Classification Models Analysis of Internal Human Population Migration in Districts

KŘUPKA JIŘÍ, KAŠPAROVÁ MILOSLAVA
Institute of System Engineering and Informatics
Faculty of Economics and Administration, University of Pardubice
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC
Jiri.Krupka@upce.cz, Miloslava.Kasparova@upce.cz

Abstract: The paper presents a design and an analysis of classifiers for modelling of internal human migration in districts. Classification models are realized by means of supervised methods. Fuzzy inference systems and a hierarchical structure of fuzzy inference system seem to be preferable in terms of modelling. This hierarchical structure can be realized for a great number of fuzzy rules effectively. The population migration is solved for districts in the Czech Republic. Economic and demographic indicators that affect a size of migration are defined. Dependences between indicators are searched by correlation analysis.

Key-words: Internal human migration, fuzzy inference system, hierarchical structure

1 Introduction

A role of classification is to classify objects, events and real-life situations into classes. Each of the reviewed objects is unique, original and its classification means a certain degree of generalization. Let's define a system for the particular objects i.e. input and output variables, elements (objects) and their mutual relations. Defining and collecting the data of input/output variables cannot be generalized, even though this stage influences the classification result.

Our case deals with a demography system which calls for defining an input parameters (characteristics, variables) vector and an output variables vector for each object. Demography investigates human population reproduction. It encompasses the study of the size, structure and distribution of populations, and the way populations change over the time due to births, deaths, migration and ageing. Changes in the population number and the population increase are basic topics of demography. A natality, mortality and a spatial mobility (migration) influence the status of the population number directly.

Demography teams up with population geography that deals with migrations and a population distribution. Population evolution is a result of natural reproduction population (births, deaths) but also migration results.

Demographic events form the demographic reproductions. The birth and the death are the most significant demographic events. Derived processes are the natality and the mortality. Abortions are special types of death. An abortion rate is a derived process. Next events influence the demographic reproduction vicariously. For example solemnizations of marriages and divorces have an impact on the natality. Illnesses affect the mortality. Events are registered, studied and modified in processes of the natality, the mortality, the nuptiality, the divorce rate and the abortion rate. Analysis and searching of periodicity and important characteristics of their evolution follow then.

The migration means a change of a permanent residence. It is possible to separate an internal and international migration. The international migration is defined as the change of habitual abode outside the state boundary. The internal migration is the change of permanent residence outside an administrative unit, usually a municipality. This migration is registered by a document called "Report on migration", see more in [8].

Many factors influence the size of the population migration. They are e.g. job opportunities, an environment, the nuptiality, the natality, the mortality, etc.

2 Problem Formulation

Goals of this paper are: to define factors that affect the internal human population migration size in 76 districts in the Czech Republic (CR); to determine a factors intensity on the migration; to create a classification model of a migration rate (MR) in the districts where the MR expresses a number of migrants per 1 000 people to date (July, 1) in the year. There are two groups of indicators [8]: basic demographic indicators (a crude marriage rate, a crude birth rate, a crude abortion rate, a crude death rate, a crude divorce rate) and selected

economic indicators (an unemployment rate and a gross average monthly wage).

The crude birth rate (CBR) is a number of live births per 1 000 people to the date in the year t. The calculation is the following (1):

$$CBR_t = (BIR_t / MYP_t)\ 1000, \quad (1)$$

where: BIR_t is a number of live born people in the year t; MYP_t is mid-year population it means a number of population to date 1st July in the year t.

The crude marriage rate (CMR) is a number of marriages per 1 000 people to the date in the year t. The calculation is the following (2):

$$CMR_t = (MAR_t / MYP_t)\ 1000, \quad (2)$$

where: MAR_t is a number of marriages in the year t.

The crude abortion rate (CAR) is a number of abortions per 1 000 people to the date in the year t. The calculation is the following (3):

$$CAR_t = (ABO_t / MYP_t)\ 1000, \quad (3)$$

where: ABO_t is a number of abortions in the year t.

The crude death rate (CDR) is a number of deaths per 1 000 people to the date in the year t. The calculation is the following (4):

$$CDR_t = (DEA_t / MYP_t)\ 1000, \quad (4)$$

where: DEA_t is a number of deaths in the year t.

The crude divorce rate (CDiR) is a number of divorces per 1 000 people to the date in the year t. The calculation is the following (5):

$$CDiR_t = (DIV_t / MYP_t)\ 1000, \quad (5)$$

where: DIV_t is a number of divorces in the year t.

Other examples are e.g. a prevalence and an incidence which are indicators of morbidity [8].

The calculation of the unemployment rate (UR) is the following (6):

$$UR_t = U_t / (E_t + U_t), \quad (6)$$

where: U_t is a number of the unemployed in the year t; E_t is a number of employees in the year t.

The calculation of the gross average monthly wage (W) is the following (7):

$$W_t = Wa_t / (ARN_t\ a_t), \quad (7)$$

where: Wa_t are wages without other personal costs in the year t; ARN_t is an average registration number of employees in the year t; a_t is a number of months in the year t.

Every district is an object $\mathbf{o}_j$ and everyone is described by p indicators (characteristics). A vector of measurement $\mathbf{o}_j$ contains values z_{jp} of p characteristics in formula (8) that it is the following:

$$\mathbf{o}_j = \{z_{j1}, z_{j2}, \ldots, z_{jp}\} \quad (8)$$

for j-th object $\mathbf{o}_j$, (j = 1, 2, …, n). The input set of the objects which are determined for the clustering can be expressed by an objects matrix $\mathbf{O}(n \times p)$ where n is a number of objects (districts), for j =1, 2, ..., n and p is a number of characteristics, for i = 1, 2, ..., p.

3 Model Definition

The problem of classification of MR is composed of two phases: first is divided into data collection and data pre-processing and the second is classification. In the first phase: we used an expert evaluation (EE) and certain cluster analysis methods; we defined 5 clusters of the districts and 5 classes for MR and the demographic and economic indicators. During the second phase we created classification models for MR evaluation in the districts. These models use the demographic and economic indicators values for the year of 2004.

A general scheme of this problem is depicted in the Fig.1. The SPSS and MS Excel software is used for data pre-processing.

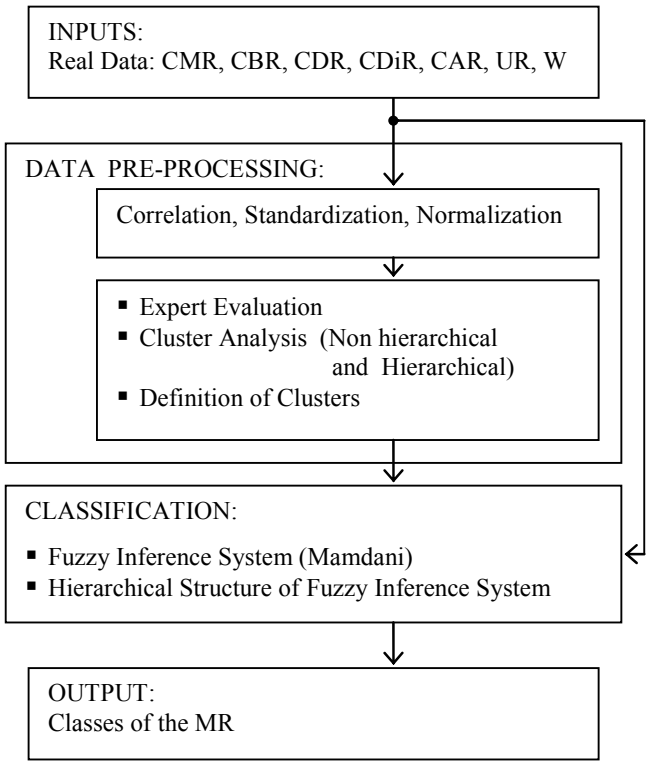

Fig.1 Scheme of the classification model creation

The indicators in the year of 2004 (independent variables) CMR, CBR, CDR, CDiR, CAR, UR and W were selected as factors that influence size of migration rate MR (dependent variable) in 76 districts in the CR. The matrix $\mathbf{O}(76 \times 8)$ is defined for a classification model and includes values of 8 indicators for 76 districts. The

basic descriptive characteristics of variables (indicators) as average, minimal and maximal value etc. of matrixes are in the Table 1. The correlation analysis deals with interdependences of these indicators and a dependence of a migration on them.

Table 1 Description of variables for classification

Variable	Min	Max	Mean	Std. Deviation
CMR	4.06	6.25	4.91	0.46
CBR	8.16	11.11	9.58	0.64
CDR	8.95	12.31	10.48	0.76
CDiR	2.16	4.8	3.16	0.60
CAR	2.79	6.92	4.13	0.94
UR	2.75	19.86	10.16	3.15
W	12625	17106	14441.41	1033.41
MR	-7.63	34.19	1.95	6.14

The most widely-used type of a correlation coefficient is Pearson correlation coefficient ρ_{ij} [17]. Its range, dependences and the other attributes of ρ_{ij} are described e.g. in [11], [17].

In the data matrix the top correlation was found between variables CAR and CDiR (ρ_{ij} = 0.648). Soft linear dependence was between variables CMR and CDiR (ρ_{ij} = 0.551), CDiR and CBR (ρ_{ij} = 0.459), CBR and CMR (ρ_{ij} = 0.431), CMR and CAR (ρ_{ij} = 0.481, CBR and CAR (ρ_{ij} = 0.409), CBR and MR (ρ_{ij} = 0.344), too. These correlations of variables were positive. No correlation was found between variables W and CAR (ρ_{ij} = 0.000). The soft linear dependence (ρ_{ij} = 0.344) was achieved between variables MR and CBR. These data were standardized and normalized.

The second goal of this part is the creation of 5 classes for the MR which are represented by lexical (linguistic) values: very small (VS), small (S), middle (M), high (H) and very high (VH). Two approaches to creation classes (clusters) were used. The first one is based on the EE, see more in [8] and the second one uses the cluster analysis.

The cluster analysis [19] is an exploratory data analysis tool for solving classification problems. The object is to sort cases (people, events, etc.) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters.

An existence of n objects is an initial condition for the usage of the cluster analysis. The task of clustering is then to divide the set of objects in the matrix **O** into the disjunctive clusters.

The decision making about the object clustering in cluster is realized on the basis of the similarity by application of metric [5], [6], [11]. A sum of the square errors to centre (centre of gravity) of clusters E [11] is chosen as a criterion of the quality of clustering. It is defined in this way:

Let $\Omega = \{M_1, M_2, \ldots, M_h \ldots, M_k\}$ is the clustering of objects set in k clusters $M_1 = \{\mathbf{o}_{11}, \mathbf{o}_{12}, \ldots, \mathbf{o}_{1n_1}\}$, $M_2 = \{\mathbf{o}_{21}, \mathbf{o}_{22}, \ldots, \mathbf{o}_{2n_2}\}$, …, $M_k = \{\mathbf{o}_{k1}, \mathbf{o}_{k2}, \ldots, \mathbf{o}_{kn_k}\}$ where $\mathbf{o}_{hj}$ is object j of h-th cluster M_h. Then E is determined in the formula (9):

$$E = \sum_{h=1}^{k} \sum_{j=1}^{n_h} d^2\left(\mathbf{o}_{hj}, T_h\right), \qquad (9)$$

where: $d^2(\mathbf{o}_{hj}, T_h)$ is the square of the Euclidian metric of object $\mathbf{o}_{hj}$ to the centre T_h of cluster M_h ; T_h is the centre of cluster M_h; it is determined by the vector of mean values of characteristics i of objects in cluster M_h in formula $T_h = (t_{h1}, t_{h2}, \ldots, t_{hp})$, for its characteristics i, where i = 1, 2, …, p, is (10):

$$t_{hi} = 1/n_h \sum_{j=1}^{n_h} z_{hji}\,, \qquad (10)$$

where: n_h is a number of objects in cluster M_h; z_{hji} is a characteristic i of object j in cluster M_h.

We have used following methods for a generation of 5 clusters of districts:

- non-hierarchical clustering methods with sold number of clusters (McQueen [11], Forgy [3], Jancey [7]). The total sum of the square errors E was chosen as a quality criterion of a set of objects resolution. On the basis of results of E (in these methods) the Jancey method (E = 21.32) was chosen as the best method for data pre-processing;
- hierarchical clustering method[1] (HiCM), see more in [1], [5], [6], [11] was realized in programme SPSS.

On the basis of the best results [8] we used the Jancey non-hierarchical cluster method and divided the set of objects into 5 clusters and computed centre of gravity for each cluster. On the basis of the centre of gravity value of MR we defined a lexical value for classification classes (VS, S, M, H, VH), see the Table 2.

Table 2 Definition of classification classes for clusters

Cluster	Value of center of gravity		Class of MR
	[CMR; CBR; CDR; CDiR; CAR; UR; W]	MR	
1st	[4.66; 8.93; 11.20; 2.92; 3.94; 9.76; 15276.70]	2.10	S
2nd	[4.66; 9.24; 10.04; 2.89; 3.43; 8.39; 15076.00]	0.49	VS
3rd	[5.31; 10.23; 10.39; 3.86; 5.50; 10.74; 14508.53]	2.22	M
4th	[4.80; 9.69; 10.15; 2.81; 3.66; 13.62; 14023.29]	2.49	H
5th	[5.01; 9.52; 10.97; 3.21; 4.01; 7.85; 13563.13]	2.58	VH

The clusters of districts in the CR were created on the basis of introduced methods. We defined the centre of

[1] We used Average-linkage method.

gravity for these clusters and extracted lexical variables (values) for the classification classes of MR on the basis of the clusters centre of gravity.

4 Modelling of Internal Human Population Migration Classifiers

Classification models have been used for modelling of the internal human population migration. This part is focused on a design of fuzzy classification models because formerly designed classification models based on non hierarchical cluster analysis, neural networks and regression trees are described in [8].

These fuzzy models were created in the programme Clementine[2] [18] and MATLAB.

Parameters for a definition model of MR evaluation can be expressed by incompleteness and disproportion. Classification deals with knowledge and data characterized by uncertainty. This was realized by means of fuzzy inference system (FIS) [9], [16]. The heuristic approach for the creation of FIS (it means the shape and number of membership function of inputs and output variables, and the base of fuzzy rules (BFRs) was used because an exact general method for definition of their number does not exist [10]. The definition of the number of fuzzy rules (FRs) is described in [9], [10] or the method in [12], [20] can be used. The number of FRs can be also optimized by genetic algorithms and evolution strategies [2], [13].

A FIS is represented by a block with inputs x_n and output y. It is graphically interpreted and described in [10], [15]. A disadvantage of this approach to the design of FIS [4], [14], [15] is an exponential growth of the number FRs in BFRs and FIS can be realized ineffectively.

This problem can be removed by a hierarchical structure of FIS [4], [14], [15]. In the hierarchical structure of FIS it is necessary to determine the number of FRs for the first and other levels, see more in [15].

The first classification model considered using of FIS1 block (Mamdani FIS) which was defined for 7 input variables (CMR, CDiR, CBR, CAR, CDR, UR and W) and 1 output variable MR. It contains the fuzzification process, the inference mechanism and the defuzzification process [9], [10], [14], [16]. For input (output) variables there were defined membership functions (MFs) of fuzzy sets (VS, S, M, H, VH) where FRs are written in form IF antecedent THEN consequent. Consequently on the basis of a lot of simulations the defuzzification method of centre of gravity had to be vetoed in the FIS1 model and we used the mean of maximum method. We used min and max values from these variables for the definition of interval of universe. The 5 triangle MFs of fuzzy sets for each variable were designed by means of Jancey method, where an average value (Table 2) of an individual clusters gravity centre defines the centre of the MF. It means degree of MF is 1.

There is problem with the number of FRs. The complete BFRs would contain more thousands of FRs (3125 rules). This model was not used for its complication, an untransparency and a time-demanding creation of a simulation model.

There are techniques for an optimization of FIS1 structure [16] e.g. the number of MFs can reduce by means of an aggregation of nearby MFs. We used 3 or 4 MFs in our model. On the basis of a decision tree (algorithm C5.0) we achieved a decrease of number of input variables (CDiR, CAR, CDR, UR and W) and number of FRs (1048 rules) in new FIS (FIS2).

The hierarchical structure of FIS (HSFIS) can be realized effectively for a great number of FRs. For creating of a HSFIS it is possible to use the EE or dendrograms from the hierarchical clustering (Ward, Median, Centroid, Average-linkage and Single-linkage methods). In our case we applied the EE which was supported by the input data correlation analysis.

The 5th level HSFIS inputs were formed by two groups of variables (CBR, CDR) and (CMR, CDiR, CAR) and two autonomous input variables (UR and W). There were used auxiliary parameters P1, …, P5 into the levels of the HSFIS. Definition of MFs issued from the FIS1 model. The FRs were set for all defined combinations of input/output variables of individual FISs (it means Level 1/1, Level 1/2, …, Level 5/1) under the given HSFIS (Fig.2).

Classification assignments by means of FIS classification models (FIS2 and HSFIS) are represented in the Table 3 and a compliance between results of classification by FIS2 and HSFIS and other classification algorithms (Jancey, EE and HiCM) is in the Table 4.

Table 3 Classification assignments by means of FIS classification models (in %)

Class	Type of FIS model	
	FIS2	HSFIS
1st	3.95	26.32
2nd	84.21	32.89
3rd	1.32	18.42
4th	10.53	19.74
5th	0.00	2.63

[2] Clementine is an enterprise data mining workbench of SPSS Inc. that enables a quick development of predictive models using expertise and deploying them into operations to improve decision making. It supports all steps of standard methodology CRISP-DM (Cross-Industry Standard Process for Data Mining).

Table 4 Compliance between FIS models and other classification algorithms (in %)

Algorithm (Method)	Type of FIS model	
	FIS2	HSFIS
Jancey	30.26	30.26
EE	13.16	23.68
HiCM	9.21	11.84

The FISs (FIS1, FIS2 and HSFIS) are used for the modelling of classification models and are realized in MATLAB\Simulink (ver.7.0).

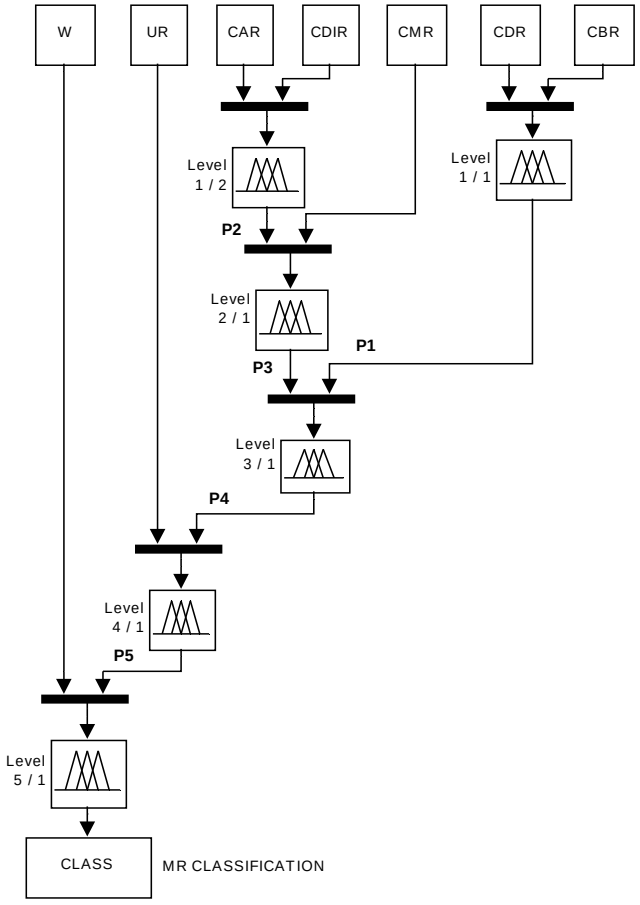

Fig.2 Classification model of MR on the basis of FIS hierarchical structure

5 Conclusion

Demographic and economic indicators that influence the size of MR were defined in the paper. By the correlation analysis specific dependences between indicators were found. The clusters of districts in the CR were created on the basis of introduced methods. We defined the centre of gravity for these clusters and extracted lexical variables for the classification classes of MR on the basis of the centre of gravity. In the second phase we create the optimization structure of FIS (it means FIS2) and HSFIS. This hierarchical structure contents a smaller number of FRs than FIS1 and FIS2. It represents an effective method for realization of classification model with a lot of input variables. Based on the acquired classification results analysis there was revealed a preference of assigning most objects into 2nd class by means of FIS2 and HSFIS (Table 3).

From the point of view of comparing classification results (Table 4) with the final classes achieved by means of HiCM, Jancey method and EE. The highest compliance was achieved between Jancey methods of classification and the classification on the basis of the FIS2 and HSFIS. It means we compare classification results on the basis Jancey methods on the one hand and type of FIS models on the other hand. The achieved compliance is 30.26 %. The lowest compliance was achieved by means of HiCM method and FIS2 (9.21 %).

Achieving better results is conditioned by defining other factors. These are e.g. a description of districts from the point of view of an environment, an area topology, a structure of the population education, job opportunities etc. The usage of a fuzzy logic appears convenient for district rating by these factors.

We finally used the fuzzy clustering for the classification of MR, too. In the fuzzy clustering [16] as distinct from the classical clustering, data elements can belong to more than one cluster, and a set of membership levels is associated with each element. These indicate the strength of the association between that data element and a particular cluster. The output of such algorithms is clustering, but not a partition.

For a generation of 5 clusters of districts by fuzzy clustering we have used Fuzzy c-Means method [16]. On the basis of maximal cluster degree of membership function it was possible to determine number of cluster in 52 objects. It was not possible to determine this maximal cluster degree of membership function into one cluster in 24 objects (e.g. 1 object belongs into 1st and 2nd cluster with the identical membership degree). Therefore it was not used these cluster results during the creation and optimization of FIS.

6 Acknowledgement

This paper has been prepared as a part of the institutional project FG/FES/2006/31.

References:

[1] Berthold, M., Hand, D.J. (eds.), *Intelligent Data Analysis: An Introduction*. Springer Verlag, Berlin, 2003.

[2] Buckley, J.J., Hayaashi, Y., Genetic Algorithm and Applications. *Fuzzy Sets and Systems*, Vol.61, 1994, pp.129-136.

[3] Forgy, E.W., *Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications*. Biometrics Soc. Meetings, Riverside, 1965.

[4] Gegov, A.E., Frank, P.M., Hierarchical Fuzzy Control Multivariable Systems. *Fuzzy Sets and Systems*, Vol.72, 1995, pp.299-310.

[5] Guidici, P., *Applied Data Mining: Statistical Methods for Business and Industry*. Wiley, West Sussex, 2003.

[6] Han, J., Kamber, M., *Data Mining: Concepts and Techniques*. Morgan Kaufmann Press, San Francisco, 2001.

[7] Jancey, R.C., Multidimensional Group Analysis. *Austral J. Botany*, Vol.14, 1966.

[8] Kašparová, M., Křupka, J., Classification and Prediction Models for Internal Population Migration in Distrists. *WSEAS Transaction on Systems*, Vol.5, WSEAS Press, Athens New York, 2006, pp.1540-1547.

[9] Kuncheva, L.I., *Fuzzy Classifier Design*. Physica-Verl., Heidelberg New York, 2000.

[10] Lee, CH.CH., Fuzzy Logic in Control Systems: Fuzzy Logic Controller- Part I and II. *IEEE Transaction on Systems, Man, and Cybernetics*, Vol.20, 1990, pp.404-433.

[11] Lukasová, A., Šarmanová, J., *Metody shlukové analýzy*. SNTL, Praha, 1985.

[12] Nozaki, K., Ishibuchi, H., Tanaka, H., Adaptive Fuzzy Rules Based Classification Systems. *IEEE Transaction on Fuzzy Systems*, Vol.4, No.3, 1996, pp.238-250.

[13] Pal, S.K., Wang, P.P., *Genetic Algorithm for Pattern Recognition*. CRC Press Inc., Boston, 1996.

[14] Pedrycz, W., *Fuzzy Control and Fuzzy Systems*. 2nd edn. Research Studies Press Ltd., London, 1993.

[15] Raju, G.V.S., Zhou, J., Adaptive Hierarchical Fuzzy Controler. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.23, No.4, 1993, pp.973-980.

[16] Ross, T.J., *Fuzzy Logic with Engineering Applications*. 2nd edn. John Wiley and Sons, Ltd., New York, 2004.

[17] Rublík, F.: *Základy pravdepodobnosti a štatistiky*. Alfa, Bratislava, 1983.

[18] SPSS Inc. *Clementine® 7.0 User's Guide*, 2002.

[19] Turban, E., Aronson J.E., Liang T.-P., *Decision Support Systems and Intelligent Systems*. 7th edn. Pearson Prentice Hall, New Jersey, 2005.

[20] Wang, L.X., Mendel, J.M., Generating Fuzzy Rules by Learning from Examples. *IEEE Transaction on Systems, Man, and Cybernetics*, Vol.22, 1992, pp.1414-1427.