Comparison of Cluster Analysis Algorithms for Heavy Metals Contamination in Atmospheric Precipitation

MILOSLAVA KAŠPAROVÁ[1], JIŘÍ KŘUPKA[1], JAROMÍRA CHÝLKOVÁ[2]
[1]Faculty of Economics and Administration, University of Pardubice, Institute of System Engineering and Informatics
[2] Faculty of Chemical Technology, University of Pardubice, Institute of Environmental and Chemical Engineering
Studentská 84, 532 10 Pardubice
CZECH REPUBLIC
Miloslava.Kasparova@upce.cz, Jiri.Krupka@upce.cz, Jaromira.Chylkova@upce.cz

Abstract: The submitted article addresses the problematic of the cluster analysis. Real data sets concerning the atmospheric precipitation in chosen localities of the Czech Republic were used. The designed models use monthly data from years 2000 to 2007. Presented results refer to a pair of heavy metals – cadmium and lead. The data was processed by the help of the cluster analysis with presenting the results received through the Self organizing map, Two Step and K-Means methods. On the basis of the achieved results it is possible to decide about "quality" and "contamination" in the locality.

Key-Words: Heavy metal contamination, atmospheric precipitation, correlation, cluster analysis algorithms

1 Introduction

Air, water and soil belong to the main components of the environment, therefore, significant attention has been paid to its quality and pollution on national, as well as international level. The chemical structures of atmospheric precipitation and atmospheric deposition have been observed in the territory of the Czech Republic (CR) for a long time. Stations of the Czech Hydro-Meteorological Institute (CHMI) [2] measure – in the most of the cases – the level of precipitations in weekly interval. The change from the monthly interval to the weekly one happened in 1996 in accordance with the EMEM international methodology. Since the year 1997, a weekly precipitation taking of the bulk type has been introduced. Bulk represents precipitations with more closely indefinable content of dust fall-out on the analysis of heavy metals – lead [Pb], cadmium [Cd], nickel [Ni] and manganese [Mn].

Models designed in this work use data received from monthly measuring of the "bulk" type. These are real data coming from three stations which occupy themselves with the measurement of the chemical composition of precipitations and with the measurement of atmospheric depositions in the Trutnov region. These concern Station 1 (Hříběcí) with the height of 842 meters above sea level (measuring is provided by the Research Aquiculturing Institute (RAI) T.G.M.); Station 2 (Modrý potok) in the height of 1010 meters above sea level, the local administrator is CHMI; and Station 3 (Rýchory) with the height of 1003 meters above sea level (measuring is provided by RAI. These concern measuring from year 2000 to 2007. Concerning precipitations, conductibility, pH, anions, cations and the group of elements showed in Table 1 are analyzed.

Table 1 Measured Elements (in alphabetic order)

Chemical Matter	Description / Attribute	Value	
		Min	Max
Al^{3+} [µg/l]	Aluminum / (x_{10})	0.1	213
Ca^{2+} [µg /l]	Calcium cations / (x_7)	0.03	15400
Cd^{2+} [µg /l]	Cadmium / (x_{12})	0.01	21.2
Cl^- [µg /l]	Chloride anions / (x_{14})	0.08	6880
cond [µS/cm]	Conductibility / (x_2)	1.04	312
F^- [µg /l]	Fluoride anions / (x_{13})	0.01	390
Fe^{3+} [µg /l]	Iron / (x_{22})	0.1	270
K^+ [µg /l]	Potassium cations / (x_5)	0.02	9310
Mg^{2+} [µg /l]	Magnesium cations / (x_{21})	0.01	1120
Mn^{2+} [µg /l]	Manganese / (x_8)	0.02	375
Na^+ [µg /l]	Sodium cations / (x_4)	0.02	2270
NH^+_4 [µg /l]	Ammonium cations / (x_6)	0.01	13900
Ni^{2+} [µg /l]	Nickel / (x_{20})	0.01	40.1
NO^-_3 [µg /l]	Nitrous anions / (x_{15})	0.13	36300
Pb^{2+} [µg /l]	Lead / (x_{11})	0.2	39
pH	pH / (x_3)	3.69	7.85
Rain [mm]	Precipitations / (x_1)	2.1	567
SO^{2-}_4 [µg /l]	Sulphur anions / (x_{19})	0.02	16700
Zn^{2+} [µg /l]	Zinc / (x_9)	0.03	9277

Alkaline metals form a homogenous group of very reactive elements. Despite the fact that sodium [Na] and potassium [K] are chemically similar, they cannot be found in nature in one place, which is caused especially by the differences in their proportions [4, 22].

The features of the alkaline soils elements are dependent on their atomic number, similarly as the alkaline metals elements. There is about 0.13% of magnesium [Mg] in the sea water while 100 million tons of magnesium is produced by the help of electrolysis. Similarly as magnesium, also calcium [Ca] occurs in minerals in the form of insoluble carbonates, sulphates and silicates. The estimate number of their total representation depends on the geochemical model.

Heavy metals occur naturally in the ecosystem with large variations in concentration. Living organisms require varying amounts of "heavy metals." Excessive levels can be damaging to the organism. heavy metals such as mercury [12], plutonium, and Pb are toxic metals that have no known vital or beneficial effect on organisms. Certain elements that are normally toxic are, for certain organisms or under certain conditions, beneficial. Examples include vanadium [V], tungsten [W], and even Cd. Motivations for controlling heavy metal concentrations in gas streams are diverse. Some of them are dangerous to health or to the environment (e.g. Hg, Cd, As, Pb, Cr). Within the European community [28] the 13 elements of highest concern are As, Cd, Co, Cr, Cu, Hg, Mn, Ni, Pb, Sn, and Tl, the emissions of which are regulated in waste incinerators. Some of these elements are actually necessary for humans in minute amounts (Co, Cu, Cr, Ni) while others are carcinogenic or toxic, affecting, among others, the central nervous system (Hg, Pb, As), the kidneys or liver (Hg, Pb, Cd, Cu) or skin, bones, or teeth (Ni, Cd, Cu, Cr).

The amount of precipitation is monitored every day by the help of a rain gauge. The amount of other elements is measured in a one-week interval. For finding the value of conductibility, method of conductometric analysis is used; pH-meter is used for finding the levels of pH, for getting to know the amount of fluoride, chloride, sulfur and nitrate, anions ion chromatography is used. To determinate the amount of sodium, magnesium, calcium, potassium and ammonium cations, Zn^{2+} and Fe^{3+}, the atomic absorption spectrometry with flame atomisation is used. The amount of Cd^{2+} (further in the text as Cd only), Pb^{2+} (further in the text as Pb only), Ni^{2+} and Mn^{2+} are determined by the help of atomic absorption spectrometry with electrotheral atomization.

2 Problem Formulation

With the development of the civilization, mining and manufacturing of metals extends. Simultaneously, the contamination of the environment – air, soil and water sources increases and metals are received also by organisms. If their intake exceeds the organism abilities to eliminate these elements, they can effect toxically. Some elements cannot be eliminated by organisms at all. These are usually elements which occur in the crust of the Earth and in the sea water in a relatively low concentration, so that living organisms did not need during their evolution any mechanisms for their processing. These elements accumulate in the tissues and so that they can be extremely dangerous. Significant contaminants of the environment are Pb and Cd [22].

Objectives of this paper are:

- By selected methods (algorithms) of cluster analysis to design clusters of observations on the base of selected attributes and to evaluate them;
- By achieved results to determine contamination of localities (environs of stations 1, 2, and 3) in given district of the CR by contents of heavy metals in precipitation.

Basic scheme of modelling is in Fig.1.

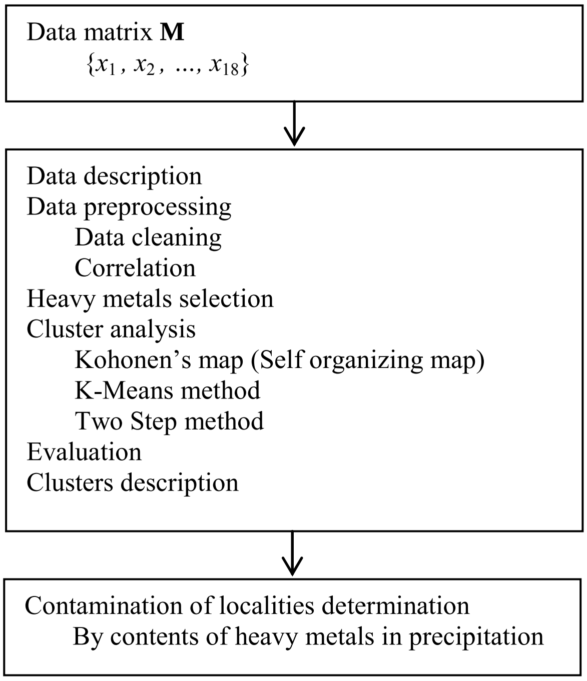

Fig.1 Basic scheme of modelling

2.1 Data Collection and Data Pre-processing

For our work we achieved 284 monthly observations described by 18 variables (attributes, characteristic) – the first fifteen attributes from Table 1, and years, months and localities; it means we had data matrix $\mathbf{M}(284\times 18)$ at disposal.

There are selected given chemical matters in Table 1 (x_1, x_2, ..., x_{15}) and additional attributes: year x_{16} = {2000, 2001, … , 2007}, month x_{17} = {1, 2, … , 12}and locality x_{18} = {1, 2, 3} of measurement. Every object (observation) o_i for i = 1, 2, …, 284 we can described by the following vector:

$$o_i = (x_{i1}, x_{i2}, ..., x_{i18}). \qquad (1)$$

After data collection we realized data description, data cleaning, and correlation. Data cleaning techniques [14] fill in missing values, smooth noisy data, identify outliers and correct inconsistencies in the data. Methods used for dealing with missing values include ignoring the objects, filling in the missing value manually, using the attribute mean to fill in the missing value etc. [6, 14, 27]. In our case for missing values we used mean value of given attributes. We eliminated observations that showed outlier values of attributes. Final dimension of data matrix **M** is (274 × 18). For instance, mean, minimal and maximal values of attributes x_1, x_3, x_{11} and x_{12} are in Table 2.

Table 2 Amount of Pb, Cd in precipitation, value of pH and rain in localities

Locality	Attribute	Mean	Min	Max
1	Pb (x_{11}) [μg /l]	2.90	0.30	19.90
	Cd (x_{12}) [μg /l]	0.76	0.01	7.40
	pH (x_3)	4.93	3.69	7.14
	Rain (x_1) [mm]	100.22	2.50	300.00
2	Pb (x_{11}) [μg /l]	2.51	0.20	11.00
	Cd (x_{12}) [μg /l]	0.27	0.02	0.65
	pH (x_3)	5.02	4.05	7.85
	Rain (x_1) [mm]	157.49	12.50	567.00
3	Pb (x_{11}) [μg /l]	4.03	0.30	15.90
	Cd (x_{12}) [μg /l]	0.58	0.01	7.80
	pH (x_3)	4.83	3.90	7.54
	Rain (x_1) [mm]	89.87	2.10	226.00

In Fig.2 and Fig.3 we can see values of Pb and Cd by months in localities. The locality 2 (Modrý potok) is typical by small values of given elements.

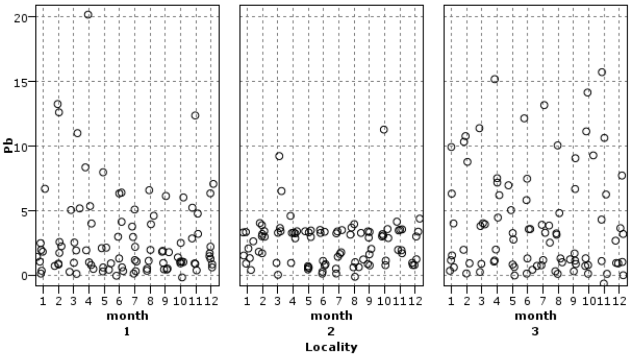

Fig.2 Values of Pb in localities

Afterwards we focused on a correlation analysis. It is a measure of the relation between two (a and j) or more variables. The most widely-used type of a correlation coefficient is Pearson correlation coefficient ρ_{aj} [9, 19]. This one ρ_{aj} can range from -1.00 to +1.00 and can be expressed by the following way: if $\rho_{aj} > 0$ it is a positive correlation of variables; if $\rho_{aj} < 0$ it is a negative correlation of variables; if $\rho_{aj} = 0$ this value represents a lack of correlation between variables; if $\rho_{aj} = 1$ it is a perfect positive correlation between variables. From the point of view of ρ_{aj} a size defines this linear dependence of variables [19]: if $\rho_{aj} \leq 0.3$ it is a small linear dependence; if $\rho_{aj} \in (0.3;0.8>$ it is a soft linear dependence; if $\rho_{aj} \in (0.8;1>$ it is a strong linear dependence. In the data matrix **M** the strong linear dependence was found between variables x_3 [pH] and x_5 [K^+] (ρ = 0.83), x_3 [pH] and x_6 [NH^+_4] (ρ = 0.83), x_4 [Na^+] and x_5 [K^+] (ρ = 0.92), x_4 [Na^+] and x_6 [NH^+_4] (ρ = 0.92), x_5 [K^+] and x_6 [NH^+_4] (ρ = 1), x_{15} [NO^-_3] and x_7 [Ca^{2+}] (ρ = 0.80). The soft linear dependence was between variables x_7 [Ca^{2+}] and x_{14} [Cl^-] (ρ = 0.77), x_7 [Ca^{2+}] and x_{13} [F^-] (ρ= 0.66), x_{14} [Cl^-] and x_{13} [F^-] (ρ = 0.67), x_7 [Ca^{2+}] and x_{10} [Al^{3+}] (ρ = 0.60), and x_6 [NH^+_4] and x_3 [pH] (ρ = 0.69).

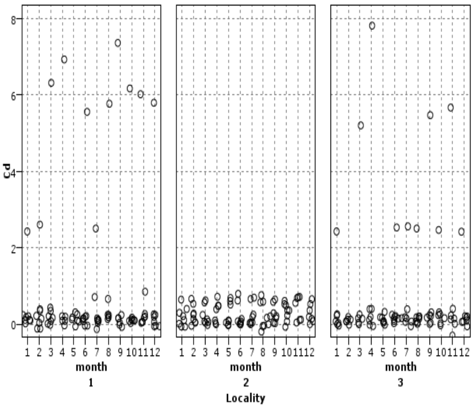

Fig.3 Values of Cd in localities

On the base of expert evaluation correlations between chemical elements x_3 [pH] and x_6 [NH^+_4]; x_4 [Na^+] and x_5 [K^+]; and x_{14} [Cl^-] and x_{13} [F^-] were only confirmed. We took given correlation into consideration by design of clusters and on the basis of consultation and [8, 22] we only focused on acquired heavy metals (Pb and Cd) in precipitation in our work.

2.2 Cluster Analysis

A cluster analysis [6, 13, 23] is an exploratory data analysis tool for solving classification problems. The object is sorted 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. Observations (where the object i is the observation i) are objects of the clustering. Every object is described by p characteristics. A vector of measurement o_i that contains p characteristics for p = 1, 2, …, 18 is in formula (1). The input set of the objects which are determined for the clustering (classification), is possible to write in a formula of objects matrix **M**.

The task of clustering is then to divide the set of objects 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, 13]. The cluster analysis distinguishes hierarchical and non-hierarchical methods. The basic division of methods is mentioned for instance in [6, 14] and application in [10].

2.2.1 Quality of Clustering Criterion

A sum of the square errors to centre of clusters E [13] is chosen as a criterion of the quality of clustering. It is defined in this way:

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

$$E = \sum_{h=1}^{k}\sum_{i=1}^{n_h} d^2(o_{hi}, T_h), \qquad (2)$$

where: $d^2(o_{hi}, T_h)$ is the square of the Euclidian metric of object o_{hi} 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 $T_h = \{t_{h1}, t_{h2}, ..., t_{hp}\}$, for its characteristics j, where $j = 1, 2, \ldots, p$, is (3):

$$t_{hj} = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hij}\,, \qquad (3)$$

where: n_h is number of objects in cluster M_h; x_{hij} is characteristic j of object i in cluster M_h.

2.2.2 Kohonen's Map

Kohonen's (Self organizing) map [11] is an artificial neural network algorithm in the unsupervised learning category. Many fields of science have adopted Self organizing map (SOM) as a standard analytical toll and many projects use it as a toll for solving hard real-word problems [3, 7, 11, 20, 25, 26] defines "elastic net" of points that are fitted to the input signal space to approximate its density function in an ordered fashion. The main applications of SOM are thus in the visualization of complex data in a two-dimensional display and creation of abstractions like in many clustering techniques [11].

I our work we used software Clementine 10.01. It is tool that offers opportunity to work with SOM and visualization and analysis of achieved results.

This method was used as an alternative cluster method [6, 17, 18, 21, 27]. It takes the input vectors in formula (1) and performs a type of spatially organized clustering, or feature mapping, to group similar records together and collapse the input space to a two-dimensional space that approximates the multidimensional proximity relationships between the clusters. SOM consists of two layers of neurons or units: an input layer and an output layer. The input layer is fully connected to the output layer, and each connection has an associated weight. Another way to think of the network structure is to think of each output layer unit having an associated centre, represented as a vector of inputs to which it most strongly responds (where each element of the centre vector is a weight from the output unit to the corresponding input unit) [21].

The parameters of SOM are represented as weights between input units and output units, or alternately, as a cluster centre associated with each output unit, more for example in [5, 11, 21].

2.2.3 Two Step Method

This method is a two-step clustering method. The first step makes a single pass through the data, during which it compresses the raw input data into a manageable set of subclusters. The second step uses a hierarchical clustering method to progressively merge the subclusters into larger and larger clusters, without requiring another pass through the data. Hierarchical clustering has the advantage of not requiring the number of clusters to be selected ahead of time.

Many hierarchical clustering methods start with individual objects as starting clusters and merge them recursively to produce ever larger clusters. Though such approaches often break down with large amounts of data, TwoStep's initial preclustering makes hierarchical clustering fast even for large datasets [21].

2.2.4 K-Means Method

K-Means method [6, 14, 27] can be used to cluster the dataset into distinct groups when you don't know what those groups are at the beginning. It tries to uncover patterns in the set of inputs. Objects are grouped so that objects within a group or cluster tend to be similar to each other, but objects in different groups are dissimilar.

This method works by defining a set of starting cluster centers derived from data. It then assigns each object to the cluster to which it is most similar, based on the object's input field values. After all cases have been assigned, the cluster centers are updated to reflect the new set of objects assigned to each cluster. The objects are then checked again to see whether they subject assignment/cluster iteration process continues until either the maximum number of iterations is reached, or the change between one iteration and the next fails to exceed a specified threshold [21].

3 Problem Solution

The objective of our work was to get three clusters. By clustering we only used two attributes (heavy metals Pb and Cd) from the data matrix **M** and we applied Kohonen's map, Two Step and K-Means method.

Achieved results (three clusters) are in Table 3, Table 5 (number of objects in clusters, mean values of used attributes and standard deviation (StDev)), and in Table 4 and Table 6, and in Fig.4, Fig.5, …, Fig.12, Fig.13. Each cluster is described by its center, which can be thought of as the prototype for the cluster [21]. For used attributes (Pb and Cd), the mean value and standard deviation for objects assigned to the cluster is given.

3.1 Kohonen's Map Application

By application Kohonen's map we can see cluster 1 (00) is characterized by small mean values of Pb (0.92 μg/l) and Cd (0.10 μg/l). Cluster 2 (10) is described by higher Pb mean value (3.28 μg/l) than in cluster 1. Cluster 3 (20) shows the highest mean value of heavy metal Pb in precipitation (7.22 μg/l). Classification of constituent objects by amount of Pb in precipitation recorded in localities shows Fig.5 and for Cd it is in Fig.6. We can see higher value of Pb belongs to cluster 3. Locality 2 (Modrý potok) is typical by smaller value of given attribute and by Kohonen's maps observation were classified into cluster 1 (42 objects) and 2 (48 objects). By mean values of Pb and Cd in clusters 1 and 2 it is possible to expect cleaner environment from heavy metals pollution view in locality 2 (Modrý potok). Observations from locality 1 (Hříběcí) chiefly belong to cluster 1 (cluster with the lowest values of Pb and Cd). However approximately 33 % of observations from this locality show considerable heavy metals pollution and belong to cluster 3 (cluster with the most polluted environment by mean values of Cd and Pb).

Table 3 Clusters by Kohonen's map

Cluster	Cluster 1 (00)	Cluster 2 (10)	Cluster 3 (20)
Number of objects	136	67	71
Mean of Pb (x_{11}) [μg /l]	0.92	3.28	7.22
StDev of Pb (x_{11}) [μg /l]	0.59	0.54	3.79
Mean of Cd (x_{12}) [μg /l]	0.10	0.50	1.40
StDev of Cd (x_{12}) [μg /l]	0.07	0.57	2.29

Mean values of Pb and Cd in clusters by Kohonen's map we can see in Fig.4.

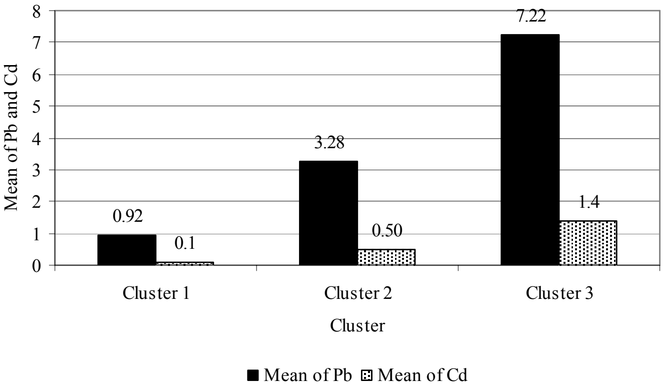

Fig.4 Mean values of Pb and Cd in clusters

In Table 4 we can see in cluster 1 any locality does not predominate. In cluster 2 locality 2 (Modrý potok) is dominant and cluster 3 contents predominantly observations from localities 1 (Hříběcí) and 3 (Rýchory). Criterion of the quality of clustering E shows rather big dispersion objects from centre of clusters (E = 362.21). The smallest dispersion of objects is in the cluster 2. Value is of Euclidian distance is 33.55. For cluster 1 this value is 64.33 and the highest dispersion of objects is in cluster 3. Euclidian distance value is 264.33.

Table 4 Count of objects in clusters by localities

	Cluster 1 (00)	Cluster 2 (10)	Cluster 3 (20)
Locality 1	56	7	31
Locality 2	42	48	3
Locality 3	38	12	37

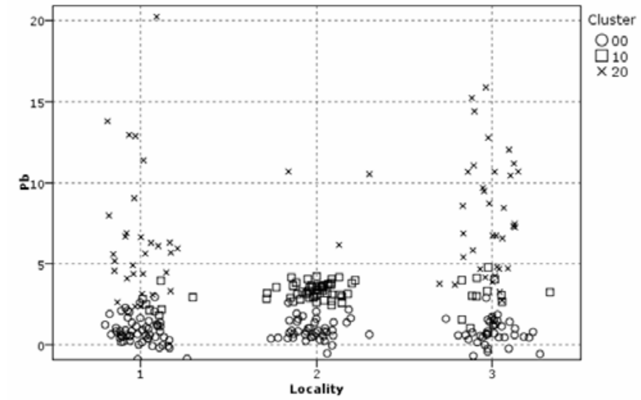

Fig.5 Classification of constituent objects by amount of Pb in precipitation recorded in localities

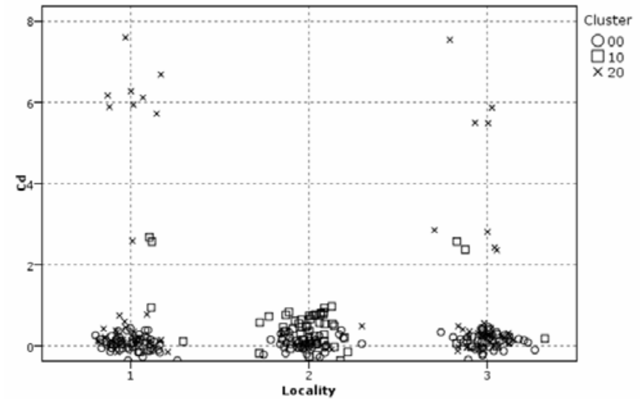

Fig.6 Classification of constituent objects by amount of Cd in precipitation recorded in localities

3.2 K-Means Method Application

Mean values of given heavy metals Pb and Cd in clusters by application of K-Means method we can see in Table 5. Cluster 1 includes 218 objects and is typical by small mean values of Cd (0.29) and Pb (1.87). Cluster 2 is characteristic by high mean value of Pb (9.33) and small mean value of Cd (0.23). The third cluster is described by high mean value of Cd (6.22) and by 3.17 of Pb. We can see that in cluster 2 value of heavy metal Pb is dominant and heavy metal Cd is dominant in cluster 3. Graphical representation of mean values of selected heavy metals Pb and Cd in clusters designed by K-Means method we can see in Fig.7.

Table 5 Clusters by K-Means method

Cluster	Cluster 1	Cluster 2	Cluster 3
Number of objects	218	44	12
Mean of Pb	1.87	9.33	3.17
StDev of Pb	1.39	3.26	1.10
Mean of Cd	0.29	0.23	6.22
StDev of Cd	0.50	0.13	0.79

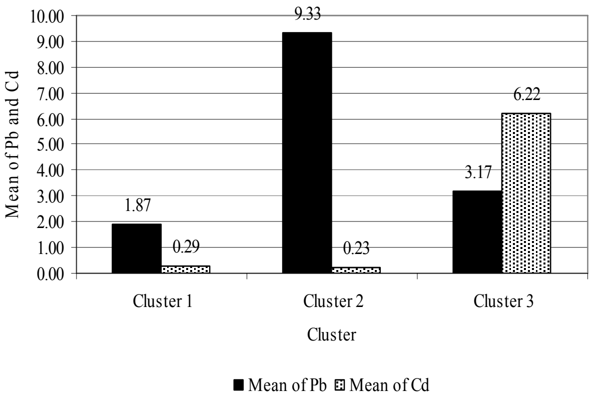

Fig.7 Mean values of Pb and Cd in clusters by K-Means method

In Table 6 we can see that preponderance of observations from all three localities is included in cluster 1; locality 2 (Modrý potok) is the least represented in cluster 2 and cluster 3 contents very small number of objects (12). Observations from locality 2 (Modry potok) are not in this cluster 3. Criterion of the quality of clustering *E* shows big dispersion objects from centre of clusters (*E* = 419.5). *E* values for all three constituent clusters are in the Fig.12.

Table 6 Count of objects in clusters by localities (K- Means method)

	Cluster 1	Cluster 2	Cluster 3
Locality 1	70	16	8
Locality 2	90	3	0
Locality 3	58	25	4

Classification of objects by amount of heavy metals Pb and Cd in precipitation recorded in given localities (locality 1: Hříběcí; lokality 2: Modrý potok and locality 3: Rýchory) are in Fig.8 and Fig.9.

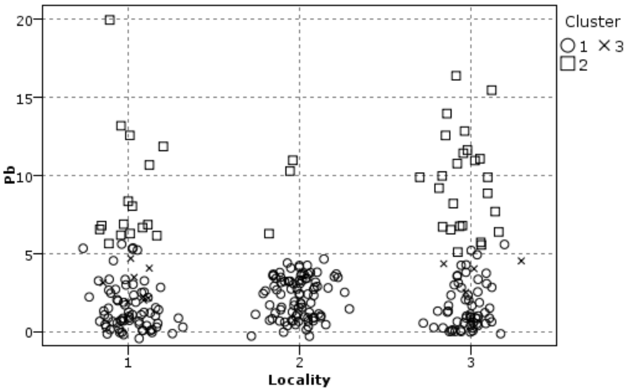

Fig.8 Classification of constituent objects by amount of Pb in precipitation recorded in localities (K-Means)

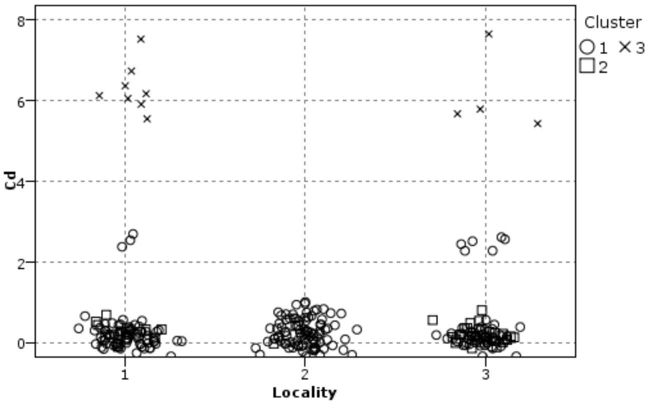

Fig.9 Classification of constituent objects by amount of Cd in precipitation recorded in localities (K-Means)

3.3 Two Step Method Application

By application of Two Step method we achieved identical results of clustering (mean values of Pb and Cd in clusters, and standard deviation are the same) as K-Means method application (Table 5). Number of observations and representations of localities in clusters are identical too (the same results as in Table 6).

Graphical representation of objects' classification by amount of heavy metal Pb in precipitation recorded in given localities is absolutely same like in Fig.8 and heavy metal Cd is like as in Fig.9, too.

3.4 Methods Comparison

For clustering we used three selected different methods. Kohonen's map is method based on neural network. K-Means is typical method of non-hierarchical cluster analysis and Two Step method works in two steps and in this algorithm is used hierarchical clustering among others. In our case we only used two inputs (two heavy metals: Pb and Cd in precipitation) for clustering.

If we compare results of clustering we can state that Kohonen's map shows the best results from point of view of methods' application (the criterion of the quality of clustering - sum of the square errors to centre of clusters *E* is 362.21). Graphical representation of this criterion *E* for all three clusters created by above mentioned methods are in Fig.10.

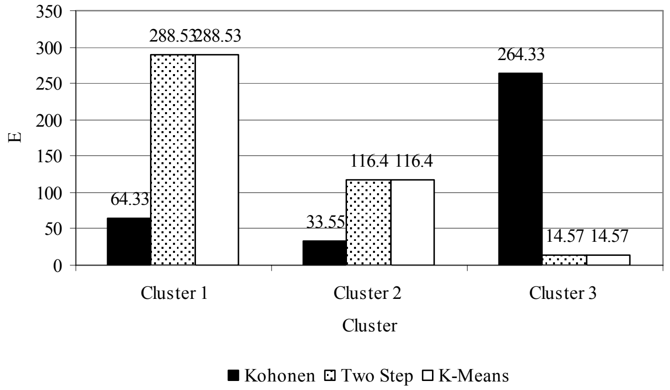

Fig.10 Evaluation of clustering by sum of the square errors to centre of clusters *E*

The lowest values of *E* (14.57) in cluster 3 created by Two Step and K-Means method result from small number of objects in this cluster 3. Number of objects in clusters by given methods is in Fig.11.

If we concern with mean values of Pb in clusters we can see (Table 7) that cluster 2 created by K-Means and Two Step method (value is 9.33; number of observations in cluster is 44) corresponds with cluster 3 created by Kohonen's map (value is 7.22; number of observations in cluster is 71); mean value of Pb in these clusters is the highest. Cluster 2 by Kohonen's map (value is 3.28) corresponds to cluster 3 created by K-Means and Two Step method (value is 3.17).

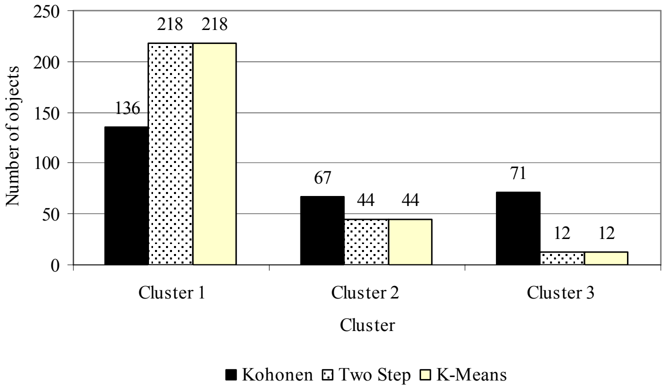

Fig.11 Number of objects in clusters by given methods

Table 7 Mean values of Pb in clusters by given methods

Method	Mean of Pb in		
	Cluster 1	Cluster 2	Cluster 3
Kohonen	0.92	3.28	7.22
K-Means	1.87	9.33	3.17
TwoStep	1.87	9.33	3.17

If we deals with mean values of Cd (Table 8) we can see that cluster 3 created by K-Means and Two Step method is typical by high mean value of this heavy metal in precipitation (value is 6.22); cluster 1 and cluster 2 have very similar values (difference is 0.06). From point of view of mean value of Cd we can see clustering by Kohonen's map shows the most different mean values of Cd.

Table 8 Mean values of Cd in clusters by given methods

Method	Mean of Cd in		
	Cluster 1	Cluster 2	Cluster 3
Kohonen	0.1	0.50	1.4
K-Means	0.29	0.23	6.22
TwoStep	0.29	0.23	6.22

We mentioned cluster 2 is typical by high content of Pb. There are 16 observations from locality 1; 25 observations from locality 3 and only 3 objects from locality 2. Cluster 3 is described by high content of Cd. There are only 12 classified objects; 8 objects belong to locality 1 and 4 objects belong to locality 3. Cluster 3 does not content any object from locality 3. It is possible to see the cluster 3 as group with relatively clean environment from point of view of given heavy metals in precipitation (mean value of Pb is 1.87 and Cd is 0.29). Predominant part of observations (218 objects) belongs to cluster 1 created by mentioned cluster methods (K-Means and Two Step).

Fig.12 and Fig.13 show graphical representations of mean values of given heavy metals in precipitation in clusters created by Kohonen's map, K-Means and Two Step methods.

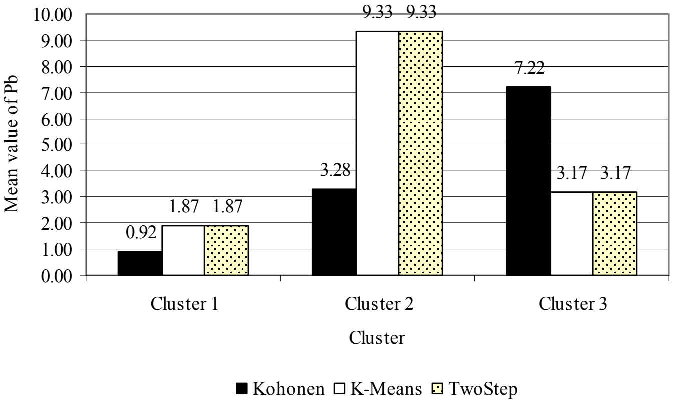

Fig.12 Mean values of Pb in clusters by given methods

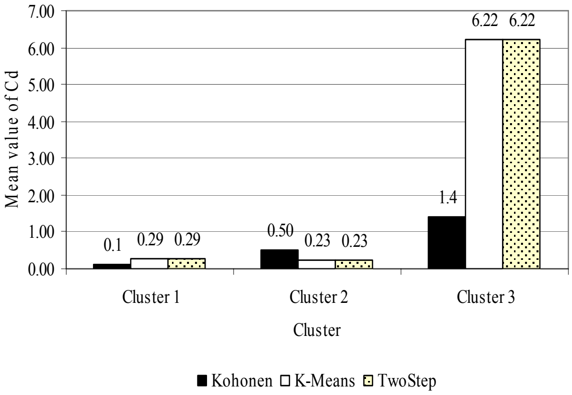

Fig.13 Mean values of Cd in clusters by given methods

4 Conclusion

On the basis of mentioned figures (Fig.2 and Fig.3) it is obvious, that in areas with low levels of monitored elements Pb and Cd we can speak about “natural” background in a certain locality.

According to [22], p. 160 and 285, average values of Cd = 0.5 [µg /l] and Pb = 5.0 [µg /l] can be deduced from 1997 levels for the region of the three named localities. Results shown on figures (Fig.5 and Fig.6) correspond to this fact and the cluster 1 (00) and cluster 2 (10) contain objects of measurement which are below mentioned limiting levels. According to the analysis it is possible to present current real levels which exceed the given “natural” background in a certain locality and the locality’s cluster belonging.

For clustering we used three different methods: Kohonen’s map, K-Means and Two Step method. Characteristics of the defined clusters (Table 3 and Table 4, Table 5, and Table 6) show the representation (occurrence) of localities in a given cluster. On the basis of the average cluster levels, it is possible to deduce a conclusion which locality is “cleaner” and so that more suitable for a healthy life.

Better results were achieved by Kohonen’s map application. Therefore, we focused on new clustering by three inputs (Pb, Cd and Rain). Results of clustering are in Table 9 and Table 10. Graphical representation of clustering is in Fig.14, Fig.15 and Fig.16.

Table 9 Clusters by Kohonen’s map (three inputs)

Cluster	Cluster 1 (00)	Cluster 2 (10)	Cluster 3 (20)
Number of objects	141	71	62
Mean of Pb	3.92	2.56	1.99
StDev of Pb	3.92	2.57	1.39
Mean of Cd	0.59	0.52	0.43
StDev of Cd	1.54	1.09	0.93
Rain_Mean	61.30	119.45	238.10
Rain_SDev	23.70	12.04	94.60

Table 10 Count of objects in clusters by localities (three inputs, Kohonen’s map)

	Cluster 1 (00)	Cluster 2 (10)	Cluster 3 (20)
Locality 1	53	25	16
Locality 2	30	27	36
Locality 3	58	19	10

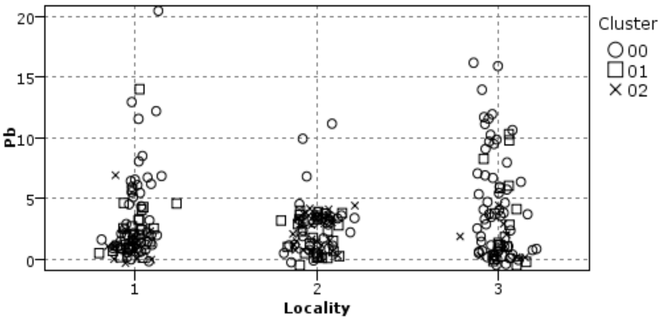

Fig.14 Classification of constituent objects by amount of Pb in precipitation recorded in localities (Kohonen’s map, inputs: Pb, Cd, Rain)

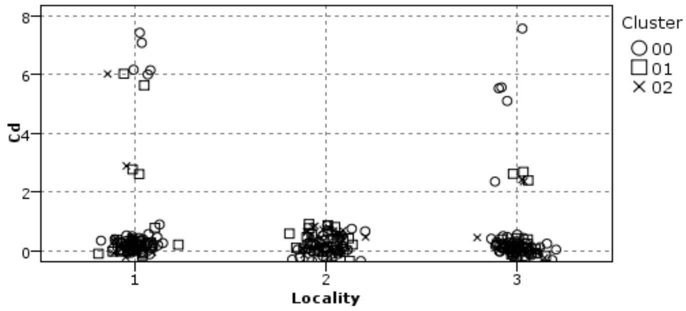

Fig.15 Classification of constituent objects by amount of Cd in precipitation recorded in localities (Kohonen’s map, inputs: Pb, Cd, and Rain)

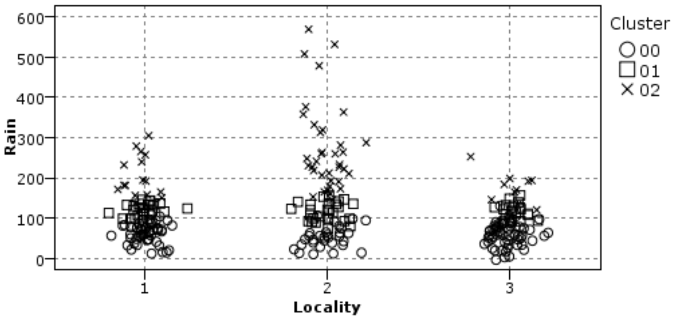

Fig.16 Classification of constituent objects by amount of Rain in precipitation recorded in localities (Kohonen's map, inputs: Pb, Cd, and Rain)

In Table 9 we can see that difference of mean values of Pb and Cd in clusters is not significant and standard deviations of inputs are high. On the basis of mean values and dispersions we can reason that it was achieved better results of Kohonen's map clustering by only two inputs Pb and Cd.

It is possible not only to compare limit average values of Pb and Cd with their values in Fig. 5 and Fig. 6 but also it is possible to determine a total contamination TC^j of soil in the j-th locality for Pd and Cd based on (4):

$$TC^j(a) = \sum_{i=1}^{k} c_i^j(a) \ , \qquad (4)$$

where: a is the heavy metal Pb or Cd; $c_i^j(a)$ is the contamination of Pb (Cd) of the i-th cluster in the j-th locality and it is defined by following way:

$$c_i^j\ (a) = n_i^j \cdot v(a)_i^j \cdot r_i^j \ , \qquad (5)$$

where: n_i^j is the number of elements in the i-th cluster (see Table 4), and $v(a)_i^j$ is the mean value of Pb or Cd in the i-th cluster, and r_i^j is the mean value of Rain in the i-th cluster in the j-th locality.
Values of n_i^j, $v(a)_i^j$, and r_i^j are in Table 11 and 12.

Table 11 Mean values of Pb and Cd in clusters by localities (in [µg / l])

	Cluster 1 (00) Pb / Cd	Cluster 2 (10) Pb / Cd	Cluster 3 (20) Pb / Cd
Locality 1	0.99 / 0.11	2.62 / 0.92	6.42 / 1.88
Locality 2	1.01 / 0.09	3.41 / 0.43	9.07 / 0.22
Locality 3	0.71 / 0.10	3.13 / 0.55	7.74 / 1.09

Table 12 Mean values of Rain in clusters by localities (in [mm])

	Cluster 1 (00)	Cluster 2 (10)	Cluster 3 (20)
Locality 1	105.11	121.77	86.516
Locality 2	153.07	167.95	52.13
Locality 3	94.07	106.47	80.16

Values of contamination c_i^j(Pb) and c_i^j(Cd) are in Table 13 and 14.

Table 13 Contamination of Pb in clusters by localities (in [mg / m^2])

	Cluster 1 (00)	Cluster 2 (10)	Cluster 3 (20)
Locality 1	5.85	2.24	17.21
Locality 2	6.51	27.52	1.42
Locality 3	2.53	3.99	22.96

Table 14 Contamination of Cd in clusters by localities (in [mg / m^2])

	Cluster 1 (00)	Cluster 2 (10)	Cluster 3 (20)
Locality 1	0.66	0.78	5.04
Locality 2	0.60	3.49	0.04
Locality 3	0.34	0.70	3.22

Values of total contamination TC^j(Pb) and TC^j(Cd) of heavy metals in localities for eight years (2000, …, 2007) are in Table 15.

Table 15 Total contamination of Pb and Cd in localities (in [mg / m^2]) based on cluster analysis

	Pb	Cd
Locality 1	25.29	6.48
Locality 2	35.45	4.12
Locality 3	29.49	4.27

On the basis of Table 15 it is possible to make inference that the locality 3 is the best of them (it means that it is cleaner, good for life etc.) because the value of TC^j(Pb) is middle and the value of TC^j(Cd) is small.

If the quality of locality is only evaluated based on/on the basis of the count of objects in clusters (Table 4) the locality 2 is the best locality, because has only three objects in the "worst" cluster (Cluster 3). However on the basis of contamination it is the most contaminated area.

In future, it would be appropriate to become occupied with an analysis of Pb and Cd in a potential risk of soil contamination [15], and in surface water contamination [16], and in a health risk assessment of air contamination [1].

Furthermore it would be possible to used classifiers based on e.g. fuzzy logic [24], artificial neural networks for comparison of localities.

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] Bozek, F. et al. Health Risk Assessment of Air Contamination Caused by Polycyclic Aromatic Hydrocarbons from Traffic. *Recent Advances in Environment, Ecosystems and Development*. WSEAS Press, 2009, pp. 104-108.

[2] Czech Hydrometeorological Institute [online], URL http://www.chmi.cz, (in Czech).

[3] Dzemyda, G., Kurasova, O. Comparative Analysis of the Graphical Result Presentation in the SOM Software. *INFORMATICA.* Vol. 13, No. 3, pp. 275–286, 2002.

[4] Greenwood, N. N., Earnshaw, A. *Chemistry of elemnts,* Praha: Informatorium, 1993. (In Czech).

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

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

[7] Haykin, S. *Neural Network, A Comprehensive Foundation.* Prentice-Hall, Inc. New Jersey, 1999.

[8] *Heavy metals in environment and the in influence* [online], URL http://hygiena.gastronews.cz/tezke-kovy-v-zivotnim-prostredi-a-jejich-vliv-na-lidsky-organismus (In Czech).

[9] Ho, R. *Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS.* Taylor & Francis Group, 2006.

[10] 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.

[11] Kohonen, T. *Self-Organizing Maps. Springer*, 2001.

[12] Lane, T. W., Morel, F. M. A biological function for cadmium in marine diatoms. *Proc. of the National Academy of Sciences*, Vol. 97, No. 9, pp. 4627-4631, 2000.

[13] Lukasová, A., Šarmanová, J. *Metody shlukové analýzy,* SNTL Nakladatelství technické literatury v Praze, 1985, (in Czech).

[14] Maimon, O., Rokach, L. *Decomposition Metodology for Knowledge Discovery and Data Mining,* World Scientific Publishing, 2005.

[15] Navratil, T. et al. Study of Charged Particles Transport Across Model and Real Phospholipid Bilayers. *Recent Advances in Environment, Ecosystems and Development.* WSEAS Press, 2009, pp. 212-217.

[16] Navratil, J. et al. Contamination of Surface Waters by Heavy Metals in the Brno Region and the Assessment of Health Hazards. *Recent Advances in Environment, Ecosystems and Development*. WSEAS Press, 2009, pp. 174-179.

[17] Olej, V., Hájek, P., Křupka, J., Obršálová, I. Air Quallity Modelling by Kohonen's Neural Networks, *WSEAS Environmental Science, Ecosystems & Development*, 2007, pp. 221-226.

[18] Pyle, D. *Business Modeling and Data Mining,* Morgan Kaufmann Publishers, 2003.

[19] Rublík, F. *Základy pravdepodobnosti a štatistiky*. Alfa, Bratislava, 1983, (in Slovak).

[20] SOM. *SOM Toolbox*, [online], URL http://www.cis.hut.fi/projects/somtoolbox/, 2009.

[21] SPSS Inc. Clementine® 12.0 *User's Guide*, 2008.

[22] Trebichavský, J., Šavrdová, D., Blohderger, M. *Noxious substances - Heavy Metals,* Kutna Hora: NSO, František Nekvasil, Expertízy a poradenství v oblasti odpadů a nerostných surovin, 1998, (in Czech).

[23] Turban, E. et al. *Decision Support Systems and Inteligent Systems*, Prentice Hall, 2005.

[24] Vaščák, J., Rutrich, M. Path Planning in Dynamic Environment Using Fuzzy Cognitive Maps. *Proc. of the 6th International Symposium on Applied Machine Inteligence and Informatics*, Herľany, Slovakia, pp. 5-9, 2008.

[25] Vesanto, J. SOM-Based Data Visualization Methods. *Intelligent Data Analysis.* Elsevier Science IOS Press. Vol. 3, No. 2, pp. 111 – 126, 1999.

[26] Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J. Self-organizing map in Matlab: the SOM toolbox. *Proceedings of the Matlab DSP Conference.* Espoo, Finland, 1999.

[27] Witten, I. H., Frank, E. *Data Mining: Practical Machina Learning Tools and Techniques*, Morgan Kaufman, 2005.

[28] Zevenhoven, R., Kilpinen, P. *Control of Pollutants in Flue Gases and Fuel Gases*. TKK, Espoo, 2001. [online], URL http://users.abo.fi/rzevenho/gasbook.html.