A Fuzzy Set Approach to Ecological Knowledge Discovery

Arkadiusz Salski

Institute of Computer Science, University of Kiel
Kiel, Germany
E-mail: asa@email.uni-kiel.de

Abstract*— Besides the problem of searching for effective methods for extracting knowledge from large databases (KDD) there are some additional problems with handling ecological data, namely heterogeneity and uncertainty of these data. A fuzzy set approach can be used to handle these problems at some stages of the knowledge discovery process. Ecological data can be defined as fuzzy sets without sharp boundaries, which reflect better the continuous nature of ecological parameters. The paper focuses on one of the important methods of data reduction, namely clustering, and on the data transformation and construction of a combining operator. Two support systems developed at the University of Kiel and their applications are presented, namely the Fuzzy Clustering System ECOFUCS and the Fuzzy Evaluation and Kriging System FUZZEKS.*

Keywords— ecological data, fuzzy clustering, fuzzy data transformation, uncertainty of ecological data.

1 Some properties of ecological data

Dealing with a very large data basis (long time series, spatial data with high resolution, etc.) is a typical problem in data mining and knowledge discovery, but there are some additional problems with handling the environmental data. The first one is the heterogeneity of ecological data sources (e.g. sources of quantitative and qualitative information). The next problem in ecological data mining is the large inherent uncertainty of these data that results not only from the presence of random variables but also from the difficult comparability of these data, approximate estimations, and imprecision and the subjectivity of the information obtained from an expert [12, 13]. This paper deals with processing both “subjective” data derived from experts and "objective" measurement data. Statistical or stochastic aspects of the uncertainty problem are not taken into account.

Some requirements for methods of searching for ecological knowledge arise from the properties of ecological data mentioned above. Special methods of data analysis should be used to handle the uncertainty and heterogeneity of these data. The fuzzy approach, as a possible way to handle uncertainty, is particularly useful for processing imprecise or uncertain data. Ecological data can be defined as fuzzy sets or fuzzy clusters without sharp boundaries. That reflects better the continuous nature of environmental parameters.

Knowledge discovery is a very complex process, which includes data cleaning, data integration, data reduction, data transformation, data mining, pattern evaluation and knowledge presentation [9]. Collaboration with an expert in the data domain can be very useful at some stages of this process. The fuzzy approach enables us to integrate the expert knowledge in the knowledge discovery process. The paper focuses on the applications of fuzzy sets and fuzzy operators at some stages of this process, namely data reduction (section 2) and constructing a combining attribute (section 3).

2 Data reduction: a fuzzy clustering approach

Clustering belongs to the most popular methods of numerosity reduction of data sets, i.e. “replacing” the data set by smaller representations such as clusters in order to reduce the size of the data set. The clustering methods are based on the principle of maximizing the intraclass similarity of objects and minimizing the interclass similarity of these objects, i.e. objects within a cluster have a high similarity, but are very dissimilar to objects in other clusters. Conventional clustering methods based on Boolean logic ignore the continuous nature of ecological parameters and the uncertainty of data. That can result in misclassification.

Fuzzy clustering methods provide additional information, namely the distribution of the membership values which can be interpreted as a similarity measure. The common fuzzy clustering methods, like the fuzzy c-means method, work only with crisp data, that means they provide the fuzzy partition only for crisp data (e.g. exact measurement data). In ecology we have often to deal with data with a semblance of accuracy. In such cases it may only be possible to obtain estimates of data scatter which can be treated in the context of fuzzy sets and used for defining fuzzy data in the form of fuzzy vectors in a high dimension, the so-called conical fuzzy vectors [5]. They are defined by the apex and the so-called panderance matrix which describes the accuracy of the data. This matrix contains spreads of data in each dimension on its diagonal. Yang [15] defined the distance between two conical fuzzy vectors, $\tilde{A}$ and $\tilde{B}$ in (1), as follows:

$$d^2(\tilde{A},\tilde{B}) = \|\bar{a}-\bar{b}\|^2 + tr\left((A-B)^T(A-B)\right) \qquad (1)$$

where:

A and B are the so-called panderance matrices of $\tilde{A}$ and $\tilde{B}$,
$\|\bar{a}-\bar{b}\|$ is the distance norm (metric) between the apexes $\bar{a}$ and $\bar{b}$, and the trace $tr\left((A-B)^T(A-B)\right)$ is the diagonal sum of $\left((A-B)^T(A-B)\right)$.

Yang proved that $d^2(\tilde{A},\tilde{B})$ defined above is a complete metric, which is an assumption for the convergence of the fuzzy c-means clustering procedure by Bezdek [3]. That means, we can define the well known objective function of the fuzzy c-means procedure for conical fuzzy vectors by

$$F\left(c\right) = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(\mu_{ij}\right)^{m} d_c^2\left(\tilde{A}_i, \tilde{B}_j\right) \qquad (2)$$

where:

$\tilde{A}_i$ is the *i*th object and $\tilde{B}_i$ is the *j*th cluster, both defined as conical fuzzy vectors,
n is the number of objects, and
c is the number of clusters.

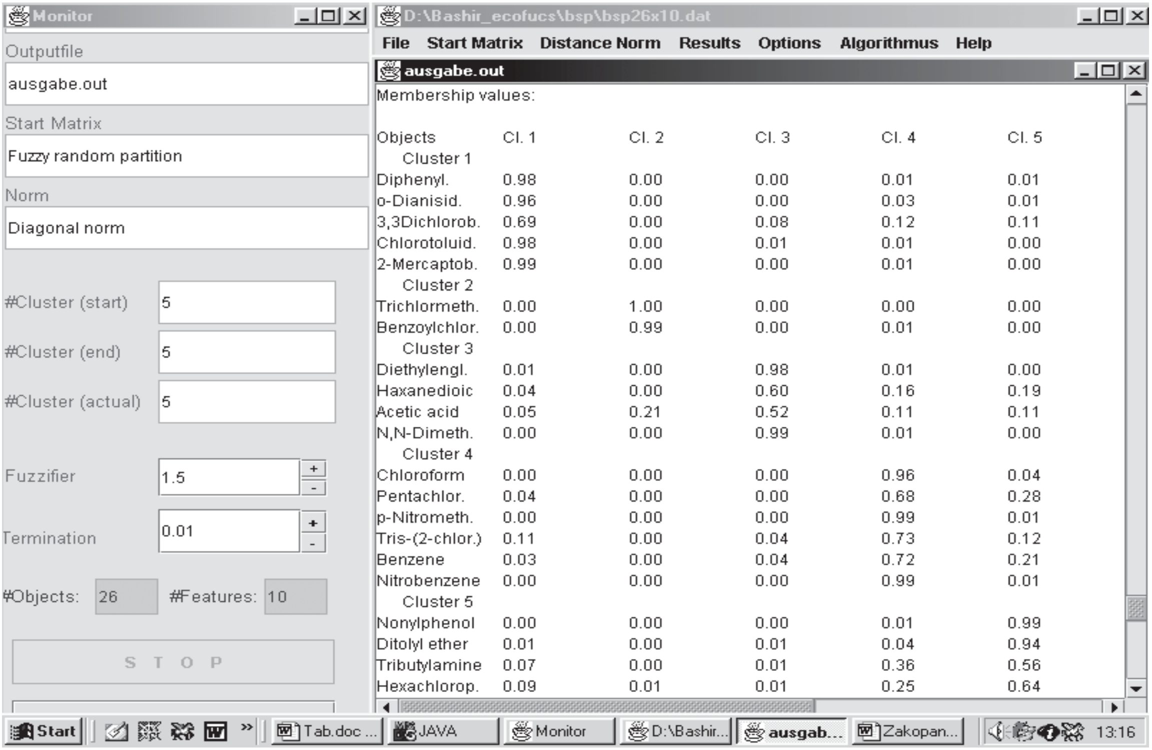
The clustering algorithm for conical vectors proposed by Yang has been extended for the diagonal norm using the so-called z-transformation of the Euclidian norm and implemented for the Fuzzy Clustering System EcoFucs v.5.1 developed at the University of Kiel [8]. The diagonal norm is a highly recommendable distance measure in the case of heterogeneous ecological data with different domain scales. In such cases we can transform data in a uniform manner before we start the fuzzy c-means procedure for conical vectors:

$$z - trans\left(\tilde{A}_i\right) = \frac{\tilde{A}_i - \tilde{V}}{\tilde{S}} \qquad (3)$$

where:

$\tilde{V} = \frac{1}{n} \sum_{j=1}^{n} \tilde{A}_j$ is the mean vector of all fuzzy conical vectors of the input data set, and
$\tilde{S}$ is the vector of spreads from the panderance matrix.

To obtain back the coordinates of the cluster centers in the real scale we have to apply the inverse transformation of the results of the fuzzy c-means procedure. EcoFucs works also with crisp data and offers four different distance norms as a measure of similarity between an object and a respective cluster (the Euclidean-, the Diagonal-, the Mahalonobis- and L1-norms) and a set of methods for calculating the start partition (WARD, conventional c-means, maximum-distance-algorithm, sharp or fuzzy random partitions). The choice of the distance norm depends on the data set. The partition efficiency indicators (entropy, partition coefficient, payoff and non-fuzziness index) available in EcoFucs can be very helpful in searching for the optimal partition and finding the objects which can serve as the representatives of each cluster (see the example below) [10].



Outputfile: ausgabe.out
Start Matrix: Fuzzy random partition
Norm: Diagonal norm
#Cluster (start): 5
#Cluster (end): 5
#Cluster (actual): 5
Fuzzifier: 1.5
Termination: 0.01
#Objects: 26
#Features: 10

```
Membership values:

Objects        Cl. 1       Cl. 2       Cl. 3       Cl. 4       Cl. 5
    Cluster 1
Diphenyl.      0.98        0.00        0.00        0.01        0.01
o-Dianisid.    0.96        0.00        0.00        0.03        0.01
3,3Dichlorob.  0.69        0.00        0.08        0.12        0.11
Chlorotoluid.  0.98        0.00        0.01        0.01        0.00
2-Mercaptob.   0.99        0.00        0.00        0.01        0.00
    Cluster 2
Trichlormeth.  0.00        1.00        0.00        0.00        0.00
Benzoylchlor.  0.00        0.99        0.00        0.01        0.00
    Cluster 3
Diethylengl.   0.01        0.00        0.98        0.01        0.00
Haxanedioic    0.04        0.00        0.60        0.16        0.19
Acetic acid    0.05        0.21        0.52        0.11        0.11
N,N-Dimeth.    0.00        0.00        0.99        0.01        0.00
    Cluster 4
Chloroform     0.00        0.00        0.00        0.96        0.04
Pentachlor.    0.04        0.00        0.00        0.68        0.28
p-Nitrometh.   0.00        0.00        0.00        0.99        0.01
Tris-(2-chlor.) 0.11       0.00        0.04        0.73        0.12
Benzene        0.03        0.00        0.04        0.72        0.21
Nitrobenzene   0.00        0.00        0.00        0.99        0.01
    Cluster 5
Nonylphenol    0.00        0.00        0.00        0.01        0.99
Ditolyl ether  0.01        0.00        0.01        0.04        0.94
Tributylamine  0.07        0.00        0.01        0.36        0.56
Hexachlorop.   0.09        0.01        0.01        0.25        0.64
```

Figure 1: The main window of EcoFucs [8] with a small part of the results of the clustering of ecotoxicological data into 5 clusters.

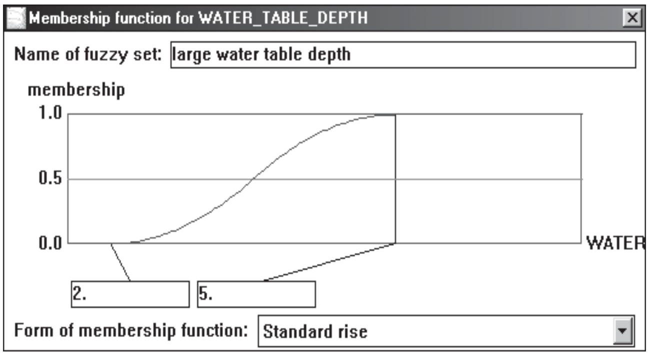
The fuzzy clustering of chemicals according to their ecotoxicological properties [7] can be mentioned here as an example of data reduction. Both the large number of existing chemicals and the costs of ecotoxicological testing procedures make it necessary to select representative further data mining and other steps of the knowledge discovery process instead of all chemicals which belong to the cluster. Compared to conventional clustering methods a chemicals which faithfully reflect the relevant properties of possibly a major group of compounds. So the main tasks of this application were to find distinguishable clusters with characteristic properties and to find chemicals representative for each cluster. These representatives can be used for fuzzy clustering technique is more appropriate to handle the uncertainty of ecotoxicological data. The degree of uncertainty of these data is very high and can arise, for example, from a mixture of quantitative and qualitative data or from the difficult comparability of these data because of different measurement or test conditions (e.g. test results for different animals). The distribution of the membership values provides information from which the degree of similarity between the properties of a particular chemical and the properties characterising particular clusters can be deduced. This is particularly important since there are many chemicals with more or less overlapping properties. That would not be recognised by conventional clustering methods.

The analysis of the membership values helps us to find the representatives of each cluster. Their membership values are close (or equal) to 1. Figure 1 presents a small example of the results of the clustering of chemicals into 5 clusters for 10 ecotoxicological features (toxicity indicators and the potentials for biodegradability and hydrolysis). Trichlormethylbenzene and 2-Mercaptobenzothiazole, for example, can be taken as the representatives of clusters 2 and 1, respectively. We can also see some chemicals (e.g. Tributylamine and Hexachloropentadiene in cluster 5) with membership values strongly divided between different clusters. These values can be interpreted as the degree of similarity to the respective clusters.

3 Data transformation and the construction of a combining operator

One of the important stages of the knowledge discovery process is the data transformation needed to combine the data. In the case of heterogeneous data we have to normalise these data in a uniform manner before we construct the combining operator [4]. In order to do this we can scale data so that they fall within the same range. We can use the membership functions of fuzzy sets to transform the data into the interval 0.0 to 1.0. The definition of the membership functions of these fuzzy sets should express the evaluation criterion formulated by an expert.

We can consider an analysis of hydrogeological spatial data as an example. Figure 2 presents the membership function of the feature "large water table depth" used for the data normalization in the analysis of the suitability of a specified land unit as a waste disposal site [2, 14]. The shape of the defined membership function corresponds to the evaluation criterion. The values of the water table depth lower than 2 m are not suitable (the membership values equal 0); the values bigger than 5 m are very suitable for a waste disposal site (the membership values equal 1).



Name of fuzzy set: large water table depth

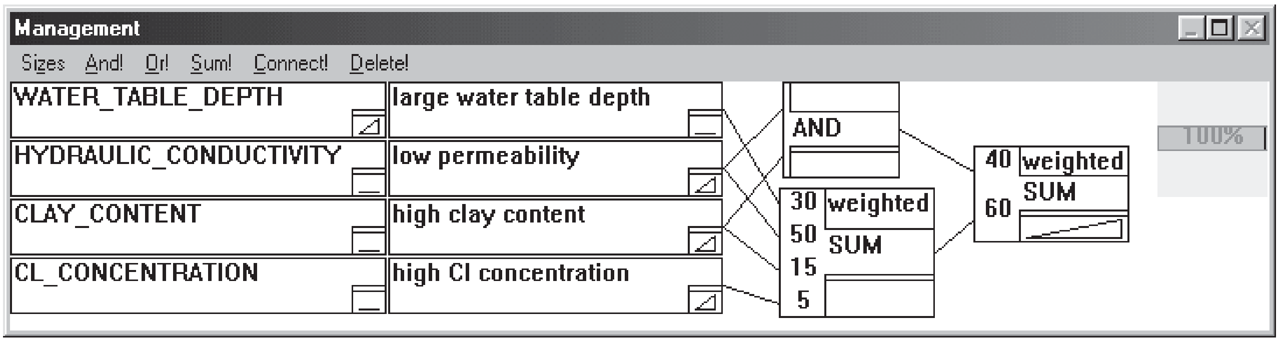
2.

5.

Form of membership function: Standard rise

Figure 2: The membership function of the fuzzy set "large water table depth" used for the data transformation (the membership function window of FUZZEKS [2]).

Four land characteristics, namely water table depth, hydrologic conductivity, clay content and Cl concentration, are taken into account in this example. The values of these four features were transformed into a common scale 0.0 to 1.0 using suitable membership functions, like "large water table depth" or "high Cl concentration". Now, we can combine these data by constructing a new combining attribute (Fig. 3), namely the joint degree of land suitability for a specified utilization (in this case, as a waste disposal site). Different logical and arithmetical operators can be used for the construction of such a combining attribute. Arithmetical operators (like the sum operator) can be weighted and that makes them particularly useful to express the degree of importance of a particular parameter. The distribution of weights is subjectively determined by a domain expert. In our example we combined the four land characteristics by means of a weighted sum operator and the "and"- operator using the Fuzzy Evaluation and Kriging System FUZZEKS developed at the University of Kiel (Fig. 3) [2].



WATER_TABLE_DEPTH — large water table depth

HYDRAULIC_CONDUCTIVITY — low permeability

CLAY_CONTENT — high clay content

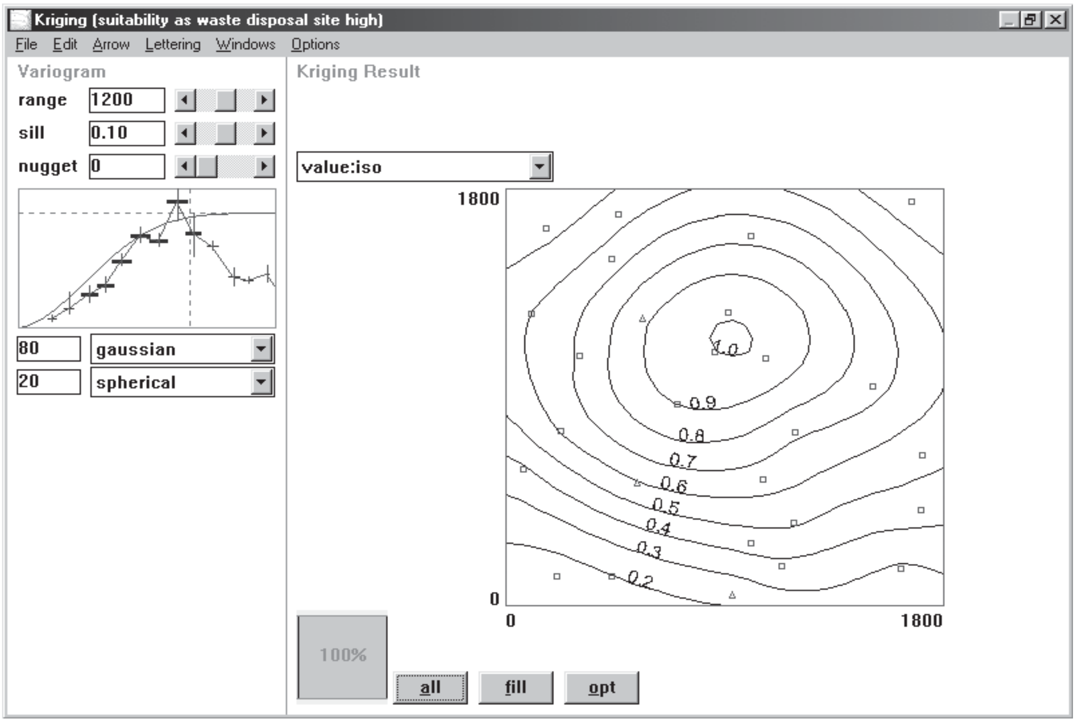
CL_CONCENTRATION — high Cl concentration

30 50 15 5 weighted SUM

40 60 weighted SUM

Figure 3: Constructing a combining operator (the management window of FUZZEKS [2]).

The calculated values of the joint attribute can be presented in the form of isolines using the fuzzy interpolation procedure, the so-called fuzzy kriging (Fig. 4). Fuzzy kriging is an extension of the conventional kriging procedure [1,6]. The application of the conventional methods of spatial interpolation is often restricted owing to insufficient amounts of data [11]. If the gathering of new data is too expensive or impossible, we can consider the use of additional imprecise data subjectively estimated by an expert. The fuzzy kriging procedure utilizes exact (crisp) measurement data as well as imprecise estimates obtained from an expert. These imprecise data can be defined as fuzzy numbers and taken as additional input data for the kriging procedure implemented in FUZZEKS. To simplify the preparation of the input data file a special ASCII-file format was implemented, combining both exact (crisp) and fuzzy data (fuzzy numbers) in one unified form [2]. Fuzzeks supports a user in the preparation of the so-called experimental variogram and in the interactive fitting of the crisp theoretical variogram, which is a basis for the interpolation procedure, to the fuzzy experimental variogram (see the small window left in Fig. 4).


Kriging (suitability as waste disposal site high)

Variogram

range 1200

sill 0.10

nugget 0

80 gaussian

20 spherical

Kriging Result

value:iso

Figure 4: The presentation of the interpolated values of the joint attribute in the form of isolines (on the right in the main window of FUZZEKS,); interactive fitting of a theoretical variogram to an experimental variogram (small window left).

The logical structure of this fuzzy kriging procedure is shown in Fig. 5. The zigzag lines mark the stages with fuzzy data input in the form of fuzzy numbers. At two stages fuzziness is introduced into the calculation. First, fuzziness in the input values causes fuzziness in the experimental variogram. An expert takes the experimental variogram and its fuzziness into account when fitting the crisp theoretical variogram. Second, the fuzzy input values are used at the final step of kriging. Therefore, the kriging results are also fuzzy.

The main fuzzy kriging estimation is a linear combination of the input values and can be calculated using the extension principle and the α -cut-representation of fuzzy sets:

$$Z^*(x)_\alpha = \sum_{i=1}^{n} \delta_i(x) Z(x_i)_\alpha \qquad (4)$$

where:

$Z^*(x)_\alpha$ is is the α -cut of the interpolated value $Z^*(x)$ at the position *x*,

$Z(x_i)_\alpha$ are the α -cuts of the input values $Z(x_i)$, and

$\delta_i(x)$ are the crisp minimizing parameters.

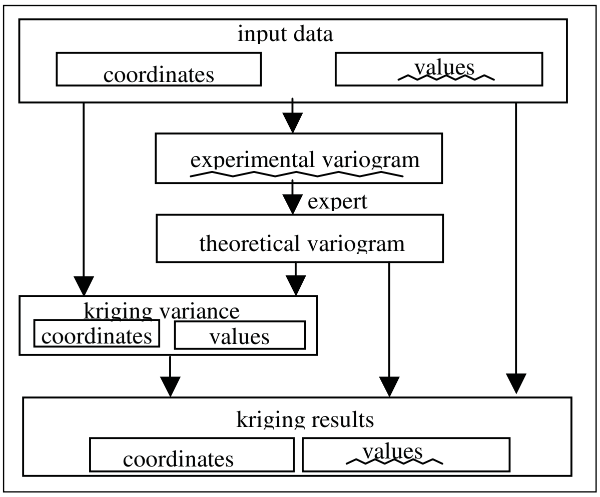


Fig. 5. Logical structure of fuzzy kriging (zigzag lines indicate the fuzziness of data).

The kriging estimation (4) formulated by α-cut-representation of fuzzy sets has been used for the implementation of the fuzzy kriging procedure for the Fuzzy Kriging and Evaluation System FUZZEKS.

4 Final remarks

The fuzzy approach can support users in discovering interesting knowledge in uncertain ecological data at different stages of this process. The searching for representatives of groups of ecological objects by means of clustering methods can be very useful not only in numerosity reduction but also in dimensionality reduction of the data set. Fuzzy clusters of objects without sharp boundaries reflect better the continuous nature of ecological features. That enables a better interpretation of the data structure.

Heterogeneous and imprecise ecological data and vague expert knowledge can be integrated more effectively using a fuzzy approach. The data transformation by means of fuzzy sets can be used for this integration. The membership functions of fuzzy sets can describe the evaluation criteria (e.g. "high clay content") defined by an expert. The construction of a combining attribute in the next step reduces the dimensionality and the size of the data set.

And finally, the intention of the author was to draw the reader's attention to a large field for applications, namely ecology. The paper illustrates it briefly using some short application examples. The development of the easy to use tools (like EcoFucs and FUZZEKS developed at the University of Kiel) can be very helpful for the promotion of fuzzy methods in ecological applications.

References

[1] A. Bárdossy, I. Bogardi and W. E. Kelly, Geostatistics utilizing imprecise (fuzzy) information. *Fuzzy Sets and Systems*, 31/3: 311-327, 1989.

[2] F. Bartels, *Ein Fuzzy-Auswertungs- und Krigingsystem für raumbezogene Daten*. Diplomarbeit, Inst. für Informatik und Praktische Mathematik, Universität Kiel, 1997.

[3] J. C. Bezdek, A convergence theorem for the fuzzy c-means clustering algorithms. *IEEE Trans. PAMI*, PAMI-2(1): 1-8, 1980.

[4] Burrough P.A., Macmillan R.A. and Van Deursen, Fuzzy classification methods for determination land suitability from soil profile observations and topography. *Journal of Soil Science,* 43: 193-210, 1992.

[5] Celmiņš, A., Least Squares model fitting to fuzzy vector data. *Fuzzy Sets an Systems*, 22: 245-269, 1987.

[6] Diamond, P., Fuzzy kriging. *Fuzzy Sets and Systems*, 33/3: 315-332, 1989.

[7] Friderichs M., Fränzle O. and Salski A., Fuzzy clustering of existing chemicals according to their ecotoxicological properties. *Ecological Modelling*, 85/1: 27-40, 1996.

[8] Gömann, M., *Erweiterung des Fuzzy-Clustering-Systems EcoFucs*, Studienarbeit, Institut für Informatik, Christian-Albrechts-Universität zu Kiel, 2002.

[9] Han J. and Kamber M., *Data Mining. Concepts and Techniques*. Morgan Kaufman Publishers, San Francisco, 2001.

[10] Krenawi M., *Entwicklung eines plattformunabhängigen Systems zur Fuzzy-Clusteranalyse*. Diplomarbeit, Inst. für Informatik und Praktische Mathematik, Universität Kiel, 2001.

[11] Piotrowski, J.A., Bartels, F., Salski, A. and Schmidt, G., Geostatistical regionalization of glacial aquitard thicknessin northwestern Germany, based on fuzzy kriging., *Mathematical Geology*, 28/4: 437-452, 1996.

[12] Salski, A., Ecological modeling and data analysis. In: H.-J. Zimmermann: *Practical application of fuzzy technologies. The Handbooks of Fuzzy Sets*, Kluwer, , pp. 247-265, 1999.

[13] Salski, A., Ecological applications of fuzzy logic. In: F. Recknagel (ed): *Ecological Informatics,* Springer, pp. 3-14, 2002.

[14] Salski, A., Bartels, F., A fuzzy approach to land evaluation. *IASME Transactions,* 5/2: 774-780, 2005.

[15] Yang, M.-S. and Liu, H-H, Fuzzy clustering procedures for conical fuzzy vector data, *Fuzzy Sets and Systems*, 106: 189-200, 1999.

ISBN: 978-989-95079-6-8