Visual Analysis of the Multidimensional Meteorological Data

Gintautas Dzemyda

Institute of Mathematics and Informatics, Akademijos St. 4, 2021 Vilnius, Lithuania
dzemyda@ktl.mii.lt

Abstract. A method for the visualization of correlation-based data has been applied for analysis of the set of meteorological and environmental parameters that describe the air pollution. A visual presentation of data stored in the correlation matrix makes it possible for ecologists to discover additional knowledge hidden in it. The method consists of two stages: building of a system of vectors based on the correlation matrix and visualization of these vectors. Sammon's mapping and the self-organizing map were applied for visualization of the vectors.

1 Introduction

Any set of environmental objects (cases, vectors) may often be characterized by common parameters (variables, features). A combination of values of all the 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. A problem of the analysis of correlations arises here. This problem as well as a lot of real correlation matrices became classical (see [1]). However, recent research and technology development applications produce correlation matrices and discover knowledge via their analysis, too. Correlations of meteorological and environmental parameters and their analysis appear in various studies. The references cover air pollution, vegetation of coastal dunes, groundwater chemistry, minimum temperature trends, zoobenthic species-environmental relationships, analysis of large environmental and taxonomic databases.

The goal of this paper is the illustration of a possibility to apply the visualization method to the analysis of correlation matrices of parameters that are of an environmental and ecological nature. The analysis is based on the correlation matrix of parameters that describe the air pollution. A visual presentation of data stored in the correlation matrix makes it possible for ecologists to discover additional knowledge hidden in the matrix and to make proper decisions about the interlocation of parameters and about their groups (clusters).

2 The Method of Visual Analysis of Correlation Matrices

One of the most popular methods of analysing correlations is the principal component analysis [2]. However, it does not show an interlocation of variables –

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3036, pp. 652–656, 2004.

only their location around the zero-correlation. It means that we need more sophisticated means for the analysis of correlations.

A method for visualizing a set of parameters $x_1, ..., x_n$ characterized by their correlation matrix has been proposed in [1]. The method consists of two stages: building and visualization of a system of multidimensional vectors $Y_1, ..., Y_n \in S^n$ corresponding to the parameters $x_1, ..., x_n$. S^n is a subset of the n-dimensional Euclidean space R^n containing vectors of unit length.

There exist lots of methods that can be used for reducing the dimensionality of data (see, e.g., [3]). We apply here two popular and effective methods: Sammon's mapping [4] that is a nonlinear projection method closely related to the metric multidimensional scaling, and the self-organizing map (SOM) [3] that is a neural network.

Using Sammon's mapping we reduce the dimensionality of vectors $Y_1, ..., Y_n \in S^n$ by computing a correspondent system of two-dimensional vectors $Z_1, ..., Z_n \in R^2$. The self-organizing map (SOM) is a class of neural networks that are trained in an unsupervised manner. 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. 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 order numbers of vectors $Y_1, ..., Y_n$. Some cells may remain empty (see Fig. 2a). One can make a decision visually on the distribution of the vectors $Y_1, ..., Y_n$ in the n-dimensional space in accordance with their distribution among the cells of the table. However, the table does not answer the question, how much the vectors of the neighboring cells are close in the n-dimensional space. This question may be answered by a combined mapping, i.e., by applying Sammon's mapping to visualize the n-dimensional vectors that are the numerical characteristics of neuron-winners.

3 Meteorological and Environmental Data Set

The experiment was carried out using the correlation matrix of 10 meteorological and environmental parameters that describe the air pollution in Vilnius city [5]: x_1, x_2, 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 correlation matrix is presented in Table 1.

4 Results of the Analysis

In Fig. 1a, we present Sammon's mapping results of the vectors $Y_1, ..., Y_n$ calculated on the basis of the correlation matrix of parameters $x_1, ..., x_n (n = 10)$. Fig. 1a shows the distribution of vectors $Z_s \in R^2$, $s = \overline{1, n}$, obtained after the application of Sammon's algorithm to the vectors $Y_1, ..., Y_n$. In fact, we observe the distribution of parameters $x_1, ..., x_n$ on a plane. We do not present legends

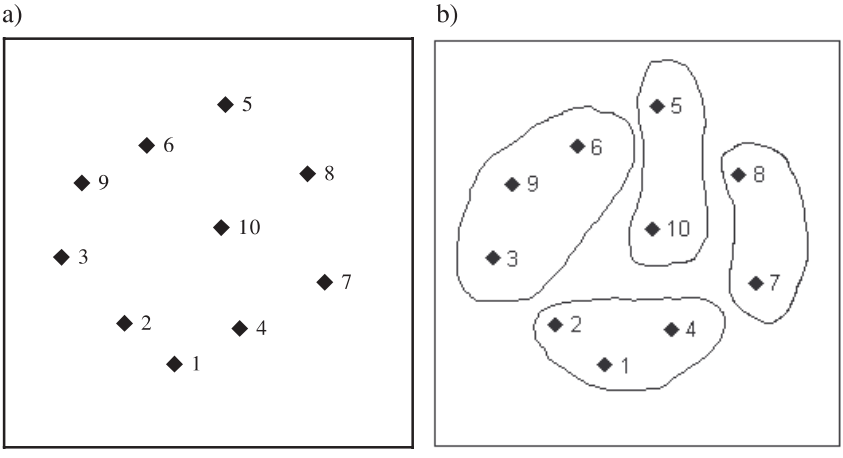

Fig. 1. Sammon's mapping results: a) mapping of 10 parameters; b) final conclusions on the clusters of parameters.

and units for both axes in the figure, because we are interested in observing the interlocation of points corresponding to the parameters only. The parameters are almost uniformly distributed after a direct compression of 10-dimensional points to a plane by Sammon's mapping, but more similar parameters are located nearer.

The SOM of size 4×4 was used in the experiments. The mapping results are presented in Fig. 2a. They indicate that there are at least three clusters of parameters. This estimate may be considered as the lower bounds for the number of clusters. What is the upper bound? The combined mapping should be used in search for the answer.

The results of combined mapping are presented in Fig. 2b. The figure shows the distribution of two-dimensional vectors obtained after an application of

Table 1. Correlation matrix of meteorological and environmental parameters

$i\backslash j$	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

Sammon's algorithm to the neurons-winners in the SOM from Fig. 2a. We can visually observe four clusters in Fig. 2b.

5 Conclusions

Comparing the results in Figures 1 and 2 on the parameters of air pollution, we can conclude that the meteorological and environmental parameters that describe the air pollution in Vilnius city form four clusters: $\{x_1, x_2, x_4\}$, $\{x_3, x_6, x_9\}$, $\{x_5, x_{10}\}$, $\{x_7, x_8\}$. These clusters are separated by curves in Fig. 1b. Data points in this figure repeat these of Fig. 1a.

One can see that the interlocation of parameters is similar in Figures 1a and 2a. The only difference is that clusters of parameters are more explicit in Fig. 2a. The clusters are more explicit in Fig. 2b as compared with Fig. 2a.

The analysis allows us to conclude that visualization is a powerful tool in data analysis. Its extension to the analysis of correlation matrices widened the range of applications. Most of problems with the correlation-based data sets may be solved in this way. In particular, the environmental problem was analysed successfully. The conclusions on the similarity of the measured parameters as well as on the possible number of clusters of similar parameters are drawn analysing the visual data presentation. This becomes possible due to the given method.

a)

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

b)

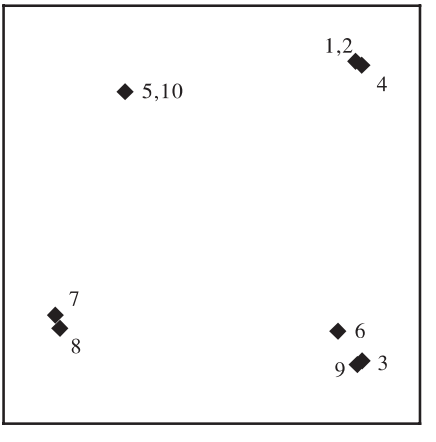

Fig. 2. Distribution of parameters characterizing the air pollution: a) 4×4 SOM; b) combined mapping (4×4 SOM + Sammon's mapping)

References

1. Dzemyda, G.: Visualization of a set of parameters characterized by their correlation matrix. Computational Statistics and Data Analysis **36(10)** (2001) 15–30

2. Jolliffe, I.T.: Principal Component Analysis. Springer (1986)
3. Kohonen, T.: Self-Organizing Maps. 3nd ed. Springer Series in Information Sciences, Springer–Verlag, Vol. 30 (2001)
4. Sammon, J.W.: A nonlinear mapping for data structure analysis. IEEE Transactions on Computers **18** (1969) 401–409
5. Zickus, M.: 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