Czech Pattern Recognition Workshop 2000, Tomáš Svoboda (Ed.)
Peršlák, Czech Republic, February 2–4, 2000
Czech Pattern Recognition Society

Visual Analysis of Correlation Matrices

Gintautas Dzemyda

Institute of Mathematics and Informatics
Akademijos St. 4, Vilnius 2600, Lithuania
dzemyda@ktl.mii.lt

Abstract *We have developed here the approach for visualization of a set of parameters characterized by their correlation matrix. The approach integrates two methods for data mapping: Sammon's mapping and the self-organizing map (SOM). They are based on different principles, and, therefore, they supplement each other when used jointly. Some (sometimes sufficient) knowledge on a set of parameters may be obtained by using individual methods. In most cases, however, the necessity and efficiency of their joint use is unquestionable, – this allows us to observe the same data set from various standpoints and to extend our knowledge on the object of investigation.*

1 Introduction

Any set of similar objects (cases, vectors) may be often characterized by common parameters (variables, features). The term "object" may cover, e.g. people, equipment, or produce of manufacturing. Any parameter may take some values. A combination of values of all parameters characterizes a concrete object from the whole set. The values obtained by any parameter depend on the values of other parameters, i.e. the parameters are correlated. There exist groups (clusters) of parameters characterizing different properties of the object. The correlation matrix of parameters may be calculated by analysing the objects that compose the set. The problem is to discover knowledge on the groups (clusters) of parameters in the correlation matrix.

The following real correlation matrices (see references in [6]) became classic: the matrix of 8 physical parameters measured on schoolgirls; the matrix of 11 parameters characterizing the development of agriculture in two Canadian provinces, the matrix of 33 parameters of a tractor driver; the matrix of 24 psychological tests on pupils of the 7th and 8th forms in Chicago; the matrix of 11 frequencies that influence human mentality; the matrix of 10 geological parameters. However, recent research and technology development applications produce correlation matrices and discover knowledge via their analysis, too (see, e.g., [16] for correlation between 7 environmental variables, describing the distribution of ground by commune and by activity type in the intercity transport design in Belgium, and [11] for the matrix of 6 parameters characterizing the ethylene production technology).

The approach investigated in this paper is oriented to the analysis of correlation matrices and to the visual presentation of a set of parameters using the nonlinear Sammon's mapping and Kohonen's artificial neural network – the self-organizing map. The performance of our approach is illustrated by the visual analysis of four correlation matrices. The first two matrices have the known "ideal" partition of parameters into groups. The third and forth matrices store the correlation of environmental parameters that describe the air pollution in Vilnius city [22] and the development of coastal dunes and their vegetation in Finland [10]. These two problems are very urgent because they are of ecological nature: a visual presentation of data stored in the correlation matrices makes it possible for ecologists to discover additional knowledge hidden in the matrices and to make proper decisions.

2 Representation of a Set of Parameters by a Set of Vectors of S^n

Denote the correlation matrix of parameters $x_1, \dots, x_n$ by $R = \{r_{x_i x_j},\ i,j = \overline{1,n}\}$. Here $r_{x_i x_j}$ is a correlation coefficient of parameters x_i and x_j. A specific character of the problem of parameter clustering lies in the fact that the parameters x_i and x_j are related more strongly if the absolute value of the correlation coefficient $|r_{x_i x_j}|$ is higher, and less strongly if the value of $|r_{x_i x_j}|$ is lower (see [3]). The minimal relationship between the parameters is equal to 0. The relationship is maximal when the correlation coefficient is equal to 1 or -1.

Let S^n be a subset of an n-dimensional Euclidean space R^n containing vectors of unit length, i.e. S^n is a unit sphere, $\|Y\| = 1$ if $Y \in S^n$. The application of two matrices

1) $|R| = \{|r_{x_i x_j}|,\ i,j = \overline{1,n}\}$ and
2) $R^2 = \{r^2_{x_i x_j},\ i,j = \overline{1,n}\}$

in clustering of parameters is investigated in [5] both theoretically and experimentally: a system $\widetilde{Y}$ of vectors $Y_1, \dots, Y_n \in S^n$, corresponding to the system of parameters $x_1, \dots, x_n$, was analysed on the basis of the matrix of cosines between pairs of vectors from the system $\widetilde{Y}$ by using a modification of the k-means algorithm of Späth [20] (subroutine KMEANS) adapted to analyse the cosine matrix. Let us describe this approach more in detail.

In order to apply functionals that describe the quality of vector clustering (e.g. the sum of interior dispersions of clusters) to parameter clustering, it is necessary to determine a system of vectors $Y_1, \dots, Y_n \in S^n$ corresponding to the system of parameters $x_1,\dots,x_n$ so that $\cos(Y_i,Y_j) = |r_{x_ix_j}|$ or $\cos(Y_i,Y_j) = r^2_{x_ix_j}$. Then the clustering of vectors $Y_1,\dots,Y_n$ should be performed. Bearing in mind that $Y_1,\dots,Y_n \in S^n$, it suffices to know cosines between any pair Y_a and Y_b of vectors if we need to compute their Euclidean distance $\rho^2(Y_a,Y_b) = 2[1-\cos(Y_a,Y_b)]$. The distance between the weight centre of a cluster and any vector from this cluster may be computed in a similar manner by the known cosines between all the pairs of vectors from the chosen cluster, too (see [5]).

If only the matrix of cosines $K = \{\cos(Y_i,Y_j),\ i,j=\overline{1,n}\}$ is known, it is possible to restore the system of vectors $Y_s = (y_{s1},\dots,y_{sn}) \in S^n$, $s=\overline{1,n}$, as follows: $y_{sk} = \sqrt{\lambda_k}\alpha_{sk}$, $k=\overline{1,n}$, where λ_k is the k-th eigenvalue of the matrix K, the vector $(\alpha_{1k},\dots,\alpha_{nk})$ is a normalized eigenvector corresponding to the eigenvalue λ_k.

The system of vectors $Y_1,\dots,Y_n \in S^n$ does exist, if the matrix of their scalar products is non-negative definite. The matrix $R^2 = \{r^2_{x_ix_j},\ i,j=\overline{1,n}\}$ is non-negative definite (see [5]). However, the non-negative definiteness of the matrix $|R| = \{|r_{x_ix_j}|,\ i,j=\overline{1,n}\}$ does not follow from that of the matrix R: the matrix $|R|$ may not be non-negative definite if it has just one negative element.

Remark 1. The system of vectors $Y_1,\dots,Y_n \in S^n$, corresponding to the system of parameters $x_1,\dots,x_n$, does exist if

a) $\cos(Y_i,Y_j) = r_{x_ix_j},\ i,j=\overline{1,n},$
when all $r_{x_ix_j} \geq 0,\ i,j=\overline{1,n},$

b) $\cos(Y_i,Y_j) = r^2_{x_ix_j},\ i,j=\overline{1,n},$
when there exists just one $r_{x_ix_j} < 0$.

3 Background for Visual Presentation of a Set of Parameters $x_1,\dots,x_n$

The goal of this section is to analyse possibilities of mapping a set of vectors $Y_1,\dots,Y_n \in S^n$, that corresponds to the set of parameters $x_1,\dots,x_n$, on a plane trying to preserve the relative distances between $Y_1,\dots,Y_n \in S^n$. This leads to the possible visual observation of a layout of parameters $x_1,\dots,x_n$ on the plane.

3.1 Sammon's mapping

There exist a lot of methods that can be used for reducing the dimensionality of data. The analysis of relative performance of the different algorithms in reducing the dimensionality of multidimensional vectors starting from the paper by Biswas, Jain, and Dubes [2] indicates Sammon's projection [19] to be still one of the best methods of this class (see also [1] and [8]).

Sammon's projection is a nonlinear projection method to map a high dimensional space onto a space of lower dimensionality. In our case, the initial dimensionality is n, and the resulting one is 2. Denote the distance between vector Y_i and vector Y_j in the original space (in our case in S^n) by d^*_{ij}, and the distance between the same vectors in the projected space by d_{ij}. Sammon's algorithm tries to minimize the distortion of projection:

$$E = \frac{1}{\sum\limits_{\substack{i,j=1\\ i<j}}^{n} d^*_{ij}} \sum_{\substack{i,j=1\\ i<j}}^{n} \frac{(d^*_{ij}-d_{ij})^2}{d^*_{ij}}.$$

In fact, Sammon's mapping is closely related to the metric multidimensional scaling (MDS) (see [12] for details on the metric and nonmetric MDS). It, too, tries to optimise a cost function that describes how well the pairwise distances in a data set are preserved. The only difference between Sammon's mapping and the metric MDS is that the errors in distance preservation are normalized with the distance in the original space (see the distortion of projection E above).

3.2 The self-organizing map

The self-organizing map (SOM) proposed by Kohonen (see, e.g., [7], [13], [14], [18]) is a class of neural networks that are trained in an unsupervised manner using competitive learning. It is a well-known method for mapping a high dimensional space onto a low dimensional one. We consider here a mapping onto a two-dimensional grid of neurons. The method allows putting complex data into order based on its similarity and shows a map from which the features of the data can be identified and evaluated.

A variety of realizations of SOM have been developed (see, e.g., [14], [15], [17], [21]). All of them produce different results to some extent. Therefore, we present below additional details on our realization of SOM used in the experiments.

Usually, the neurons are connected to each other via rectangular or hexagonal topology. In this paper, we consider the rectangular case, only (see Figure 1 for example of SOM of size 4×4: circles denote neurons; indices of neurons are given inside the circles). The rectangular SOM is a two-dimensional array of neurons $M = \{m_{ij},\ i=\overline{1,k_x},\ j=\overline{1,k_y}\}$. Here k_x is the number of rows, and k_y is the number of columns (in Figure 1, both k_x and k_y are equal to 4). All neurons adjacent to a given neuron can be defined as its neighbours of a first order, then the neurons adjacent to the first order neighbour, excluding those already considered, as neighbours of a second order, etc. For example, the first order neighbours of m_{23} are $m_{12}, m_{13}, m_{14}, m_{22}, m_{24}, m_{32}, m_{33}, m_{34}$; the remaining neurons are the second order neighbours. The dimension of the vectors, which will be presented as inputs to train the network, is n. Each component of the input vector is connected to every individual neuron. Thus, there is a connection between the neuron of the network and every component of the input vector. The weights of these connections form an n-dimensional synaptic weight vector (the codebook vector). Thus, any neuron is entirely defined by its location on the grid (number of row i and column j) and by the codebook vector, i.e. we can consider a neuron as an n-dimensional vector $m_{ij} = (m^1_{ij}, m^2_{ij},\dots,m^n_{ij})$. In this way, each vector (neuron) m_{ij} represents a part of S^n because $Y_1,\dots,Y_n \in S^n$, but in most cases the vector m_{ij} itself does not belong to S^n, i.e. $m_{ij} \notin S^n$.

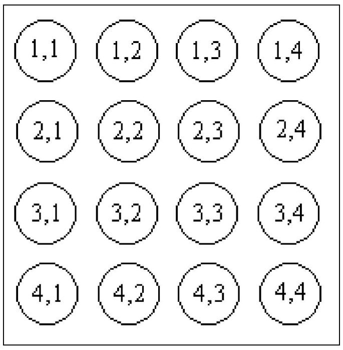

Figure 1: The rectangular SOM.

The map is trained in an unsupervised manner using competitive learning. Learning starts from the vectors m_{ij} initialized randomly. The starting values of m_{ij} are selected so that cosines between their pairs be positive – just like between the pairs of vectors from the training set $\{Y_1, \dots, Y_n\}$. At each learning step, an input vector Y is drawn from the training set $\{Y_1, \dots, Y_n\}$ and passed to the neural network. The Euclidean distance from this input vector to each vector m_{ij} is calculated and the vector (neuron) $m_c \in \{m_{ij}, i = \overline{1, k_x}, j = \overline{1, k_y}\}$ with the minimal Euclidean distance to Y is designated as a winner. Denote the row, where m_c is located, by i_c, and the column by j_c, i.e. c is a combination of two numbers – i_c and j_c. The components of the vector m_i are adapted according to the rule

$$m_{ij} \leftarrow m_{ij} + h_{ij}^c (Y - m_{ij}),$$

where h_{ij}^c is the learning rate, which is maximal for the winning neuron, and decreases with the neighbourhood order and the learning steps. After a large number v of learning steps, the network has been organized and n-dimensional input vectors have been mapped – each input vector is related to the nearest neuron (vector) m_{ij}.

Let us introduce a term "learning iteration". The learning iteration consists of n learning steps: the input vectors from Y_1 to Y_n are passed to the neural network in consecutive order. The whole learning process consists of v iterations ($v = 200$ was used in the experiments). In our case, such a partition of the learning process into learning iterations is sensible because of a small number n of input vectors $Y_1, \dots, Y_n$.

$$h_{ij}^c = \frac{\alpha}{\alpha \eta_{ij}^c + 1}, \quad \alpha = \max(\frac{v + 1 - e}{v},\ 0.01),$$

where η_{ij}^c is the neighbourhood order between the neurons m_c and m_{ij}; e is the number of current iteration ($e \in [1,\ v]$). The set of possible neighbours of m_c is restricted: we recalculate the vector m_{ij} if

$$\eta_{ij}^c \leq \max[\alpha \max(k_x,\ k_y),\ 1].$$

Note that $0 < h_{ij}^c \leq 1$, $h_{ij}^c = 1$ in the first learning iteration when $i = i_c$ and $j = j_c$, only.

Using the SOM-based approach above we can draw a table with cells corresponding to the neurons. The cells corresponding to the neurons-winners are filled with the numbers of vectors $Y_1, \dots, Y_n$. Some cells may remain empty. One can decide visually on the distribution of vectors $Y_1, \dots, Y_n$ in the n-dimensional space in accordance to their distribution among the cells of the table. However, the table doesn't answer the question, how much the vectors of the neighbouring cells are close in the n-dimensional space.

3.3 Combination of the self-organizing map and Sammon's mapping

Two methods for data mapping are discussed above. They are based on different approaches to mapping the data set. We try below to apply them together.

The self-organizing map provides structured information about the set of the analysed vectors – several elements (neurons) of the two-dimensional rectangular grid are activated (become winners), while the remaining elements are not activated. The activated elements of the grid may be considered as points on the plane. The number of row and column characterizes any of these elements, i.e. the location of these elements is fixed on the plane by the nodes of the rectangular grid. But the elements are characterized by n-dimensional vectors, too. A natural idea comes to apply the distance-preserving projection method to additional mapping of vectors-winners in SOM. Sammon's mapping may be used for such purposes.

Following Sammon [19] who suggested that clustering could be used as a front-end to his mapping algorithm, Kaski in [12] makes an assumption (without its theoretical or experimental background) that an especially useful combination seems to be first to reduce the amount of data by SOM, and then to display the reference vectors (winners) with some distance-preserving projection method (e.g. Sammon's mapping) to gain an additional insight. We will follow this scheme and experimentally compare the results of Sammon's mapping of the vectors that correspond to the parameters characterized by their correlation matrix and of the winners in SOM.

4 Visual Presentation of Sets of Parameters Having Known "Ideal" Partition

4.1 Data sets

The experiments were carried out on the basis of two correlation matrices with the known "ideal" partition of parameters into groups.

The first experiment was carried out using the correlation matrix R_8 of 8 physical parameters measured on 305 schoolgirls [4], [9]: height, arm span, length of forearm, length of lower leg, weight, bitrochanteric diameter, chest girth, chest width. Wide investigations of these classical test data divided parameters into two groups: $A_1 = \{x_1, \dots, x_4\}$ and $A_1 = \{x_5, \dots, x_8\}$. The parameters of the first group characterize shapeliness, while the parameters of the second group characterize plumpness of girls. It is an "ideal" partition of parameters.

The second experiment was carried out using the correlation matrix R_{24} of 24 psychological tests on 145 pupils of the 7th and 8th forms in Chicago [4], [9]. There are five groups of tests:

1) spatial perception $\{x_1, \ldots, x_4\}$,
2) verbal tests $\{x_5, \ldots, x_9\}$,
3) the rapidity of thinking $\{x_{10}, \ldots, x_{13}\}$,
4) memory $\{x_{14}, \ldots, x_{19}\}$,
5) mathematical capabilities $\{x_{20}, \ldots, x_{24}\}$.

The tests of the fifth group characterize a general development of the tested person. They do not characterize separate parts of his intellect. Thus, classifying all the tests into four groups the algorithms distribute the tests of the fifth group among the other four groups. The investigations in [4] substantiate considering the partition

$A_1 = \{x_1, \ldots, x_4, x_{20}, x_{22}, x_{23}\}$,
$A_2 = \{x_5, \ldots, x_9\}$,
$A_3 = \{x_{10}, \ldots, x_{13}, x_{21}, x_{24}\}$,
$A_4 = \{x_{14}, \ldots, x_{19}\}$

as an "ideal" one.

Elements of both the matrices R_8 and R_{24} are positive. Therefore, their values were not squared for analysis (see Remark 1).

4.2 Visualization by using Sammon's mapping

In Figure 2, we present Sammon's mapping results of vectors $Y_1, \ldots, Y_n \in S^n$ calculated on the basis of matrices R_8 (Figure 2a) and R_{24} (Figure 2b) using the approach of Section 2. In fact, Figure 2 shows the layout of parameters $x_1, \ldots, x_n$ on the plane. The indices of parameters are given at the points showing a place of the parameter on the plane.

Mapping of parameters on the basis of R_8 gave good results, – we can visually observe two clusters (see Figure 2a). But it is impossible to evaluate the number of clusters in Figure 2b (parameters are characterized by R_{24}). However, more correlated parameters are shown to be nearer to one another.

The experiments show that, in general, it is not sufficient to use Sammon's mapping (or any other method of this class) for visualization of a set of parameters.

4.3 Visualization by using SOM

Two sizes of SOM were used in the experiments: 3×3 and 4×4.

In Tables 1 and 2, the mapping results are presented (Tables 1a and 2a for R_8, and Tables 1b and 2b for R_{24}). A cell of the table means a neuron. Indices of the input vectors (in fact, parameters) nearest to the neuron m_{ij} are listed in the i-th row and j-th column, i.e. only the cells corresponding to the neurons-winners are not empty.

Parameters were distributed among the cells of the tables (among the nodes of SOM). The results in Table 1b may serve as a good example of application of SOM for clustering: we see four clusters which are separated by empty cells and contain the parameters like in the "ideal" partition given in Section 4.1. Results in Tables 1a, 2a, and 2b need an additional analysis.

4.4 Combined mapping

Sammon's mapping has been applied to the winners in SOM (to the vectors m_{ij} that correspond to non-empty cells in Tables 1 and 2). The results are presented in Figures 3 and 4 (Figures 3a and 4a for R_8, and Figures 3b and 4b for R_{24}).

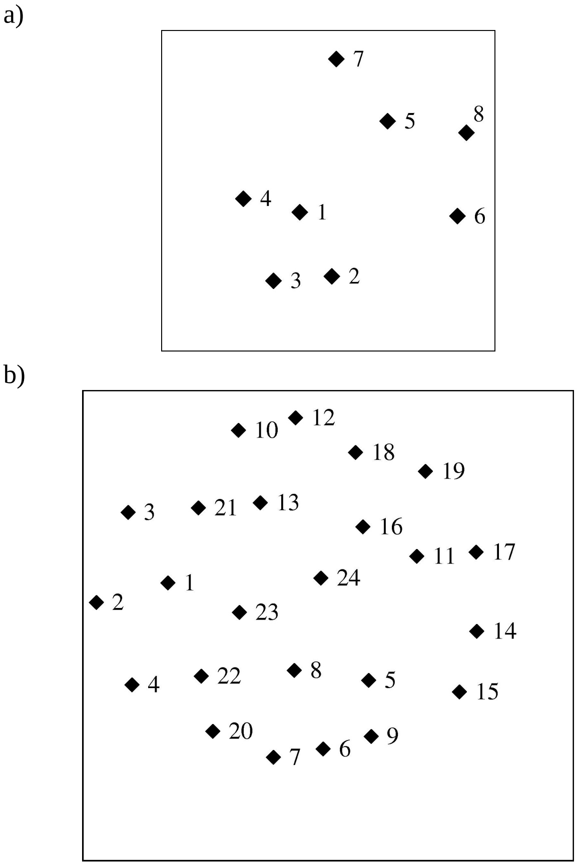

Figure 2: Sammon's mapping results: a) 8 parameters, b) 24 parameters.

The results of combined mapping give the answer to the questions, which remained open after the application of SOM alone. In Figures 3 and 4, we can observe interlocation of clusters. It means that we can visually determine, which groups of parameters are more neighbouring, and which are less ones. From Table 2b one can make a decision that the parameters x_1,x_{13} and x_{21} are similar. Figure 3b, however, refutes such a proposition. Figures 3a, 4a, and 4b give a possibility to visually observe the interlocation of parameters both inside the clusters and on the whole.

In addition to the results of Figure 3b obtained applying Sammon's mapping to the winners of 3×3 SOM, Figure 4b indicates that some clusters of parameters have a tendency of division into subclusters (the results of Figure 4b are obtained applying Sammon's mapping to the winners of 4×4 SOM). The conclusion on possible cluster division with growing the dimension of SOM cannot be applied to the results in Figures 3a and 4a. However, the tendency of cluster division may be observed in the respective Tables 1a and 2a. This shows the advantage of simultaneous presentation of both the SOM (table) and the results of combined mapping (figure) to the investigator.

Table 1: 3×3 SOM.

a)

8		2,3
5,6,7		1,4

b)

14,15,16,17,18,19		10,11,12,13,21,24
1,2,3,4,20,22,23		5,6,7,8,9

Table 2: 4×4 SOM.

a)

		4	2,3
7			1
5	6		8

b)

10,11,12,24		17,18	14,15,16
13,21		19	
1			
2,3,4	20,22,23		5,6,7,8,9

5 Visual Presentation of a Set of Environmental Parameters

5.1 Data sets

The experiments were carried out on the basis of two correlation matrices of environmental parameters.

The first experiment was carried out using the correlation matrix R_{10} of 10 meteorological and environmental parameters that describe the air pollution in Vilnius city [22]:

- x_1, x_2, and x_3 are the concentrations of carbon monoxide CO, nitrogen oxides NO_x , and ozone O_3;
- x_4 is the vertical temperature gradient measured at a 2–8 m height;
- x_5 is the intensity of solar radiation;
- x_6 is the boundary layer depth;
- x_7 is the amount of precipitation;
- x_8 is the temperature;
- x_9 is the wind speed;
- x_{10} is the stability class of atmosphere.

The second experiment was carried out using the correlation matrix R_{16} of 16 environmental parameters that describe the development of coastal dunes and their vegetation in Finland [10]:

- x_1 is the distance from the water line;
- x_2 is the height above the sea level;
- x_3 is the soil pH;

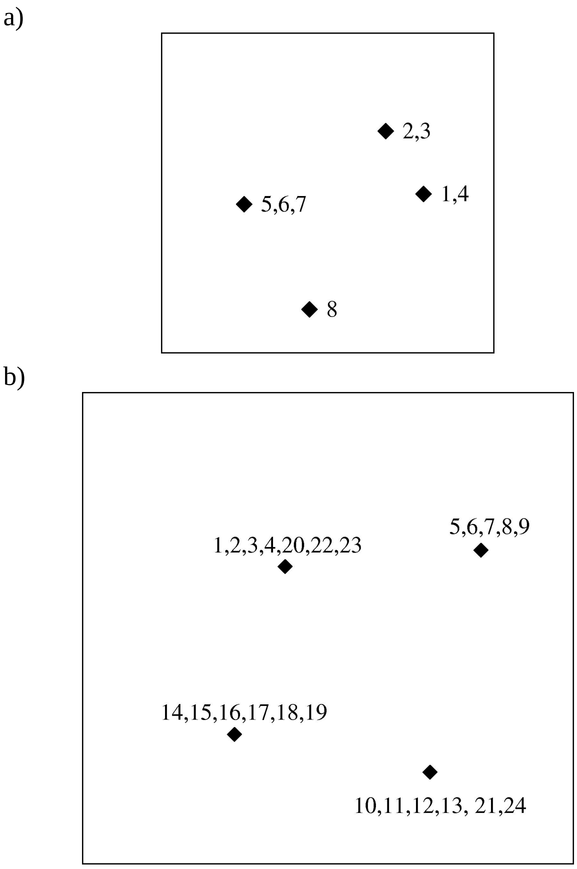

Figure 3: Combined mapping (3×3 SOM + Sammon's mapping) a) 8 parameters, b) 24 parameters.

- x_4, x_5, x_6, and x_7 are the contents of calcium (Ca), phosphorous (P), potassium (K), and manganese (Mg);
- x_8 and x_9 are the mean diameter and sorting of sand;
- x_{10} is the northernness in the Finnish coordinate system;
- x_{11} is the rate of land uplift;
- x_{12} is the sea level fluctuation;
- x_{13} is the soil moisture content;
- x_{14} is the slope tangent;
- x_{15} is the proportion of bare sand surface;
- x_{16} is the tree cover.

The measurements of parameters were performed in different sample plots and a correlation matrix was computed. Matrices R_{10} and R_{16} contain both positive and negative correlation coefficients. Therefore, their squared values were used in forming the set of vectors $Y_1, \ldots, Y_n \in S^n$ correspondent to the set of parameters $x_1, \ldots, x_n$ (see Remark 1).

5.2 Visualization by using Sammon's mapping

In Figure 5, we present Sammon's mapping results of vectors $Y_1, \ldots, Y_n \in S^n$ calculated on the basis of matrices R_{10} (Figure 5a) and R_{16} (Figure 5b) using the approach of Section 2.

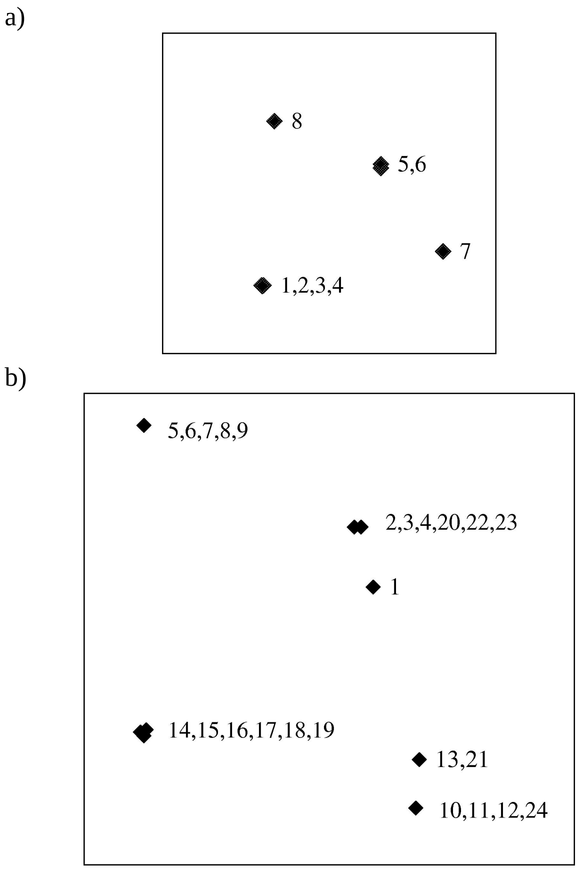

Figure 4: Combined mapping (4 × 4 SOM + Sammon's mapping): a) 8 parameters, b) 24 parameters.

5.3 Visualization by using SOM

The SOM of size 4×4 was used in the experiments. In Tables 3a and 3b, the mapping results are presented (Table 3a for R_{10}, and Table 3b for R_{16}). Table 3a indicates that there are at least four clusters of parameters in the first set of parameters. Table 3b indicates at least three clusters in the second set. These are the lower bounds for the number of clusters. What are the upper bounds? The combined mapping should be used in search for the answer.

Table 3: 4×4 SOM.

a)

7	1,2		4
			10
3,9	6		5,8

b)

8,9	14		1,2,16
15	5		3
			7
10,11,12	13		4,6

5.4 Combined Mapping

The results of combined mapping are presented in Figure 6 (Figure 6a for R_{10}, and Figure 6b for R_{16}). We can visually observe four clusters in Figure 6a and at least four clusters in Figure 6b. However, in Figure 6b the fourth cluster $\{x_1, x_2, x_3, x_{16}\}$ may be divided into two ones – the soil pH x_3 can form a separate cluster.

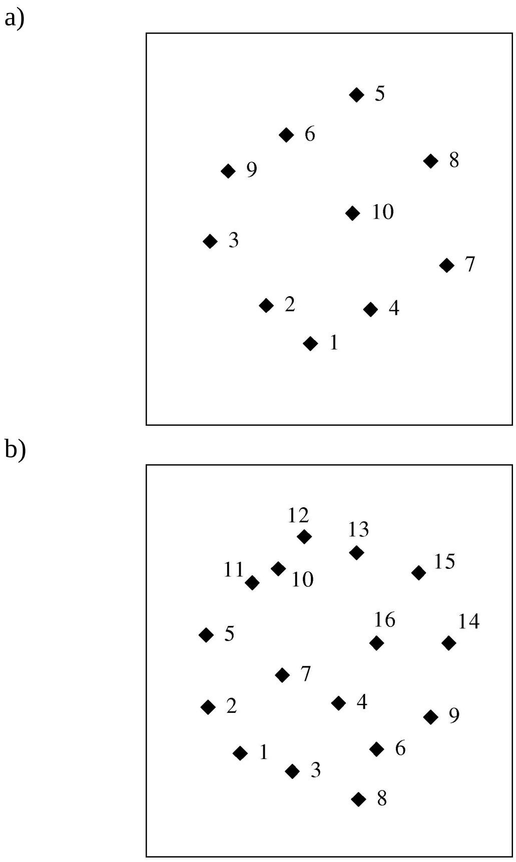

Figure 5: Sammon's mapping results: a) 10 parameters, b) 16 parameters.

6 Conclusions

We have developed here the approach for visualization of a set of parameters characterized by their correlation matrix. The proposed approach integrates two methods for data mapping: Sammon's mapping and the self-organizing map (SOM). They are based on different principles, and, therefore, they supplement each other when used jointly. Some (sometimes sufficient) knowledge on a set of parameters may be obtained by using individual methods (see Figure 2a – Sammon's mapping, and Table 1b – SOM). In most cases, however, the necessity and efficiency of their joint use is unquestionable, – this allows us to observe the same data set from various standpoints and to extend our knowledge on the object of investigation.

In general, if we have a data matrix $Z = \{z_{ij},\ i = \overline{1,t},\ j = \overline{1,n}\}$ for the analysis, where n is the number of parameters, t is the number of objects (i.e. we observe t different combinations of values of parameters $x_1, \ldots, x_n$), and $t >> n$, i.e. the value of t may be hundreds or thousands, our approach to decrease the dimensionality t via mapping contains the following items:

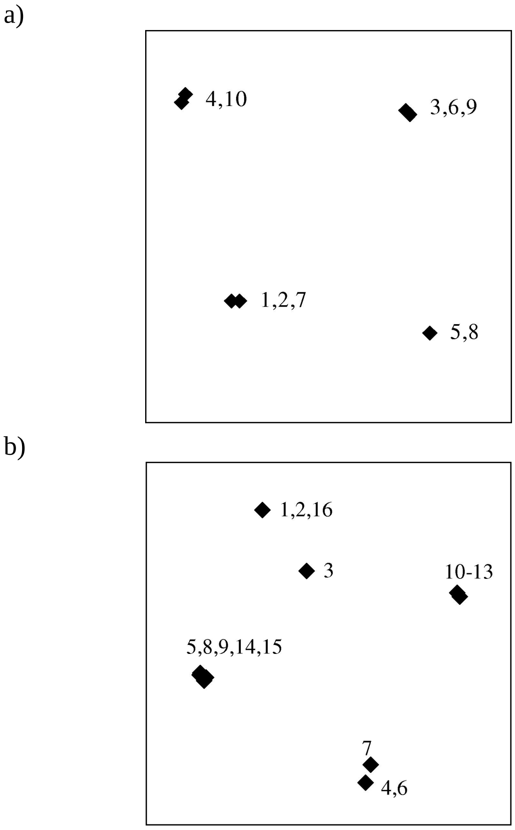

Figure 6: Combined mapping (4 × 4 SOM + Sammon's mapping) a) 10 parameters, b) 16 parameters.

1. Building a correlation matrix R of dimensions $n \times n$ for n parameters on a basis of the matrix Z.

2. Finding a set of vectors $Y_1, \ldots, Y_n \in S^n$ correspondent to the set of parameters $x_1, \ldots, x_n$.

3. Graphical presentation of the set of vectors $Y_1, \ldots, Y_n \in S^n$ using Sammon's mapping, SOM, or the combined mapping.

References

[1] J.C. Bezdek and N.R. Pal. An index of topological preservation for feature extraction. *Pattern Recognition*, 28:381−391, 1995.

[2] G. Biswas, A.K. Jain and R.C. Dubes. Evaluation of projection algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(6): 701−708, 1981.

[3] E.M. Braverman and I.B. Muchnik. *The Structural Methods for Empirical Data Processing*. Nauka, Moscow. (in Russian), 1983.

[4] G. Dzemyda. Clustering of parameters on the basis of correlations via simulated annealing. *Control and Cybernetics, Special Issue on Simulated Annealing Applied to Combinatorial Optimization*, 25(1):55−74, 1996.

[5] G. Dzemyda. Clustering of parameters on the basis of correlations: A comparative review of deterministic approaches. *Informatica*, 8(1):83−118, 1997.

[6] G. Dzemyda. *Knowledge Discovery Seeking a Higher Optimization Efficiency. Research Report Presented for Habilitation.* Mokslo Aidai, Vilnius, ISBN 9986–479–28–2, 1997.

[7] L. Fausett. *Fundamentals of Neural Networks.* Prentice Hall, 1994.

[8] A. Flexer. Limitations of self-organizing maps for vector quantization and multidimensional scaling. In M.C.Mozer, M.I.Jordan and T.Petsche (Eds.), *Advances in Neural Information Processing Systems 9*. pp.445–451. MIT Press/Bradford Books, Cambridge, MA, 1997.

[9] H.H. Harman. *Modern factor analysis*, 3 rd ed. University of Chicago Press, Chicago, 1976.

[10] P. Hellemaa. *The Development of Coastal Dunes and Their Vgetation in Finland*. Dissertation. Fennia 176: 1, Helsinki. ISSN 0015–0010, 1998.
http://renki.lib.helsinki.fi/elbanco/julkaisut/mat/maant/vk/hellemaa/

[11] Su-Nan Huang and Hui-He Shao. Application of pattern recognition to ethylene optimization. *Engineering Applications of Artificial Intelligence*, 7(3):329−333, 1994.

[12] S. Kaski. *Data Exploration Using Self-Organizing Maps.* In *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No.* 82. Espoo, Finish Academy of Technology. ISBN 952–5148–13–0, 1997.

[13] T. Kohonen. *Self-Organization and Associative Memory*, 3rd Edition. Springer-Verlag, Heidelberg, 1989.

[14] T. Kohonen. *Self-Organizing Maps* (2nd Ed.). Springer Series in Information Sciences, Vol. 30, 1997.

[15] T. Kohonen, J. Hynninen, J. Kangas and J. Laaksonen. *SOM_PAK: The Self-Organizing Map Program Package*. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland, 1996.

[16] M. Mouchart, L. Simar, Y. Baeyens, E. Barbieux, B. Frank, B. Jucquois, M.-P. Kestemont, Soiliou Daw Namoro, A.-M. Rutgeerts and E. Scheihing. *Module 3. Modal Choice and Flow Models for the Intercity Transport of Persons in Belgium*, 1996.
http://www.stat.ucl.ac.be/ISconrech/transport.

[17] F. Murtagh and M. Hernindez-Pajares. The Kohonen self-organizing map method: an assessment. *Journal of Classification*, 12:165−190, 1995.

[18] H. Ritter, K. Schulten and T. Martinetz. *Neural Computation and Self-Organizing Maps: An Introduction*. Addison-Wesley, 1991.

[19] J.W. Sammon, Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18: 401–409, 1969.

[20] H. Späth. *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*. Ellis Horwood, Chichester, 1980.

[21] A. Zell, G. Mamier, M. Vogt, N. Mache, R. Hübner, S. Döring, K.-U. Herrmann, T. Soyez, M. Schmalzl, T. Sommer, A. Hatzigeorgiou, D. Posselt, T. Schreiner, B. Kett, G. Clemente, J. Wieland, M. Reczko, M. Riedmiller, M. Seemann, M. Ritt, J. DeCoster, J. Biedermann, J. Danz, C. Wehrfritz, R. Werner, M. Berthold and B. Orsier. *SNNS. Stuttgart Neural Network Simulator*, *User Manual*, *Version 4.1*. Report No. 6/95, Institute for Parallel and Distributed High Performance Systems, University of Stuttgart, 1995.

[22] M. Zickus. *Influence of Meteorological Parameters on the Urban Air Pollution and Its Forecast.Thesis Presented for the Degree of Doctor in Physical Sciences*, 1998. http://vilnair.gamta.lt/thesis/content.html

Appendix 1. Correlation matrix $R_8 = \{r_{x_i x_j}, i, j = \overline{1,8}\}$ of physical parameters

$i \backslash j$	1	2	3	4	5	6	7	8
1	1.000	0.846	0.805	0.859	0.473	0.398	0.301	0.382
2	0.846	1.000	0.881	0.826	0.376	0.326	0.277	0.415
3	0.805	0.881	1.000	0.801	0.380	0.319	0.237	0.345
4	0.859	0.826	0.801	1.000	0.436	0.329	0.327	0.365
5	0.473	0.376	0.380	0.436	1.000	0.762	0.730	0.629
6	0.398	0.326	0.319	0.329	0.762	1.000	0.583	0.577
7	0.301	0.277	0.237	0.327	0.730	0.583	1.000	0.539
8	0.382	0.415	0.345	0.365	0.629	0.577	0.539	1.000

Appendix 2. Correlation matrix $R_{24} = \{r_{x_i x_j}, i, j = \overline{1,24}\}$ of psychological tests

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1.0	.318	.403	.468	.321	.355	.304	.332	.326	.116	.308	.314	.489	.125	.238	.414	.176	.368	.270	.365	.369	.413	.474	.282
2	.318	1.0	.317	.230	.285	.234	.157	.157	.195	.057	.150	.145	.239	.103	.131	.272	.005	.255	.112	.292	.306	.232	.348	.211
3	.403	.317	1.0	.305	.247	.268	.223	.382	.184	.075	.091	.140	.321	.177	.065	.263	.177	.211	.312	.297	.165	.250	.383	.203
4	.468	.230	.305	1.0	.227	.327	.335	.391	.325	.099	.110	.160	.327	.066	.127	.322	.187	.251	.137	.339	.349	.380	.335	.248
5	.321	.285	.247	.227	1.0	.622	.656	.578	.723	.311	.344	.215	.344	.280	.229	.187	.208	.263	.190	.398	.318	.341	.435	.420
6	.355	.234	.268	.327	.622	1.0	.722	.527	.714	.203	.353	.095	.309	.292	.251	.291	.273	.197	.251	.435	.263	.386	.431	.433
7	.304	.157	.223	.335	.656	.722	1.0	.619	.585	.246	.232	.181	.345	.236	.172	.180	.228	.159	.226	.451	.314	.396	.405	.437
8	.332	.157	.382	.391	.578	.527	.619	1.0	.532	.285	.300	.271	.395	.252	.175	.296	.255	.250	.274	.427	.362	.357	.501	.388
9	.326	.195	.184	.325	.723	.714	.585	.532	1.0	.170	.280	.113	.280	.260	.248	.242	.274	.208	.274	.446	.266	.483	.504	.424
10	.116	.057	.075	.099	.311	.203	.246	.285	.170	1.0	.484	.585	.408	.172	.154	.124	.289	.317	.190	.173	.405	.160	.262	.531
11	.308	.150	.091	.110	.344	.353	.232	.300	.280	.484	1.0	.428	.535	.350	.240	.314	.362	.350	.290	.202	.399	.304	.251	.412
12	.314	.145	.140	.160	.215	.095	.181	.271	.113	.585	.428	1.0	.512	.131	.173	.119	.278	.349	.110	.246	.355	.193	.350	.414
13	.489	.239	.321	.327	.344	.309	.345	.395	.280	.408	.535	.512	1.0	.195	.139	.281	.194	.323	.263	.241	.425	.279	.392	.458
14	.125	.103	.177	.066	.280	.292	.236	.252	.260	.172	.350	.131	.195	1.0	.370	.412	.341	.201	.206	.302	.183	.243	.242	.304
15	.238	.131	.065	.127	.229	.251	.172	.175	.248	.154	.240	.173	.139	.370	1.0	.325	.345	.334	.192	.272	.232	.246	.256	.165
16	.414	.272	.263	.322	.187	.291	.180	.296	.242	.124	.314	.119	.281	.412	.325	1.0	.324	.344	.258	.388	.348	.283	.360	.262
17	.176	.005	.177	.187	.208	.273	.228	.255	.274	.289	.362	.278	.194	.341	.345	.324	1.0	.448	.324	.262	.173	.273	.287	.326
18	.368	.255	.211	.251	.263	.197	.159	.250	.208	.317	.350	.349	.323	.201	.334	.344	.448	1.0	.358	.301	.357	.317	.272	.405
19	.270	.112	.312	.137	.190	.251	.226	.274	.274	.190	.290	.110	.263	.206	.192	.258	.324	.358	1.0	.167	.331	.342	.303	.374
20	.365	.292	.297	.339	.398	.435	.451	.427	.446	.173	.202	.246	.241	.302	.272	.388	.262	.301	.167	1.0	.413	.463	.509	.366
21	.369	.306	.165	.349	.318	.263	.314	.362	.266	.405	.399	.355	.425	.183	.232	.348	.173	.357	.331	.413	1.0	.374	.451	.448
22	.413	.232	.250	.380	.341	.386	.396	.357	.483	.160	.304	.193	.279	.243	.246	.283	.273	.317	.342	.463	.374	1.0	.503	.375
23	.474	.348	.383	.335	.435	.431	.405	.501	.504	.262	.251	.350	.392	.242	.256	.360	.287	.272	.303	.509	.451	.503	1.0	.434
24	.282	.211	.203	.248	.420	.433	.437	.388	.424	.531	.412	.414	.458	.304	.165	.262	.326	.405	.374	.366	.448	.375	.434	1.0

Appendix 3. Correlation matrix $R_{10} = \{r_{x_i x_j}, i, j = \overline{1,10}\}$ of meteorological and enviromental parameters that describe the air pollution in Vilnius city

i\j	1	2	3	4	5	6	7	8	9	10
1	1.00	0.78	-0.28	0.66	0.07	-0.33	-0.05	-0.09	-0.35	0.38
2	0.78	1.00	-0.37	0.63	-0.01	-0.31	-0.05	0.24	-0.38	0.37
3	-0.28	-0.37	1.00	-0.10	0.24	0.28	-0.11	0.18	0.64	0.04
4	0.66	0.63	-0.10	1.00	0.06	-0.45	-0.14	-0.06	-0.33	0.58
5	0.07	-0.01	0.24	0.06	1.00	-0.08	-0.05	0.09	-0.07	0.17
6	-0.33	-0.31	0.28	-0.45	-0.08	1.00	0.07	-0.10	0.60	-0.52
7	-0.05	-0.05	-0.11	-0.14	-0.05	0.07	1.00	-0.01	0.04	-0.11
8	-0.09	0.24	0.18	-0.06	0.09	-0.10	-0.01	1.00	0.01	0.23
9	-0.35	-0.38	0.64	-0.33	-0.07	0.60	0.04	0.01	1.00	-0.27
10	0.38	0.37	0.04	0.58	0.17	-0.52	-0.11	0.23	-0.27	1.00

Appendix 4. Correlation matrix $R_{16} = \{r_{x_i x_j}, i, j = \overline{1,16}\}$ of environmental parameters that describe the development of coastal dunes and their vegetation in Finland

i\j	1	2	3	4	5	6	7	8	9	10	11	12	13	114	15	16
1	1.00	0.72	-0.60	-0.23	-0.02	-0.33	-0.38	-0.12	0.21	0.17	0.20	0.07	-0.02	0.02	-0.20	0.61
2	0.72	1.00	-0.36	-0.17	-0.09	-0.20	-0.22	-0.31	0.23	0.12	0.17	0.07	0.05	0.11	0.16	0.52
3	-0.60	-0.36	1.00	0.41	0.29	0.60	0.70	0.08	-0.36	-0.23	-0.22	-0.20	-0.26	-0.08	0.22	-0.39
4	-0.23	-0.17	0.41	1.00	0.20	0.79	0.70	-0.25	0.10	-0.42	-0.46	-0.26	-0.29	-0.10	-0.02	-0.07
5	-0.02	-0.09	0.29	0.20	1.00	0.17	0.47	0.35	-0.40	-0.31	-0.34	-0.29	-0.13	-0.36	0.01	0.02
6	-0.33	-0.20	0.60	0.79	0.17	1.00	0.69	-0.13	-0.02	-0.24	-0.28	-0.08	-0.19	-0.06	-0.02	-0.04
7	-0.38	-0.22	0.70	0.70	0.47	0.69	1.00	0.01	-0.20	-0.50	-0.52	-0.39	-0.47	-0.14	0.13	-0.06
8	-0.12	-0.31	0.08	-0.25	0.35	-0.13	0.01	1.00	-0.60	0.12	0.07	0.07	-0.05	-0.06	-0.15	-0.19
9	0.21	0.23	-0.36	0.10	-0.40	-0.02	-0.20	-0.60	1.00	0.27	0.30	0.25	0.30	0.02	-0.13	0.30
10	0.17	0.12	-0.23	-0.42	-0.31	-0.24	-0.50	0.12	0.27	1.00	0.96	0.91	0.69	0.18	-0.24	0.14
11	0.20	0.17	-0.22	-0.46	-0.34	-0.28	-0.52	0.07	0.30	0.96	1.00	0.76	0.64	0.21	-0.16	0.16
12	0.07	0.07	-0.20	-0.26	-0.29	-0.08	-0.39	0.07	0.25	0.91	0.76	1.00	0.67	0.15	-0.31	0.11
13	-0.02	0.05	-0.26	-0.29	-0.13	-0.19	-0.47	-0.05	0.30	0.69	0.64	0.67	1.00	-0.05	-0.06	-0.01
14	0.02	0.11	-0.08	-0.10	-0.36	-0.06	-0.14	-0.06	0.02	0.18	0.21	0.15	-0.05	1.00	-0.13	-0.02
15	-0.20	0.16	0.22	-0.02	0.01	-0.02	0.13	-0.15	-0.13	-0.24	-0.16	-0.31	-0.06	-0.13	1.00	-0.19
16	0.61	0.52	-0.39	-0.07	0.02	-0.04	-0.06	-0.19	0.30	0.14	0.16	0.11	-0.01	-0.02	-0.19	1.00