Soil data clustering by using K-means and fuzzy K-means algorithm

Elma Hot and Vesna Popović-Bugarin, *member, IEEE*

***Abstract* — A problem of soil clustering based on the chemical characteristics of soil, and proper visual representation of the obtained results, is analysed in the paper. To that aim, K-means and fuzzy K-means algorithms are adapted for soil data clustering. A database of soil characteristics sampled in Montenegro is used for a comparative analysis of implemented algorithms. The procedure of setting proper values for control parameters of fuzzy K-means is illustrated on the used database. In addition, validation of clustering is made through visualisation. Classified soil data are presented on the static Google map and dynamic Open Street Map.**

***Keywords* — clustering, data mining, K-means, fuzzy K-means, pedologic map.**

I. Introduction

IN Montenegro, in the period 1958-1988, a detailed soil map with scale of 1:50 000 was made. Unfortunately, as in other Former Republics of Yugoslavia, the enormous effort and the work was not properly presented to the wide professional community and land users, since the data and data map were available only in a hard copy version. The goal of this paper is to investigate the possibility of using data mining tools for soil classification based on the available data and to illustrate adequate visualisation, in order to make this data understandable to wider society.

Data mining (DM) is a set of techniques that aims to discover implicit useful information from big data. Information discovery is usually performed by identifying patterns and establishing relationships. Data Mining allows focusing on the most important information in the data.

DM includes: clustering, anomaly detection, association rule learning, classification, regression and summarization and sequence or path analysis and forecasting. DM is the computer-assisted process of "digging" through enormous databases in order to analyse and extract the meaning of the data.

In this paper focus is on clustering. Clustering has found applications in many research areas such as mathematics, engineering, economics, marketing, machine learning, pattern recognition, genetics, bioinformatics, psychology, biology, data compression and information retrieval. Clustering is a process of grouping similar sets of data. This grouping is unsupervised; it is done without using known structures in the data. Clustering aims to make clusters with data samples which are more similar to each other than to data samples that belong to the other clusters. Each cluster is defined by a central point, a centroid. Similarity of data in one cluster is measured by using different criteria. Thus, there are lots of different methods which can solve the general task of clustering [1].

Two types of K-means algorithm are analysed in this paper and the obtained results are discussed. The standard K-means algorithm divides a data sample into exclusive clusters. The initial values of clusters' centroids are randomly selected from the available data. Updating centroids and clustering of data is then repeated until convergence is reached or for a defined number of iterations. A new centroid for a cluster is calculated based on each data sample that belongs to that cluster. The first issue of application of K-means-type algorithms is that the number of clusters should be known in advance. Thus, before discovering knowledge from big data, we have to know how many cluster we expect in a database. The second issue is that this kind of algorithms is very sensitive to the initial clusters' centroids. Usually, initial centroids are chosen randomly.

In fuzzy K-means clustering data samples belong to every estimated cluster, with a certain belonging degree. Hence, a result of this algorithm are not exclusive clusters, but clusters with fuzzy borders. A fuzzifier is a parameter which defines fuzziness of fuzzy K-means clusters. This method is used for clustering data in cases where clusters are not clearly defined and one cannot estimate clear borders among data samples.

In this paper, standard and fuzzy K-means clustering of soil data is implemented for a database and results are graphically presented. The results of fuzzy K-means clustering for different values of fuzzifier are tested, and a procedure for the selection of fuzzifier for a database is proposed. In addition to simple graphical representation, maps are a proper way of presenting the results of clustering of soil. Thus, the results of K-means clustering of soil are present on the static Google map and dynamic Open Street Map.

Paper received May 24, 2016; accepted June 12, 2016. Date of publication July 20, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Miroslav Lutovac.

This paper is a revised and expanded version of the paper presented at the 23rd Telecommunications Forum TELFOR 2015 [21].

This research is supported in part by FP7 Project Fore Mont and Project for establishment of pilot Montenegrin Centre of Excellence in Bio-informatics - BIO-ICT (Contract No. 01-1001).

Elma Hot, Faculty of Electrical Engineering Podgorica, University of Montenegro, Džordža Vašingtona bb, 81000 Podgorica (Phone: +382 20 245 839, Fax: +382 20 245 873, e-mail: elma_hot@live.com).

Vesna Popović-Bugarin, Faculty of Electrical Engineering Podgorica, University of Montenegro, Džordža Vašingtona bb, 81000 Podgorica (Phone:+382 20 245 839, Fax: +382 20 245 873, e-mail: pvesna@ac.me).

The paper is organized as follows. In Section II K-means clustering is presented; in Section III fuzzy K-means algorithm is reviewed. The performances of analysed algorithms in the case of clustering data samples with chemical characteristic of soil are analysed in Section IV through simulation results. A conclusion is drawn in Section V.

II. K-Means clustering

K-means (KM) clustering is a widely used partitioning method. This method aims to make K mutually exclusive clusters of *n* data samples characterised by *d* parameters. Each cluster *K* is defined with one central point (centroid) determined by a certain combination of parameters contained in each data sample.

KM is known as a method of vector quantization, since it is based on the location of points and their mutual distances [2]. Namely, data samples described by *d* parameters can be presented as points in a d-dimensional space, where their coordinates are determined by the values of *d* parameters. A data sample belongs to a cluster defined with a centroid which is the closest one to the considered sample (point). The closest centroid is chosen after calculating the distance of each data from each centroid. Each data sample can belong to exactly one cluster. Hence, KM clustering is also called hard clustering.

A. K-means clustering algorithm

The KM algorithm aims to distribute a set *X* of *n* data samples into *K* clusters. Each data sample is defined by *d* parameters. We consider data samples as points in a d-dimensional space for better visualization. Input to the algorithm is the number of clusters *K*. The initial values of centroids $c_1^1, c_2^1, ..., c_K^1$, $c_i \in R^d$, are chosen randomly from the available data samples. After the calculation of the distance of each data sample from set X to each clusters' centroid, each data sample is declared to be a member of its closest cluster. A set of data samples that belong to a cluster defined by centroid c_i is denoted as c_i, $1 \le i \le K$. In each iteration, a centroid is estimated as a mean value of *d* corresponding parameters of all data samples which are a member of a corresponding cluster. Calculating *K* new centroids in each iteration is equivalent to changing a clusters' position in a d-dimensional space, till optimal clusters positions are reached. The processes of clustering and updating centroids are repeated until convergence has been reached or for a specified number of iterations.

One drawback of KM clustering appears when a point is equally close to more than one centroid. In this case, the algorithm will not converge, since this point belonging will oscillate among a few different clusters, resulting in different clusterings. However, this rarely happens in practice [3], [4].

III. Fuzzy K-Means clustering

Fuzzy K-Means (FKM) clustering method is a modification of the standard KM clustering. As in KM clustering, initial centroids are chosen randomly. Each iteration in FKM clustering also starts with calculating the distance of each data sample to each centroid. However, in FKM there is a belonging degree, which is inversely proportional to that distance. A data sample belongs to every cluster with a certain degree [5]. Hence, borders among clusters are fuzzy. FKM clustering is also referred to as soft clustering.

In FKM clustering, all data samples affect the calculation of new centroids. The impact of a data sample on the calculation of clusters' centroids is proportional to the degree of its belonging to that cluster. The other part of the FKM algorithm is the same as in the KM algorithm.

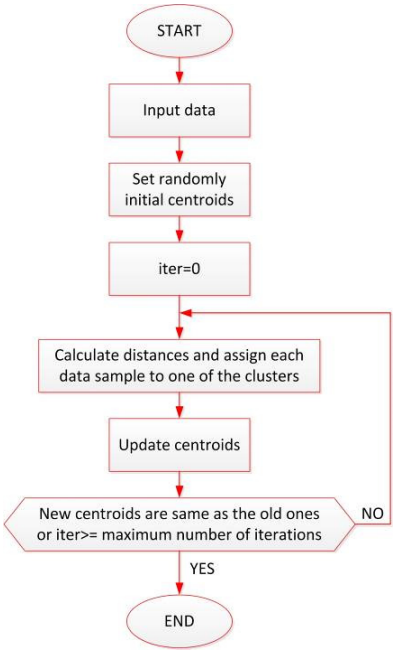

Fig. 1. K-means clustering algorithm.

A. Fuzzy K-means clustering algorithm

Input to the FKM algorithm is the number of clusters *K*. For *n* data samples the algorithm gives as a result an *n*×*K* matrix *W*, with elements:

$$w_m(i,j) = \frac{1}{\sum_{k=1}^{n}\left(\left|\frac{x_i - c_i}{x_{i-c_k}}\right|\right)^{\frac{2}{m-1}}}. \quad (1)$$

$w_{i,j}$ is the degree to which element x_i belongs to cluster c_j, $0 \le w_{i,j} \le 1$. $m \in R$ is the fuzzifier, it defines the level of cluster fuzziness, and $m \ge 1$. In the absence of a priori knowledge of the data fuzziness, it is recommended to set the fuzzifier according to the database and expected results of clustering. By adjusting the fuzzifier *m*, border between clusters can be more fuzzy or more clear. The procedure for its selection is illustrated in the next section through the simulation results of FKM algorithm for different values of *m*.

New centroids for *K* cluster are calculated on the basis of all the data samples:

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}. \quad (2)$$

The procedure repeats until reaching convergence or for a specified maximum number of iterations. Both algorithms

converge when $\left|c_k^t - c_k^{t+1}\right| < \delta\left|c_k^t\right|$, where t is the number of iteration and δ is a sensitivity threshold. In all our experiments, this threshold is 0.01.

FKM clustering is computationally more complex than KM clustering. KM calculates a distance and chooses the smallest one in order to find to which cluster a data belongs, while FKM performs additional Kxd multiplications for each data sample (1), (2). However, the FKM algorithm has a better result for a data set with overlapped clusters than the KM algorithm. Moreover, in a case where data samples are equally close to more than one centroid, the FKM algorithm will not oscillate, unlike KM algorithm. FKM algorithm will give an equal belonging degree of these data samples to more than one cluster.

IV. Simulation Results

Algorithms for KM and FKM are implemented in Java and adapted for a soil database. A database of soil samples sampled in Montenegro is used for clustering; 2526 data samples are used. The goal is to estimate to which soil type a sample belongs, using KM and FKM methods of clustering. Based on the knowledge of the types of soil in Montenegro, the number of clusters is chosen. The number of clusters is essential to proper clustering. Thus, it is important to have a correct value of this parameter. Since this parameter is not always known in advance, the cases of clustering with a smaller or higher number of clusters is also investigated. The conclusion is that, the final centroids are quite similar to initial centroids in case of using a number of clusters smaller than a correct one. On the other hand, in the case of using a larger number of clusters than a correct one, clusters with a small number, zero or nearly zero, of elements appears. This was an additional confirmation of defined number of soil types in Montenegro. Moreover, these conclusions can be used as a guide for determination of a correct number of clusters in the case when this information is not known in advance.

In Montenegro there are four to six types of soil. In our simulations, data samples with only six parameters are used for clustering. The used parameters represent the numerical values of chemical characteristics of soil samples. Consequently, this clustering of soil can be considered as basic one. Hence, the optimal number of clusters for this basic clustering is four.

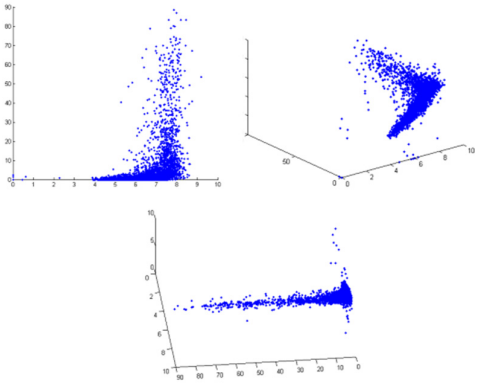

Fig. 2. Visualisation of soil database, each point is a data sample with three parameters.

For the verification and visualization of the performance of the algorithm, clustering of soil samples based on two or three parameters in three or four clusters was done and is presented in this section. MATLAB is used for a graphical representation of results.

Data samples with three parameters from soil database are presented in a 3-D space before clustering (Fig. 2.).

Clustering soil samples is done by KM and FKM algorithms. The maximum number of iterations is 100 in both algorithms.

The mean value of the number of iterations needed for achieving the convergence of KM algorithm in the case of clustering in four clusters based on six parameters is 7 in 500 runs. Fig. 3 shows how the parameters of centroid converge, in the case of clustering in four clusters based on six parameters. Convergence is reached after 8 iterations.

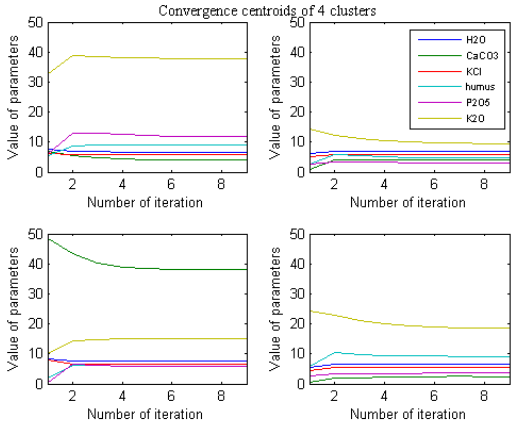

Fig. 3. Convergence of centroids of K-means clusters in the case of clustering in four clusters based on six parameters.

The mean value of the number of iterations for the FKM algorithm is 21 in 500 runs for the case of clustering in four clusters based on six parameters. In the case of clustering in four clusters based on six parameters, presented in Fig. 4, the convergence is reached after 14 iterations.

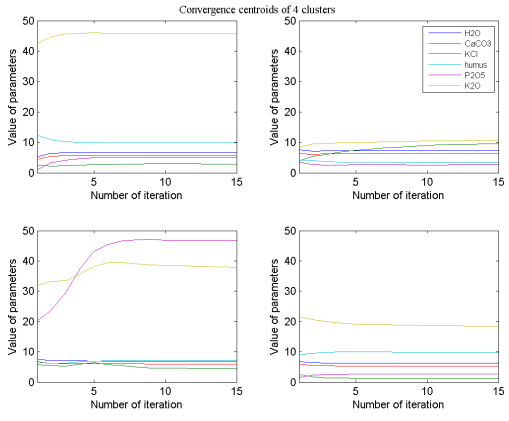

Fig. 4. Convergence of centroids of fuzzy K-means clusters in the case of clustering in four clusters based on six parameters.

A. *Results of K-means Clustering*

Results of K-means clustering based on two parameters for three and four clusters are presented graphically in Fig. 5.a) and b), respectively. Each data sample belongs to only one cluster. Hence, each colour of data sample presents one cluster.

Clustering based on three parameters in four clusters of the same data is presented in Fig. 6.

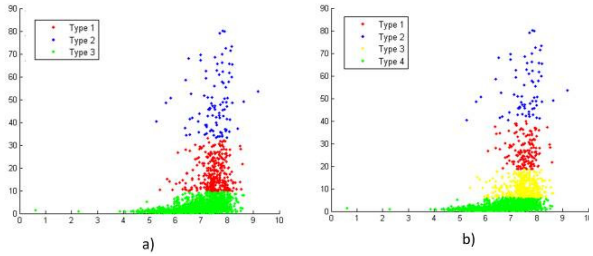

Fig. 5. K-means clustering based on two parameters. Each colour presents one cluster a) K=3; b) K=4.

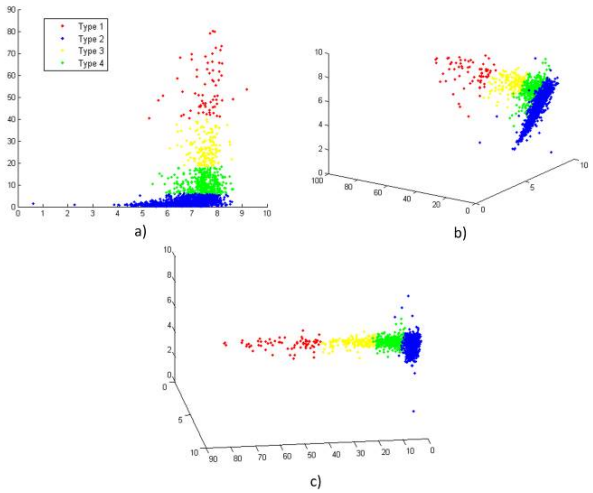

Fig. 6. Clustering in four clusters (types) based on three parameters. Each colour presents one cluster. a), b), c) show different projections in three-dimensional space;

B. *Results of Fuzzy K-means Clustering*

Validation of FKM clustering results is also done by its visualisation. In FKM clustering each data sample belongs to each cluster, with a different belonging degree. Hence, each FKM cluster was drawn on a different graph, while the colours of samples depend on the belonging degree of each sample to that cluster. This way of visualisation allows a proper way of presenting the propagation of clusters. Dark red presents a belonging degree near to 100%. Yellow and green present data samples witch partly belong to a plotted cluster. Light blue to dark blue colours present belonging degrees whose values are less than 30%.

Parameter m is the fuzzifier and it defines the level of cluster fuzziness [6]. FKM clustering of soil database was done with different values of m. Results of FKM clustering with different fuzzifiers are presented in Fig. 8-9.

Fig. 8.a) presents four FKM clusters of soil database, where fuzzifier m is 1.1. Dark red is a dominant colour around centroids and the borders between clusters are clear. Hence, fuzziness of clusters is missing, so results are similar to KM clusters.

The second example is FKM clustering where fuzzifier is 2 (Fig. 8.b). The appearance of all colours, from dark red to dark blue, means that all belonging degrees from 0 to 100% appear. Different clusters are visible but borders are still fuzzy.

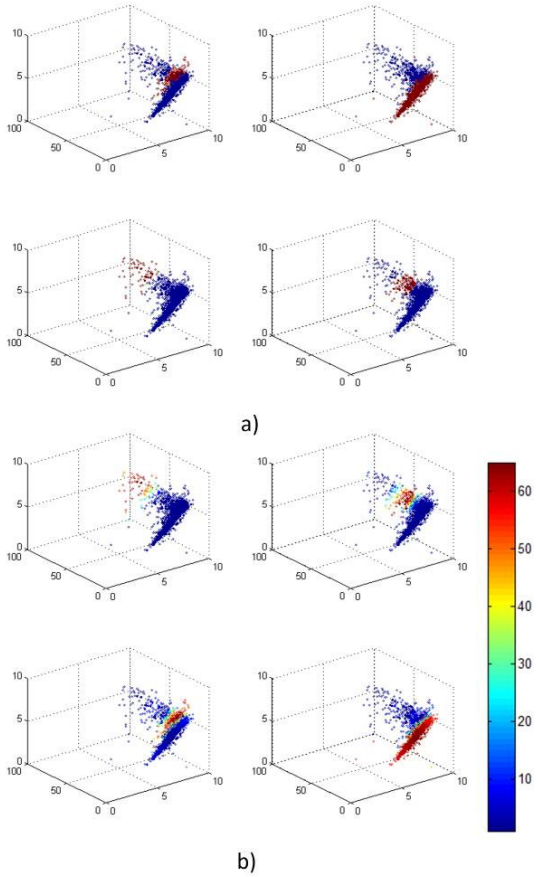

Fig. 7. Fuzzy K-means clustering based on three parameters, each graph presents one cluster, colour depends on belonging degree, K=4 a) m=1.1; b) m=2.

FKM clusters with a fuzzifier 3 are shown in Fig. 9.a) Clusters are visible, but dark red colours are missing. This means that belonging degrees near to 100% do not exist. All belonging degrees in this case are smaller, while a blue colour is dominant. Hence, fuzziness of clusters is significant. A border between clusters is less visible than in the case where the value of m is smaller.

The results of FKM clustering of the same database with a fuzzifier 5 are shown in Fig. 9.b). Clusters are not visible; all belonging degrees are equal. Hence, this clustering is unsuccessful and pointless. The belonging degree of each sample tends to be equal for every cluster when a fuzzifier takes values higher than 3.

A conclusion from these examples is that a fuzzifier value 2 is optimal for this database. Clusters are visible and borders of clusters are fuzzy enough. These examples show the choosing procedure of fuzzifier for a specific database.

Having in mind the previous conclusion, FKM clustering of soil database with a fuzzifier 2 is performed.

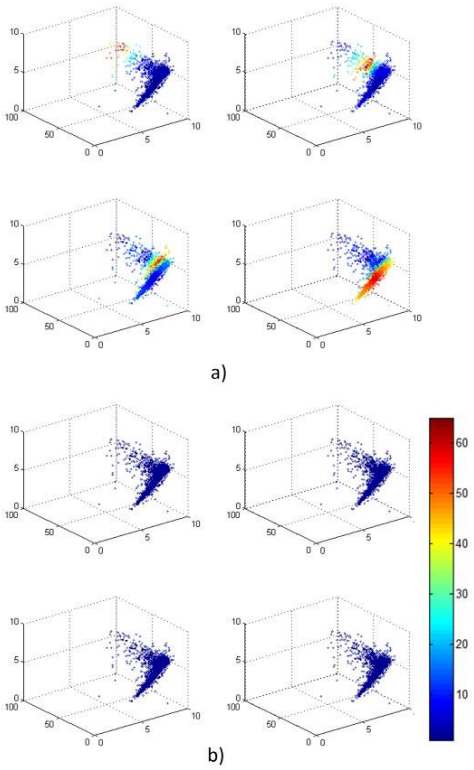

Fig. 8. Fuzzy K-means clustering based on three parameters; each graph presents one cluster; colour depends on belonging degree, K=4 a) *m*=3; b) *m*=5.

The result of the FKM clustering of soil database based on two parameters and three clusters is presented in Fig. 9. Clustering is done based on two parameters, so coordinates present the position of soil samples in 2D space. Belonging degrees are inversely proportional to the distance between samples and centroids in 2D space. Increasing the distance between a sample and centroid reduces the belonging degree and colours change proportionally from red to blue. Based on these graphs, a conclusion is made that the FKM clustering algorithm made a successful clusterization of soil data.

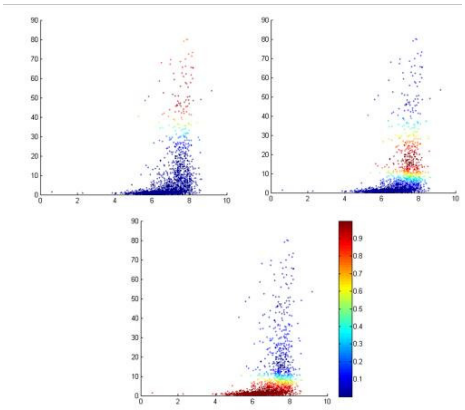

Fig. 9. Fuzzy K-means clustering in three clusters, based on two parameters. Each graph presents one cluster, while colour depends on belonging degree.

Fig. 10 presents the result of FKM clustering in four clusters based on two parameters of the same database. A conclusion that clusterization of soil data is successful can be made, as in the previous case.

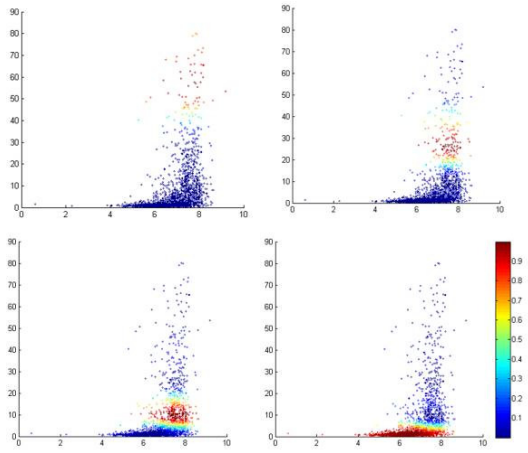

Fig. 10. Fuzzy K-means clustering in four clusters, based on two parameters. Each graph presents one cluster, while colour depends on belonging degree, K=4.

Soil samples in used database besides physical and chemical characteristics, have the coordinates of location. The coordinates of soil samples from database are given in meters in the coordinate system MGI 1901 / Balkans zone 6. First, coordinates are converted to longitude and latitude.

Graphical environment for presenting the results of clustering on the static Google map of Montenegro is implemented in Java (Fig. 11.). Soil samples are labelled with markers and the colour of marker depends on a marked sample's soil type. It allows searching the map by municipalities and soil types, and adjusting the zoom. Different types of maps are available: roadmap, satellite, hybrid and terrain.

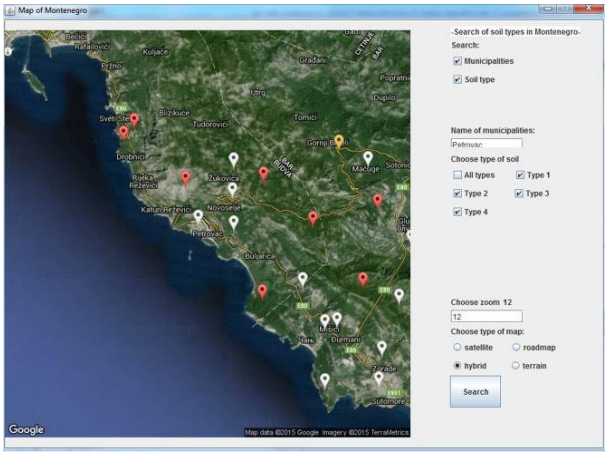

Fig. 11. K-means clustering of soil samples presented on Google static map.

The second method of presenting the results of soil clustering is on dynamic maps. Dynamic maps are implemented using R programing language. In addition, a Leaflet is used. It is a set of open-source JavaScript libraries for interactive maps. An Open Street dynamic map is shown in Fig. 12. Markers present all soil samples from the database with coordinates. The colour of marker presents the results of KM clustering. Each colour is related to a different cluster. Four clusters are depicted.

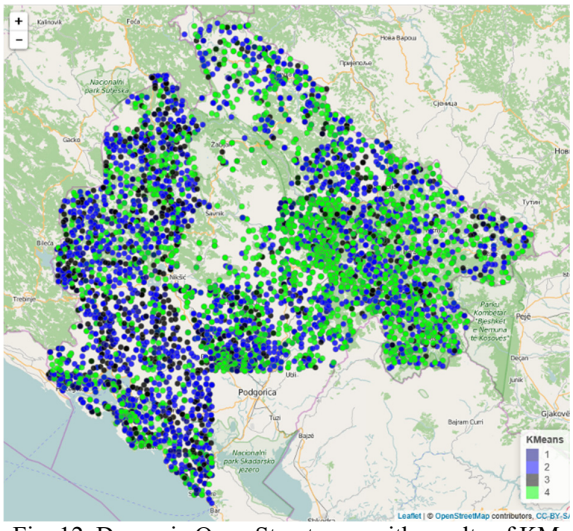

Fig. 12. Dynamic Open Street map with results of KM clustering, markers of different colours presents different KM clusters.

This map can be considered as a basic pedologic map, because clustering is made based on only six chemical parameters and only four clusters were made. Mapping of KM results more precisely is possible by using more soil parameters and clustering in more clusters. On this map in Fig. 12. it is visible that two types of soil are dominant, blue and green colours. Comparing this to pedologic maps of soil of Montenegro made by experts, it can be seen that two dominant soil types are the same (at the same parts of Montenegro) on both maps.

Using maps for presenting soil data and clustering results allows marking data with markers, polygons, raster images, images of soil profiles etc.

V. Conclusion

The problem of soil data clustering and visualization is analysed in the paper. Data mining techniques, KM and FKM, are adapted for this purpose. The visualisation of KM and FKM results is used for the validation of results.

Results obtained by using KM are presented on the Static Google map and dynamic Open Street Map of Montenegro. Presented soil data and data mining result on maps are a proper way of presenting data to scientists, land users and people who want to get information about soil in Montenegro. Our future work will be dedicated to improving data mining techniques and publishing all results through a WEB application.

References

[1] J. C. Bezdek, R. Ehrilich, W Fill, "FCM: The Fuzzy C-means Clustering Algorithm", *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984.

[2] "Vector Quantization and Clustering", Courses of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

[3] J. Macqueen, "Some Methods for Classification and Analysis of Multivariate Observations" *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, page 281-297. University of California Press, (1967)

[4] Andrew Ng," CS229 Lecture notes", Machine Learning Course Materials

[5] S. Ghosh, S. K. Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", *((IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 4, no.4, 2013.

[6] E. Hot, V. Popović-Bugarin, Ana Topalović, Mirko Knežević, "Generating thematic pedologic maps by using data mining and interpolations," submitted for *3nd International Conference on Electrical, Electronic and Computing Engineering IcETRAN 2016,* Zlatibor, Serbia, June 2016

[7] Md. K. I. Rahmani, N. Pal, K. Arora, "Clustering of Image Data Using K-Means and Fuzzy K-Means", *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 7, 2014*

[8] R. Suganya, R. Shanthi, "Fuzzy C- Means Algorithm- A Review", *International Journal of Scientific and Research Publications,* vol. 2, Issue 11, November 2012 1 ISSN 2250-3153

[9] E. Hot, V. Popović-Bugarin, "Analysis Of Fuzzy K-Means Clustering Method Using Database Of Soil Samples Sampled In Montenegro," *Information Technology IT 2016*, Žabljak, Montenegro, February 2016

[10] J. Balkovič, Z. Rampasekova, V. Hutar, J. Sobocka and R. Skalsky, "Digital Soil Mapping from Conventional Field Soil Observations," *Soil & Water Res.*, 8, 2013 (1): 13–25

[11] S. Har-Peled, B. Sadri, "How Fast is the k-means Method?*", January 2, 2010

[12] Singaravelu.S, A.Sherin and S.Savitha "Agglomerative Fuzzy K-Means Clustering Algorithm", *Nehru E-Journal A Journal of Nehru Arts and Science College (NASC) Research Article*

[13] S. Chattopadhyay, D. Kumar Pratihar, S. C. De Sarkar, "A Comparative Study of Fuzzy C-Means Algorithm and Entropy-Based Fuzzy Clustering Algorithms", *Computing and Informatics*, Vol. 30, 2011, 701–720

[14] L. G. Vendrusculo, A. L. Kaleita, "Terrain Analysis and Data Mining Techniques Applied to Location of Classic Gully in a Watershed", 2013 ASABE Annual International Meeting

[15] L. Rokach, O. Maimon, "Clustering Methods", In *The Data Mining and Knowledge Discovery Handbook*, pages 321–352. 2005.

[16] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics* 3: pp. 32-57*,* 1973

[17] D. Rajesh, "Application of Spatial Data Mining for Agriculture," *International Journal of Computer Applications*, (0975 – 8887) Volume 15– No.2, February 2011

[18] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine,* Vol 17, No 3, 1996

[19] B. Fustic, G. Djuretic, "The Soils of Montenegro," University of Montenegro and Biotechnical Institute, Podgorica, Montenegro, 2000.

[20] S. Ghosh, S. K. Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 4, no.4, 2013.

[21] E. Hot and V. Popović-Bugarin, "Soil data clustering by using K-means and fuzzy K-means algorithm," *Telecommunications Forum Telfor (TELFOR), 2015 23rd*, Belgrade, 2015, pp. 890-893.