University of Miskolc,
Hungary

microCAD 2002
International Scientific Conference
7-8 March 2002

C: KÉMIA
Section C: Chemistry

PRINCIPAL COMPONENT ANALYSIS (PCA) AND PRINCIPAL COMPONENT REGRESSION (PCR) FOR EVALUATION OF MONITORING DATA

A. Lengyel[1,3], K. Héberger[2], L. Paksy[4], O. Bánhidi [5,6], R. Rajkó[7]

1. University of Miskolc, Institute of Chemistry, Department of Analytical Chemistry E-mail: akmla@gold. uni-miskolc.hu
2. Institute of Chemistry, Chemical Research Center, Hungarian Academy of Sciences, H-1025 Budapest, Pusztaszeri út 59/67, Hungary
3. University of Miskolc, Research Co-ordination Centre for Materials Science and Mechatronics
4. University of Miskolc, Institute of Chemistry, Department of Analytical Chemistry E-mail: akmpl@gold. uni-miskolc.hu
5. University of Miskolc, Department of Analytical Chemistry, Department of Analytical Chemistry, E-mail: banhidio@freemail.hu
6. DAM STEEL Special Steelmaking, Ltd. Miskolc, 3540. Vasgyári u 43
7. University of Szeged, E-mai: rajko@sol.cc.u-szeged .hu

1. INTRODUCTION

In the last decade PCA and PCR became a powerful tool to assess environment quality, to determine spatial and temporal trends and to discriminate between natural and anthropogenic pollution sources. There is no possibility to survey all of the trials here. Below we summarized the most characteristic investigations and those, which are in closest connection with our study.

2. EXPERIMENTAL CONDITIONS

2.1 Principal Component Analysis (PCA)

The variables are ordered as columns of the input matrix, whereas the observations are ordered in rows. First, the correlation matrix of the variables is calculated. PCA reduces the dimensionality of the data by revealing several underlying components. The underlying components are represented by new variables called principal components. Their values are the component scores. The principal components are, in fact, linear combinations of the original variables and vice versa. The linear coefficients of the latter linear combinations are called the component loadings, *i.e.* the correlation coefficients between the original variables and the principal components.

The principal components are uncorrelated, and they account for the total variance of the original variables. The first principal component accounts for the maximum of the total variance, the second one is uncorrelated with the first one and accounts for the maximum of the residual variance, and so on until the total variance is accounted for. For a practical problem it is sufficient to retain only a few components accounting for a large percentage of the total variance.

In summary, PCA decomposes the original matrix into several products of multiplication by loading (variables) and score (observations) vectors.

PCA shows which kind of variables and observations are similar in nature.

2.2 Data sources:

a.) PCA has been used for data collected on the Hungarian section of river Sajó (160 km) in five sampling places: (SP 1) -(SP 5) between 1986-1993. The latter place is the meeting point with river Tisza.

The following variables were measured: water flow (output) (WATER), amount of the total suspended particles (TSP), chemical oxygen demand (COD), amount of the nitrate anion (NITRATE) and the amount of phosphate anion (PHOSPH). The first two variables are considered as natural, the third and fourth as anthropogenic ones. Phosphate amount can be classified into both.

Data were measured by the Northern-Hungary Environmental Agency, depending on sampling places (environmental load) either on a weekly or a fortnightly basis between years 1986-1993.

The total number of measuring cases at the five sampling places during four seasons is 1460 (resulting in 7300 data).

b.) The following variables were used: concentration of nitrogen monoxide (NO), concentration of nitrogen dioxide (NO_2), concentration of sulfur dioxide (SO_2), concentration of carbon monoxide (CO), concentration of ozone (O_3), velocity of wind (WIND), direction (DIR) of wind, temperature (TEMP), moisture content in air (HUM), exposure of sun (SUN). The samples were arranged in the rows as taken by time. All of these properties were continuously measured by a monitoring station. The monitoring station is located at the most frequented square of Miskolc, where two main roads with heavy traffic cross. Near this cross bus and coach stations are situated. One of the market places of the city works in this area. Consequently, the streets are very crowded and considerably polluted. The date was chosen when the weather was fine not only this chosen day but during some day before.

The frequency of sampling was half an hour (30 min). The sampling and analysis were continuous. Three cases (samples) were missing (No. 7, 43, 44) caused by sampling error, so we have 45 cases for calculation. In this paper we named the variables with the short chemical formula if the parameter was a chemical compound. Otherwise, we used a maximum four character long sign for the meteorological parameters (see above).

3. DISCUSSION and CONCLUSION

a.) Classify of contamination of River Sajó

1) By means of Fig. 1-4 the effect of seasonal changes is much easier to survey for every component to be studied using only two principal components than average values, standard deviations etc. (Sampling place No. 4 is Miskolc)
2) The dominance of the output (WATER) and the Total Suspended Particles (TSP) can be observed at each sampling place. The 2nd group contains the nitrate-anion concentration and the Chemical Oxygen Demand (COD). It seems that the amount of phosphate has an individual effect.
3) Above Miskolc, the biggest anthropogenic source of this area, the most significant change can be expected from the water output (high loading values comparing to those of the polluting components in the PC1). At Miskolc (see Fig. 5), the outliers

Fig 1

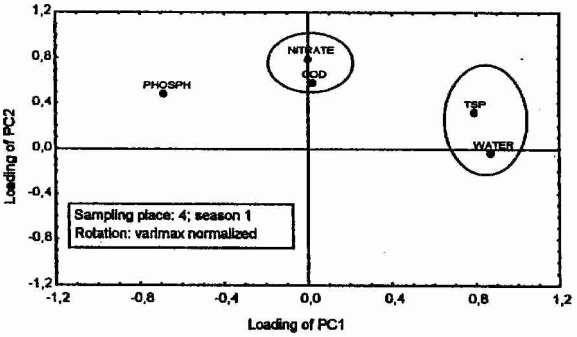


Fig. 2

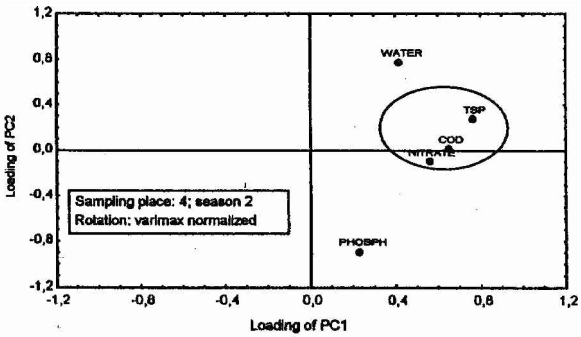


Fig. 3

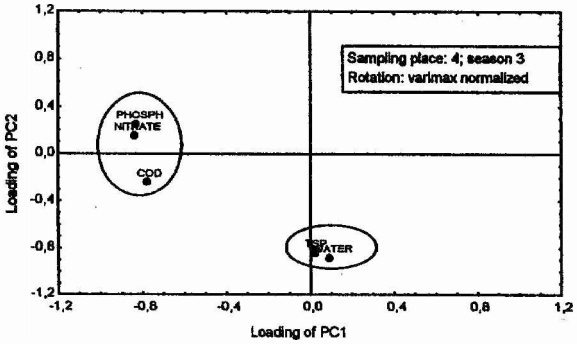


Fig. 4

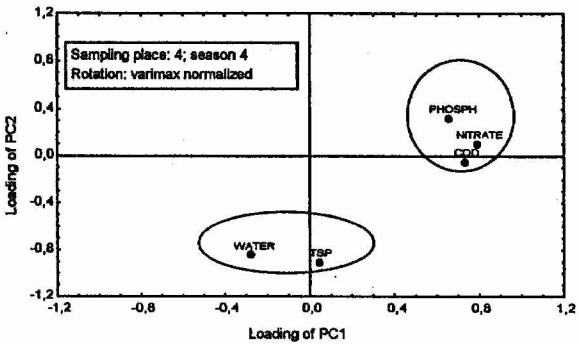


Fig. 5

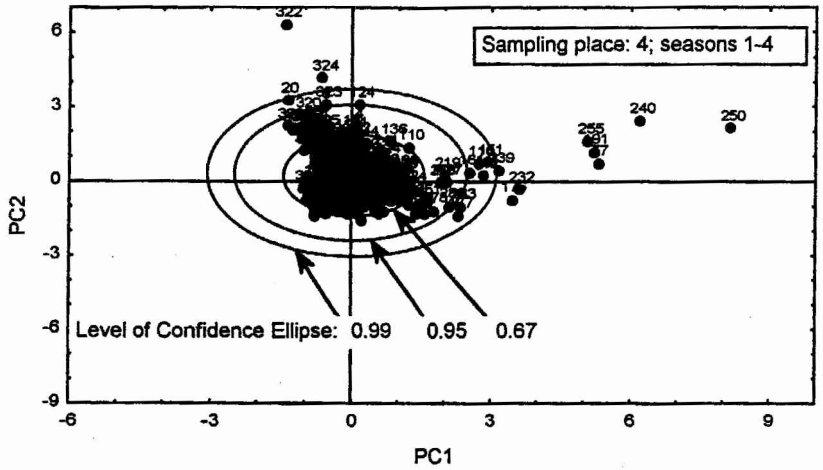

Sampling place: 4; seasons 1-4
Level of Confidence Ellipse: 0.99 0.95 0.67
PC1
PC2

Fig. 6

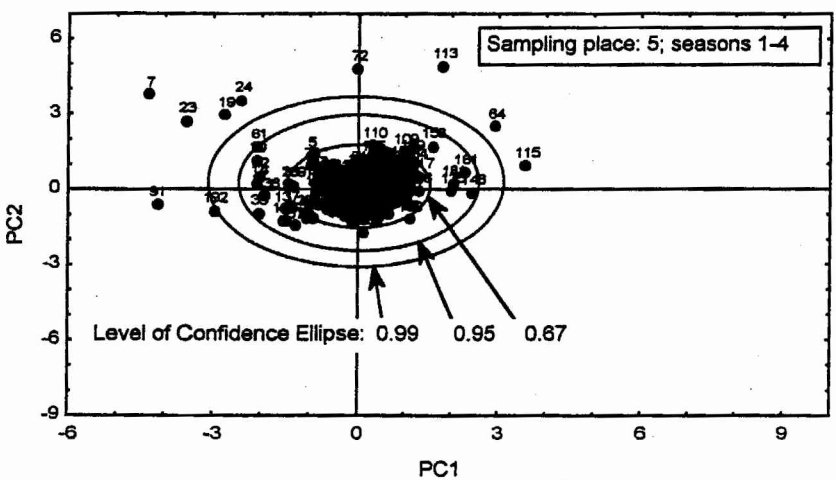

Sampling place: 5; seasons 1-4
Level of Confidence Ellipse: 0.99 0.95 0.67
PC1
PC2

Fig. 7

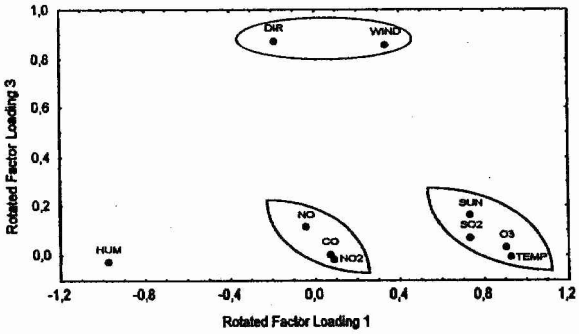


Fig. 8

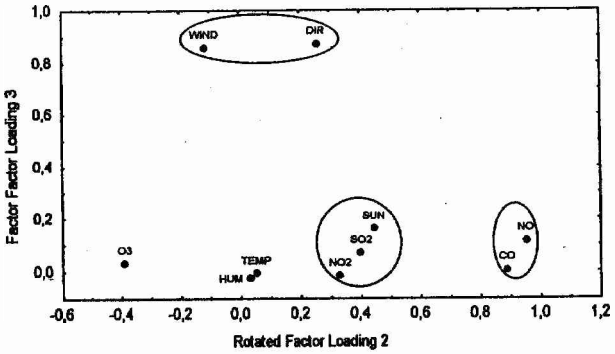


Fig. 9

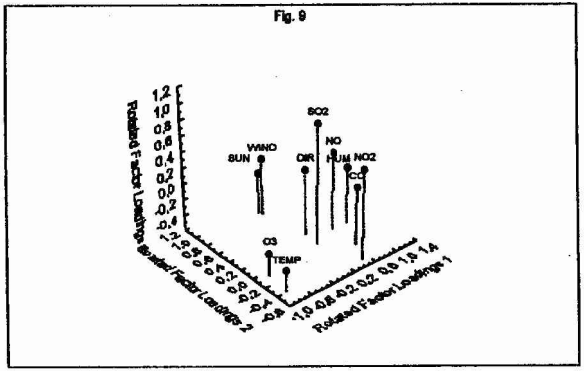
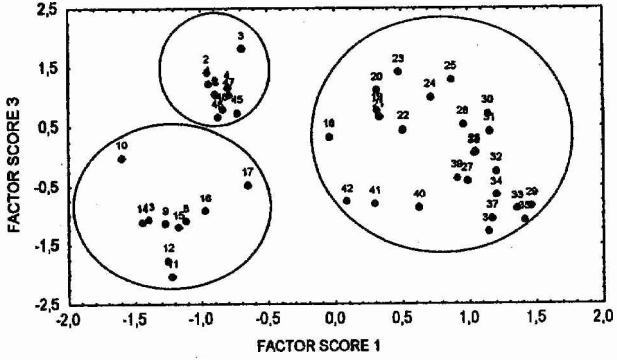


Fig. 10



4) are probably due to the WATER output and independently from it, to the chemical compounds (NITRATE, PHOSPHATE and several others influencing COD). Below Miskolc, a distribution of score values can only be observed at the score plot (Fig. 6) without significant direction concerning the outliers (Sampling place No. 5 is Kesznyéten). This way, the loading values as well as the scores can help to gain a quick and accurate demonstration of the state of the river in various seasons or sampling places without detailed studies of the several (100-1000 etc.) data points.

b.) Investigation of arising of ozone from NO_x near ground

1.) The first PC correlates with O_3, SO_2, HUM, TEMP and SUN, the PC 2 with monoxides: NO and CO; the PC 3 with wind velocity and direction whereas PC4 with NO_2.
2.) On the loading plots (Fig. 7-8) the point for humidity is an outlier; it is separated clearly from the other points and clusters. Two definite clusters can be observed: (i) pollutant as CO, NO and (ii) SO_2 and sun exposure.
3.) Examining the loadings from the third direction show four clusters (Fig. 9). The closest resemblance can be observed between moisture content and temperature (cf. Figure 7: their points were the farthest in Figure 7.) The two factors exert reverse effect. The other clusters have already been observed in Figures 7 and 8. The presence of NO_2 in any cluster is arbitrary (apparent) only. It is well separated from the other factors in the fourth dimension.
4.) There is a strong difference between the daily and night temperature as well as between sun exposure during day and night. Consequently the humidity also changes during the day considerably. Score plots may show whether these changes become apparent (Fig. 10). The first PC separates day and night clearly. Three definite clusters can be observed: cluster 1 is on the left side (No. 45-48 and 1-8: from 22^h to 4^h). Cluster 2 is located on the right side (No. 18-42: from 09^h to 21^h). The 3^{rd} one contents the early morning data (No. 9-17). Consequently, the PCA automatically separate data, which represent the chemical conditions for the arising of ozone by $NO_2 + O_2 = NO + O_3$ reaction (cluster 2).
5.) Predictive models was built using MLR and PCR. The MLR results are summarized below:

$$O_3 = 92.53 - 0.2396*NO - 0.7361*HUM - 0.4836*NO_2 - 0.02250*DIR$$

$R = 0.98894$; $F(4,20)=222.30$; $p<.00000$; $S = 1.83$

Slightly better model can be derived using PCR:

$$O_3 = 39.60 - 8.638*F1 - 0.7339*F2 - 3.439*F3 - 6.069*F4$$

$R = .99024$; $F(4,20)=252.44$; $p<.00000$; $S = 1.72$

4. REFERENCES

[1] P. Zhang, N. Dudley, A.M. Ure, D.Littlejohn, Anal. Chim. Acta, 258 (1992) 1-10.

[2] A. Melloul, M. Collin, J. Hydrol. (Amsterdam), 140 (1992), 49-73.

[3] J.M. Barradas, E. C. Fonseca, E. Ferreira da Silva, H. G. Pereira, Appl. Geochem., 7 (1992), 563-72.

[4] R. Henrion, G. Henrion, G.C. Onuoha, Chemometrics Intell. Lab. Syst., 16(1992)

[5] Y-S. Yu, S. Zou, Water Resour. Bull., 29 (1993), 797-806

[6] I. Pardo, Arch. Hydrobiol., 132 (1994), 95-114

[7] T. Berg, O. Roeyset, E. Steinnes, Environ. Monit. Assess., 31 (1994), 259-73

[8] K. Gurunathan, S. Ravichandran, IAHS Publ., 219 (1994), 343-6. 87-94.

[9] K. Héberger: *Chemometrics Intell. Lab. Syst.* **47** (1999) 41-49.

[10] K. Héberger, M. Görgényi, *J. Chromatogr. A*, **845** (1999) 13-20.

[11] P. Geladi and L. Hadjiiski and P. Hopke: *Chemometrics Intell. Lab. Syst.*, **47** (1999) 165-173.

[12] A. Lengyel, L. Paksy, O. Bánhidi: Proceedings of microCAD'99 Conference, Miskolc, 4-5. March, 2000. Section A. pp.65-69.

[13] A.M.C. Davies: *Spectroscopy Europe* 8/6 (1996), pp. 26-28.

[14] A. Lengyel, L. Paksy, O. Bánhidi: IX Italian-Hungarian Symposium on Spectrochemistry, October 11-15, 1999, Siena, Italy.

[15] K. G. Paterson, J. L. Sagady, D. L. Hooper, Environ. Sci. Technol. **33** (1999) 635-641.

[16] E. Alvarez, F. de Pablo, C. Tomas, Rivas, , Int. J. Biometeorol. 44 (2000) 44-51.

[17] D.K. Pissimanis, V.A. Notaridou, N.A. Kaltsounidis, P.S. Viglas, Theor. Appl. Climatol. **65** (2000) 49-62.

[18] M. Trainer, D.D. Parrish, P.D. Goldan, J. Roberts, F.C. Fehsenfeld, Atmos. Environ. **34** (2000) 2045-2061.

[19] D. Klaus, A. Poth, M.Voss, Atmosfera **14** (2001) 171-188.

[20] K. Hébeger and R. Rajkó: *QSAR in Environmental Sciences VII.* Eds. Fei Chen and G. Schüürmann, SETAC Press, Pensacola, Florida USA 1997 Chapter 29 pp. 425-433.

[21] R. Rajkó and K. Héberger: Chemometrics Intell. Lab. Syst., **57** 1-14 (2001).

[22] K. Héberger and R. Rajkó: Variable Selection for Environmental Data using Pair-Correlation Method SAR and QSAR in Environmental Research (in press).