Proceedings of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia, June 12-14, 2006 (pp114-119)

Modelling of Internal Population Migration in Districts of the Czech Republic by Selected Methods

KAŠPAROVÁ MILOSLAVA
System Engineering and Informatics Institute
Faculty of Economics and Administration
University of Pardubice
Studentská 95, 532 10 Pardubice
Czech Republic
http://www.upce.cz

Abstract: Population migration denotes any human movement from one locality to another. Population migration in districts of the Czech Republic is solved in the paper. Firstly economic and demographic indicators that affect size of migration are defined. Dependences between indicators are searched by correlation analysis. The goal is to model population migration by these indicators. Regression analysis and neural networks are used for modeling of a migration rate in districts of the Czech Republic. The achieved results are analysed.

Key – words: population migration, indicators, regression analysis, neural networks

1 Introduction

Demography investigates human population reproduction. It encompasses the study of the size, structure and distribution of populations, and the way populations change over time due to births, deaths, migration and ageing. Changes of the population number and the increase of population are basic topics of demography. A natality, mortality and a spatial mobility – migration influence directly the status of the population number.

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

Demographic events form the demographic reproductions. The birth and the death are most significant demographic events. Derived processes are the natality and the mortality. Abortions are special type of death. An abortion rate is derived process. Next events influence the demographic reproduction vicariously. For example solemnization 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 aboration rate. Analysis and searching of periodicity and important characteristics of their evolution follow then.

Migration means the change of the permanent residence. It is possible to separate an internal migration and an international migration. The international migration is defined as the change of habitual abode outside the state boundary. Internal migration is the change of permanent residence outside an administrative unit usually municipality. This migration is registered by document called "Report on migration" [1].

Many factors influence the size of the population migration. They are for example job opportunities, environment, nuptiality, natality, mortality, etc.

2 Problem Formulation

Goals of this paper are:

- To define factors that affect the population migration size in 76 districts in the Czech Republic (CR) in years 2002 – 2004;
- to determine a factors intensity on the migration;
- to design a model for prediction of the migration rate (*MR* – number of migrants per 1 000 people to date 1st July in the year *t*) in districts of the CR on the basis of existing influences.

2. 1 Basic Demographic Indicators

A crude marriage rate, a crude birth rate, a crude abortion rate, a crude death rate, a crude divorce rate are counted among basic demographic indicators.

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

$$CBR_t = \frac{BIR_t}{MYP_t} \cdot 1000 , \quad (1)$$

where: BIR_t a number of live born people in the year *t*;

MYP_t mid-year population; a number of population to date 1st July in the year *t*.

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

$$CMR_t = \frac{MAR_t}{MYP_t} \cdot 1000 , \quad (2)$$

where: MAR_t a number of marriages in the year *t*;

MYP_t mid-year population; a number of population to date 1st July in the year *t*.

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

$$CAR_t = \frac{ABO_t}{MYP_t} \cdot 1000 , \quad (3)$$

where: ABO_t a number of abortions in the year *t*;

MYP_t mid-year population; a number of population to date 1st July in the year *t*.

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

$$CDR_t = \frac{DEA_t}{MYP_t} \cdot 1000 , \quad (4)$$

where: DEA_t a number of deaths in the year *t*;

MYP_t mid-year population; a number of population to date 1st July in the year *t*.

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

$$CDiR_t = \frac{DIV_t}{MYP_t} \cdot 1000 , \quad (5)$$

where: DIV_t a number of divorces in the year *t*;

MYP_t mid-year population; a number of population to date 1st July in the year *t*.

Other examples are e.g. a prevalence and incidence which are indicators of morbidity [1].

2. 2 Selected Economic Indicators

An unemployment rate and a gross average monthly wage are basic economic indicators measured in districts of the CR.

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

$$u_t = \frac{U_t}{(E_t + U_t)}, \quad (6)$$

where: U_t a number of unemployees in the year *t;*

E_t a number of employees in the year *t*.

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

$$W_t = \frac{Wa_t}{(ARNa)_t} , \quad (7)$$

where: Wa_t wages without other personal costs in the year *t*;

ARN_t average registration number of employees in year *t*;

a_t a number of months in year *t.*

Calculation of wages in CR follows Act Nr. 586/1992 Coll., Act Nr. 589/1992 Coll. and Act Nr. 592/1992 Coll.

3 Problem Solution

Indicators in the years of 2002, 2003, 2004 (independent variables) described in chapter 2.1 and 2.2: *CMR, CBR, CAR, CDR, CDiR,, u* and *W* were selected as factors that influence size of migration rate *MR* (dependent variable)

in 76 districts of the CR. The basic descriptive characteristics of variables (indicators) as average, minimal and maximal value etc. of matrix are in the Table 1. The data matrix contented 228 objects (districts). Correlation analysis deals with interdependences of these indicators and a dependence of a migration on them.

3. 1 Correlation Analysis

A correlation is a measure of the relation between two (*i* and *j*) or more variables. The most widely-used type of correlation coefficient is Pearson correlation coefficient ρ_{ij} [2]. It is calculated as (8):

$$\rho_{ij} = \frac{cov(\xi_i, \xi_j)}{\sigma_i \sigma_j} \tag{8}$$

The Pearson correlation coefficient ρ_{ij} can range from -1.00 to +1.00 and can be expressed by the following way:

- if $\rho_{ij} > 0$ it is a positive correlation of variables;
- if $\rho_{ij} < 0$ it is a negative correlation of variables;
- if $\rho_{ij} = 0$ this value represents a lack of correlation between variables;
- if $\rho_{ij} = 1$ it is a perfect positive correlation between variables.

From the point of view of ρ_{ij} size defines this linear dependence of variables [3]:

- if $\rho_{ij} \leq 0.3$ it is a small linear dependence;
- if $\rho_{ij} \in (0.3;0.8>$ it is a soft linear dependence;
- if $\rho_{ij} \in (0.8;1>$ it is a strong linear dependence.

The other attributes of the Pearson correlation coefficient ρ_{ij} are described for example in [2, 3, 4].

Table 1: Description of variables

Variable	Corre-lation	Min	Max	Mean	Std. Dev.
u	-0.37	2.75	23.51	10.13	4.08
W	0.3	11910	2011	14841.36	1470.07
CMR	0.10	3.88	6.40	4.9	0.47
CDiR	0.07	1.91	4.80	3.13	0.6
CBR	0.24	8.03	11.21	9.35	0.61
CAR	-0.02	2.58	7.19	4.22	0.91
CDR	0.22	8.04	13.06	10.7	0.82
MR	1.00	-7.63	34.19	2.01	5.1

The top correlation was found between variables *CAR* and *CDiR* (ρ_{ij} = 0,641). Soft linear dependence was between variables *W* and *CDiR* (ρ_{ij} = 0.403), *CDiR* and *CBR* (ρ_{ij} = 0.436), *CMR* and *CDiR* (ρ_{ij} = 0.399), *CBR* and *CMR* (ρ_{ij} = 0.387), *W* and *CBR* (ρ_{ij} = 0.355), *CMR* and *CAR* (ρ_{ij} = 0.427), too. These correlations of variables were positive. The correlations between the dependent variable *MR* and independent variables are in the Table 1. The soft linear dependence is achieved between variables *MR* and *u* (ρ_{ij} = -0.37).

3. 2 Modelling of Internal Population Migration

Regression analysis and neural networks (NNs) were used for modelling of the migration. The methods results were compared. The modelling was realized in the Clementine[1] software.

3.2.1 Regression Analysis

The purpose of multiple regression [2,4,5] is to predict a single variable from one or more independent variables. Multiple regression with many predictor variables is an extension of linear regression with two predictor variables. A linear transformation of the *x* variables is done so that the sum of squared deviations of the observed and predicted $\hat{y}$ is a minimum. The $\hat{y}$ is expressed by the following equation (9):

$$\hat{y}_i = \hat{\beta}_0 x_{0i} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_m x_{mi} + \varepsilon_i . \tag{9}$$

Values of $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_m$ are estimates of unknown regression parameters; *x* is independent variable, $\hat{y}$ is prediction of dependent variable and ε is random error.

The model training is based on a vector linear parameters optimization so that predicted values of training data were consistent with actual values as much as possible. A method of lest squares [6] is the most used. It minimizes residual sum of squares of the predicted and the actual values. It is Residual Sum of Squares (RSS) criterion.

$$RSS(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 . \tag{10}$$

Regression Analysis Results

The data matrix comprised 228 objects and was divided in two parts. Two thirds of objects were used for creation of the regression model and one third for the verification of achieved results. For the regression model creation there were used methods of Enter[2], Stepwise[3],

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

[2] Enter Method [10] is the default method, which enters all the input fields into the equation directly. No field selection is performed in building the model.

[3] The Stepwise [10] method of field selection builds the equation in steps to the contrary of the Enter method. The initial model is the simplest model possible, with no input fields in the equation. At each step, input fields that have not yet been added to the model are evaluated, and if the best of those input fields adds significantly to the predictive power of the model, it is added. In addition, input fields that are currently in the model are reevaluated to

Backwards[4] and Forwards[5]. The results in Table 2 were achieved by these methods.

Table 2: Results of prediction by multiple regression

Prediction Method	MAE in [‰]	RSS
Enter	2.97	2 726.64
Stepwise	3.11	2 822.91
Backwards	3.11	2 822.91
Forwards	3.11	2 822.91

The best result was achieved with Enter method. It achieved $\hat{MR}$ 2.97 of mean absolute error (MAE). The calculation of MAE is the following (11):

$$MAE = \frac{1}{76}\sum\left|\left(MR_i - \hat{MR}_i\right)\right|. \quad (11)$$

Regression model parameters are (12):

$$\hat{MR} = \begin{array}{l} -0{,}4825\,u + 0{,}0009655\,W - 1{,}428\,CMR \\ +\,0{,}6392\,CDiR + -0{,}6913\,CAR + 3{,}331 \\ CBR + 1{,}337\,CDR - 0{,}9242\,t + 1\,806{,}3 \end{array} \quad (12)$$

3.2.2 Multilayer Perceptron Neural Networks

Multilayer perceptron neural networks (MPNN) [7, 8] were used by the population migration modeling. There are typically three parts in a NN: an input layer with units representing the input fields, one or more hidden layers, and an output layer with a unit or units representing the output fields. The units - neurons are connected with varying connection strengths or weights. The network learns by examining individual records, generating a prediction for each record, and making adjustment to the weights whenever it makes an incorrect prediction. This process is repeated many times, and the NN continues to improve its predictions until one or more of the stopping criteria have been met [9].

The training of a MPNN uses a method called Backpropagation of error [9]. For each record presented to the NN during training, information (in the form of input fields) feeds forward through the NN to generate a prediction from the output layer. This prediction is compared to the recorded output value for the training record, and the difference between the predicted and actual outputs is propagated backward through the NN to adjust the connection weights to improve the prediction for similar patterns [9]. For training in Clementine are used following methods: Quick, Dynamic, Multiple, Prune and Exhaustive Prune.

Quick method uses rules of thumb and characteristics of the data to choose an appropriate shape for the NN. Dynamic method created an initial topology, but modifies the topology by adding and/or removing hidden units as training progresses. Multiple networks are trained in pseudo-parallel fashion. Each specified NN is initialized, and all NNs are trained. When the stopping criterion is met for all NNs, the NN with the highest accuracy is returned as the final model. Prune method starts with a large NN and gradually prunes it by removing unhelpful neurons from the input and hidden layers. Pruning proceeds in two stages: pruning the hidden neurons and pruning the input neurons. The Exhaustive Prune method is a special case of the Prune method. This method is usually the slowest, but it often yields better results than other methods [9], more about methods in [9]. These methods were used for prediction of the population migration.

Neural Networks Results

Data matrix was divided into a training set and testing set. From training set it is chosen a part of date for validation. According to [8] a possible size of validation set the range varies from 10 % to 50 %. The size of testing set is 76 objects. It is one third of objects from the total size of the data matrix.

There were carried a lot of tests that differ in setting stop criterions (time between 2 and 10 minutes for NN training) and size of validation sets (25 % - 50 %). Better results were achieved by Quick method, Prune method and Exhaustive prune method. The best results are in Table 3. These results were achieved by 5 minutes of NN training time and by size of validation set 50 %.

Table 3: The best results of prediction

Method of NN training	Topology of NN	CPA in [%]	MAE in [‰]	Standard Deviation
Ex. Prune	8-8-4-1	94.45	2.28	2.54
Quick	8-3-1	93.54	2.34	2.51
Prune	3-2-1	95.35	2.40	2.60

determine if any of them can be removed without significantly detracting from the model. If so, they are removed. Then the process repeats, and other fields are added and/or removed. When no more fields can be added to improve the model, and no more can be removed without detracting from the model,
the final model is generated.

[4] The Backwards method of field selection is similar to the Stepwise method in that the model is built in steps. However, with this method, the initial model contains all of the input fields as predictors, and fields can only be removed from the model. Input fields that contribute little to the model are removed one by one until no more fields can be removed without significantly worsening the model, yielding the final model [10].

[5] The Forwards method is essentially the opposite of the Backwards method. With this method, the initial model is the simplest model with no input fields, and fields can only be added to the model. At each step, input fields not yet in the model are tested based on how much they would improve the model, and the best of those is added to the model. When no more fields can be added or the best candidate field does not produce a large enough improvement in the model, the final model is generated [10].

The best result of prediction was extracted from the Exhaustive Prune method. The topology of NN is the following: 8 neurons are there in an input layer. Hidden layers create 8 and 4 neurons and the output layer with one neuron represents the output field. Time of NN training was 5 minutes. The overall predicted accuracy (CPA) [9] of NN training is 94.45. Calculation of *PA* is the following (13):

$$PA = \left(0.5 - \left| \frac{(MR_i - \hat{MR}_i)}{MR_{\max} - MR_{\min}} \right| \right) 100 \text{ (in [\%])}. \qquad (13)$$

The PA is calculated for each object, and the CPA is the average of the values for all records in the training data [9].

The NN achieved $\hat{MR}$ 2.28 MAE .

Because the best result of prediction was achieved by the Exhaustive Prune method, the new 30 tests were realized on the basis of the same stopping critera setting. The average values of MAE and CPA are in the Table 4.

Table 4: Average MAE and average CPA from 30 tests

Method of NN training	Average MAE in [‰]	Average CPA in [%]
Exhaustive Prune	2.56	94.09

The graphical representation of MAEs from 30 tests is in Fig. 1.

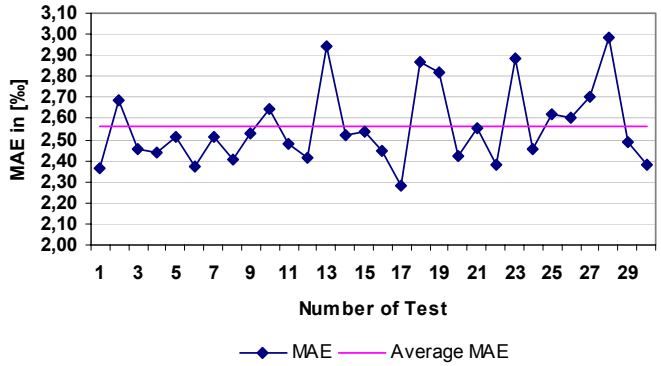

Fig. 1 Graphical representation of MAEs

The average CPA and all CPAs are in Fig. 2.

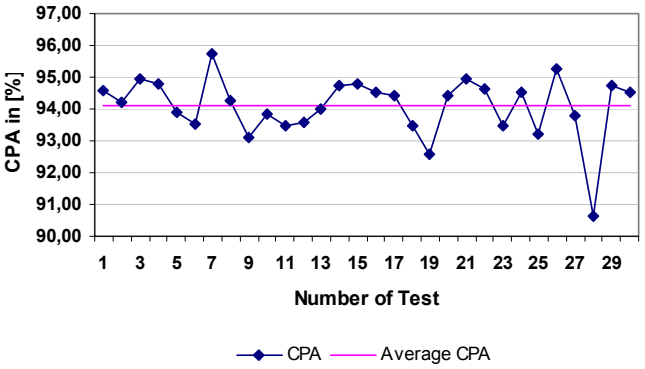

Fig. 2 Graphical representation of CPAs

Importance of Inputs Variables

An incidence of three most significant variables was measured by the realization of 30 tests by Exhaustive Prune method. This method is based on the principle of removing unhelpful neurons from the input and hidden layers. The achieved results are in the Table 5.

Table 5: Frequency of variables

Variable	Importance 1	Importance 2	Importance 3
u	19	6	2
W	8	0	2
CBR	2	12	2
CDR	1	2	3
CMR	0	0	2
CAR	0	0	1

Economic indicators became important inputs. As the most significant input the unemployment rate *u* was selected approximately in 64 % of tests. The gross average monthly wage *W* was the most significant approximately in 27 % of tests. The demographic indicator - the crude birth rate *CBR* was found as the second important input. Graphical representation of the variables frequency is in Fig. 3.

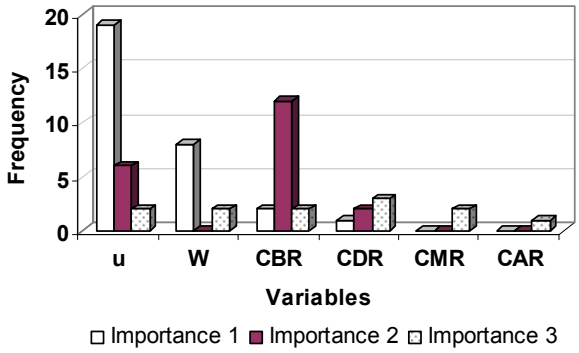

Fig. 3 Frequency of variables

4 Conclusions

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 multiple regression and NNs were used for prediction of *MR* in districts of the CR on the basis of existing indicators. The usage of NNs appears as preferable. But these methods don't demonstrate such results that would approximate to reality. The difference between the best results of methods is 0.69. The Fig. 4 shows the comparison of predicted and actual values of *MR* in 10 districts in the CR. Predicted values in Fig. 4 were achieved by NNs.

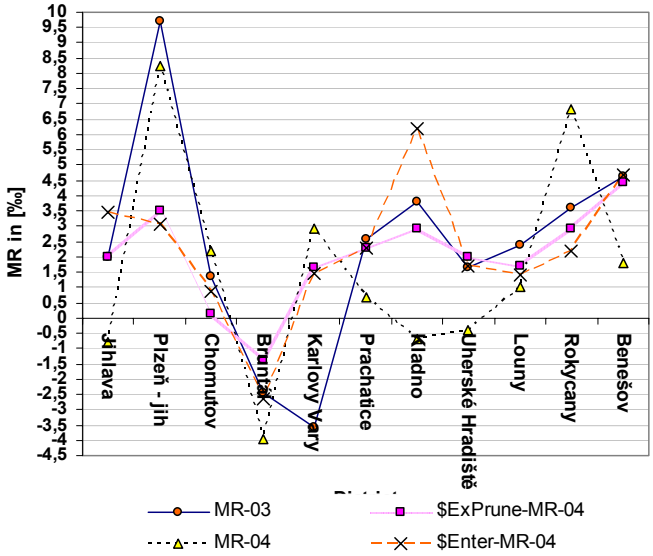

Fig. 4 Predicted and actual values ccomparison of *MR* in 10 districts in the CR

Achieving of better results is conditioned by defining other factors. These are for example 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.

References:

[1] *Demografický informační portál Česká verze* [online].URL<http://www.demografie.info> cit. [2006-04-14].

[2] Meloun, M. - Militký, J., *Kompendium statistického zpracování dat: metody a řešené úlohy včetně CD,* Praha: Academia Praha, 2002.

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

[4] Stockburger, D. W., *Multiple Regression with Many Predictor Variables* [online]. URL<http://badame.vse.cz/mirrors/multibook/mlt07.htm> cit. [2006-04-14].

[5] Berthold, M. - Hand, D. J., *Intelligent Data Analysis: An Introduction*, Springer Verlag, 2003.

[6] Mařík, V. - Štěpánková, O. - Lažanský, J. a kol., *Umělá inteligence (4)*, Praha: Academia Praha, 2003.

[7] Rusell, S. J. - Norvig, P., *Artificial Intelligence: A Modern Approach,* Prentice Hall, 2002.

[8] Novák, M. a kol., *Umělé neuronové sítě, teorie a aplikace,* C. H. Beck, 1998.

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

[10] SPSS Inc. Clementine Overview. Clementine [CD-ROM]. Ver. 7.0.