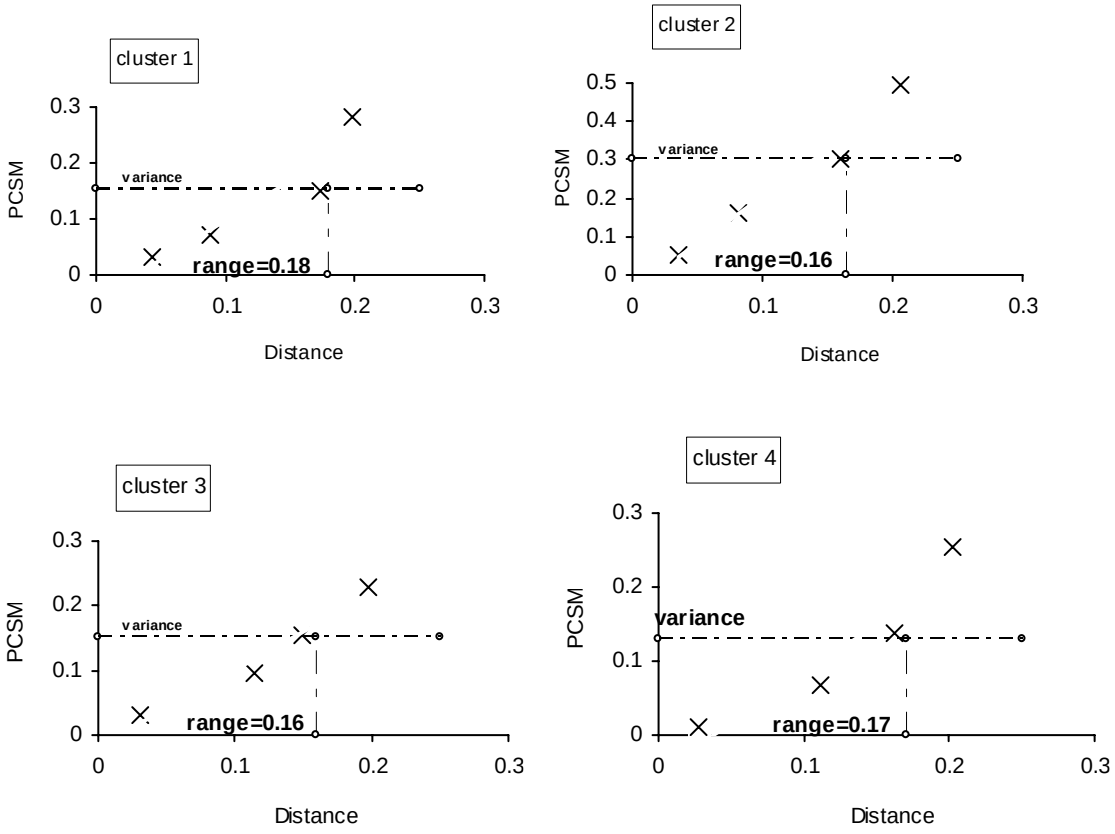


**Fig. 8.** Experimental PCSMs.

Observed number of data within each cluster describes the propagations of the clusters. These propagations may be quantified using probabilities which give the weights of zones in the site considered. The probabilities can be computed as follows:

$P(C_1) = 9/27 = 0.33$ $P(C_2) = 6/27 = 0.22$

$P(C_3) = 7/27 = 0.26$ $P(C_4) = 5/27 = 0.19$

The prior probabilities have been obtained from the membership matrix using five intervals. The mean values of the intervals are defined as [0.1 0.3 0.5 0.7 0.9]. Table 2 indicates these probabilities for each cluster.

According to prior discrete probabilities, following mean and joint probabilities have been computed:

$C_1 \rightarrow P_X = 0.1(0.629) + 0.3(0.111) + 0.5(0.148) + 0.7(0.074) + 0.9(0.037) = 0.255$

$C_2 \rightarrow P_X = 0.1(0.704) + 0.3(0.074) + 0.5(0.037) + 0.7(0.074) + 0.85(0.111) = 0.263$

$C_3 \rightarrow P_X = 0.1(0.741) + 0.3(0.074) + 0.5(0.074) + 0.7(0.0) + 0.9(0.111) = 0.233$

$C_4 \rightarrow P_X = 0.1(0.592) + 0.3(0.148) + 0.5(0.037) + 0.7(0.0) + 0.9(0.222) = 0.322$

$$C_1 \rightarrow P(X \cap C_1) = P(X \setminus C_1)P(C_1) = 0.255 * 0.33 = 0.084$$

$$C_2 \rightarrow P(X \cap C_2) = P(X \setminus C_2)P(C_2) = 0.263 * 0.22 = 0.058$$

$$C_3 \rightarrow P(X \cap C_3) = P(X \setminus C_3)P(C_3) = 0.233 * 0.26 = 0.060$$

$$C_4 \rightarrow P(X \cap C_4) = P(X \setminus C_4)P(C_4) = 0.322 * 0.19 = 0.061$$

**Table 2.** Prior probabilities.

| Levels | Cluster 1 | | Cluster 2 | | Cluster 3 | | Cluster 4 | |
|---|---|---|---|---|---|---|---|---|
| | $N^*$ | $P(p)^{**}$ | $N$ | $P(p)$ | $N$ | $P(p)$ | $N$ | $P(p)$ |
| $\mu < 0.2$ | 17 | 0.629 | 19 | 0.704 | 20 | 0.741 | 16 | 0.592 |
| $0.2 \le \mu \le 0.4$ | 3 | 0.111 | 2 | 0.074 | 2 | 0.074 | 4 | 0.148 |
| $0.4 \le \mu \le 0.6$ | 4 | 0.148 | 1 | 0.037 | 2 | 0.074 | 1 | 0.037 |
| $0.6 \le \mu \le 0.8$ | 2 | 0.074 | 2 | 0.074 | 0 | 0.0 | 0 | 0.0 |
| $\mu > 0.8$ | 1 | 0.037 | 3 | 0.111 | 3 | 0.111 | 6 | 0.222 |

***: Number of data **: Prior probability**

Results show that cluster 1 has the biggest contribution. Because of this, the further analyses may be performed using this cluster. For this case study, a testing data set was sampled from the aquifer. The first observation in this set is found within cluster1. Posterior probabilities $P_p^{'}(p)$ and posterior mean probability $\bar{p}(X)$ have been determined as follows:

$$P_p^{'}(0.1) = \frac{0.629(0.1)}{0.255} = 0.247$$

$$P_p^{'}(0.3) = \frac{0.111(0.3)}{0.255} = 0.131$$

$$P_p^{'}(0.5) = \frac{0.148(0.5)}{0.255} = 0.290$$

$$P_p^{'}(0.7) = \frac{0.074(0.7)}{0.255} = 0.203$$

$$P_p^{'}(0.9) = \frac{0.037(0.9)}{0.255} = 0.130$$

$$\bar{p}(X) = 0.1(0.247) + 0.3(0.131) + 0.5(0.290) + 0.7(0.203) + 0.9(0.130) = 0.469$$

The procedure presented above has been carried out for each test data. For example, for 5$^{th}$ observation, which is beyond cluster 1, the following calculations have been performed:

$$P_p^{'}(0.1) = \frac{0.025(1-0.1)}{0.418} = 0.054$$

$$P_p^{'}(0.3) = \frac{0.095(1-0.3)}{0.418} = 0.159$$

$$P_p^{'}(0.5) = \frac{0.414(1-0.5)}{0.418} = 0.495$$

$$P_p^{'}(0.7) = \frac{0.340(1-0.7)}{0.418} = 0.244$$

$$P_p^{'}(0.9) = \frac{0.120(1-0.9)}{0.418} = 0.029$$

$$\bar{p}(CX) = 0.9(0.054) + 0.7(0.159) + 0.5(0.495) + 0.3(0.244) + 0.1(0.029) = 0.484$$

$$\bar{p}(X) = 0.516$$

Table 3 indicates the results of the calculations and mean probabilities for entire test data set. The results are also shown in Figure 9. The figure gives the effect of observations which are beyond the cluster considered.

**Table 3.** Prior and posterior distributions for cluster 1.

| | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | $\bar{p}(X)$ | $\bar{p}(CX)$ |
|---|---|---|---|---|---|---|---|
| P(p) | 0.6290 | 0.1110 | 0.1480 | 0.0740 | 0.0370 | 0.2550 | 0.7450 |
| Post.1 **(X)**[a] | 0.2467 | 0.1306 | 0.2902 | 0.2031 | 0.1306 | 0.4687 | 0.5313 |
| Post.2 **(X)** | 0.0526 | 0.0836 | 0.3096 | 0.3034 | 0.2508 | 0.6232 | 0.3768 |
| Post.3 **(CX)**[b] | 0.1257 | 0.1553 | 0.4108 | 0.2416 | 0.0666 | 0.4936 | 0.5064 |
| Post.4 **(X)** | 0.0255 | 0.0944 | 0.4162 | 0.3426 | 0.1214 | 0.5880 | 0.4120 |
| Post.5 **(CX)** | 0.0556 | 0.1604 | 0.5051 | 0.2495 | 0.0295 | 0.5073 | 0.4927 |
| Post.6 **(X)** | 0.0110 | 0.0948 | 0.4978 | 0.3442 | 0.0523 | 0.5664 | 0.4336 |
| Post.7 **(X)** | 0.0019 | 0.0502 | 0.4394 | 0.4254 | 0.0830 | 0.6075 | 0.3925 |
| Post.8 **(CX)** | 0.0044 | 0.0896 | 0.5597 | 0.3251 | 0.0212 | 0.5538 | 0.4462 |
| Post.9 **(X)** | 0.0008 | 0.0485 | 0.5054 | 0.4109 | 0.0344 | 0.5859 | 0.4141 |
| Post.10 **(CX)** | 0.0001 | 0.0248 | 0.4313 | 0.4910 | 0.0528 | 0.6143 | 0.3857 |
| Post.11 **(X)** | 0.0 | 0.0121 | 0.3510 | 0.5595 | 0.0774 | 0.6404 | 0.3596 |
| Post.12 **(X)** | 0.0 | 0.0057 | 0.2741 | 0.6115 | 0.1087 | 0.6647 | 0.3353 |
| Post.13 **(X)** | 0.0 | 0.0026 | 0.2062 | 0.6440 | 0.1472 | 0.6872 | 0.3128 |
| Post.14 **(X)** | 0.0 | 0.0011 | 0.1500 | 0.6561 | 0.1928 | 0.7081 | 0.2919 |
| Post.15 **(CX)** | 0.0 | 0.0027 | 0.2570 | 0.6743 | 0.0661 | 0.6607 | 0.3393 |
| Post.16 **(X)** | 0.0 | 0.0012 | 0.1944 | 0.7144 | 0.0900 | 0.6786 | 0.3214 |
| Post.17 **(X)** | 0.0 | 0.0005 | 0.1433 | 0.7369 | 0.1193 | 0.6950 | 0.3050 |

**(X)[a]: in the cluster, (CX)[b]: beyond the cluster**

### 3.3 Discussion

The main motivation of this paper is integrating the soft and probabilistic computing in the same ground. In addition, geoscientists need to evaluate uncertainties and make decisions based on limited information. The results show that the methodology presented in this paper provides a new tool for handling the geological uncertainties. The case studies indicate that there is a close relationship between uncertainty and spatial variability. The posterior mean probabilities are changed in connection with the spatial positions of the observations. As can be seen from the case studies that the average probability is approaching 1 as more and more testing data are used.

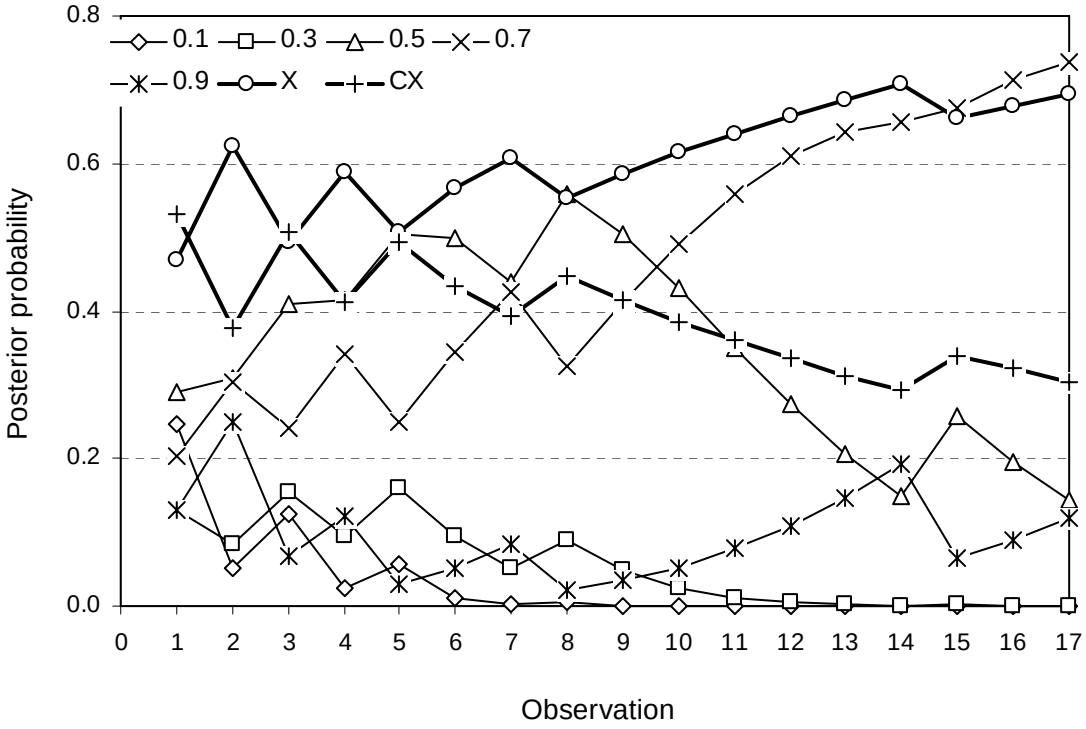


**Fig. 9.** Posterior distributions for cluster 1.

## 4 Conclusions

The conditional probability has been employed for appraising uncertainty in the geological sites based on fuzzy and geostatistical approaches. By the proposed methodology, first spatial data have been classified using fuzzy clustering and spatial measures have been performed by PCSM functions. Finally, the posterior probabilities, which allow to characterize the geological uncertainties, have been computed based on testing observations. The results from simulation and real data indicate that the posterior mean probability decreases as the testing observation is beyond control domain.

## References

[1] Anderson EL, and Hattis D, Uncertainty and variability, Risk Analysis, Vol.19, No.1, 1999.
[2] Ayyub BM, and McCuen R, Probability, Statistics, and Reliability for Engineers and Scientists, CRS Press, Boca Raton, 2003.
[3] Ayyub BM, and Klir GJ, Uncertainty Modeling and Analysis in Engineering and the Sciences, CRC Press, Boca Raton, 2006.

[4] Bardossy Gy, and Fodor J, Traditional and new ways to handle uncertainty in geology, Natural Resources Research , Vol.10, No.3, 169-187, 2001.
[5] Bardossy Gy, and Fodor J, Evaluation of Uncertainties and Risks in Geology, Springer-Verlag, Heidelberg, 2004.
[6] Bezdek JC, Ehrlich R, and Full W, FCM: the fuzzy c-means clustering algorithm, Computers & Geosciences, Vol.10, No.2-3, 191-203, 1984.
[7] Deutsch CV, and Journel AG, GSLIB: Geostatistical Software Library, Oxford Press, 1998.
[8] Goovaerts P, Geostatistics for Natural Resources Evaluation, Oxford University Press, New York, 1997.
[9] Höppner F, Klawonn F, Kruse R, Runkler T, Fuzzy Cluster Analysis, John Wiley & sons, 1999.
[10] Jang J-SR, Sun CT, and Mizutani E, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice Hall, London, 1997.
[11] Jantzen J, Foundations of Fuzzy Control, John Wiley and Sons, Chichester; 2007.
[12] Klir GJ, Uncertainty and Information: Generalized Information Theory, John Wiley & Sons, Hoboken, 2006.
[13] Ortiz JC, and Deutsch CV, Calculation of uncertainty in the variogram. Mathematical Geology, Vol. 34, 169-183.
[14] Pal NI, and Bezdek JC, On cluster validity for the fuzzy c-means model, IEEE Trans. Fuzzy Systems., Vol.3, No.8, 370-379, 1995.
[15] Şen Z, Point cumulative semivariogram for identification of heterogeneities in regional seismicity of Turkey: Mathematical Geology, Vol.30, No.7, 767-787, 1998.
[16] Tutmez B, and Hatipoglu Z, Spatial estimation model of porosity, Computers & Geosciences, Vol.33, 465-475, 2007.
[17] Tutmez B, An uncertainty oriented fuzzy methodology for grade estimation, Computers & Geosciences, Vol 33, No. 2, 280-288, 2007.
[18] Tutmez B, Tercan AE, and Kaymak U, Fuzzy modelling for reserve estimation based on spatial variability, Mathematical Geology, Vol.39, No.1, 2007.
[19] Webster R, and Oliver MA, Geostatistics for Environmental Scientists, John Wiley & Sons, Chichester, 2001.