Air Quality Modelling by Kohonen's Neural Networks

* VLADIMÍR OLEJ, PETR HÁJEK, JIŘÍ KŘUPKA, ** ILONA OBRŠÁLOVÁ
* Institute of System Engineering and Informatics
** Institute of Public Administration and Law
Faculty of Economics and Administration
University of Pardubice
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC

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.

Key-Words: - Air quality, Modelling, Kohonen's Self-organizing Feature Maps, K-means algorithm, 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 be direct or as 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 increase.

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 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 supervised methods (if classes $\omega_{i,j}^t \in \Omega$ are known) or unsupervised methods (if classes $\omega_{i,j}^t \in \Omega$ are not 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 vectors $\mathbf{p}_i^t$ characterizes the districts $o_i^t \in O$. Further, the paper presents the basic concepts of the Kohonen's Self-organizing Feature Maps (KSOFM) which are intended for unsupervised learning.

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

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 secondary in the air which harmfully influent the environment directly, after the 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 the air quality development. The interaction of both types of parameters can cause the increase of air pollution and influence the human health this way. The design of 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= SV, SV is the speed of wind.
	x_7= SmV, SmV is the direction of wind.
	x_8= T_3, T_3 is the temperature 3 meters above the Earth's surface.
	x_9= RV, RV is a relative air humidity.
	x_{10}= T, T is air pressure.
	x_{11}= SZ, SZ is solar radiation.

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

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

where $o_i^t \in O$, $O=\{o_1^t, o_2^t, \ldots, o_i^t, \ldots, o_n^t\}$ 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 $o_i^t \in O$, $\mathbf{p}_i^t = (x_{i,1}^t, x_{i,2}^t, \ldots, x_{i,k}^t, \ldots, x_{i,m}^t)$ is the i-th pattern, $\mathbf{x}^t = (x_1^t, x_2^t, \ldots, x_k^t, \ldots, x_m^t)$ 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 the government order of the Czech Republic No: 350/2002 Coll. (No: 429/2005 Coll.) which set the limits of pollutants, the conditions and the procedure of air quality's monitoring, evaluation and management. These limits are set for health protection and vegetation and ecosystems protection separately. The dispersion conditions depend on the horizontal and vertical air flow especially [1] (Table3).

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	There is a trap layer that 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 a weak or no air flow.

3 Model Design for the Classification of Air Quality Development

The model realizes the air quality modelling. Data pre-processing makes the suitable environmental interpretation of results possible. The KSOFM assign objects to clusters. Subsequently, the clusters are labelled with classes $\omega_{i,j}^t \in \Omega$.

Modelling air quality represents a classification problem. It is generally possible to define it this way: Let F($\mathbf{x}$) 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 $\mathbf{x} \in A$, $\hat{x} = F(\mathbf{x}) \in B$, $F: A \rightarrow B$. The problem defined this way is possible to model by unsupervised methods (if classes $\omega_{i,j}^t \in \Omega$ are not known). The districts in the city of Pardubice (Czech Republic) have no class $\omega_{i,j}^t \in \Omega$ assigned. However, the descriptions of classes $\omega_{i,j}^t \in \Omega$ are known (Table2, Table3). 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. Model for classification objects $o_i^t \in O$ into classes $\omega_{i,j}^t \in \Omega$ is presented in Fig. 1.

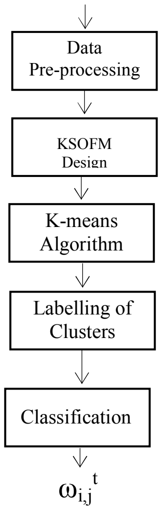


Fig. 1 Model for classification of objects o_i^t into classes $\omega_{i,j}^t$

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 the representatives (Codebook Vectors), and they are organized into topological structure (most often as a two-dimensional grid), which designates the neighbouring network neurons. 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

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

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. 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 cooperation 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. 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

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

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 $\mathbf{p}_i^t$.

4 Analysis of the Results

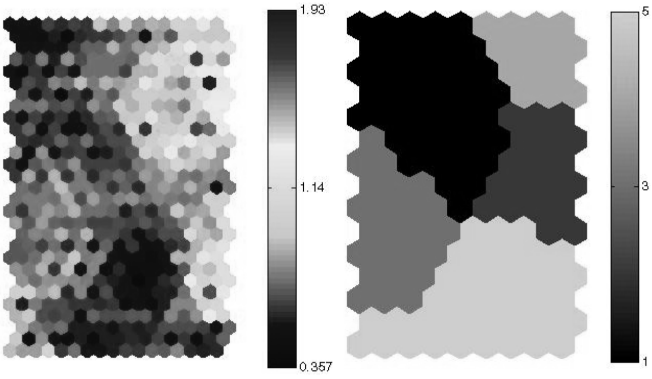
The goal of the air quality modelling is the classification of the districts $o_i^t \in O$ in time t into classes $\omega_{i,j}^t \in \Omega$ according to air quality. The input parameters of the designed KSOFM are based on a number of experiments and are specified in Table 4. Using the KSOFM as such can detect the data structure (Fig 2a). The U-matrix shows the square Euclidean distances d between representatives $\mathbf{w}_{i,j}$.

Table 4 Input parameters of the KSOFM

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

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 can be applied to the adapted KSOFM in order to find clusters as presented in Fig. 2b.

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 $c_1^t, c_2^t, \dots, c_i^t, \dots, c_q^t$. The number of clusters q=5 is determined by indexes evaluating the quality of clustering [5].



2a 2b

Fig. 2a U-matrix of square Euclidean distances, Fig. 2b Clustering of the KSOFM by K-means algorithm

Clustering 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 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. 3 to Fig. 5). The interpretation of parameters results from the air quality and dispersion conditions [1].

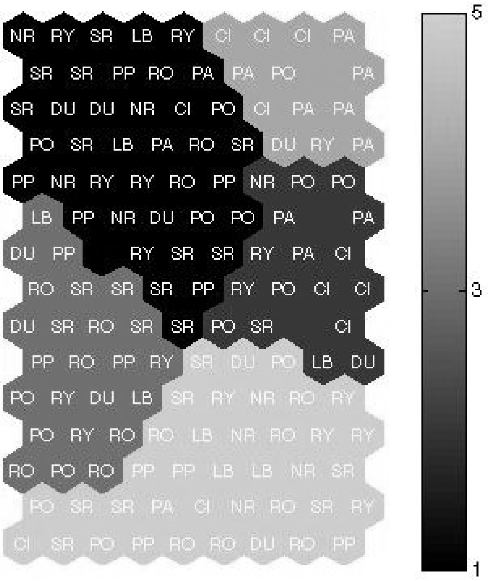


Fig. 3 Clustering of the KSOFM by K-means algorithm (districts)

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).

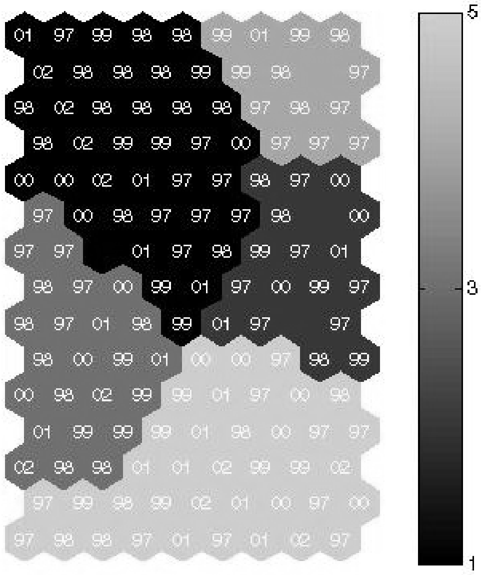


Fig. 4 Clustering of the KSOFM by K-means algorithm (years)

Legend: Years 1997 (97), 1998 (98), 1999 (99), 2000 (00), 2001 (01), 2002 (02).

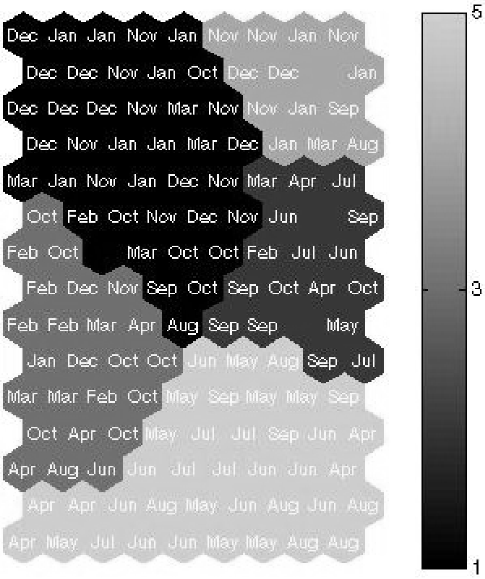


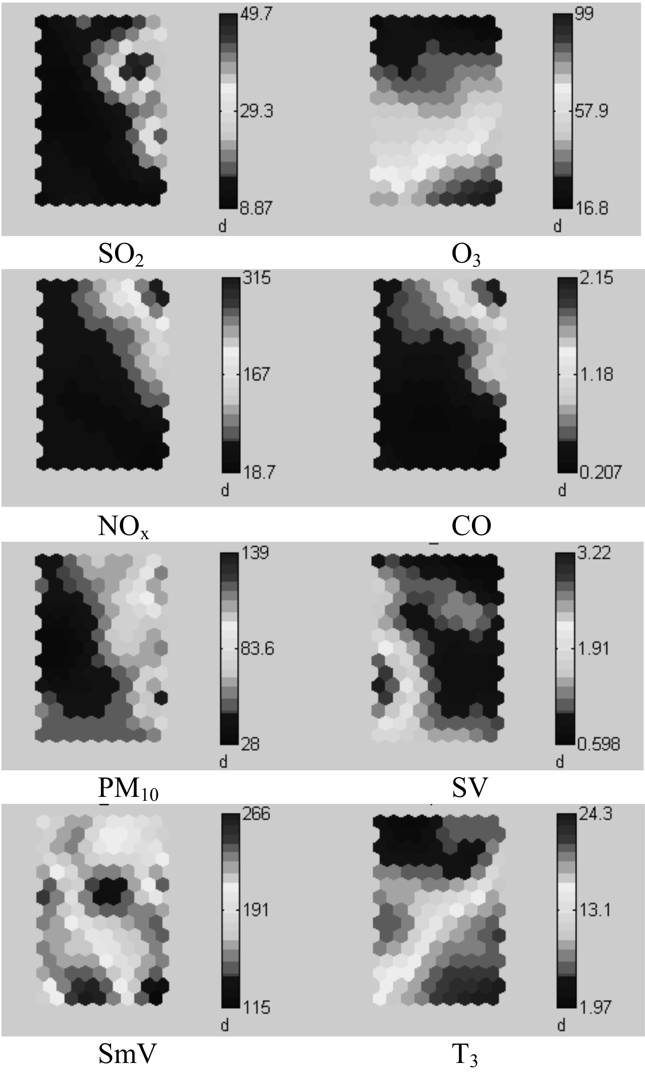
Fig. 5 Clustering of the KSOFM by K-means algorithm (months)

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 by the parameters are presented in Table 5. Further, the interpretation of clusters results from the parameters values (Fig. 6).

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 ■	Excellent quality, slightly unfavourable dispersion conditions, very healthy environment.	$\omega_{i,2}^t$
2 ■	Favourable quality, slightly unfavourable dispersion conditions, acceptable environment.	$\omega_{i,3}^t$
3 ▨	Good quality, good dispersion conditions, healthy environment.	$\omega_{i,1}^t$
4 ▨	Bad quality, unfavourable dispersion conditions, environment dangerous for the whole population.	$\omega_{i,4}^t$
5 □	Satisfactory quality, , slightly unfavourable dispersion conditions, environment dangerous for sensitive people.	$\omega_{i,5}^t$

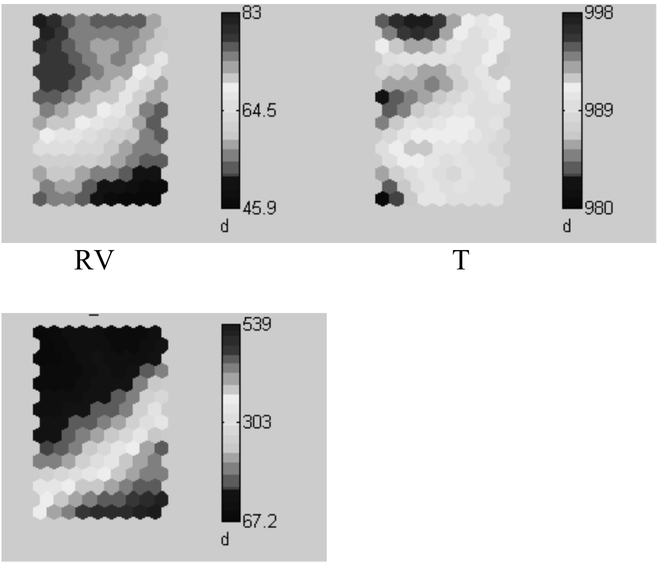


SO_2 O_3

NO_x CO

PM_{10} SV

SmV T_3



RV T

SZ

Fig. 6 Values of parameters $x_1^t, x_2^t, \dots, x_{11}^t$ for the KSOFM representatives

The locality and month (season) influent mostly 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.

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 $\omega_1^t, \omega_2^t, \dots, \omega_5^t$, 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) is presented in Fig. 7.

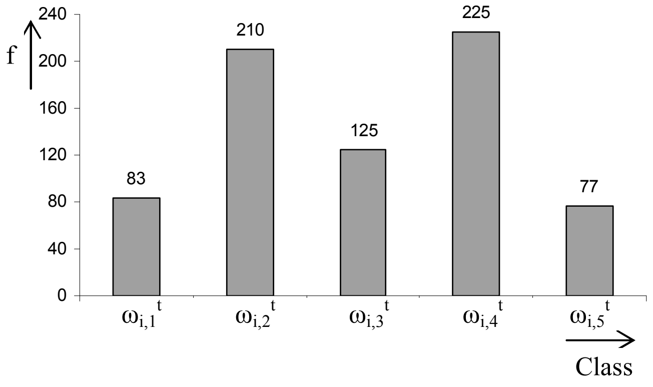


Fig. 7 Classification of the districts o_i^t into classes $\omega_{i,j}^t$

5 Conclusion

Considering the unknown assignment of the districts o_i^t to classes $\omega_{i,j}^t$, the modelling of air quality was realized by the combination of the KSOFM and K-means algorithm which is a representatives of unsupervised methods. This method makes the finding of well separated clusters and their suitable interpretation possible. The measurements of the air quality parameters were realized by the mobile monitoring system HORIBA. This system can not classify the measurements into the classes $\omega_{i,j}^t$. 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 model design was carried out in Matlab in MS Windows XP operation system.

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-Velag 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.