Air Quality Modelling by Kohonen’s Self-organizing Feature Maps and LVQ Neural Networks

PETR HÁJEK, VLADIMÍR OLEJ
Institute of System Engineering and Informatics
Faculty of Economics and Administration
University of Pardubice
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC
Petr.Hajek@upce.cz, Vladimír.Olej@upce.cz

Abstract: - The paper presents a design of parameters for air quality modelling and the classification of districts into classes according to their pollution. Further, it presents a model design, data pre-processing, the designs of various structures of Kohonen’s Self-organizing Feature Maps (unsupervised methods), the clustering by K-means algorithm and the classification by Learning Vector Quantization neural networks (supervised methods). Therefore, the model generates well-separated clusters and has good generalization ability as well.

Key-Words: - Air quality, modelling, Kohonen’s self-organizing feature maps, K-means algorithm, Learning Vector Quantization neural networks, classification.

1 Introduction

The air pollution involves the spectrum of activities causing the emission of substances or energy into the atmosphere. In other words, air pollution represents a result of materials emissions in solid, liquid or gaseous state from different sources into the air, which negatively influence the quality and composition of air [1]. The influence can either be direct or a result of chemical changes. Air protection stands for the set of technical and administrative measures [1], which aim at the direct or indirect reduction of the rapid air pollution growth. The technical measures involve technological, material, optimization or restriction measures. The legislative, administrative, economic, control and other measures are samples of the administrative ones. The importance of air protection goes up as the air pollution increases.

The air quality modelling (classification of the districts $o_i^t \in O$ into the classes $\omega_{i,j}^t \in \Omega$ according to air pollution) can be realized by various methods. For example, fuzzy inference systems [2], unsupervised (supervised) methods [3,4] and neuro-fuzzy systems [2] are suitable for air quality modelling. Neural networks [3,4] seem to be appropriate due to their ability to learn, generalize and model non-linear relations. Their output is represented for example by an assignment of the i-th district $o_i^t \in O$, $O=\{o_1^t, o_2^t, \dots, o_i^t, \dots, o_n^t\}$ in time t to the j-th class $\omega_{i,j}^t \in \Omega$, $\Omega=\{\omega_{1,j}^t, \omega_{2,j}^t, \dots, \omega_{i,j}^t, \dots, \omega_{n,j}^t\}$. The air quality modelling is considered a problem of classification, which can be realized by various models of neural networks.

Classification can be realized by unsupervised methods (if classes $\omega_{i,j}^t \in \Omega$ are not known) or supervised methods (if classes $\omega_{i,j}^t \in \Omega$ are known). The paper presents the parameters design for air quality modelling. Only those parameters were selected which show low correlation dependences. Therefore, data matrix **P** is designed where $\mathbf{p}_i^t$ vectors characterize the districts $o_i^t \in O$. Further, the paper presents the basic concepts of the Kohonen’s self-organizing feature maps (KSOFM) and Learning Vector Quantization (LVQ) neural networks.

The contribution of the paper lies in the model design for air quality evaluation. The model realizes the advantages of both the unsupervised methods (combination of the KSOFM and K means algorithm) and supervised methods (LVQ neural networks). The final part of the paper includes the analysis of the results and a presentation of the classification of districts $o_i^t \in O$ into classes $\omega_{i,j}^t \in \Omega$.

2 Parameters Design for Air Quality Modelling

Harmful substances in the air represent the parameters of air quality modelling. They are defined as the substances emitted into the external air or generated secondarily in the air which harmfully influent the environment directly, after a physical or chemical transformation or eventually in the interaction with other substances. Except the harmful substances, other components influence the overall air pollution. For example, ozone, solar radiation, the speed or the direction of wind, air humidity and air pressure represent these components. Both the parameters concerning the harmful substances in the air and the meteorological parameters influence air quality development. The interaction of both types of parameters can cause an increase of air pollution and influence the human health this way. The design of the parameters, based on previous correlation analysis and recommendations of notable experts, can be realized as presented in Table 1.

Table 1 Parameters design for air quality modelling

Parameters	
Harmful substances	x_1= SO_2, SO_2 is sulphur dioxide.
	x_2= O_3, O_3 is ozone.
	x_3= NO, NO_2 (NO_x) are nitrogen oxides.
	x_4= CO, CO is carbon monoxide.
	x_5= PM_{10}, PM_{10} is particulate matter (dust).
Meteorological	x_6= SW, SW is the speed of wind.
	x_7= DW, DW is the direction of wind.
	x_8= T_3, T_3 is the temperature 3 meters above the Earth's surface.
	x_9= RH, RH is the relative air humidity.
	x_{10}= AP, AP is air pressure.
	x_{11}= SR, SR is solar radiation.

Based on the presented facts, the following data matrix **P** can be designed

$$\mathbf{P} = \begin{array}{c|cccccc:c}
 & x_1^t & \dots & x_k^t & \dots & x_m^t & & \omega_{i,j}^t \\
\hline
o_1^t & x_{1,1}^t & \dots & x_{1,k}^t & \dots & x_{1,m}^t & & \omega_{1,j}^t \\
\dots & \dots & \dots & \dots & \dots & \dots & & \dots \\
o_i^t & x_{i,1}^t & \dots & x_{i,k}^t & \dots & x_{i,m}^t & & \omega_{i,j}^t \\
\dots & \dots & \dots & \dots & \dots & \dots & & \dots \\
o_n^t & x_{n,1}^t & \dots & x_{n,k}^t & \dots & x_{n,m}^t & & \omega_{n,j}^t
\end{array},$$

where: - $o_i^t \in O$, $O=\{o_1^t, o_2^t, \dots, o_i^t, \dots, o_n^t\}$ are objects (districts) in time t,
- x_k^t is the k-th parameter in time t,
- $x_{i,k}^t$ is the value of the parameter x_k^t for the i-th object $o_i^t \in O$,
- $\omega_{i,j}^t$ is the j-th class assigned to the i-th object $o_i^t \in O$,
- $\mathbf{p}_i^t = (x_{i,1}^t, x_{i,2}^t, \dots, x_{i,k}^t, \dots, x_{i,m}^t)$ is the i-th pattern,
- $\mathbf{x}^t = (x_1^t, x_2^t, \dots, x_k^t, \dots, x_m^t)$ is the parameters vector.

The air quality evaluation is based on the results of weight concentrations measures of substances in the air (Table 2). The evaluation takes the possible influence of human health into account [1]. New limits specified in Government Order of the Czech Republic No: 350/2002 Coll. (No: 429/2005 Coll.) which sets the limits of pollutants, the conditions and the procedure of air quality's monitoring, evaluation and management. These limits are set for health protection, vegetation, and ecosystems protection separately. The dispersion conditions depend on the horizontal and vertical airflow especially [1] (Table 3).

Table 2 Air quality evaluation

Air quality	SO_2	NO_2	CO	O_3	PM_{10}
	1h [μg.m^{-3}]		8h [μg.m^{-3}]	1h [μg.m^{-3}]	
Very good	0-25	0-25	0-1.10^3	0-33	0-15
Good	25-50	25-50	1000-2000	33-65	15-30
Favourable	50-120	50-100	2000-4000	65-120	30-50
Satisfactory	120-250	100-200	4000-10000	120-180	50-70
Bad	250-500	200-400	10000-30000	180-240	70-150
Very bad	500-	400-	30000-	240-	150-

Table 3 Dispersion conditions

Dispersion conditions	Characteristics
Good	There is no trap layer in the height up to (1000-1500) meters above the ground that could limit the dispersion of harmful substances.
Slightly unfavourable	A trap layer limits the dispersion of harmful substances depending on the strength of wind. Yet, it does not match both the unfavourable and good dispersion conditions.
Unfavourable	The state of impossible dispersion of admixtures in the atmosphere when the limits of pollutants exceed significantly in a long time. This state corresponds to the thick trap layer in the height up to 1000 meters above the ground in combination with weak or no airflow.

3 Model Design for the Classification of Air Quality Development

Modelling air quality represents a classification problem. It is generally possible to define it this way:

Let F(**x**) be a function defined on a set A, which assigns picture $\hat{x}$ (the value of the function from a set B) to each element $\mathbf{x} \in A$, $\hat{x} = F(\mathbf{x}) \in B$, $F: A \to B$. A problem defined this way is possible to model using unsupervised methods (if classes $\omega_{i,j}^t \in \Omega$ are not known). The districts in the city of Pardubice, Czech Republic, (Fig. 1) have no class $\omega_{i,j}^t \in \Omega$ assigned.

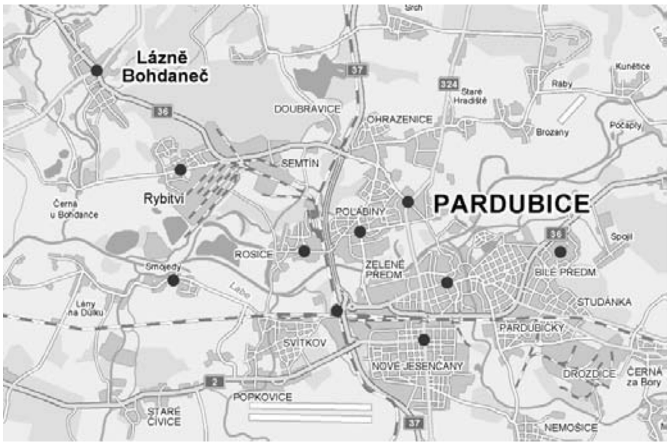


Fig. 1 The map of the districts (black points)

However, the descriptions of classes $\omega_{i,j}^t \in \Omega$ are known (Table 2, Table 3). Therefore, it is suitable to realize the modelling of air quality by unsupervised methods. Data pre-processing is carried out by means of data standardization. Thereby, the dependency on units is eliminated. Based on the analysis presented in [4] the combination of KSOFM and K-means algorithm is a suitable unsupervised method for air quality modelling. The LVQ neural networks use its results as the inputs in order to get a good generalization ability of the model. Model for classification objects $o_i^t \in O$ into classes $\omega_{i,j}^t \in \Omega$ is presented in Fig. 2.

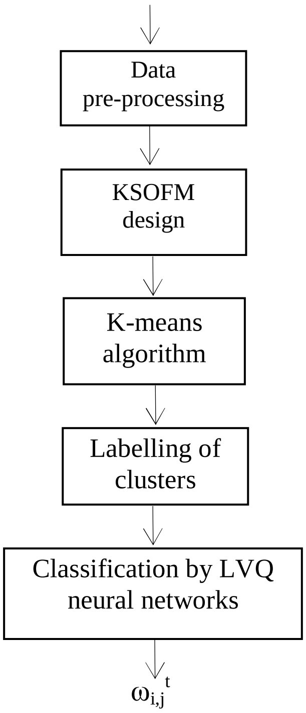


Fig. 2 Model for classification of objects $o_i^t \in O$ into classes $\omega_{i,j}^t \in \Omega$

The KSOFM [4] are based on competitive learning strategy. The input layer serves the distribution of the input patterns $\mathbf{p}_i^t$, i=1,2, … ,n. The neurons in the competitive layer serve as representatives (Codebook Vectors), and they are organized into topological structure (most often as a two-dimensional grid, Fig. 3), which designates the neighbouring network neurons.

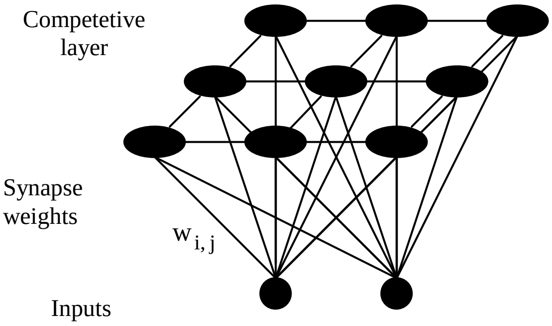


Fig. 3 Example of Kohonen's self-organizing feature map

First, the distances d_j are computed between pattern $\mathbf{p}_i^t$ and synapse weights $\mathbf{w}_{i,j}$ of all neurons in the competitive layer according to the relation

$$d_j = \sum_{i=1}^{n} (\mathbf{p}_i{}^t - \mathbf{w}_{i,j})^2 , \tag{1}$$

where j goes over s neurons of competitive layer, j=1,2, … ,s, $\mathbf{p}_i^t$ is the i-th pattern, i=1,2, … ,n, $\mathbf{w}_{i,j}$ are synapse weights. The winning neuron j* (Best Matching Unit, BMU) is chosen, for which the distance d_j from the given pattern $\mathbf{p}_i^t$ is minimum. Best Matching Unit and neighbourhood are given in Fig. 4

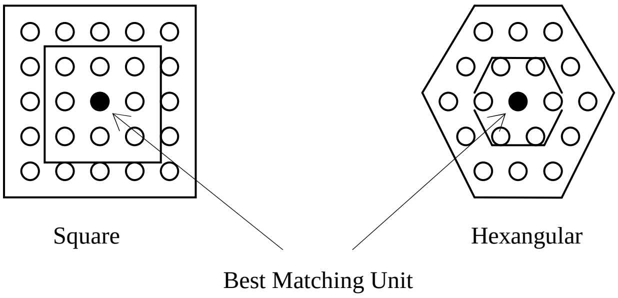


Fig. 4 Activity of the neurons and neighbourhood

The output of this neuron is active, while the outputs of other neurons are inactive. The aim of the KSOFM learning is to approximate the probability density of the real input vectors $\mathbf{p}_i^t \in R^n$ by the finite number of representatives $\mathbf{w}_{i,j} \in R^n$, where j=1,2, … ,s. When the representatives $\mathbf{w}_{i,j}$ are identified, the representative $\mathbf{w}_{i,j*}$ of the BMU is assigned to each vector $\mathbf{p}_i^t$. In the learning process of the KSOFM, it is necessary to define the concept of neighbourhood function, which determines the range of co-operation among the neurons, i.e. how many representatives $\mathbf{w}_{i,j}$ in the neighbourhood of the BMU will be adapted, and to what degree. Activity of the neurons and neighbourhood are shown in Fig. 5.

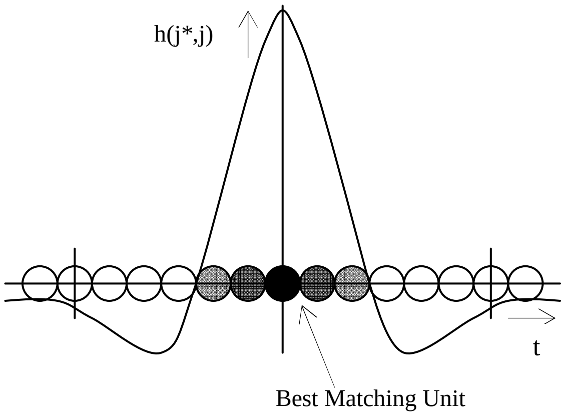


Fig. 5 Example of Best Matching Unit and neighbourhood

Gaussian neighbourhood function is in common use, which is defined as

$$h(j^*,j) = e^{\left(-\frac{d_E^2(j^*,j)}{\lambda^2(t)}\right)}, \qquad (2)$$

where h(j*,j) is neighbourhood function, $d^2_E(j^*,j)$ is Euclidean distance of neurons j* and j in the grid, λ(t) is the size of the neighbourhood in time t'. After the BMUs are found, the adaptation of synapse weights $\mathbf{w}_{i,j}$ follows. The principle of the sequential learning algorithm [4] is the fact, that the representatives $\mathbf{w}_{i,j*}$ of the BMU and its topological neighbours move towards the actual input vector $\mathbf{p}_i^t$ according to the relation

$$\mathbf{w}_{i,j}(t'+1) = \mathbf{w}_{i,j}(t') + \eta(t') \times h(j^*,j) \times (\mathbf{p}_i{}^t(t') - \mathbf{w}_{i,j}(t')) \text{ ,} (3)$$

where η(t')∈(0,1) is the learning rate. The batch-learning algorithm of the KSOFM [4] is a variant of the sequential algorithm. The difference consists in the fact that the whole training set passes through the KSOFM only once, and only then the synapse weights $\mathbf{w}_{i,j}$ are adapted. The adaptation is realized by replacing the representative $\mathbf{w}_{i,j}$ with the weighted average of the input vectors $\mathbf{p}_i^t$.

In [4] there are presented several versions of learning which refers to the structures of LVQ1, LVQ2, LVQ3 and OLVQ1 (Optimized Learning Vector Quantization) neural networks. They differ in the process of searching for the optimum boundaries between classes $\omega_{i,j}{}^t \in \Omega$. The LVQ neural networks (Fig. 6) are the supervised versions of the KSOFM.

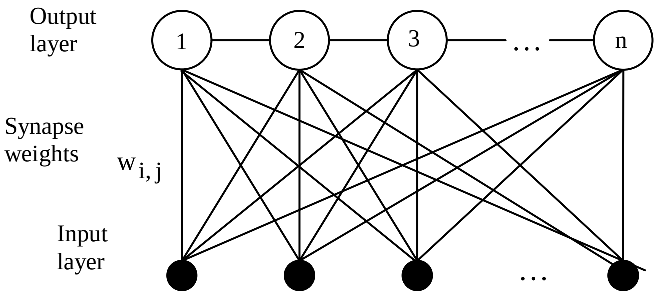


Fig. 6 Learning Vector Quantization neural network

Let there is a LVQ1 neural network and a known number of classes $\omega_{i,j}{}^t \in \Omega$. Classes $\omega_{i,j}{}^t \in \Omega$ are assigned to all patterns $\mathbf{p}_i^t$ in the process of the LVQ initialization. Then, the goal of the learning process is finding the winning neuron j*. The difference to the KSOFM consists in the fact that the process of learning finishes if $\mathbf{p}_i^t$ and $\mathbf{w}_{i,j*}$ belong to the same class $\omega_{i,j}{}^t \in \Omega$.

Further, let the input vector $\mathbf{p}_i^t$ belong to the class $\omega_{i,p}{}^t$ and its representative $\mathbf{w}_{i,j*}$ is a centre of the class $\omega_{i,q}{}^t$. In the process of learning only the synapse weights $\mathbf{w}_{i,j*}(t)$ are adapted as follows

$$\mathbf{w}_{i,j*}(t'+1) = \mathbf{w}_{i,j*}(t') + \eta(t') \times (\mathbf{p}_i^t(t') - \mathbf{w}_{i,j*}(t')), \qquad (4)$$

if $\mathbf{p}_i^t(t')$ and $\mathbf{w}_{i,j*}(t')$ belong to the same class, $\omega_{i,q}{}^t = \omega_{i,p}{}^t$,

$$\mathbf{w}_{i,j*}(t'+1) = \mathbf{w}_{i,j*}(t) - \eta(t') \times (\mathbf{p}_i^t(t') - \mathbf{w}_{i,j*}(t')), \qquad (5)$$

if $\mathbf{p}_i^t(t')$ and $\mathbf{w}_{i,j*}(t')$ belong to different classes, $\omega_{i,q}{}^t \neq \omega_{i,p}{}^t$,

$$\mathbf{w}_{i,j}(t'+1) = \mathbf{w}_{i,j}(t') \text{ for } j \neq j^*, j=1,2, \dots ,M. \qquad (6)$$

The OLVQ1 neural network represents an optimized version of the LVQ1 neural network where an individual learning rate $\eta_{j*}(t)$ is assigned to each $\mathbf{w}_{i,j*}$

$$\mathbf{w}_{i,j*}(t'+1) = \mathbf{w}_{i,j*}(t') + \eta_{j*}(t') \times (\mathbf{p}_i^t(t') - \mathbf{w}_{i,j*}(t')), \qquad (7)$$

if $\mathbf{p}_i^t(t')$ and $\mathbf{w}_{i,j*}(t')$ belong to the same class, $\omega_{i,q}{}^t = \omega_{i,p}{}^t$,

$$\mathbf{w}_{i,j*}(t'+1) = \mathbf{w}_{i,j*}(t) - \eta_{j*}(t') \times (\mathbf{p}_i^t(t') - \mathbf{w}_{i,j*}(t')), \qquad (8)$$

if $\mathbf{p}_i^t(t')$ and $\mathbf{w}_{i,j*}(t')$ belong to different classes, $\omega_{i,q}{}^t \neq \omega_{i,p}{}^t$,

$$\mathbf{w}_{i,j}(t'+1) = \mathbf{w}_{i,j}(t') \text{ for } j \neq j^*, j=1,2, \dots ,M. \qquad (9)$$

Within the process of the LVQ2 neural network's learning two codebook vectors $\mathbf{w}_{i,j}$ (a centre of the class $\omega_{i,r}{}^{t}$) and $\mathbf{w}_{i,k}$ (a centre of the class $\omega_{i,s}{}^{t}$), which are the nearest neighbours to $\mathbf{p}_i{}^{t}$ are updated simultaneously

$$\mathbf{w}_{i,j}(t'+1) = \mathbf{w}_{i,j}(t') + \eta(t') \times (\mathbf{p}_i{}^{t}(t') - \mathbf{w}_{i,j}(t')), \qquad (10)$$

if $\mathbf{p}_i{}^{t}(t')$ and $\mathbf{w}_{i,k}(t')$ belong to the same class,
moreover, $\mathbf{p}_i{}^{t}(t')$ and $\mathbf{w}_{i,j}(t')$ belong to different classes,
$\omega_{i,s}{}^{t} = \omega_{i,p}{}^{t} \neq \omega_{i,r}{}^{t}$,

$$\mathbf{w}_{i,k}(t'+1) = \mathbf{w}_{i,k}(t) - \eta(t') \times (\mathbf{p}_i{}^{t}(t') - \mathbf{w}_{i,k}(t')), \qquad (11)$$

if $\mathbf{p}_i{}^{t}(t')$ and $\mathbf{w}_{i,j}(t')$ belong to the same class,
moreover, $\mathbf{p}_i{}^{t}(t')$ and $\mathbf{w}_{i,k}(t')$ belong to different classes,
$\omega_{i,s}{}^{t} \neq \omega_{i,p}{}^{t} = \omega_{i,r}{}^{t}$.

The learning algorithm of the LVQ3 neural network ensures that $\mathbf{w}_{i,l}$ (a centre of the class $\omega_{i,u}{}^{t}$) continues approximating the class distributions

$$\mathbf{w}_{i,j}(t'+1) = \mathbf{w}_{i,j}(t') + \eta(t') \times (\mathbf{p}_i{}^{t}(t') - \mathbf{w}_{i,j}(t')), \qquad (12)$$

if $\mathbf{p}_i{}^{t}(t')$ and $\mathbf{w}_{i,k}(t')$ belong to the same class,
moreover, $\mathbf{p}_i{}^{t}(t')$ and $\mathbf{w}_{i,j}(t')$ belong to different classes,
$\omega_{i,s}{}^{t} = \omega_{i,p}{}^{t} \neq \omega_{i,r}{}^{t}$,

$$\mathbf{w}_{i,k}(t'+1) = \mathbf{w}_{i,k}(t) - \eta(t') \times (\mathbf{p}_i{}^{t}(t') - \mathbf{w}_{i,k}(t')), \qquad (13)$$

if $\mathbf{p}_i{}^{t}(t')$ and $\mathbf{w}_{i,j}(t')$ belong to the same class,
moreover, $\mathbf{p}_i{}^{t}(t')$ and $\mathbf{w}_{i,k}(t')$ belong to different classes,
$\omega_{i,s}{}^{t} \neq \omega_{i,p}{}^{t} = \omega_{i,r}{}^{t}$,

$$\mathbf{w}_{i,l}(t'+1) = \mathbf{w}_{i,l}(t) - \delta \times \eta(t') \times (\mathbf{p}_i{}^{t}(t') - \mathbf{w}_{i,l}(t')), \qquad (14)$$

if $\mathbf{p}_i{}^{t}(t')$,$\mathbf{w}_{i,k}(t')$ and $\mathbf{w}_{i,l}(t')$ belong to the same class, $l \in (j,k)$, $\omega_{i,p}{}^{t} = \omega_{i,s}{}^{t} = \omega_{i,u}{}^{t}$ and δ is parameter [4].

4 Analysis of the Results

The input parameters of the designed KSOFM are based on a number of experiments and are specified in Table 4.

Table 4 Input parameters of the KSOFM

Parameter	Init.	h(j*,j)	Initial λ(t')	Final λ(t')	η(t')	Epochs
Value	Linear	Bubble	10	1	0.01	10000

The structure KSOFM (unsupervised method, if classes $\omega_{i,j}{}^{t} \in \Omega$ are not known) and LVQ neural network (supervised method, if classes $\omega_{i,j}{}^{t} \in \Omega$ are known) are shown in Fig. 7.

The goal of the air quality modelling is the classification of the districts $o_i{}^{t} \in O$ in time t into classes $\omega_{i,j}{}^{t} \in \Omega$ according to air quality.

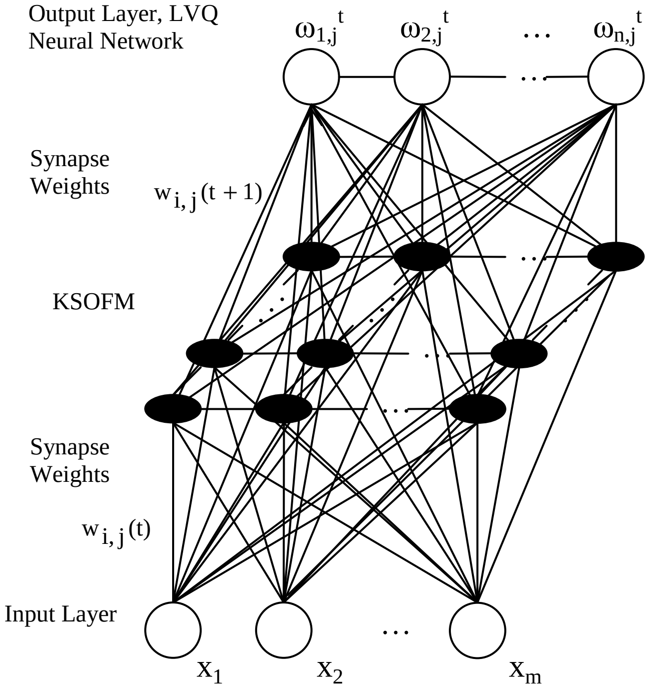
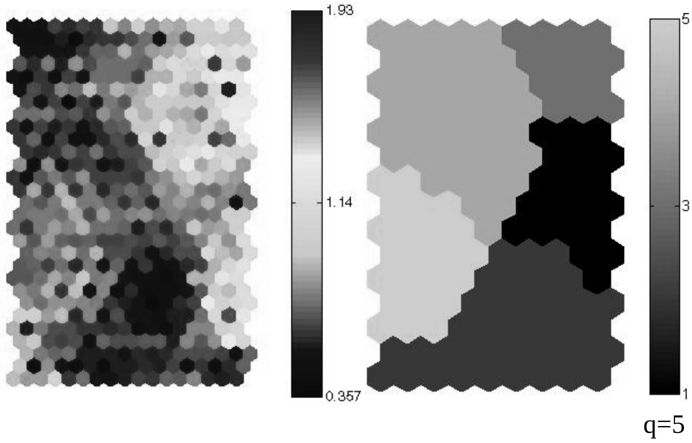


Fig. 7 Structure KSOFM and LVQ neural network for classification

Using the KSOFM as such can detect the data structure is presented in Fig. 8a. The U-matrix shows square Euclidean distances d between representatives $\mathbf{w}_{i,j}$. The K-means algorithm can be applied to the adapted KSOFM in order to find clusters as presented in Fig. 8b.



8a 8b

Fig. 8a U-matrix of square Euclidean distances
Fig. 8b Clustering of the KSOFM by K-means algorithm

The quality of the KSOFM results can be measured with quantization and topographic errors. The quantization error (QE) is computed as an Euclidean distance of the input vector $\mathbf{p}_i^t$ and the representative $\mathbf{w}_{i,j*}$ of its BMU. The topographic error (TE) is a quotient of all the input vectors for which the first and second BMUs are neighbours in the map. The TE measures the rate of the KSOFM topology preservation. We achieved the values of QE=1.6795 and TE=0.034722.

The K-means algorithm belongs to the non-hierarchical algorithms of cluster analysis, where patterns $\mathbf{p}_1^t,\mathbf{p}_2^t, \dots ,\mathbf{p}_i^t, \dots ,\mathbf{p}_n^t$ (n=720) are assigned to clusters $c_i \in C$, $C=\{c_1^t,c_2^t, \dots ,c_i^t, \dots ,c_q^t\}$. The number of clusters q=5 is determined by indexes evaluating the quality of clustering. The following clustering quality indexes are defined, separation index (S), Xie-Beni index (XB) and Dunn index (DI) [5]. On the contrary, to the S index and the XB index, the DI index uses a hard partition clustering results. The DI index can be defined as follows:

$$\mathrm{DI}(i) = \min_{i \in C}\left\{ \min_{k \in C, k \neq i}\left\{ \frac{d_{\min}(\mathbf{v}_i,\mathbf{v}_k)}{\max\limits_{l \in C}\{d_{\max}(\mathbf{v}_i,\mathbf{v}_k)\}} \right\}\right\}, \qquad (15)$$

where k,l,i are cluster indexes, $\mathbf{v}_i,\mathbf{v}_k$ are centres of the i-th and k-th clusters $c_i,c_k \in C$, $d_{\min}$ is the minimum and $d_{\max}$ is the maximum Euclidean distance. Many experiments were carried out. The mean values of the DI for the designed clustering method are presented in Fig. 9.

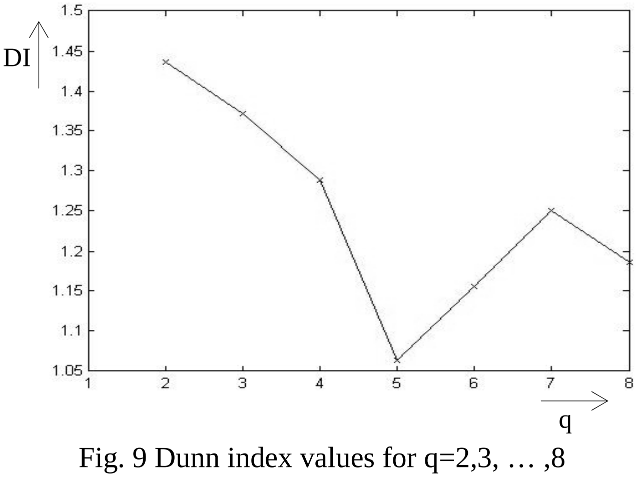


Fig. 9 Dunn index values for q=2,3, … ,8

Clustering [6] process is realized in two levels. In the first level, n objects are reduced to representatives $\mathbf{w}_1,\mathbf{w}_2, \dots ,\mathbf{w}_p$ by the KSOFM; the p representatives are clustered into q clusters. Clusters $C=\{c_1^t,c_2^t, \dots ,c_i^t, \dots ,c_q^t\}$ can be interpreted on the basis of parameters values $\mathbf{p}_i^t=(x_{i,1}^t,x_{i,2}^t, \dots ,x_{i,k}^t, \dots ,x_{i,m}^t)$ for the representatives of the KSOFM (Fig. 10 to Fig. 12). The interpretation of parameters results from the air quality and dispersion conditions are defined in [1].

As an example, the parameters $x_1^t,x_2^t, \dots ,x_{11}^t$ for the crossroad Palacha-Pichlova are presented in appendix.

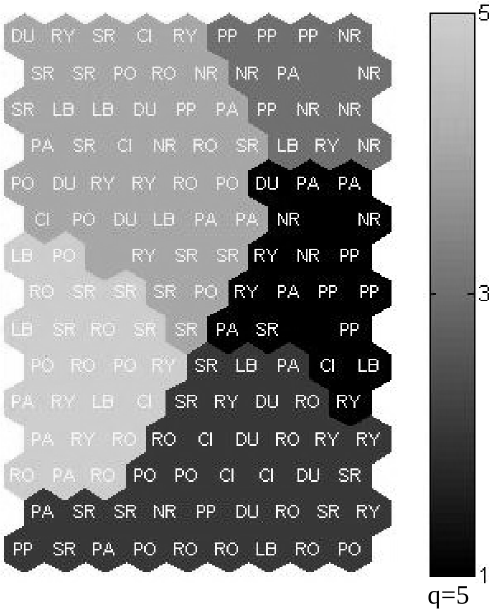

DU RY SR CI RY PP PP PP NR
SR SR PO RO NR NR PA NR
SR LB LB DU PP PA PP NR NR
PA SR CI NR RO SR LB RY NR
PO DU RY RY RO PO DU PA PA
CI PO DU LB PA PA NR NR
LB PO RY SR SR RY NR PP
RO SR SR SR PO RY PA PP PP
LB SR RO SR SR PA SR PP
PO RO PO RY SR LB PA CI LB
PA RY LB CI SR RY DU RO RY
PA RY RO RO CI DU RO RY RY
RO PA RO PO PO CI CI DU SR
PA SR SR NR PP DU RO SR RY
PP SR PA PO RO RO LB RO PO
5
3
1
q=5

Fig. 10 Clustering of the KSOFM by K-means algorithm (districts)

Legend: Bus stops: (Cihelna (CI), Dubina (DU), Polabiny (PO), Rosice (RO), Rybitví (RY), Srnojedy (SR)), crossroads: (Palacha-Pichlova (PP), Náměstí Republiky (NR)), Lázně Bohdaneč (LB), chemical factory of Paramo (PA).

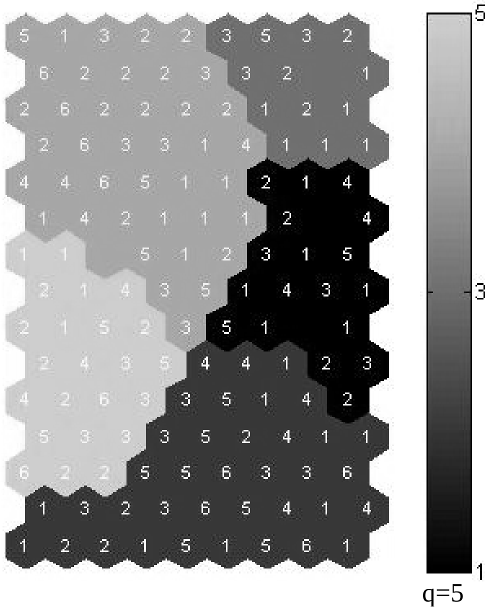


Fig. 11 Clustering of the KSOFM by K-means algorithm (years)

Legend: Years 2001 (1), 2002 (2), 2003 (3), 2004 (4), 2005 (5), 2006 (6).

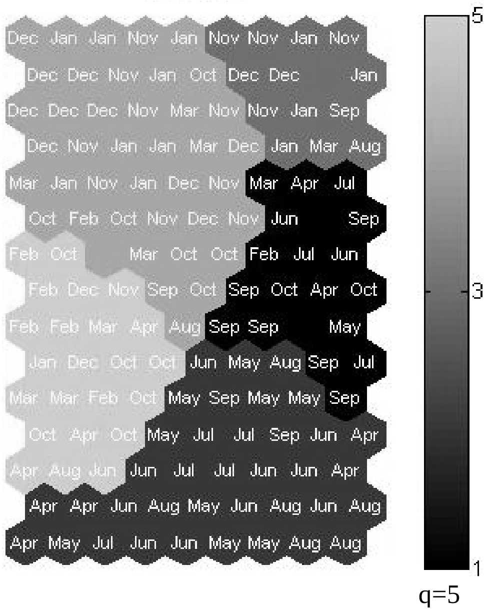


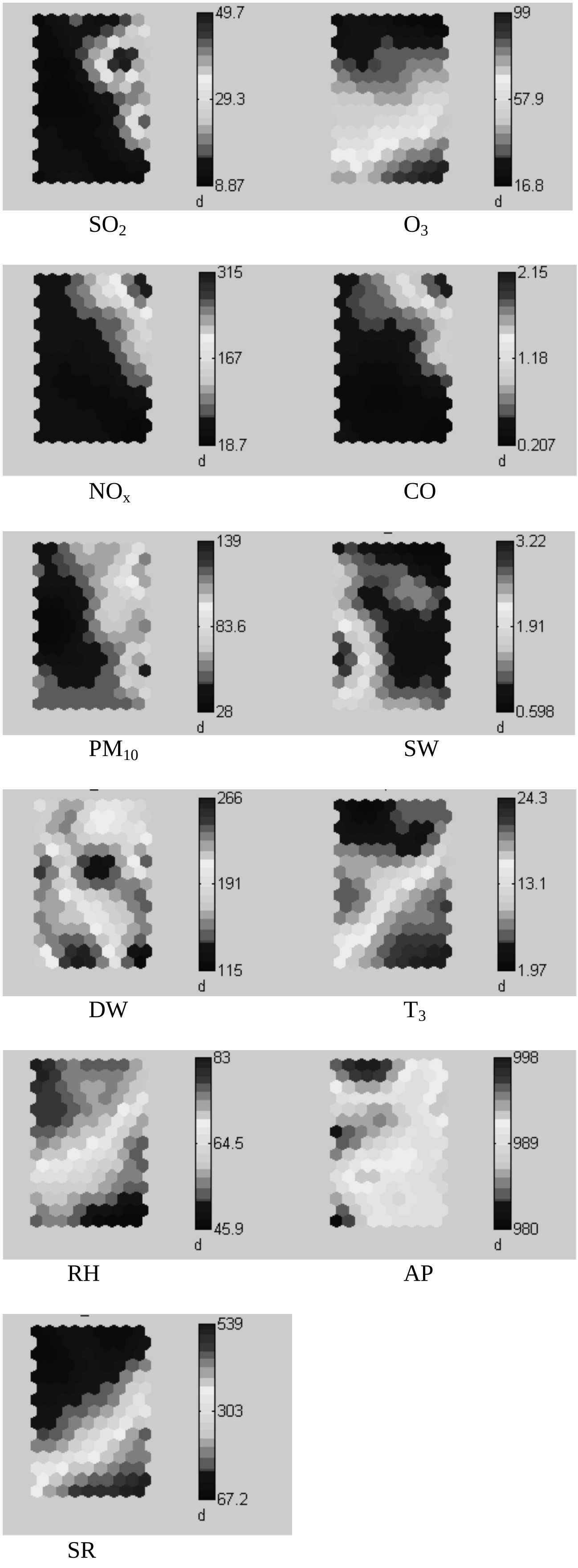
Fig. 12 Clustering of the KSOFM by K-means algorithm (months)

Legend: Months: January (Jan), February (Feb), March (Mar), April (Apr), May (May), June (Jun), July (Jul), August (Aug), September (Sep), October (Oct), November (Nov), December (Dec).

Characteristics of clusters C={$c_1^t, c_2^t, \dots, c_i^t, \dots, c_q^t$} by the parameters are presented in Table 5. Further, the interpretation of clusters results from the parameters values is shown in Fig. 13.

Table 5 Labelling of clusters with classes according to air quality

Cluster	Parameters of harmful substances in the air and dispersion conditions	$\omega_{i,j}^t$ j=1,2, … ,5
1 ■	Satisfactory quality, slightly unfavourable dispersion conditions, environment dangerous for sensitive people.	$\omega_{i,4}^t$
2 ■	Excellent quality, slightly unfavourable dispersion conditions, very healthy environment.	$\omega_{i,1}^t$
3 ■	Bad quality, unfavourable dispersion conditions, environment dangerous for the whole population.	$\omega_{i,5}^t$
4 ■	Good quality, good dispersion conditions, healthy environment.	$\omega_{i,2}^t$
5 ■	Favourable quality, slightly unfavourable dispersion conditions, acceptable environment.	$\omega_{i,3}^t$



SO_2 O_3

NO_x CO

PM_{10} SW

DW T_3

RH AP

SR

Fig. 13 Values of parameters $x_1^t, x_2^t, \dots, x_{11}^t$ for the KSOFM representatives

The interpretation leads to the labelling of clusters with the classes $\omega_{i,j}^t \in \Omega$. The classes are set based on the air quality (Table 2, Table 3). All clusters are labelled with classes $\omega_{i,j}^t \in \Omega$, j=5, where the class ω_1^t represents the least polluted air and the class ω_5^t represents the most polluted air. The frequencies f of the classes (the classification of the districts $o_i^t \in O$ in time t into the classes $\omega_{i,j}^t \in \Omega$ according to their air quality) by KSOFM are presented in Fig. 14.

The locality and month (season) have major impact on the development of air quality in the city of Pardubice. A general name can be assigned to each of the clusters. They can be called green zones or crossroads as an example. The year has an insignificant influence on the partition of clusters (there are no fluctuations in years). The influence of the month is significant with some clusters, however it is small with the others.

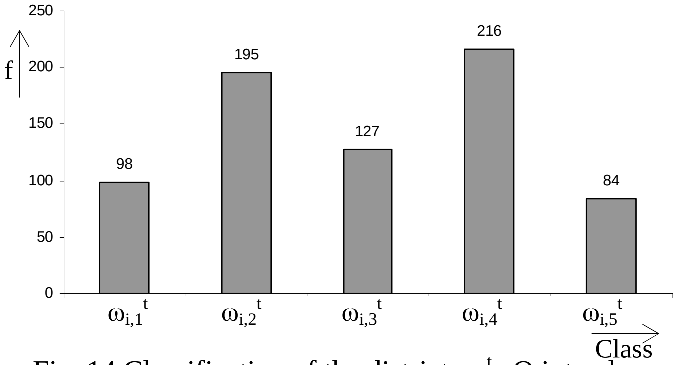


Fig. 14 Classification of the districts $o_i^t \in O$ into classes $\omega_{i,j}^t \in \Omega$ by KSOFM

We designed a number of the KSOFM structures with various input parameters in the process of modelling. The specific characteristic of the KSOFM lies in the fact that it makes possible to realize the representation, which preserves the topology and characteristics of the training set. For this purpose, the neurons are ordered in a regular, mostly two-dimensional or one-dimensional structure. This structure represents the output space, where the distance of neurons is computed as the Euclidean distance of their vectors' coordinates. The projection preserving the topology of the adapted KSOFM has the following important feature. Any pair of patterns $\mathbf{p}_i^t$, which are nearby in the input space, evokes the responses of the KSOFM neurons, which are also nearby in the output space.

Learning Vector Quantization neural networks use these results as their inputs. The dataset is divided into the training and testing set [7] as presented in Fig. 15.

The input parameters of the LVQ neural networks' structure are presented in Table 6, where α is the number of codebook vectors, NN is the number of neighbours used in the K-Nearest Neighbour (KNN) classification, $\beta \in (0,1)$ is the width of the window and $\delta \in (0,1)$ is the stabilizing constant factor. Again, we designed loads of the LVQ1, LVQ2, LVQ3 and OLVQ1 structures with various input parameters. Finally, we obtained the best results with the input parameters presented in Table 7.

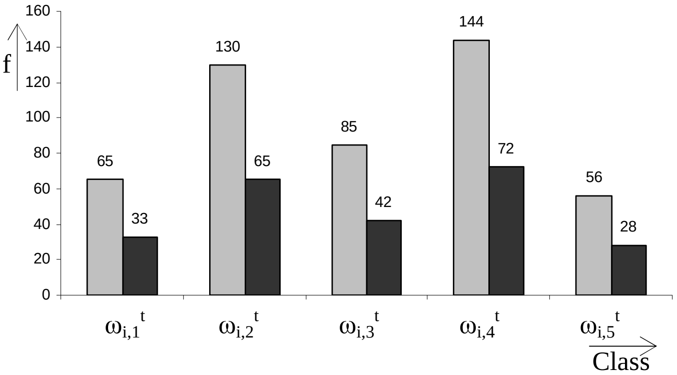


Fig. 15 Frequencies f of municipalities in classes $\omega_{i,j}^t \in \Omega$ in training (grey) and testing set (black)

Table 6 Input parameters of the LVQ neural networks

Structure	α	NN	η(t)	β	δ	Epochs
LVQ1	200	5	0.05	-	-	10000
LVQ2	200	5	0.05	0.3	-	10000
LVQ3	200	5	0.05	0.3	0.1	10000
OLVQ1	200	5	-	-	-	10000

Table 7 Classification accuracy ε[%] on testing data by the LVQ neural networks

	OLVQ1	LVQ1	LVQ2	LVQ3
ε_{max}[%]	88.33	89.17	89.13	91.25
ε_a[%]	86.08	88.27	87.97	89.43
SD[%]	1.22	0.65	1.05	1.45

The LVQ3 neural network has the best results of all the LVQ neural networks concerning the testing set, see Table 7. The LVQ3 neural network has the maximum classification accuracy ε_{max}=91.25[%], the average classification accuracy ε_a=89.43[%] and the standard deviation SD=1.45[%]. We did not obtain better results even after the application of the LVQ2 and LVQ3 training algorithms on the results of the LVQ1 neural network. The frequencies f of the classes in testing set (the classification of the districts $o_i^t \in O$ in time t into the classes $\omega_{i,j}^t \in \Omega$ according to their air quality) by LVQ3 are shown in Fig. 16.

The LVQ neural networks adjust the synapse weights **w** in order to minimize the number of misclassifications coming from the classes $\omega_{i,j}^t \in \Omega$ overlap. The classification problem works with the set of input patterns $\mathbf{p}_i^t$ assigned to one of the classes $\omega_{i,j} \in \Omega$. The classifier chooses one of the classes $\omega_{i,j}^t \in \Omega$ for the given pattern $\mathbf{p}_i^t$. The classification of the pattern $\mathbf{p}_i^t$ is based on the label of its nearest synapse weight **w**, which represents an assignment to a class $\omega_{i,j}{}^t \in \Omega$. Contrary to the problem of vector quantization, it is not important which of the neurons is the winner. What matters is the fact that the winner should belong to one of the neurons representing the correct class $\omega_{i,j}{}^t \in \Omega$. Questionable situations can rise exactly in the areas where the classes neighbour.

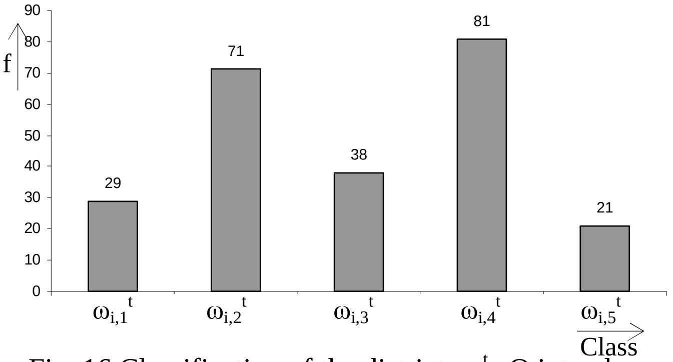


Fig. 16 Classification of the districts $o_i{}^t \in O$ into classes $\omega_{i,j}{}^t \in \Omega$ by LVQ3

The results of the designed model show the possibility of evaluating air quality of the given districts in months and years to come. The visualization of the air quality by the KSOFM makes it possible to monitor the structure of air quality in space and the relations between the designed parameters. Further, the model presents an easier conception of the air quality for the public administration managers. The generalization of the gained knowledge (LVQ neural networks) makes it also possible to classify the districts not involved in the training process.

5 Conclusion

The air quality modelling has been focused on the air quality parameters prediction [8,9] and modelling by multi-agents systems [10,11] so far while classification of the district has been realized only for wind parameters [12,13,14].

Considering the unknown assignment of the districts $o_i{}^t \in O$ to classes $\omega_{i,j}{}^t \in \Omega$, the modelling of air quality was realized by the combination of the KSOFM and K-means algorithm which is a representative of unsupervised methods. This method allows finding of well-separated clusters and their suitable interpretation. The measurements of the air quality parameters were realized by the mobile monitoring system HORIBA. This system cannot classify the measurements into the classes $\omega_{i,j}{}^t \in \Omega$. High correlation dependencies between parameters NO and NO_2 was detected within the process of data pre-processing. The NO_x was used as their representative following this fact. The analysis of results shows that the districts in the city of Pardubice can be classified into five classes. Each of the classes is evaluated with the air quality and dispersion conditions. The air quality can be classified as excellent, good, favourable, satisfactory, bad and very bad. The dispersion conditions can be classified as favourable, slightly unfavourable and unfavourable.

The outputs of the KSOFM are used as the inputs of the LVQ neural networks. The LVQ neural networks structures were designed and studied for the classification of municipalities into classes $\omega_{i,j}{}^t \in \Omega$ due to its high maximum classification accuracy ε_{max}[%] and average classification accuracy ε_a[%] with a low standard deviation SD[%]. The results obtained from the measurement with the mobile monitoring system HORIBA will be verified and made available to the public administration authorities in the future.

The gained results represent the recommendations for the state administration of the city of Pardubice in the field of air quality development. They can also serve as a basis for the municipal crisis management in crises situations. The model design was carried out in Matlab and LVQ_PAK in MS Windows XP operation system.

Future work will be focused on the modelling of parameters for air quality classification of districts into classes according to their pollution. The modelling on the basis of the KSOFMs and intuitionistic fuzzy sets [15,16,17] seems to be suitable. At this time, there are several generalizations of fuzzy set theory for various objectives. Intuitionistic fuzzy sets theory represents one of the generalizations, the notion introduced by K. Atanassov [15]. Therefore, the design of air quality classification will be presented by intuitionistic fuzzy fuzzy relations and their compositions will be designed on the basis of KSOFMs.

Acknowledgement

This work was supported by the scientific research project of Ministry of Environment, Czech Republic under Grant No: SP/4i2/60/07 with title Indicators for Valuation and Modelling of Interactions among Environment, Economics and Social Relations.

References:

[1] *State Policy of Environment in Czech Republic 2004-2010*, Praha, Ministry of Environment, 2004, (in Czech).

[2] V. Olej, *Modelling of Economics Processes on the basis Computational Intelligence*, Scientific Monograph, Hradec Králové, M&V, 2003, (in Slovak).

[3] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd edition, New Jersey, Prentice-Hall, Inc., 1999.

[4] T. Kohonen, *Self-Organizing Maps*, 3rd edition, New York, Springer-Verlag Berlin Heidelberg, 2001.

[5] B. Stein, S. Meyer zu Eissen, F. Wissbrock, On Cluster Validity and the Information Need of Users, *Proc. of the International Conference on Artificial Intelligence and Applications (AIA 03)*, Benalmádena, Spain, 2003, pp.216-221.

[6] V. Olej, P. Hájek, J. Křupka, I. Obršálová, Air Quality Modelling by Kohonen's Neural Network, *Proc. of the the 5th WSEAS International Conference on Environment, Ecosystems and Development (EED'07)*, WSEAS Press, Tenerife, Canary Islands, Spain, 2007.

[7] V. Kvasnička and all., *Introduction to Neural Networks*, Iris, Bratislava, 1997.

[8] M. Kolehmainen, H. Martikainen, T. Hiltunen, J. Ruuskanen, Forecasting Air Quality Parameters Using Hybrid Neural Network Modelling, *Environmental Monitoring and Assessment*, Issue 1, Vol.65, 2000, pp.227-286.

[9] E. Kalapanidis, N. Avouris, Applying Machine Learning Techniques in Air Quality Prediction, *Proc. of the ACAI*, Chania, Grece, 1999, pp.58-64.

[10] E. Kalapanidis, N. Avouris, Air Quality Management using a Multi-Agents Systems, *International Journal of Computer Aided Civil and Infrastructure Engineering*, Issue 2, Vol.17, 2001, pp.119-130.

[11] K. Machová, P. Illiáš, Movement Optimisation of Cooperating Ant Colony: A Study in Agent-based Social Simulation, *Studies in Informatics and Control*, Vol.16, No.4, 2007, pp. 401-412.

[12] M. Z. Boznar, P. Mlakar, Use of Neural Network in the Field of Air Pollution Modelling, *Proc. of the 25th NATO/CCMS Technical Meeting on Air Pollution Modelling and its Application*, Kluwer Academic, New York, 2002, pp.375-382.

[13] Ch. Vongmahadlek, B. Satayopas, Applicability of RAMS for a Simulation to Provide Inputs to an Air Quality Model: Modeling Evaluation and Sensitivity Test, *WSEAS Transactions on Environment and Development*, No.8, Vol.3, 2007, pp.129-138.

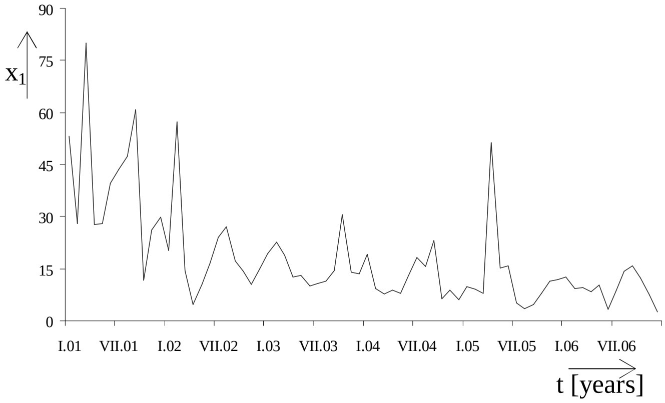
[14] R. San Jose, J. L. Perez, R. M. Gonzalez, Air Quality CFD and Mesoscale Modelling Simulations: Madrid Case Study, *WSEAS Transactions on Environment and Development*, No.10, Vol.2, 2006, pp.1291-1296.

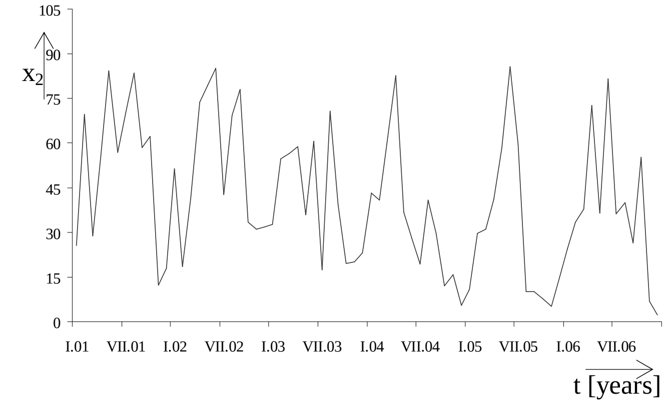
[15] K. Atanassov, Intuitionistic Fuzzy Sets, *Fuzzy Sets and Systems*, No.20, 1986, pp.87-96.

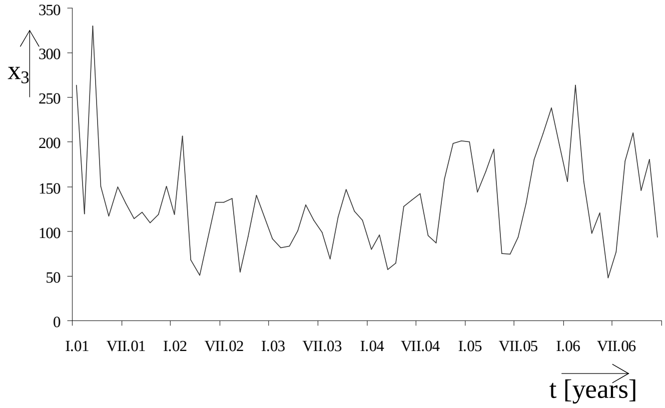
[16] K. Atanassov, *Intuitionistic Fuzzy Sets*, Springer-Verlag Berlin Heidelberg, 1999.

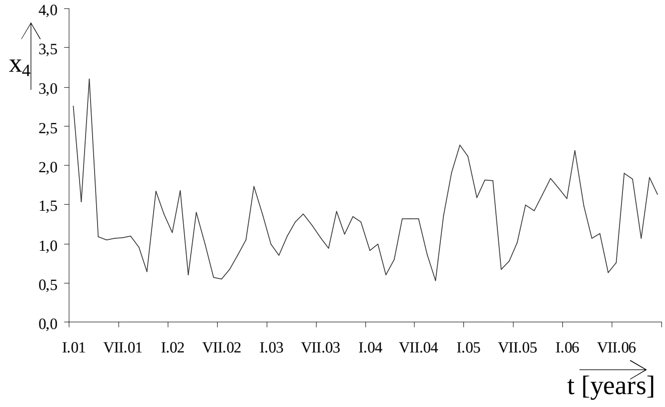
[17] K. Atanassov, Two Theorems for Intuitionistic Fuzzy Sets, *Fuzzy Sets and Systems*, No.110, 2000, 267-269.

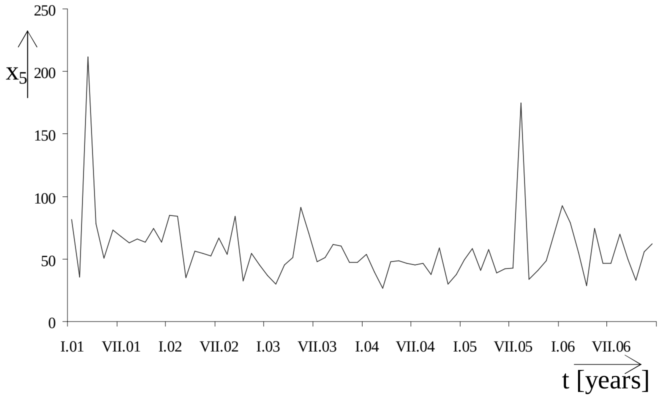
Appendix

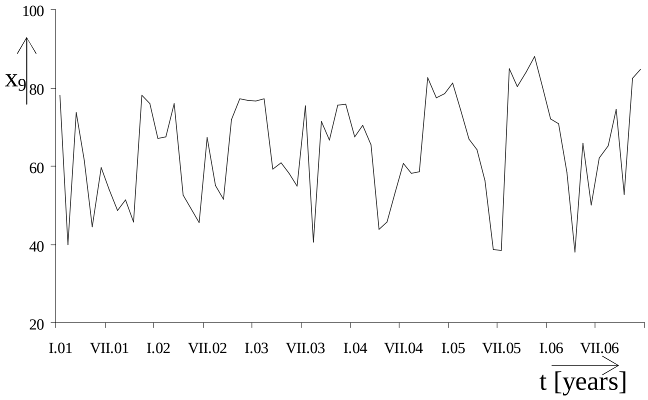


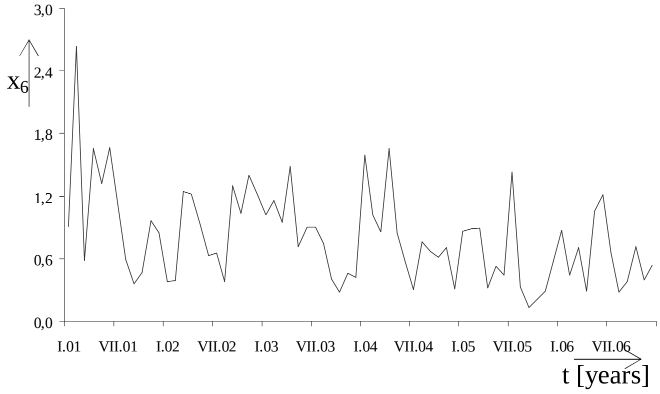


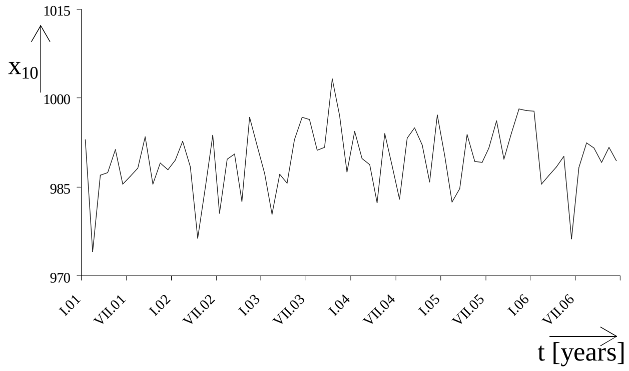


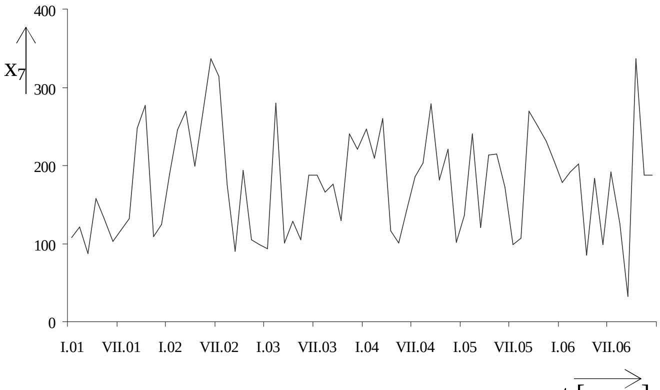


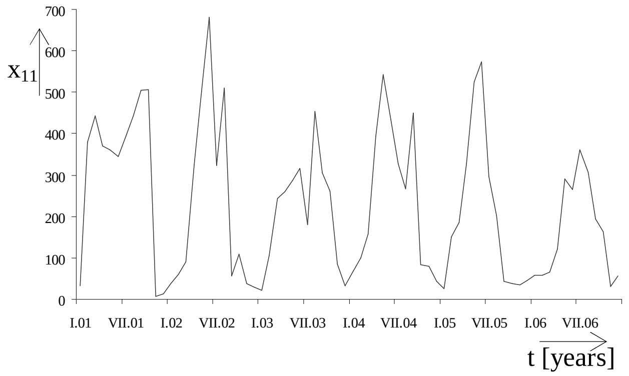

x_5
250
200
150
100
50
0
I.01 VII.01 I.02 VII.02 I.03 VII.03 I.04 VII.04 I.05 VII.05 I.06 VII.06
t [years]

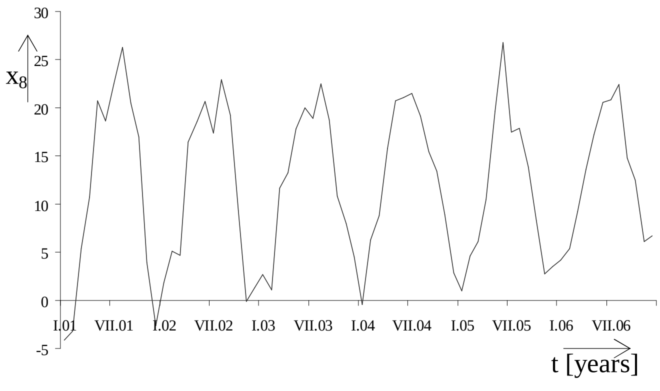

x_9
100
80
60
40
20
I.01 VII.01 I.02 VII.02 I.03 VII.03 I.04 VII.04 I.05 VII.05 I.06 VII.06
t [years]


x_6
3,0
2,4
1,8
1,2
0,6
0,0
I.01 VII.01 I.02 VII.02 I.03 VII.03 I.04 VII.04 I.05 VII.05 I.06 VII.06
t [years]


x_{10}
1015
1000
985
970
I.01 VII.01 I.02 VII.02 I.03 VII.03 I.04 VII.04 I.05 VII.05 I.06 VII.06
t [years]


x_7
400
300
200
100
0
I.01 VII.01 I.02 VII.02 I.03 VII.03 I.04 VII.04 I.05 VII.05 I.06 VII.06
t [years]


x_{11}
700
600
500
400
300
200
100
0
I.01 VII.01 I.02 VII.02 I.03 VII.03 I.04 VII.04 I.05 VII.05 I.06 VII.06
t [years]


x_8
30
25
20
15
10
5
0
-5
I.01 VII.01 I.02 VII.02 I.03 VII.03 I.04 VII.04 I.05 VII.05 I.06 VII.06
t [years]