COMPARISON OF TRADITIONAL AND FUZZY UNSUPERVISED CLASSIFICATION ON THE BASIS OF VEGETATION INDEX

ISTVÁN HAVASI [1], *DÁVID BENŐ* [2]

Introduction

Land management as a complex planning task requires the knowledge of more sectors such as agriculture, industry, environmental protection and urban management. Joint modelling of these is a very difficult task, especially the knowledge of the phenomena occurring in nature, since several of them occur at the same time, and certain model parameters are not known at all. Natural sciences took more methods from classical mathematical statistics; however, not each of them is capable of describing the continuality and uncertainty in nature. For this reason there was great demand on the development of various procedures – known as "soft computation methods" today – such as polyvalent logic, which was already used by ancient Greeks as well. In the second half of the twentieth century a quantum leap of artificial intelligence and computer technique gave a great impulse to the development of soft computation methods, and thereby artificial neural network, genetic algorithm and fuzzy logic appeared.

The purpose of our present article is to demonstrate the application of fuzzy logic to unsupervised classification, moreover to outline a concrete scientific investigation in which conventional k-mean classification is compared to a Fuzzy-C-Mean (FCM) procedure by means of a vegetation index computed from a Landsat 7 image. In connection with this we studied how the size of each class would be altered using both methods and what difference is induced by varying the number of clusters for each procedure in question.

1. About fuzzy logic briefly

Since 1950 expert systems have been being developed which deduce on the basis of data and a knowledge base using Boolean algebra. This traditional binary logic works by means of two values: true and false. In natural sciences we often meet such kind of phenomena which can be defined badly; their operation and modelling by exact methods cannot be solved at all. Many authors made an attempt at the development of polyvalent logic; the theory of continuum of infinite value set was worked out by [1] with reference to the literature [2].

The Hungarian meaning of fuzzy is "*életlen*", as a consequence of that, in these systems the belonging to a set is described by membership functions (Gauss, triangle, trapeze and sigmoid). These functions represent each linguistic variable, e.g. weed coverage as a linguistic one in a given region could be less weedy, average weedy and highly weedy. According to the afore-mentioned example the belonging to a given set, the quantity of weeds on a unit area ("What degree is the weed coverage?") is given by a function. This operation is called fuzzification. In the next step, formation of a rule system is realized among each linguistic variable, that is to say we perform logical operations with them, and draw conclusions (and, or, negation). As a result of this an aggregate consisting of member functions is reached which is the basis of defuzzification. Having finished the

[1]: associate professor, department head gbmhi@uni-miskolc.hu,
[2]: research fellow, gbmbd@uni-miskolc.hu
Department of Geodesy and Mine Surveying University of Miskolc, H-3515 Miskolc-Egyetemváros

defuzzification we obtain a concrete value which can be considered as a final result of fuzzy analysis.

2. Fuzzy analysis in land management

In land management fuzzy set theory is essentially used for classification. Its purpose is to reduce a complex system to well-separated classes when data must be grouped [3].

There are two different but complementary approaches to forming fuzzy sets, membership functions: automatic clustering procedures such as ***FCM*** (***F***uzzy ***C***-***M***ean); and ***SI*** (***S***emantic ***I***mport) model which is based on expert knowledge [3].

2.1. Clustering procedures. Among clustering procedures FCM is one of the most widely used methods, and when it is applied the following objective function is optimized:

$$J(\boldsymbol{X},\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{i,k})^{m}\left\|\underline{x_k}-\underline{v_i}\right\|_A^2 .$$

where $\boldsymbol{U} = [\mu_{i,k}]$, the membership rate of $\underline{x_k}$ to cluster i;

$\left\|\underline{x_k}-\underline{v_i}\right\|_A^2$ is an applied range norm where „A” is the range;

V is a matrix including the central points of clusters, the positions of which must be computed; and **m** is a weight index [4].

The above procedure is an unsupervised one, here the number of classes is not known at the beginning of the analysis.

An example of the above method is the research of Illés [5] in which he investigated site conditions of Észak-Hanság. His purpose was to create such a soil model or classification which can give the experts (forestry, environmental) working on the site reliable information on soil conditions of the field in question. In the first step the author developed the soil physical, soil chemical and relief database, and he executed the necessary analysis using these data. He started the establishment of membership functions with clustering, and matched a distribution function to the obtained result to produce the so-called membership truth value of each element in the fuzzy sets. With the help of this classification he investigated the relationship between soil features and environmental variables by setting up correlation equations. The maps as final results demonstrated the site conditions well inside the interpretation range, however, beyond that it was not reliable [5].

2.2. The SI model. A ***S***emantic ***I***mport model is such an empirical or expert one which specifies the membership function on the basis of the available knowledge and experience obtained from investigations. Chang and Burrough were the first to use fuzzy sets and logic based on them in soil assessment. In their scientific work they applied the SI approach, and published several kinds of membership functions which can be used well in soil science [6].

A Hungarian-related example is the work of Vid Honfi [7] from which useful information can be obtained for the suitability of land use on the basis of ***G***old ***C***rown (***GC***) value and slope conditions. When membership functions were established the author considered the recommendations of National Agrarian-environmental Protection Program and Balaton Law, moreover in connection with ***G***old ***C***rown value the experiences of agrarian experts farming on the site.

Among the membership functions produced by the above-discussed methods special fuzzy operations can be performed or rule systems can be formed. In the above-mentioned SI example Honfi drew the conclusion for land use with the help of written rules in a form "if…then…".

For example:

If ***GC*** value = ***good*** and slope = ***plane***, then suitability = ***arable land***,
If **GC** value = ***average*** and slope = ***sloping***, then suitability = ***pasture***, [7].

When fuzzy logic is applied – in contrast to unsupervised classification – the outputs (the deductions) are predetermined fuzzy sets which can also be described by membership functions.

In fuzzy logic, similarly to Boolean algebra, logical operators can be described by mathematical operations, however, while traditional logic works with discrete values (0 and 1), now fuzzy sets use numbers between 0 and 1. As a consequence of this the meaning of operators is also different (Table I).

Table I. The meaning of logical operators

Operator	**Boole**	**Fuzzy**
and	multiplication	minimum
or	addition	maximum

When a fuzzy deduction is done a function aggregation is performed from which a defuzzification gives the final value of an analysis.

3. Application of fuzzy unsupervised classification to NDVI data

3.1. Background and computation of NDVI. The ***NDVI,*** as a vegetation index, expresses the photosynthesis product of vegetation, that is to say, it is in connection with the quantity of the produced chlorophyll. On the basis of studying the spectral reflection curves it can be stated that vegetation reflects the beams of visible light to a smaller degree, however, reflection becomes stronger in proportion to its state of development and chlorophyll content in the near infrared range. That is the reason why we can express the development degree of vegetation if we demonstrate the difference between the measured data in the visible and near infrared range. The larger this value is, the more developed the vegetation is. In practice, data of red band of visible range and near infrared band are applied. Practice also proved that it is better to use a normalized difference instead of a single one, since it eliminates divergences caused by various illumination, sloping and aspect. Therefore the normalized vegetation index is applied (***NDVI*** – ***N***ormalized ***D***ifference ***V***egetation ***I***ndex). If the values perceived in the near infrared range are denoted by **NIR** and those derived from the red range by ***RED,*** then the formulae of ***NDVI*** is:

$$\textbf{NDVI = (NIR-RED)/(NIR+RED)}.$$

In case of a Landsat 7 satellite image, where band 3 is red and band 4 is near infrared, the formulae of ***NDVI*** will be the following:

$$\textbf{NDVI = (4-3)/(4+3)}.$$

The values computed in this way vary between -1 and +1. Water surfaces, clouds and snow represent negative numbers; bare soil, rock and artificial surfaces are about 0; while vegetation demonstrates positive numbers.

3.2. Classification using concrete data sets. In the first phase of our investigation we obtained a Landsat 7 satellite image connected to the territory of Hortobágy from the server of ftp://ftp.glcf.umd.edu/glcf/Landsat/ (Figure 1). In the above ftp server each band of a Landsat satellite image is saved in separated tiff files in WGS84 system. We produced the NDVI layer from red and infrared bands of this Landsat image with ArcGIS 9.2 software (Figure 2), and then it was imported in MATLAB programme in which we transformed the alphanumerical data into a one-dimensional matrix, that is to say into a vector.

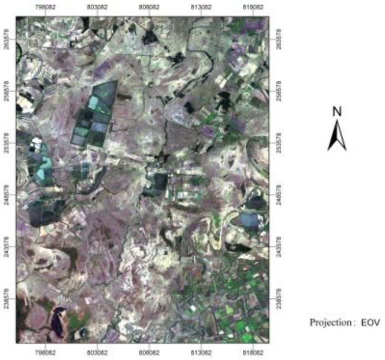

Figure 1. The test area

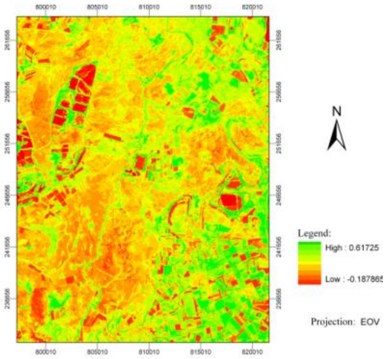

Figure 2. The NDVI layer

We applied ***F***uzzy ***C-M***ean clustering (***FCM***) for fuzzy unsupervised classification, while k-mean procedure is used for traditional unsupervised classification (it is included as a built-in function in the MATLAB programme). The vector produced from ***NDVI*** data and numbers of classes (chosen from ***5*** to ***8***) were given by us as input. After running both methods we obtained a vector each, from which later the GIS layers can be produced. Having run the k-mean algorithm we obtained a vector of classified data automatically. In case of FCM algorithm we had to complete a script, using the programme language of this software, which on the basis of membership functions describes the classified data base in a vector form. (***This operation is the defuzzification itself.***)

The next step was to produce the usable layer of classified data in various GIS software. To solve it the obtained data set was transformed into a matrix according to the raw and column arrangement matching to the original raster file. To apply this matrix in a GIS system it was converted into ASCII format giving the number of cells, cell size and coordinates of the left lower pixel. The name of a class is a number which increases with NDVI value, that is to say **1** denotes the smallest vegetation index value; **2** is the next one and so on.

3.3 Interpretation of data derived from classification. Having executed the classifications of two kinds we compared the obtained results with the original data set using several cluster numbers when the evaluation was performed. Figure 3 illustrates the results of both tested methods (***FCM*** and ***k-mean***).

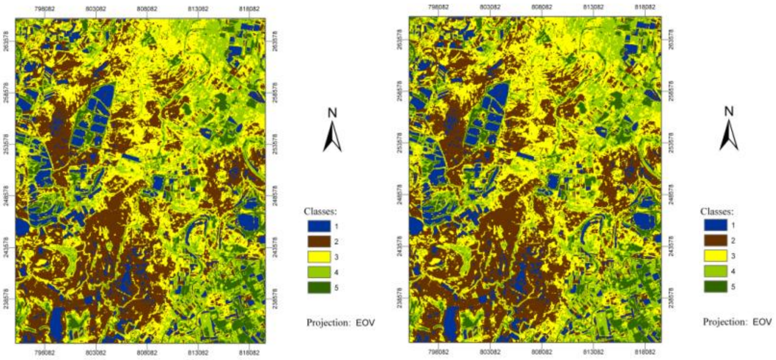

Figure 3. Classification result, cluster number 5: FCM (left), k-mean (right)

To follow our assessment we produced so-called class descriptive files (signature files) from the ASCII ones imported to ArcGIS which contain the averages and covariance values class by class. Later, using these data we formed dendrograms which describe the relations among classes and statistical divergences.

It turned out from the dendrograms that in case of the same cluster number there is no significant difference between statistical measures of the tested two methods. The ***differences of pixel numbers*** are illustrated in Table II. (The total pixel number within one layer is 1,143,954.)

Table II. The obtained differences in pixel numbers for the same cluster number in both tested procedures

Cluster number /piece/		**Differences in pixel numbers**	
5	6	***5982***	13650
7	8	*13541*	25028

The change in pixel number was also evaluated graphically. We stated that image points were evenly located, individually, or maximum three of them form a pixel group.

In the next step we examined what changes are like in the classifications of various cluster numbers. (Because of the similarity of the above-mentioned two methods we did not differentiate between ***FCM*** and ***k-mean*** methods).

It was easily recognizable that the classes 2 and 3 were just as near each other (3.14 – 3.16). The class 1 was also close to these (5.15 – 5.22). In case of cluster numbers 7 and 8 the classes 4 and 5 were very near one another (3.41 and 3.45). In case of classification of 6 cluster number the classes 5 and 6 were close to each other; the class 4, however, was found a bit farther (4.3) from the just mentioned two ones, thus two large groups were formed: 1, 2, 3 and 4, 5, 6. In case of the classification of 8 cluster number the classes 7 and 8 were near one another (2.9). The class 6 was found to be not too far from them (4.08). In this case 3 larger class-groups were formed: 1, 2, 3; 4, 5; and 6, 7, 8.

Conclusions

The above-discussed contributes to the classification of data originating from remote sensing in ***environmental protection*** and ***agriculture***. The uncertainty factor also appears in our investigation, which is an important feature of spatial data. The main purpose of our research was to decide upon the alternative use of the described two methods. Further research to establish the optimal cluster number is planned.

Acknowledgement

"The described work was carried out as part of the ***TÁMOP - 4.2.1.B - 10/2/KONV – 2010-0001 project*** in the framework of the New Hungarian Development Plan. The realization of this project is supported by the European Union, co-financed by the European Social Fund."

References

[1] Zadeh, L.: A Fuzzy sets. Information and Control, 1965, 8(3) pp. 338-353.
[2] Kóczi T. L., Tikk D.: Fuzzy rendszerek. 2001.
http://ottomat.hu/archivum/Fuzzy/fuzzy_rendszerek_.pdf
[3] McBratney, A. B., Odeh, I. O. A.: Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements, and fuzzy decisions. Geoderma 1997, 77, pp. 85-113.
[4] McBratney, A. B., Moore A. H.: Application of fuzzy sets to climatic classification. Agricultural and Forest Meteorology 1985, 35, pp. 165-185. in John Triantifilis (1990) application of continuous methods of classification in lower Namoi valley. http://www.pedometrics.org/paper/john_t.pdf
[5] Illés G., Kovács G., Bidkó A., Heil B.: Modelling the site conditions of Észak-Hanság with the use of fuzzy classification and GIS tools. Acta Agraria Kaposvarinsis, Vol. 7 No 3.
[6] Chang, L., Burough, P. A.: Fuzzy reasoning: a new quantitative aid for land evaluation. Soil Survey and Land Evaluation, 1987, **7**, pp. 69–80.
[7] Honfi V.: Optimalization of land use with the help of a fuzzy-based model. Acta Agraria Kaposvarinsis, 2006, Vol. 10 No 3, pp. 279-287.