Water quality assessment using SVD-based principal component analysis of hydrological data

Petr Praus

Department of Analytical Chemistry and Material Testing, VSB-Technical University Ostrava, 17. listopadu 15, 708 33 Ostrava, Czech Republic

Abstract

Principal component analysis (PCA) based on singular value decomposition (SVD) of hydrological data was tested for water quality assessment. Using two case studies of waste- and drinking water, PCA via SVD was able to find latent variables which explain 80.8% and 83.7% of the variance, respectively. By means of scatter and loading plots, PCA revealed the relationships among samples and hydrochemical parameters which were also confirmed by factor analysis (FA).

In the case of wastewater, these latent variables clearly displayed changes of water composition over time. Drinking water samples were clustered into four groups which were characterised by their typical water composition. On the basis of these results PCA was found to be a suitable technique for water quality assessment.

Keywords: water quality, wastewater, drinking water, principal component analysis, singular value decomposition, factor analysis

Introduction

Real hydrochemical data sets contain not only important information useful for quality assessment and/or treatment technology but also confusing noise. Mostly, measured variables are not normally distributed, often co-linear or autocorrelated, containing outliers, erroneous or nonsense values. In order to reveal mutual dependence or logical structures of data, there are several chemometric procedures generally called as data mining techniques. Some of them are based on the reduction of data dimensionality, such as principal component analysis (Lavine, 2000; Jolliffe, 2002), factor analysis (Malinowski and Howery, 1980; Malinowski, 1991), independent component analysis (Comon, 1994), independent factor analysis (Attias, 1998), generative topographic mapping (Bishop et al., 1998), etc.

PCA is used to search new abstract orthogonal principal components (eigenvectors) which explain most of the data variation in a new coordinate system. Each principal component (PC) is a linear combination of the original variables and describes a different source of variation (information). The largest or 1st PC is oriented in the direction of the largest variation of the original variables and passes through the centre of the data. The 2nd largest PC lies in the direction of the next largest variation, passes through the centre of the data and is orthogonal to the first PC, and so forth.

Classical PCA is based on the decomposition of a covariance/correlation matrix (Geladi and Kowalski, 1986) by eigenvalue (spectral) decomposition (EVD) or by the decomposition of real data matrixes using SVD. Compared with EVD, SVD is a more robust, reliable, and precise method with no need to compute the input covariance/correlation matrix. From a numerical point of view, SVD is well known for its stability and convergence, even in the ill conditioned problems.

In general, SVD decomposes an arbitrary Matrix A (n x p) into three matrices:

$$A = U\,S\,V^T \qquad (1)$$

where:

U (n x n) and V^T (p x p) are orthogonal and normalised matrices, i.e., $U^T U=I$ and $V^T V=I$

S (n x p) is a diagonal matrix with singular values in decreasing order

U columns are the left singular vectors

V^T rows are the right singular vectors.

Computing the SVD consists of finding the eigenvalues and eigenvectors of A A^T and A^T A, respectively. The U columns are eigenvectors of A A^T and the V^T rows are the eigenvectors of A^T A. The powerful property of SVD is compressing the information contained in A into the first few singular vectors which are mutually orthogonal and their importance rapidly decreases after the first columns/rows. The importance of each singular vector is given by the squares of nonnegative diagonal (singular) values of S.

SVD has found a wide range of various applications in molecular dynamic and gene expression analysis (Wall et al., 2003), information retrieval in a technique called Latent Semantic Indexing (Berry et al., 1995), image processing (Zhang et al., 2005), hearing noise filtering (Maj et al., 2001), spectral analysis (Safavi and Abdollahi H., 2001), and so forth.

Multivariate statistical methods, encompassing cluster analysis, PCA, FA and discriminant analysis, have been successfully used in hydrochemistry for many years. Quality assessment of surface water (Simeonov et al., 2003; Vega et al., 1998; Wunderlin et al., 2001), groundwater (Reghunath et al., 2002), and environmental research (Ceballos et al., 1998; Lambrakis et al., 2004) employing multicomponent techniques are well described in the literature.

☎ +420-59-732-3370; fax: +420-59-732-3370;
e-mail: petr.praus@vsb.cz

Received 7 June 2005; accepted in revised form 12 August 2005.

The aim of this paper is to demonstrate PCA using SVD (PCA/SVD) of real hydrological data matrixes for mining information which is important for wastewater and drinking water quality assessment. PCA/SVD was tested in two case studies:

- Wastewater screening before a treatment process.
- Classification of drinking water quality monitored in a city area.

Experimental

Hydrological data sets

The laboratory samples provide information intended for wastewater monitoring and water quality mapping on an industrial city area. The use of PCA/SVD is demonstrated in two case studies:

- The results of chemical and physical parameters of wastewater, including municipal and domestic wastes, taken monthly on the inlet of a small biological wastewater treatment (BWWT) plant during five years. The determined parameters were selected with respect to the technology control.
- The results of chemical and physical analyses of drinking water samples taken from the city water network. Microbiological parameters analysed contained zero values and thus were not included in the testing data set.

Water analyses

Water analyses, including sampling and preservation, were carried out according to the actual standard ISO and EN methods. The measured parameters in wastewater were pH, biochemical oxygen demand within 5d (BOD), chemical oxygen demand by dichromate (CODCr), ammonium, nitrate, nitrite, phosphate, total phosphorus (TP), total nitrogen (TN), total suspended solids (TSS), and total dissolved solids (TDS).

The parameters determined in drinking water include pH, ammonium, nitrate, nitrite, colour, turbidity, calcium, electrical conductivity (EC), alkalinity, chemical oxygen demand by permanganate (CODMn), iron, and free chlorine (FC) carried out by means of the DPD method.

Total nitrogen was determined by the Kjeldahl procedure and total phosphorus by peroxodisulphate oxidation. Ammonium, nitrate, nitrite, phosphate, free chlorine, and colour were determined by UV-VIS spectrometry (DR 4200, HACH). Turbidity was measured by nephelometry (Turbiquant 1500 IR, Merck). Calcium, alkalinity, and both types of chemical oxygen demand methods were determined volumetrically. The concentrations of iron were measured by flame atomic absorption spectrometry (Spectra AA200, Varian). TDS and TSS were determined gravimetrically after sample filtration through the 0.85 µm membrane filters. Electrochemical measurements were used for the determinations of dissolved oxygen (DO) (Oxi 320, WTW), conductivity (Jenway 4310), and pH (pH 197, WTW).

Multivariate computations

The data matrices were prepared and processed in Excel 97. Their rows were constructed from the variables analysed in waters. There were no missing values in the data sets. SVD of Matrix *A* was executed using the standard MATLAB command svds(A,k) which computes the k largest values and associated singular vectors of Matrix A. Factor analysis and other statistical calculations were performed by the software packages STATGRAPHIC Plus 5.0. The factor loadings were calculated using the Varimax rotation method.

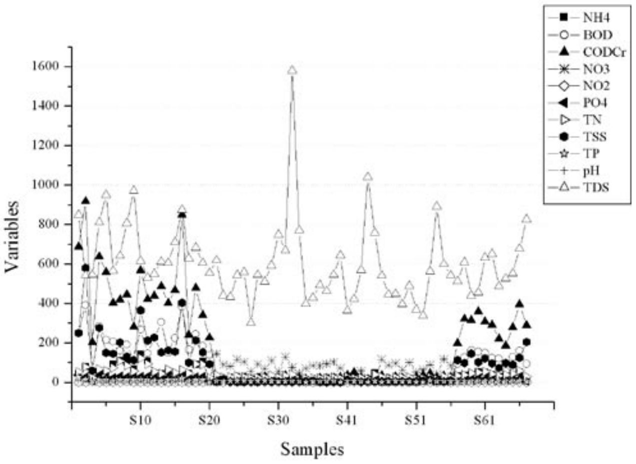

Figure 1a
Plot of the wastewater variables

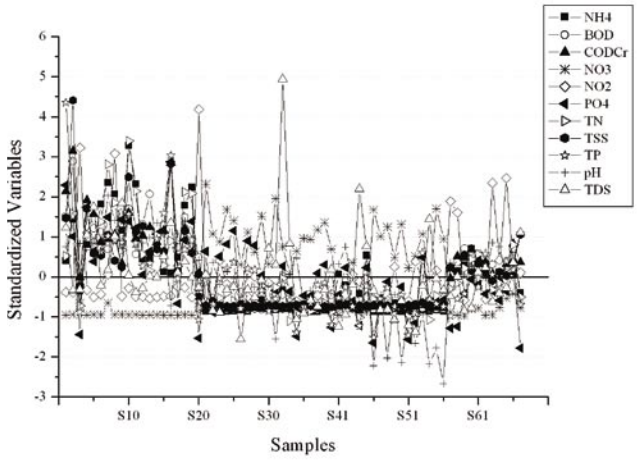

Figure 1b
Plot of the standardised wastewater variables

Before computation, the testing data were standardised in order to avoid misclassifications arising from different orders of magnitude of tested variables. Therefore the data were mean (average) centred and scaled by the standard deviations: $(x_i - \bar{x})/s$

Results and discussion

Case study I – SVD analysis of the municipal wastewater data

Domestic wastewater samples (n = 68) were taken at the inlet of a small BWWT plant. This type of wastewater consists of fall-outs of household, humans, and commercial institutions. Information about the variability of the amount of organic and inorganic wastes is necessary for monitoring and optimising BWWT processes. PCA/SVD was performed on the data matrices which summarised the given above chemical and physical determinations. The time fluctuation of waste water composition is demonstrated in Fig. 1a. Because of the very different scales of individual variables the data were centred and scaled by the standard deviations (Fig. 1b) and thus transformed data were treated by PCA via SVD.

SVD splits the data matrix into several mutually independent singular vectors which describe the most variation in wastewater

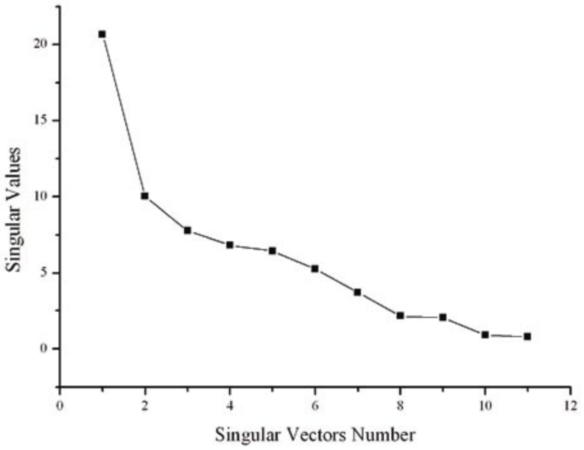

Figure 2
Scree plot of the wastewater singular values

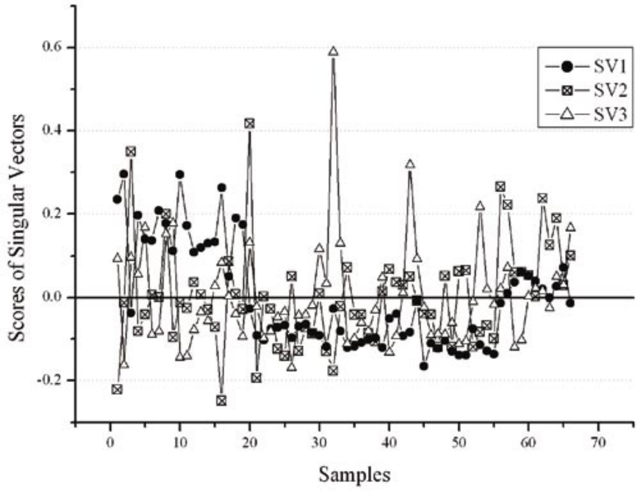

Figure 3a
Time dependence of the wastewater singular values

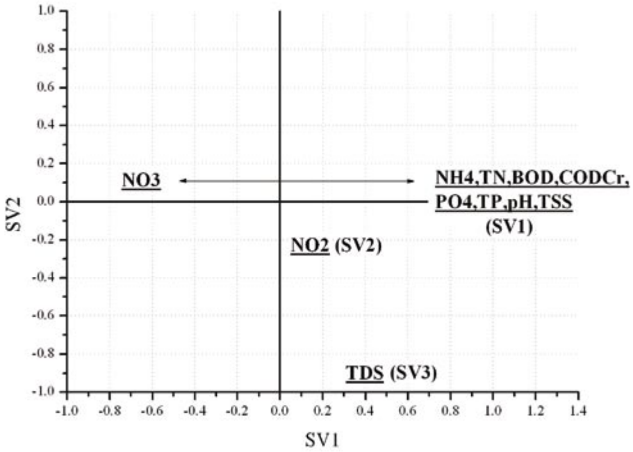

Figure 3b
Loadings plot of the wastewater singular values SV1 and SV2

TABLE 1
Some statistics of the first singular vector

Sample range	Skewness	Kurtosis	Variance	Average
S1 to S20	-0.4223	3.4694	0.006585	0.1635
S21 to S57	0.4762	2.9665	0.001211	-0.09396
S58 to S68	-0.003170	1.8657	0.0008249	0.02572
Note: No autocorrelation was detected. Normality was confirmed by statistical tests.				

composition. The number of singular vectors (SVs) can be estimated from a scree plot demonstrated in Fig. 2. As it is shown in this figure, the singular values sharply decrease within the first three singular vectors and then slowly stabilise for the remaining ones which contain a great deal of noise and therefore are not useful. Taking into account the SVD theory, the singular values correspond to the square roots of the eigenvalues. That is why the variance of SVs can be expressed according to the equation:

$$\text{var.} = \frac{s_k^2}{\sum_1^n s_i^2} \qquad (2)$$

where:

s_k is a singular value

SV1 to SV3 are 58.7%, 13.8%, and 8.3% (in sum 80.8%) of the data variance, respectively. Figure 3a shows that SV1 contains the greatest amount of information and in contrast with SV2 and SV3, which are higher in noise content, is smooth and exhibits the significant decrease of SV1 in the period from the samples 21 to 58.

The composition of the revealed singular vectors is shown from the loading plot in Fig. 3b. This plot is constructed from loadings calculated as correlations of the three selected singular vectors and all variables. The 1st SV contains mostly the parameters indicating organic wastes, such as BOD, CODCr, TN, ammonia, TP, phosphate, and nitrate which negatively correlate with other parameters as results of the ammonia biological oxidation (nitrification). The 2nd SV reveals the nitrite influence as a nitrification intermediate product. The 3th SV is closely connected with TDS mainly including inorganic salts.

The individual SVs can be statistically tested as well as the original variables. SV1 exhibits normal distribution within the three seasons (Table 1). One can assume that the distinct decrease of SV1 in Fig. 3a was caused by the temporary reduction of domestic wastes likely from food or similar type of industry and/or civic amenities in this area. On the other hand, SV2 and SV3 behave in quite a different way. Their graphs show several accidental extremes. After their exclusion from the data both SVs were normally distributed.

It is shown that these three singular vectors extracted from 10 variables can give unbiased information about the significant seasonal changes in wastewater composition. Outliers of all singular vectors can be easily detected for further inspection.

The PCA/SVD outputs were confirmed by factor analysis. In FA, each variable can be expressed as a linear combination of latent common factors and a single specific factor:

$$y_i = \sum_{i=1}^{n} \alpha_{ij} F_j + \beta_i e_i \qquad (3)$$

where:

y_i are the original variables

F_j and e_i are the common and specific (error) factors, respectively

α_{ij} and β_i are their factor loadings

FA separates a correlation matrix into two matrices: a common factor portion and a specific factor portion. The main difference between PCA and FA is that while PCA is concerned with the total variation as expressed in the correlation matrix, FA is concerned with a correlation in the common factor portion. The goal of FA is not only to reduce the data dimensionality as well

TABLE 2
The wastewater factor loadings after the Varimax rotation

Parameters	Factor 1	Factor 2	Factor 3
NH_4	0.89674	0.00338	0.05005
BOD	0.92257	0.05436	0.13427
CODCr	0.92414	0.08960	0.16778
NO_3	-0.78544	0.38009	-0.13785
NO_2	0.14924	-0.81050	0.02769
PO_4	0.59404	0.644383	0.21996
TN	0.87734	0.18772	0.07517
TSS	0.88812	0.12694	0.08705
TP	0.72241	0.47901	0.26946
pH	0.76232	-0.20536	0.01123
TDS	0.11555	0.04054	0.97975

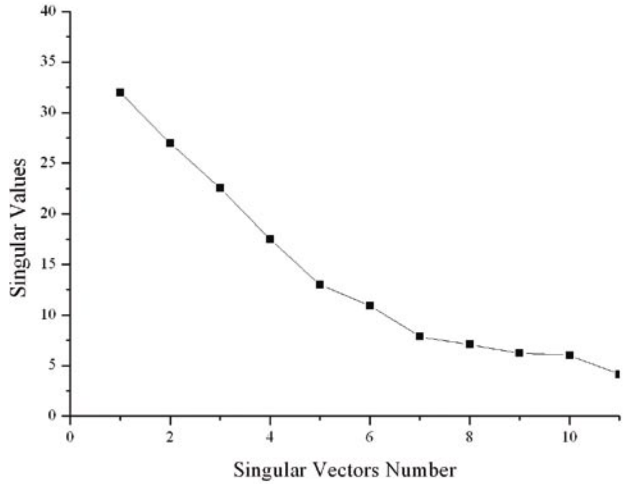

Figure 4
Scree plot of the drinking water singular values

as PCA but also to interpret the revealed common factors. The methods of factor computations including the detailed explanation of FA are described in the literature cited above.

It is obvious from the factor loadings in Table 2 that the results of FA and PCA are in a good agreement with each other.

Case study II – SVD analysis of the drinking water data

Drinking water samples (n = 253) were taken for the sake of regular screening of drinking water quality in a supply system. Drinking water is produced by treatment of ground- and surface water which are delivered and mixed together within the water network. Water composition was monitored using 12 chemical and physical variables.

The singular values and singular vectors were computed from this data matrix. The scree plot of singular values is displayed in Fig. 4. The four most significant singular vectors explain about 83.7% of variance: 33.4% by SV1, 23.7% by SV2, 16.6% by SV3, and 10% by SV4. Mapping of samples is demonstrated in the loading plots of Figs. 5a; b. The four groups (I to IV) are clearly visible in Fig. 5b. For understanding of the sample clustering, the loading plots were prepared and they are demonstrated in Figs. 6a, b.

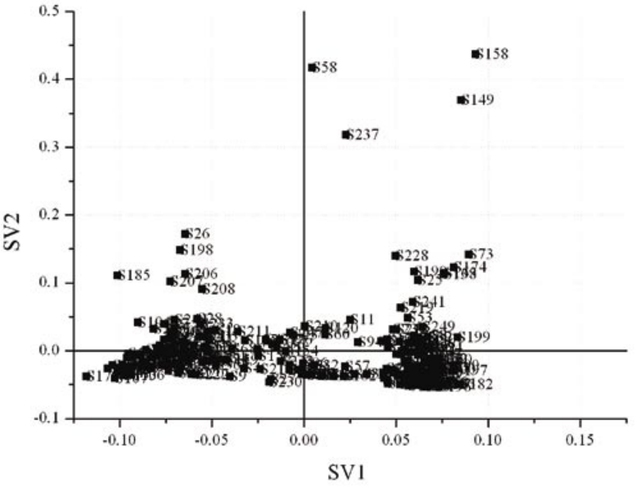

Figure 5a
Scatter plot of the drinking water singular values SV1 and SV2

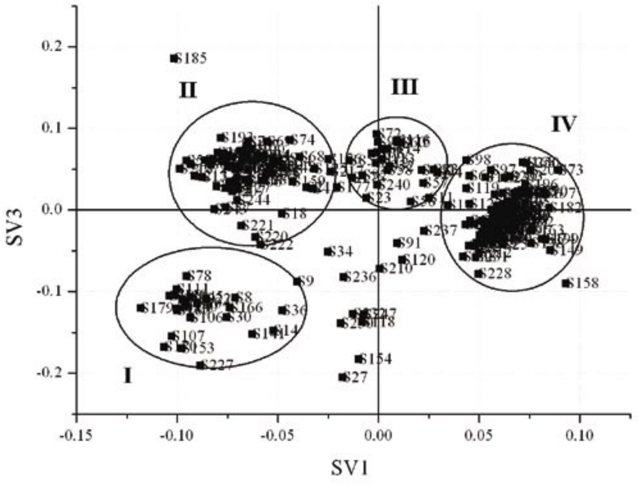

Figure 5b
Scatter plot of the drinking water singular values SV1 and SV3

It is shown in Fig. 6b that the four distinct groups of samples can be characterised by the corresponding groups of parameters. The samples of Group I are typical in terms of their higher concentrations of nitrite and lower values of pH. This water composition exists in several sources of groundwater from which treated water is delivered into the network. Group II contains the water samples with higher conductivity resulting from the higher concentrations of salts, such as chloride, sulphate, and calcium, which originate from the rest of the groundwater sources. Group III indicates samples of lower levels of free chlorine and the samples clustered into the group IV contain higher amount of iron and related parameters, such as turbidity and colour which give information about the pipeline system corrosion.

The PCA/SVD loadings were confirmed by FA. The extracted factors are summarised in Table 3. The 1st factor can be called as the salt factor because it contains conductivity and inorganic salts. The 2nd factor is connected with dissolved iron ions and related turbidity and colour as given above. The 3rd factor is composed of nitrite and pH which is caused by the presence of drinking water treated from groundwater sources. The 4th factor represents free chlorine. It is obvious that all factors are consistent with the groups of parameters revealed by PCA/SVD.

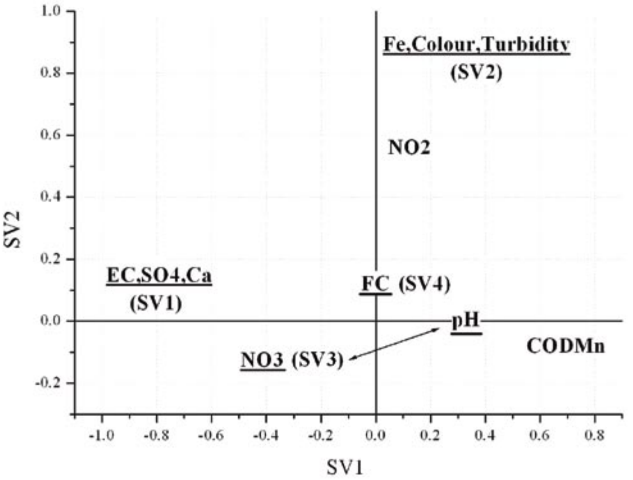

Figure 6a
Loadings plot of the drinking water singular values SV1 and SV2

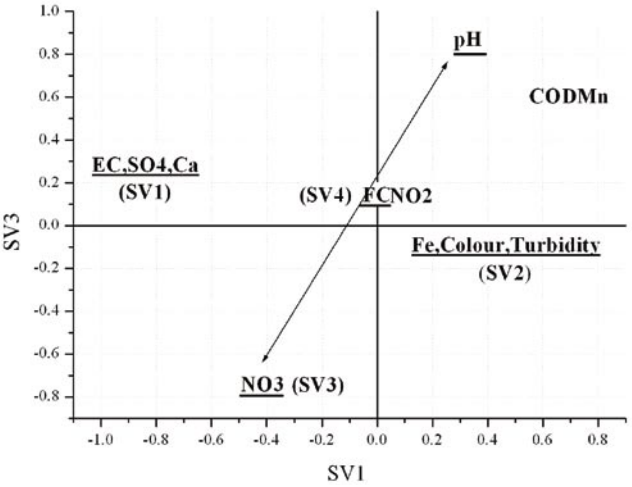

Figure 6b
Loadings plot of the drinking water singular values SV1 and SV3

Conclusion

In this paper PCA based on the SVD algorithm was applied for the multivariate analysis of the real hydrological data sets. It was found that PCA/SVD is able to extract latent variables -singular vectors from noisy hydrological data. The dependence of singular vectors on time can give information about seasonal changes of water composition as was demonstrated in the case of wastewater. Moreover, the singular vectors can be presented in form of the scatter and loading plots which allow mapping of water quality in different localities of the supply system. The revealed connections among variables were verified by very similar results of factor analysis.

Unlike statistical parametric tests (t-tests or F-tests) which require the normal distribution of variables, no such assumption is necessary for PCA. PCA is a data analytical, rather than statistical, method and can indicate associations between samples and/or variables. It was demonstrated that PCA/SVD easily provides an unbiased view of water composition and thus can be used as a very useful tool for water quality assessment.

TABLE 3
The drinking water factor loadings after the Varimax rotation

Parameters	Factor 1	Factor 2	Factor 3	Factor 4
Colour	0.00055	0.90947	0.03269	0.02806
Calcium	0.93978	-0.03456	-0.10802	-0.08339
Free chlorine	0.03804	-0.07620	-0.00867	0.89414
Chloride	0.92829	-0.07078	-0.14143	0.13849
CODMn	-0.30410	-0.08331	0.75551	0.09602
Nitrate	0.15870	-0.13011	-0.88938	0.01008
Nitrite	0.03326	0.42720	0.08796	0.64029
Fe	-0.00280	0.94463	-0.02836	-0.00931
pH	0.06732	-0.02514	0.91882	-0.00956
Sulphate	0.96746	-0.00603	-0.12072	-0.00366
Conductivity	0.97867	0.02952	-0.04778	0.05154
Turbidity	-0.06947	0.90664	0.00538	0.13276

Acknowledgement

The author kindly thanks Pavel Praks (VSB-Technical University Ostrava) for the programming of SVD in MATLAB. This work was supported by the Ministry of Education, Youth and Sport of the Czech Republic (MSM 6198910016).

References

ATTIAS H (1998) Independent factor analysis. *Neural Comput.* **11** (4) 803-851.

BERRY MW, DRMAČ Z and JESSUP ER (1995) Matrices, vector spaces and information retrievel. *SIAM Rev.* **41** (2) 335-362.

BISHOP CM, SVENSÉN M and WILLIAMS CKI (1998) GTM: The generative topographic mapping. *Neural Comput.* **10** (1) 215-234.

CEBALLOS BSO, KÖNIG A and OLIVEIRA JF (1998) Dam eutrophication: A simplified technique for a fast diagnosis of environmental degradation. *Water Res.* **32** (11) 3477-3483.

COMON P (1994) Independent component analysis, a new concept? *Signal Process.* **36** (3) 287-314.

GELADI P and KOWALSKI BR (1986) Partial least square regression: A tutorial. *Anal. Chim. Acta* **185** 1-17.

GELADI P (2003) Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochim. Acta Part B* **58** (5) 767-782.

JOLLIFFE IT (2002) *Principal Component Analysis* (2nd edn.). Springer-Verlag, New York.

LAMBARKIS N, ANTONAKOS A and PANAGOPOULOS G (2004) The use of multicomponent statistical analysis in hydrogeological environmental research. W*ater Res.* **38** (7) 1862-1872.

LAVINE BK (2000) Clustering and classification of analytical data. In: Meyers RA (ed.) *Encyclopaedia of Analytical Chemistry* (.). John Wiley & Sons, Chichester.

MAJ JB, MOONEN M and WOUTERS J (2002) SVD-based optimal filtering technique for noise reduction in hearing aids using two microphones. *Eurasip JASP* **4** 432-443.

MATHER P (1976) *Computational Methods of Multivariate Analysis in Physical Geography.* John Wiley & Sons, New York.

MALINOWSKI ER (1991) *Factor Analysis in Chemistry* (2nd edn.). John Wiley & Sons, New York.

MALINOWSKI ER and HOWERY DG (1980) *Factor Analysis in Chemistry.* John Wiley & Sons, New York..

REGHUNATH R, MURTHY TRS and RAGHAVAN BR (2002) The utility of multivariate statistical techniques in hydrochemical studies: An example from Karnataka, India. *Water Res.* **36** (10) 2437-2442.

SIMEONOV V, STRATIS JA, SAMARA C, ZACHARIADIS G, VOUTSA D, ANTHEMIDIS A, SOFONIOU M and KOUIMTZIS TH (2003) Assessment of the surface water quality in Northern Greece. *Water Res.* **37** (17) 4119-4124.

SAFAVI A and ABDOLLAHI H (2001) Thermodynamic characterization of weak association equilibria accompanied with spectral overlapping by a SVD-based chemometric method. *Talanta* **53** (5) 1001-1007.

VEGA M, PARDO R, BARRADO E and DEBÁN L (1998) Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Res.* **32** (12) 3581-3592.

WALL ME, RECHTSTEINER A and ROCHA LMM (2003) Singular value decomposition and principal component analysis. In Berrar DP, Dubitzky W and Granzow M. (eds.) *A Practical Approach to Microarray Data Analysis.* Kluwer, Norwell. 91-109.

WUNDERLIN DA, DÍAZ MP, AMÉ MV, PESCE SF, HUED AC and BISTONI MA (2001) Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquía river basin (Córdoba-Argentina). *Water Res.* **35** (12) 2881-2894.

ZHANG D, CHEN S and ZHOU ZH (2005) A new face recognition method based on SVD perturbation for single example image per person. *Appl. Math. Comp.* **163** (2) 895-907.