Scientific paper

Chemometric Characterisation of the Quality of Ground Waters from Different Wells in Slovenia

Ernest Vončina,[a] Darinka Brodnjak Vončina,[b] Nataša Mirkovič,[a] Marjana Novič[c]

[a] *Institute of Public Health Maribor, Institute of Environmental Protection, Prvomajska 1, 2000 Maribor, Slovenia*

[b] *Faculty of Chemistry and Chemical Engineering, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia. Tel: +386(2)2294432, Fax: +386(2)2527774, E-mail: darinka.brodnjak@uni-mb.si.*

[c] *National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia*

Received 10-06-06

Paper based on a presentation at the 12th International Symposium on Separation Sciences, Lipica, Slovenia, September 27–29, 2006.

Abstract

The quality of ground water as a source of drinking water in Slovenia is regularly monitored. One of the monitoring programmes is performed on 5 wells for drinking water supply, 3 industrial wells and 2 ground water monitoring wells. Two hundred and fourteen samples of ground waters were analysed in the time 2003–2004. Samples were gathered from ten different sampling sites and physical chemical measurements were performed. The following 13 physical chemical parameters were regularly controlled: temperature, pH, conductivity, nitrate, AOX (adsorbable organic halogens), metals such as chromium, pesticides (desethyl atrazine, atrazine and 2,6-dichlorobenzamide), highly-volatile halogenated hydrocarbons (trichlorometane, 1,1,2,2-tetrachloroethene and 1,1,2-trichloroethene). For handling the results different chemometrics methods were employed, such as basic statistical methods for the determination of mean and median values, standard deviations, minimal and maximal values of measured parameters and their mutual correlation coefficients, cluster analysis (CA), the principal component analysis (PCA), the clustering method based on Kohonen neural network, and linear discriminant analysis (LDA). The study gives the opportunity to follow the quality of ground waters at different sampling sites within the defined time period. Monitoring of general pollution of ground waters and following measuring can be used to search the pollution source, to plan prevention measures and to protect from pollution, as well.

Keywords: ground waters, water quality, chemometrics, principal component analysis, classification, Kohonen neural networks

1. Introduction

The physical and chemical studies on ground waters gathered from ten different sampling sites were performed. Here, we present the data collected in the time period 2003–2004. Through this period the quality of the water was followed and the classification has been made according to sampling sites. The classification is based on 13 physical and chemical parameters. The aim of this work is to find the correlation between sampling sites and the variables obtained by chemical measurements, which can be used to construct a fast decision model for separating different water quality samples.

Chemometrics methods have been often used for the classification and comparison of different samples.[1] Seasonal, spatial and polluting effects on the quality of river water were examined by exploratory data analysis.[2–9] Some examples of chemometrics characterizations are, for instance, the differentiation of rainwater compositional data by principal component analysis (PCA),[10] application of chemometric techniques to the analysis of river water quality,[11–12] identification of sources of bottom waters in the Weddel Sea by PCA and target estimation,[13] determination of correlation of chemical and sensory data in drinking waters by factor analysis,[14] to name just a few. Chemometrics methods have been used also for evaluating environmental data of Lagoon water,[15] San Francisco Bay and Estuary,[16] and Muggia Bay in Northern Adriatic Sea.[17] They were used also for the oceanographic characterization of northern Sao Paulo cost.[18] PCA and PLS were used for the characterisation of wastewater in Melbourne (Australia).[19] An example of using Kohonen

maps is given in a paper discussing the unsupervised training, clustering and classification of multivariate biological data.[20]

The quality of the ground waters was studied through the years 2003 and 2004. The monitoring programme was performed on 5 wells for drinking water supply, 3 industrial wells and 2 ground water monitoring wells. Altogether 13 characteristic features were measured for 214 samples collected and analysed during this period. Several chemometrics methods were applied in order to visualize multivariate data and to enable a quick classification of samples, regarding the source location within the studied time period.

2. Experimental

A standard method was used for sampling.[21] Water was collected in polyethylene bottles 0.5 m below the surface. All glass and plastic ware used for sampling and analyses were rinsed with milli-Q water.

Standard analytical methods were used for the determination of 13 physico-chemical variables**.** GC/MS Hewlett Packard was used for the determination of pesticides, ion chromatograph Dionex was used for the determination of nitrates, while Strohlein apparatus and WTW conductivity meter were employed for the AOX and electrical conductivity measurements, respectively. All reagents were analytical grade. The milli-Q system was used for purifying the water.

2.1. Data Analysis

The 214 samples are characterized by 13 physical and chemical variables: (1) pH, (2) water temperature, (3) electrical conductivity, (4) nitrate content, (5) adsorbable organic halogens (AOX), (6) chromium(VI) content and (7) total chromium content, (8) desethylatrazine content, (9) atrazine content, (10) 2,6-dichlorobenzamide content, (11) trichloromethane content, (12) 1,1,2,2-tetrachloroethene content, and (13) 1,1,2-trichloroethene content. The enumerated variables are the components of the vector representation of each sample which is used in further chemometric analysis. The results of all measurements have been investigated by different chemometrics methods:[1] the basic statistical methods for the determination of mean and median values, standard deviations, minimal and maximal values of measured variables and their mutual correlation coefficients. The PCA[1, 22] and artificial neural networks[23] were applied for grouping of water samples due to measured variables. Among different neural networks the Kohonen neural networks with self organising maps[24] are the most suitable for clustering.[23, 25–28] All the calculations and plots in the following (PCA) section were done with the Teach/Me software[22] using Teach/Me Data Analysis option which is one of the applications of the Teach/Me system, providing very flexible tools for most fields of data analysis.

3. Results and Discussion

3.1. Statistical Screening of Data

After determining mean and median values, and standard deviation, the mutual correlation was sought for all measured variables. The maximal correlation coefficient of the data was found between measurements of nitrate content and electrical conductivity (r = 0.92), (Figure 1), between atrazine and 2,6-dichlorobenzamide (r = 0.89), and between 2,6-dichlorobenzamide and chromium(VI) (r = 0.85). The correlation between nitrate content and electrical conductivity is expected to be high. The correlation between atrazine and 2,6-dichlorobenzamide shows the hot spot of pollution which is caused by pesticides used for the destruction of weeds. The correlation between 2,6-dichlorobenzamide and chromium(VI) shows also the hotspot of antropogenic contaminants. It shows the overall pollution of water springs.

Cluster analysis resulted in a dendrogram shown in Figure 2, where all 214 samples are divided into a number of clusters, depending on the level of similarity based on Ward distance. Only one group of samples, namely sampling site “4” (the right-most cluster, blue colour) is well distinguished from other samples.

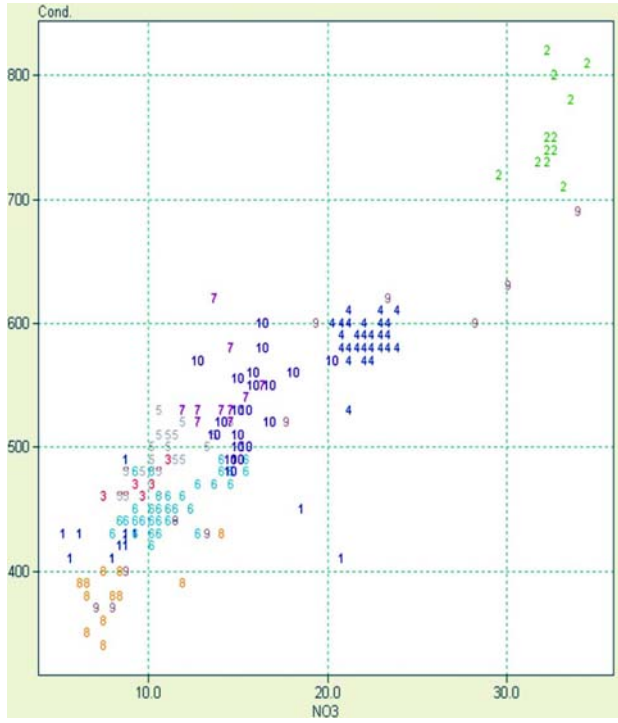

Figure 1. Correlation between nitrate content (NO_3^-) and electrical conductivity. The sampling sites (classes) are denoted by numbers from 1 to 10.

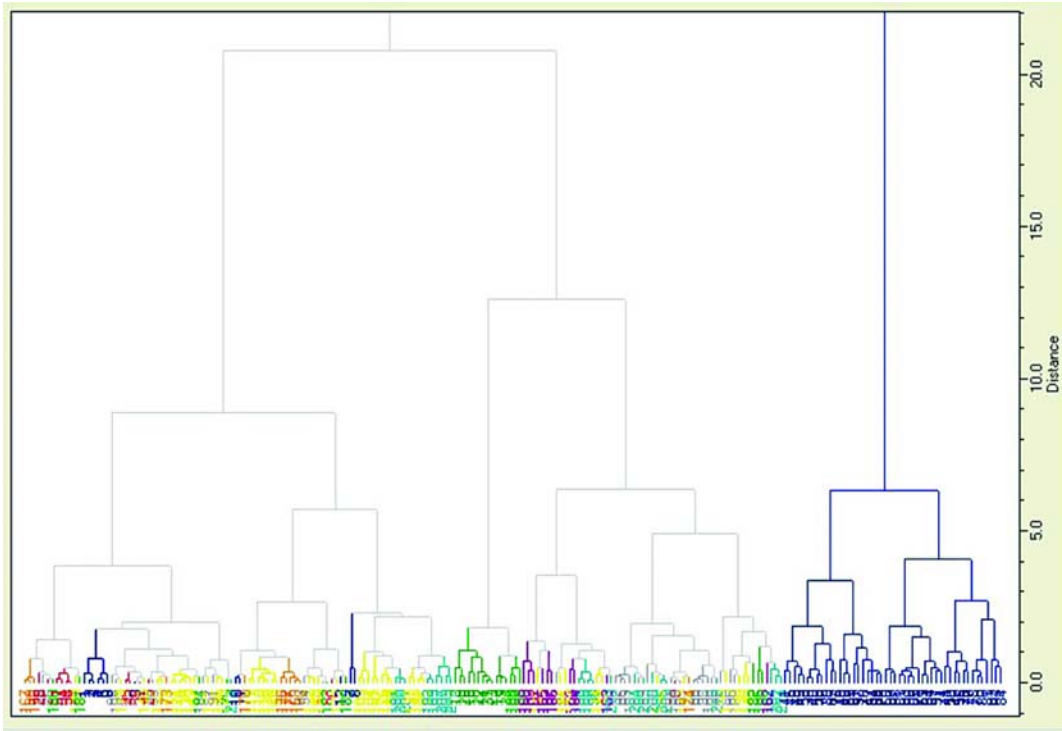

Figure 2. Dendrogram of 214 samples. Different shades correspond to 10 different sampling sites (classes).

3.2. Principal Component Analysis (PCA)

PCA was performed in order to get an overall impression about the correlation of 214 water samples, described with physical and chemical variables, with the quality of water in different sampling sites. PCA was applied on the matrix composed of 214 × 13 elements. 214 rows represent water samples composed of 13 variables. Data were additionally preprocessed in two different ways. First, "Column centering" of the data was used, which means that the mean value of each column was subtracted from individual (214) elements. Second, autoscaling of individual variables was performed, called "Column standardization". With this procedure the mean of the column elements is subtracted from individual elements and divided by the column standard deviation. Consequently, each column has zero mean and unit variance. The percentages of variances in resulting eigenvectors (PCs) for both types of preprocessing of the data are shown in Table 1.

Table 1: Comparison of variances in PCA using two different scaling procedures: autoscaling (m = 0.0, s = 1.0) and column centering of data (m = 0.0)

PC	Column standardization (Autoscaling) % variance	total	Column centering % variance	total
1	48.97	48.97	98.28	98.28
2	12.55	61.52	1.12	99.40
3	8.01	69.52	0.39	99.79
4	6.33	75.85	0.10	99.89
5	5.39	81.24	0.07	99.96
6	5.29	86.53	0.03	99.98
7	4.28	90.81	0.01	100.00
8	3.27	94.09	0.00	100.00
9	2.74	96.82	0.00	100.00
10	1.19	98.02	0.00	100.00
11	1.08	99.10	0.00	100.00
12	0.50	99.60	0.00	100.00
13	0.40	100.00	0.00	100.00

From Table 1 it can be seen that using "Column standardization" data, 62% of variance is gathered in the first two PCs. On the other hand, column centering of data resulted in PCA with very high variance in the first axis. The reason is in the selected units of individual variables, which makes the ranges between the variables incomparable. Column centering of data does not scale the data to comparable values, it only moves the average to zero. In the case that one variable is much larger than the others, the large variable has such a strong influence on the first PC axis that this axis contains the majority of variance. In our case, for example, the conductivities are between 400 and 800 (μScm^{-1}), while the values of all other variables are in the range from 0 to 45. As a consequence, the data transformed in such a way do not cluster well, because the rest of the variables do not contribute enough to the first three PC axes.

The PCA with column standardized data were further analysed for formed clusters. It was found from the score plots of the first and the second PCs that samples are well separated according to sampling sites. Clusters of samples from sampling sites 2 and 4 are especially well defined. According to the content of pollution parameters

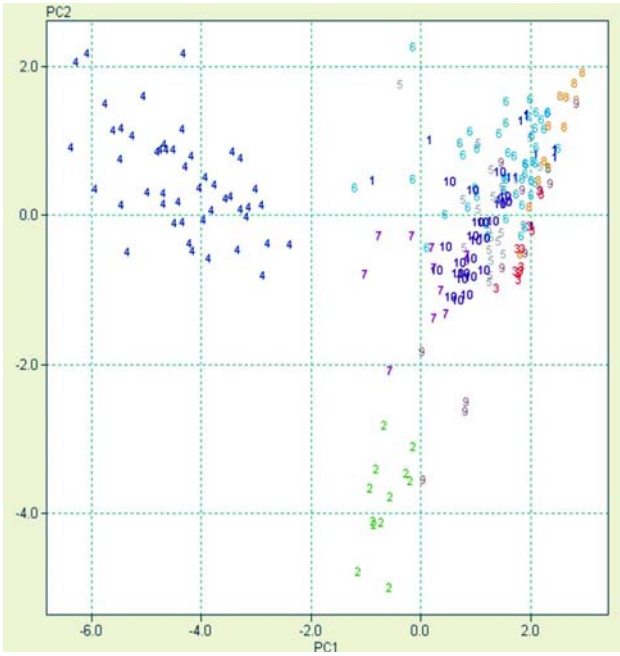

Figure 3a. PCA for all ground water samples from 10 different sampling sites denoted by class numbers from 1 to 10.

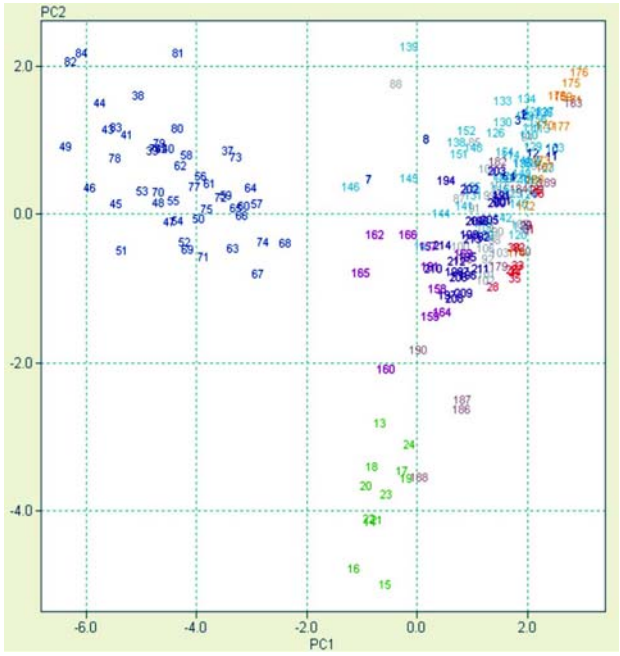

Figure 3b. PCA for 214 ground water samples from 10 different sampling sites; each sample is enumerated (1–214) and classes are associated with different shades.

from the previous mentioned sampling sites it can be concluded that these are also the most polluted sampling sites measured. The cluster of samples from site four, measured during the years 2003 and 2004, is most distant from other samples in the graph of the first two PCs (Figures 3a and 3b).

The scores and loadings plots of PCA of the water samples represented with 13 variables are shown in Figures 3 and 4, respectively. It is evident from Figure 3a that samples separated from the main central cluster and distributed in the region of larger values of PC1 were all collected from sampling sites labelled 2 and 4. The first principal component explains the properties of water samples, which are connected with high pollution. Cluster numbered "4" contains 48 samples from one of the wells. It is known that the main source of pollution is spraying with pesticides. The inspection of particular parameters shows that the content of pesticides is high throughout the whole sampling period in the previous mentioned samples. Cluster numbered "2" is also well separated from others, but the separation is obtained in the second principal component PC2, which is accounted for electrical conductivity and nitrate content (labels 3, 4, see Fig. 4).

It can be seen from Figure 4 that the first component, PC1, is associated with a group of variables such as atrazine(9), 2,6-dichlorobenzamide (10) and 1,1,2,2-tetrachloroethene (12), also variables (6) chromium(VI), (7) total chromium, and (8) desethylatrazine are present. The second component PC2 represents mainly the dependence on water temperature, electrical conductivity and nitrate content (variables 2, 3 and 4). It must be stressed that potential clusters of labelled variables in the loadings plot do not influence the clusters of samples in the score plots. The only important message from the score plots is the magnitude of the components associated with individual variables in each PC. If the absolute value is close to zero, the variable has small influence, as in case of pH, variable 1.

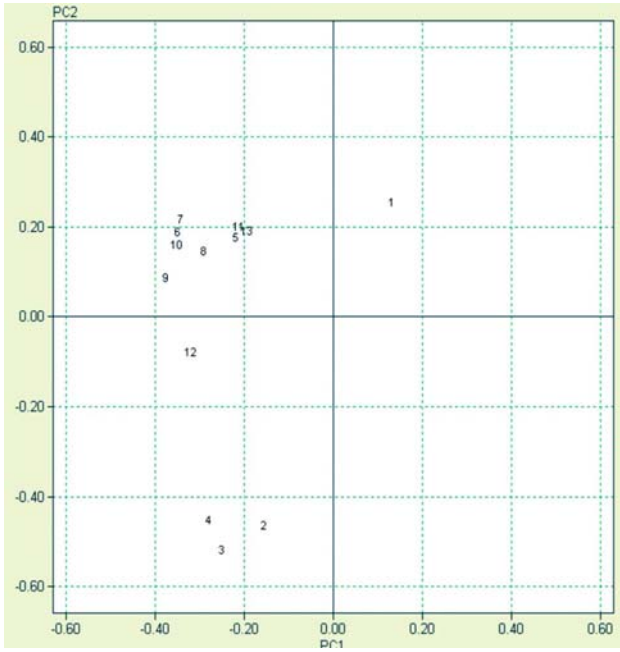

Figure 4. PCA for 214 ground water samples from 10 different sampling sites; the loadings in 13 PC axes are shown.

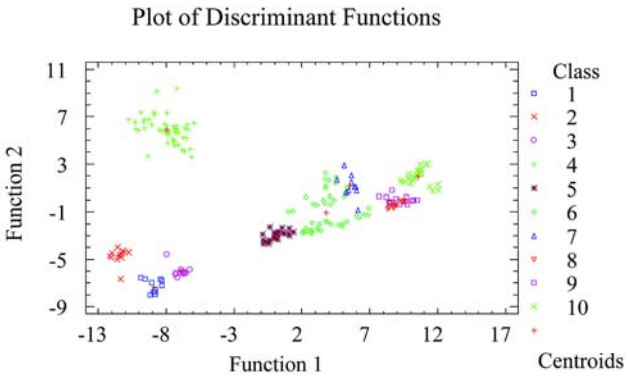

Figure 5. Linear discriminant analysis for 214 samples.

3.3. Linear Discriminant Analysis (LDA)

LDA is a supervised learning method, which can determine the classification into predetermined classes. From Figure 5 it can be seen that 94.9% (203 samples from 214 samples) are accurately classified into 10 predetermined classes.

For further evaluation of the water samples another chemometric method, Kohonen ANNs was implemented.

3.4. Kohonen Neural Network

The Kohonen ANN has been used as a nonlinear mapping method. 13 dimensional neurons were arranged in a rectangular grid. The Kohonen ANN can be used as a non-supervised method for separation and quick classification of samples. A Kohonen neural network with 625 13-dimensional neurons (25 × 25 × 13) was constructed. 214 samples were mapped in a 2D map. Data were preprocessed by scaling column-wise between 0 and 1. The analysis of formed clusters shows, similar to PCA, that samples from classes two and four are well separated from all others (Figure 6).

Regarding the different sampling sites it is evident that the quality of water is worse at the places numbered with 2 and 4. Comparing the levels 6 and 7, corresponding to the variables Cr(VI) and total Cr (Figure 7a and 7b), with the top map in Figure 6, we can see that high values of chromium coincide with the distribution of samples "4" in the Kohonen map. In general, the distribution of weights in individual levels indicates the reason for clustering of samples shown in Figure 6.

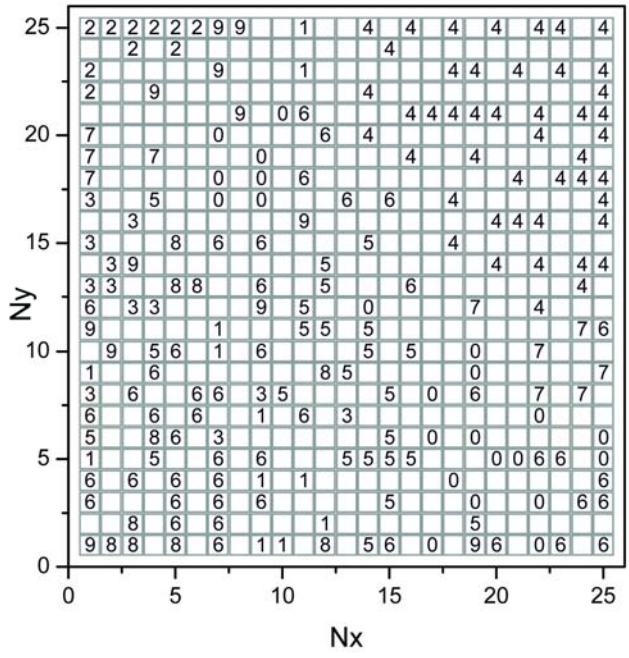

Figure 6. Kohonen map of 214 water samples.

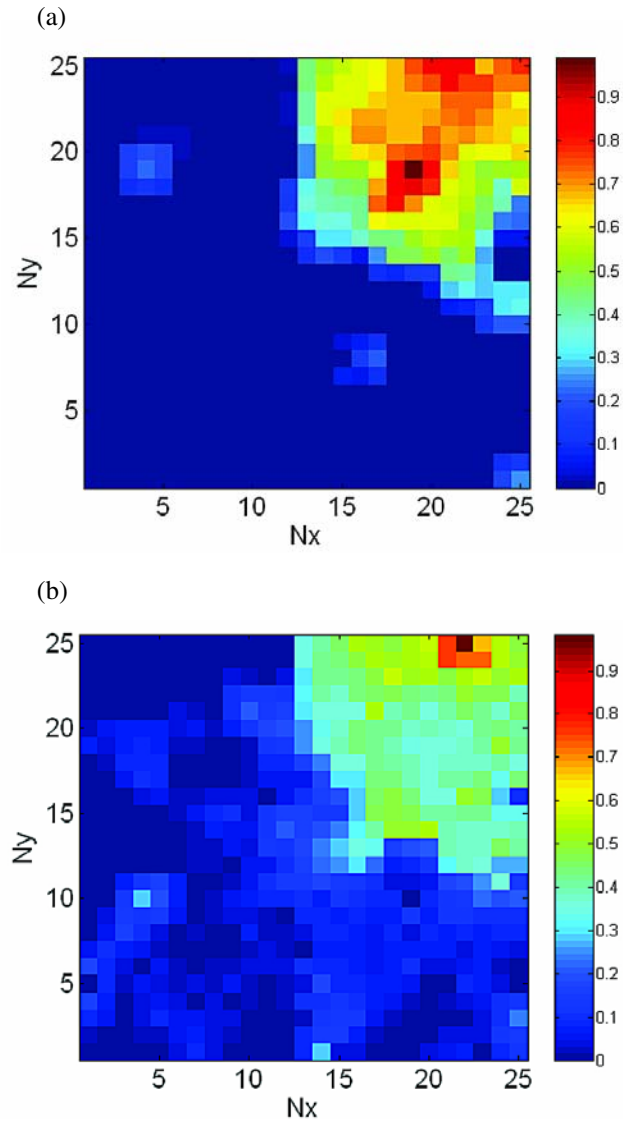

Figure 7. The weights in the 6th Kohonen level corresponding to Cr(VI) (a) and in the 7th Kohonen level corresponding to Cr total (b).

4. Conclusions

The aim of this work was to find the correlation between sampling sites and the variables obtained by chemical measurements. Groundwater quality is influenced by diverse natural and human activities. Negative effects on groundwater quality can arise from small business and industrial production processes, storage and disposal of waste material, contamination of soils by accidents and improper storage of water – hazardous materials, agriculture (input of nutrients and pesticides), leaking sewage pipelines or operation of sewage farms. Water law requires the protection of ground water by establishing the water protection zones and protection of groundwater against pollution.

There is a high risk of pollution of drinking water originating from the groundwater in central part of Slovenia because of geological condition in that area. The high permeability of covering layers results in the relatively unhindered transport of pollutants into groundwater in a relatively short time. An urban area like Ljubljana requires long-term protection of drinking water supplies through groundwater protection measures. Public water suppliers monitor their wells regularly to ensure the water meets applicable water quality standards. Responsible parties and property owners conduct groundwater monitoring at contamination sites to evaluate the extent and severity of contamination, and to monitor effectiveness of their cleanup efforts.

Monitoring is performed at and in the vicinity of water supply sources to determine the quality and trends of indicator of water quality. The program is performed on 5 wells for drinking water supply, 3 industrial wells and 2 groundwater monitoring wells. Two hundred and fourteen samples of ground waters were analysed in the time 2003–2004. Samples were gathered from ten different sampling sites and 13 physical chemical parameters were measured. The quality of ground waters have been analysed by different chemometrics methods.

Cluster analysis (CA) divided 214 samples into a number of clusters. Only one group of samples (sampling site 4) is well distinguished from others.

Principal Component Analysis (PCA) was performed after additionally pre-processed data. Column standardization (autoscaling) method gave better results than Column centering method. Clusters of samples from sampling sites 2 and 4 were especially well defined. Three samples of sampling site 9 were also separated from other nine samples of the same sampling site. The Kohonen neural network shows the same separation of mentioned sampling sites (2 and 4).

This confirms the following conclusions:

(i) Sampling site 2 lies at the industrial and partly urbanised area. This is the reason that the site 2 is rather high polluted through the whole time. The reason for the pollution could also be the irruption of surface and leaking waters which make possible quick breakthrough of polluted substances into the underground water. The nitrate concentration is higher in comparison to other sampling sites, as shown in Figure 1. Agriculture and use of mineral fertilizers are considered to be the primary cause of high nitrate concentrations.

(ii) Sampling site 4 is a well designed for drinking water supply. It lies on the area which is influenced by industrial and agricultural pollution. Because of usage of Dichlobenil in past this is the only sampling site, where presence of 2,6-dichlorobenzamide could be detected. Dichlobenil is the herbicide which was used on the not agricultural areas in the vicinity of sampling site 4. The presence of chromium(VI) is also of great concern because the values are increasing. The level of trichloroethene considerably exceeds the limit value and is still increasing, which is due to antropogenic activity.

(iii) At the sampling site 9 three samples were different from all nine samples. The sampling site is sensitive on the changes at the surface such as income of nitrogen compounds and possible irruption of surface and leaking waters.

The study gives the opportunity to follow the quality of waters at different sampling sites within a defined time period. The monitoring of the general pollution of ground waters and following of the measured parameters which are above the permitted concentration level can be used to search the source of pollution, for planning of prevention measures and to protect from pollution. The benefit of application of chemometrics methods is not only the possibility of visualization of large amounts of multivariate data, but also a possibility for a quick classification of potentially poluted unknown water samples.

5. Acknowledgements

The authors thank the Ministry of Higher Education, Science and Technology of the Republic of Slovenia, contract numbers P1-017 and P2-0006 for financial support. The Public Health Institute, Environmetal Institute, Maribor, is kindly acknowledged for completing the data about ground water samples with their results.

6. References

1. D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. S. Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier, Amsterdam, **1997.**
2. M. Vega, R. Pardo, E. Barrado, L. Deban, *Wat. Res.* **1998**, *32*, 3581–3592**.**
3. W. D. Alberto, D. M. Del Pilar, A. M. Valeria, P. S. Fabiana, H. A. Cecilia, B. M. De Los Angeles, *Water Res.* **2001,** *35,*

2881–2894.
4. D. Brodnjak-Vončina, D. Dobčnik, M. Novič, J. Zupan, *Anal. Chim. Acta* **2002,** *462*, 87–100.
5. M. E. Kotti, A. G. Vlessidis, N. C. Thanasoulias, N. P. Evrimiridis, *Wat. Res. Management* **2005,** *19*, 77–94.
6. S. Tsakovski, V. Simeonov, S. Stefanov, *Fres. Environ. Bulletin* **1999**, *8*, 28–36.
7. K. P. Singh, A. Malik, D. Mohan, S. Sinha. *Wat.Res.* **2004**, 38, 3980–3992.
8. C. Mendiguchia, C. Moreno, M. D. Galindo-Riano, M. Garcia-Vargas, *Anal. Chim. Acta* **2004,** *515*, 143–149**.**
9. C. Sarbu, H. F. Pop, *Talanta* **2005**, *65*, 1215–1220.
10. P. X. Zhang, N. Dudley, A. M. Ure, D. Littlejohn, *Anal. Chim. Acta* **1992**, *258*, 1–10.
11. F. V. Silva, M. Y. Kamogaya, M. M. C. Ferreira, J. A. Nobrega, A. R. A. Noguera, *Eclectica Quimica.* **2002**, *27,* 91–102.
12. C. Sarbu, H. W. Zwanziger, *Anal. Lett.* **2001,** *34*, 1531–1552**.**
13. R. Lindegren, M. Josefson, *Chemometr. Intell. Lab. Syst.* **1998***, 44*, 403–409.
14. A. K. Meng, I. H. Suffet, Environ. Sci. Technol. **1997** *31,* 337–345.
15. E. Marengo, M. C. Gennaro, D. Giacosa, C. Abrigo, G. Saini, M. T. Avignone, *Anal. Chim. Acta* **1995,** *317*, 53–63.
16. W. M. Jarman, G. W. Johnson, C. E. Bacon, J. A. Davis, R. W. Risebrough, R. Ramer, *Fresenius J. Anal. Chem.* **1997,** *359*, 254–260.
17. P. Barbieri, G. Adami, A. Favretto, E. Reisenhofer, *Fresenius J. Anal. Chem.* **1998,** *361,* 349–352.
18. M. M. C. Ferreira, C. G. Faria, E. T. Paes, *Chemometr. Intell. Lab. Syst*. **1999**, *47*, 289–297.
19. M. P. Kallio, S. P. Mujunen, G. Hatzimihalis, P. Koutofides, P. Minkkinen, P. J. Wilkie, M. A. Connor, *Anal. Chim. Acta* **1999**, *363*, 181–191.
20. M. F. Wilkins, L. Boddy, C. W. Morris, *Binary-Comput. Microb.* **1994,** *6*, 64–72.
21. Water quality-Sampling- Part 11: Guidance on sampling of ground waters ISO 5667–11: 1992 (E).
22. Teach/Me, SDL – Software Development Lohninger; Teach /Me DataLab 2. 002 © **1999,** Springer, Berlin. Developed by H. Lohninger and the Teach/Me people.
23. J. Zupan, J. Gasteiger, Neural Networks for Chemists: An Introduction, Verlag Chemie, Weinheim, **1993.**
24. T. Kohonen, Self-Organization and Associative Memory, Springer-Verlag, Berlin, **1988.**
25. J. Lozano, M. Novič, F. X. Rius, J. Zupan, *Chemometr. Intell. Lab. Syst.* **1995,** *28,* 61–72.
26. J. Zupan, M. Novič, I. Ruisanchez, *Chemometr. Intell. Lab. Syst.* **1997**, *38,* 1–23.
27. J. Zupan, M. Novič, J. Gasteiger, *Chemometr. Intell. Lab. Syst.* **1995**, *27,*175–187.
28. N. Majcen, K. Rajer-Kanduč, M. Novič, J. Zupan, *Anal.*

Povzetek

V Sloveniji spremljamo kakovost podzemne vode na številnih mestih, saj je podzemna voda pomemben vir pitne vode. Kot vir podatkov za kemometrično obdelavo smo izbrali monitoring v katerem je vključeno 5 vodnjakov, ki so namenjeni za oskrbo s pitno vodo, 3 industrijski vodnjaki in 2 kontrolni vrtini.

Opravili smo analizo 214 vzorcev podzemnih vod v letih 2003 in 2004. Vzorce smo odvzeli na omenjenih desetih merilnih mestih in opravili fizikalne in kemijske analize. Fizikalno kemijske preiskave vključujejo merjenje naslednjih 13 parametrov, ki so podlaga za oceno kemijskega stanja podzemnih vod: pH vrednost, temperatura vode, električna prevodnost, vsebnost nitrata, vsebnost adsorbljivih organskih halogenov (AOX), vsebnost Cr(VI) in skupnega kroma, vsebnost pesticidov kot so desetil atrazin, atrazin in 2,6-diklorobenzamid, ter vsebnost halogeniranih ogljikovodikov kot so triklorometan, 1,1,2,2-tetrakloroeten in 1,1,2-trikloroeten. Za obdelavo rezultatov meritev smo uporabili različne kemometrične metode, osnovne statistične metode za določitev povprečne vrednosti in mediane, standardnih odmikov, minimalnih in maksimalnih vrednosti merjenih parametrov in njihovih medsebojnih korelacijskih koeficientov, analizo grup (CA), metodo glavnih osi (PCA), metodo grupiranja, ki temelji na Kohonenovih nevronskih mrežah in metodo linearne diskriminantne analize (LDA). Z metodo glavnih osi in s Kohonenovimi nevronskimi mrežami smo poizkusili poiskati podobnosti med posameznimi merilnimi mesti. Študija daje možnost, da sledimo kakovost podzemne vode na posameznih merilnih mestih in v določenem časovnem obdobju. Časovno sleditev splošnega onesnaženja podzemnih vod ter rezultatov posameznih merjenih parametrov, ki presegajo dovoljene meje, lahko uporabimo za iskanje vzrokov onesnaženja in za načrtovanje preventivnih ukrepov za zaščito pred onesnaženjem.