Urban water quality evaluation using multivariate analysis

Petr Praus[1]

Hodnotenie kvality mestskej vody použitím multivariacnej analyzy

A data set, obtained for the sake of drinking water quality monitoring, was analysed by multivariate methods. Principal component analysis (PCA) reduced the data dimensionality from 18 original physico-chemical and microbiological parameters determined in drinking water samples to 6 principal components explaining about 83 % of the data variability. These 6 components represented inorganic salts, nitrate/pH, iron, chlorine, nitrite/ammonium traces, and heterotrophic bacteria. Using the PCA scatter plot and the Ward's clustering of the samples characterized by the first and second principal components, three clusters were revealed. These clusters sorted drinking water samples according to their origin - ground and surface water. The PCA results were confirmed by the factor analysis and hierarchical clustering of the original data.

Keywords*: Water quality, drinking water, principal component analysis, multivariate methods, data mining.*

Introduction

Regular drinking water monitoring is essential for supplying people with a high quality and healthy water meeting all requirements of legal regulations. Distribution systems are being usually monitored on many sampling points at which samples are regularly taken and then analysed in laboratories in accordance with a monitoring plan. The sampling frequency and the number of examined parameters are given by the government regulations, requirements of water technologists and regional health offices. The water quality evaluation should be based on the statistical analysis of the collected physical, chemical, and biological results.

The water quality is mostly characterized by many variables (parameters) which represent a water composition in specific localities and time. Real hydrological data are mostly noisy, it means that they are not normally distributed, often co-linear or autocorrelated, containing outliers or errors etc. These data sets create a n-dimensional space from which information about the water composition has to be mined. For this purpose, multivariate methods such se the cluster analysis, the principal component analysis, the factor analysis, and the discriminant analysis, are used. The quality assessment of surface water (Zeng and Rasmussen, 2005; Simeonov et al., 2003; Wunderlin et al., 2001), ground water (Reghunath et al., 2002), and the environmental research (Ceballos et al., 1998; Lambarkis et al., 2004; Praus, 2005; Bartolomeo et al., 2004) employing multicomponent techniques are well described in the literature.

The principal component analysis has been used for the data clustering and finding hidden relationships among them. Unlike other statistical methods (e.g. discriminant analysis), PCA is a robust technique which does not require normally distributed and uncorrelated variables. The aim of this paper is to use PCA for reducing the number of parameters, which must be determined during a regular monitoring, and to recognize basic features of drinking water quality. The factor analysis and the cluster analysis were used to confirm the PCA results.

Materials and methods

Water quality data

The results of 18 parameters, including chemical, physical, and microbiological ones, were determined in 126 drinking water samples taken from a city water network in North Moravia, the Czech Republic, during a half of year according to a monitoring plan. Water taps situated mostly in private and commercial buildings were selected as sampling points. Each locality was sampled at least twice during this period. Water analyses, including the sample collection and preservation, were carried out according to the actual standardized ISO and EN methods. In some cases, when alternative methods exist, the selected ones used in this study are specified by abbreviations: VIS-visual spectrometry, AAS-atomic absorption spectrometry, EDTA-titration with ethylenediaminetetraacetic acid. The determined parameters were pH, ammonium (VIS), nitrate (VIS), nitrite (VIS), colour (VIS), turbidity (VIS), temperature, calcium (EDTA), iron (AAS), electrical

[1] *doc. Ing. Petr Praus, PhD.*, Department of Analytical Chemistry and Material Testing, VSB-Technical University Ostrava, 17. listopadu 15, 708 33 Ostrava, Czech Republic
(Recenzovaná a revidovaná verzia dodaná 29. 6. 2007)

conductivity, hardness (EDTA), alkalinity, acidity, chemical oxygen demand by permanganate, total and free chlorine (VIS).

From the microbiological parameters only non zero ones, i.e. psychrophilic (cultivated at 20 ^{o}C) and mesophilic bacteria (cultivated at 37 ^{o}C), were used in this survey. However, these bacteria colony-forming units in one millilitre (CFU/ml) were deeply below the drinking water quality criteria. Coliform and faecal coliform bacteria were not detected at all.

Principal component analysis

PCA is often applied for the removal of data noise by the reduction of their dimensionality (Jolliffe, 2002). PCA searches 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)

$$PC_i = w_1x_1 + w_2x_2 + \ldots + w_nx_n \qquad (1)$$

where x_i and w_i are the original variable and the component weight, respectively. The principal component weights are used as measures of the correlation between the variables and the principal components. The largest or first PC is oriented in the direction of largest variation of the original variables and passes through the centre of the data. The second 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. The third largest PC is directed towards the next largest variance, goes through the data centre and is orthogonal to the first and second PCs, and so forth.

Classical PCA is based on the decomposition of a covariance/correlation matrix by the eigenvalue decomposition or by the singular value decomposition of real data matrices. The eigenvalues or singular values indicate variations among the observed variables (parameters).

Factor analysis

In the factor analysis (FA) each variable can be expressed as a linear combination of latent common factors and a single specific factor:

$$x_i = \sum_{i=1}^{n} \alpha_{ij} F_j + \beta_i e_i \qquad (2)$$

where 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 matrix and a specific factor matrix. The main difference between PCA and FA is that PCA is concerned with the total variation as expressed in the correlation matrix, while FA is concerned with a correlation in the common factor portion. In addition, a number of factors must be known before FA is performed. The goal of FA is not only to reduce the data dimensionality as with 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, e.g. (Malinowski and Howery, 1980; Malinowski, 1991).

Cluster analysis

Cluster analysis (CA) encompasses a number of different methods which organize objects (observations) into groups called clusters without unexplanation or interpretation. Objects within the clusters are similar whereas objects in different clusters are dissimilar. This exploratory method is used to discover the data structure not only among observations, but also among variables, arranged into a tree diagram, usually called a dendrogram. The utilized methods, algorithms, and similarity/dissimilarity measures are described elsewhere in the literature (Everitt, 2001). In this study, the commonly applied average group and the Ward's clustering methods were used. The Euclidean distance was used as a similarity measure.

Multivariate computations

The original data matrix (123x18) was prepared and processed in the MS Excel 2000. Its rows were constructed from the parameters analysed in drinking water. There were no missing values in the data set. The observations below the detection limits were replaced with values equal one half of the detection limits (Zeng and Rasmussen, 2005). The reason is that the detection limits of analytical methods are not absolute, strictly defined values and, moreover, some of them were changed during the data collection period. The principal component analysis, factor analysis, cluster analysis, and other statistical calculations were performed by the software package STATGRAPHIC Plus 5.0 (Statistical Graphics Corp., USA) and QCExpert (Trilobyte, Czech Republic). Before the computation, the testing data were standardized in order to avoid misclassifications arising from different orders of magnitude of tested variables. Therefore the original data were mean (average) centred and scaled by the standard deviations: $(x_i - \bar{x}) / s$.

Results and discussion

The PCA results of drinking water samples

Drinking water samples (n=123), taken for the sake of regular screening of drinking water quality in a city supply system, were characterized by 16 chemical and physical variables and 2 microbiological parameters (Tab. 1). From these data, 6 principal components, explaining 83 % of the total variance, was estimated on the basis of a Kaiser (1960) criterion of the eigenvalues greater or equal 1 and from a Cattel scree plot (Cattel, 1966). A scree plot shows the eigenvalues sorted from large to small as a function of the principal components number. After the sixth PC (Fig. 1), starting the elbow in the downward curve, other components can be omitted. The components weights, their eigenvalues, and variances are summarized in Table 2.

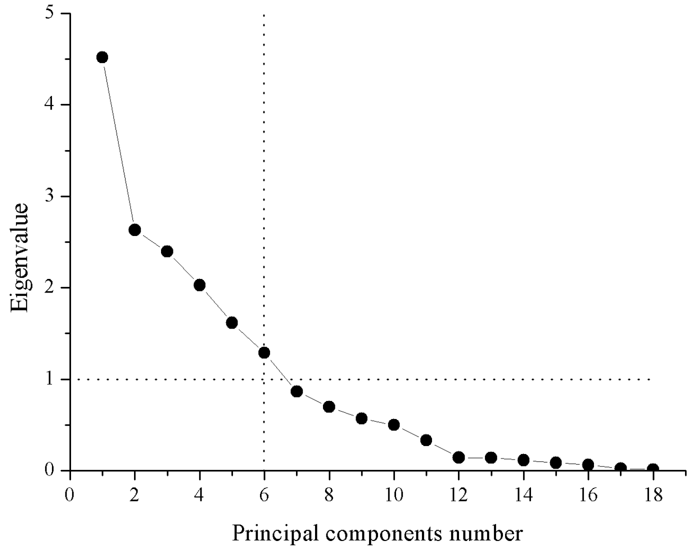

Fig. 1. Scree plot of the eigenvalues.

Tab. 1. Summary statistics of drinking water samples.

Parameter	**Median**	**Median st. dev.**	**Minimum**	**Maximum**	**Skewness**	**Kurtosis**
Acidity [mmol/l]	0.1	0	0.05	0.75	15.0484	22.9599
Alkalinity [mmol/l]	1.1	0.04	0.75	3.40	6.1111	1.6253
Ammonia [mg/l]	0.006	0	0.006	0.054	17.0816	34.8977
Ca [mg/l]	30.1	1.63	21.2	120	5.2919	0.6279
Chlorine free [mg/l]	0.05	0.005	0.020	0.26	6.0272	2.3748
Chlorine tot.[mg/l]	0.15	0.010	0.02	0.35	3.5110	0.4462
COD-Mn [mg/l]	0.77	0.041	0.30	1.9	2.2432	1.1325
Colour [mg/l]	7.3	0.70	1.40	20.0	2.7912	-0.5130
Conductivity [µS/cm]	22.9	1.58	18.8	74.8	4.3275	-1.2249
Fe [mg/l]	0.12	0.015	0.02	0.44	3.9541	0.8788
Hardness [mmol/l]	1.0	0.06	0.7	3.5	4.5301	-0.9926
Mesophiles [CFU/ml]	0	0.3	0	46	20.8225	49.9124
Nitrate [mg/l]	7.68	0.224	1.71	42.3	11.4585	12.9512
Nitrite [mg/l]	0.002	0	0.002	0.008	12.2134	17.7015
pH	7.82	0.028	6.49	8.29	-7.8288	5.2785
Psychrophiles [CFU/ml]	0	0.3	0	54	21.0270	57.0145
Temperature [°C]	10.5	0.74	5.8	19.2	2.2739	-2.7540
Turbidity [NTU]	0.45	0.020	0.13	1.07	3.2805	0.2162

Note: St. Dev.-standard deviation, COD-Mn chemical oxygen demand with permanganate,n=123.

Tab. 2. Principal component weights.

Parameter	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
Acidity	0.30067	0.42108	0.04324	0.01120	-0.05174	0.03606
Alkalinity	0.31881	-0.41704	-0.12066	-0.02343	0.10338	0.04134
Ammonia	-0.10772	0.06600	-0.05462	-0.00944	0.43977	0.33297
Calcium	0.43365	-0.15237	-0.10097	-0.04501	0.06875	0.04112
COD-Mn	-0.29351	-0.07623	-0.13840	-0.13373	0.11998	0.10814
Colour	-0.08028	-0.24234	0.44076	0.24652	0.03083	0.03319
Conductivity	0.42307	-0.19351	-0.11142	-0.01648	0.10036	0.06081
Hardness	0.43420	-0.17996	-0.09648	-0.01605	0.10530	0.07639
Chlorine free	0.04218	-0.01864	-0.18426	0.62665	-0.11232	-0.08802
Chlorine total	-0.04868	-0.04107	-0.20821	0.60734	-0.14516	-0.09225
Iron	0.04309	-0.13804	0.52467	0.18423	0.22438	0.12079
Mesophilic b.	-0.04226	-0.06378	-0.02941	-0.01218	-0.40316	0.57465
Nitrate	0.26770	0.44664	0.13679	0.02656	-0.00028	0.07331
Nitrite	-0.09440	0.11824	-0.04021	0.05758	0.42005	0.34665
pH	-0.23475	-0.44597	-0.20023	-0.09851	0.08011	-0.05883
Psychrophilic b.	-0.00049	-0.09015	-0.01514	0.00371	-0.42038	0.55675
Temperature	0.04915	-0.10005	0.14823	-0.32405	-0.37854	-0.24902
Turbidity	0.04797	-0.17656	0.54698	0.06037	-0.05075	-0.00764
Eigenvalue	4.51970	2.62920	2.39520	2.02728	1.61568	1.28716
Percent of variance	25.11	14.61	13.31	11.26	8.98	7.15
Cumulative percentage	25.11	39.72	53.02	64.29	73.26	80.41

Interpretation of the principal components

As it is obvious, PC_1 can be called as the salt component because it is mainly saturated with conductivity and hardness (including calcium). The second principal component is associated with nitrate, pH and bicarbonate expressed through acidity and alkalinity. The relation between nitrate and pH cannot be explained by nitrification because of the low concentrations of ammonium, organic substrate and, of course, the absence of nitrification bacteria in drinking water. This relation rather signalises different sources of drinking water within the city network. Really, there are ground water wells and surface water reservoirs in this area from which raw water is treated for drinking. Nitrate does not significantly contribute to conductivity because of its low concentrations.

The third component is mainly composed from colour, iron, and turbidity which are all connected with a pipeline system corrosion. In this case, turbidity is associated with iron only, not with the higher levels of disease-causing microorganisms being also indicated by high water turbidity. Thus PC_3 should be identified as the iron component. PC_4 represents free and total chlorine and can be called as the chlorine component. It is also obvious that the chlorine content does not depend on the iron concentrations in water. Therefore, the low residual chlorine concentrations in tap water can be rather ascribed to its loss in the long distribution system.

The fifth PC consists mainly of ammonia, nitrite, and temperature. Unlike nitrate, the ammonia and nitrite concentrations are very low, often around their detection limits. This principal component characterizes the traces of inorganic nitrogen. A negative sign of the temperature weight indicates its reciprocal relation to the ammonia and nitrite. It can be simply explained by their oxidation to nitrate (not nitrification through biochemical reactions) which depends on temperature.

PC_6 represents bacteria found in drinking water. Since the bacteria and chlorine components were extracted as independent ones, it means that an occurrence of psychrophiles and mesophiles was not effected by the residual chlorine concentrations. In addition, no correlation between chlorine and pH, which influences the chlorine disinfecting efficiency, was observed. Both types of bacteria belong among heterotrophic bacteria which are naturally present in the environment. Their CFUs per a millilitre of samples can be tolerated in drinking water up to the limits strictly defined by water quality regulations. In this study, no such limits were exceeded. This bacteria occurrence is very often cased by conditions of inner house distribution systems.

PCA evaluation of drinking water quality

A scatter plot in Fig. 2a, composed from the first and second principal components, demonstrates three clusters I, II, and III. The cluster I gathers the samples which are typical by the high concentrations of nitrate, low values of pH, and the high content of inorganic salts. The clusters II and III differ mainly in the concentrations of inorganic salts. Figure 2b demonstrates the two not well distinguished clusters II and a mixed cluster I+III whereas Fig. 2c shows that the tested samples can be arranged into three groups again. These clusters can be allocated to the three city parts (Fig. 3) which are supplied from the different water sources. Drinking water in the part I is coming from two ground water sources. The part II is supplied with the treated surface and ground water and drinking water occurring in the part III is produced from water of a surface reservoir.

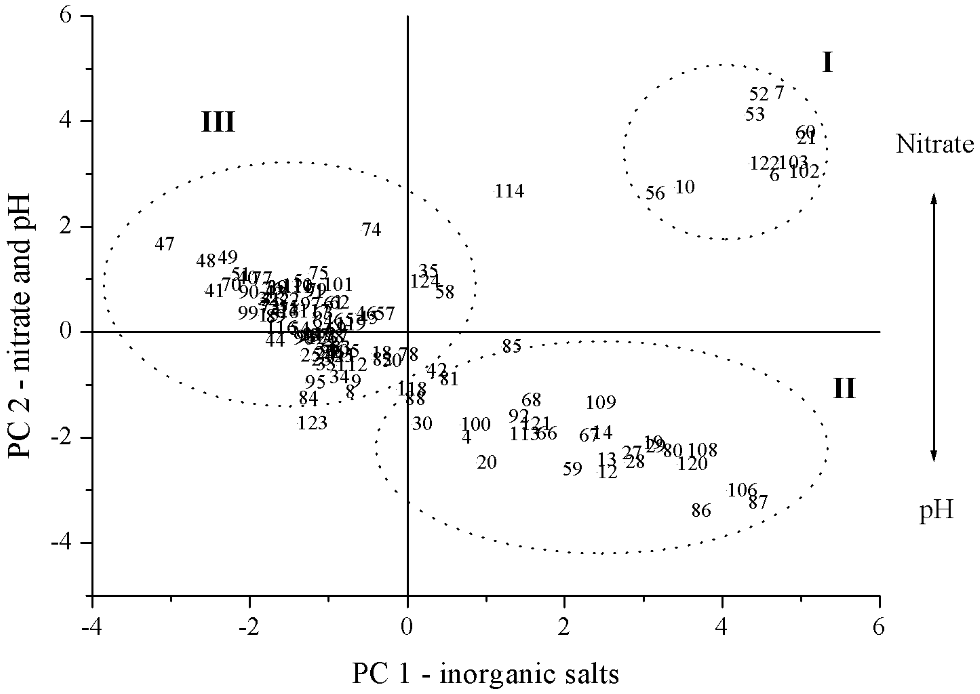

Fig. 2a. Scatter plot of the drinking water samples constructed from the first and second principal components.

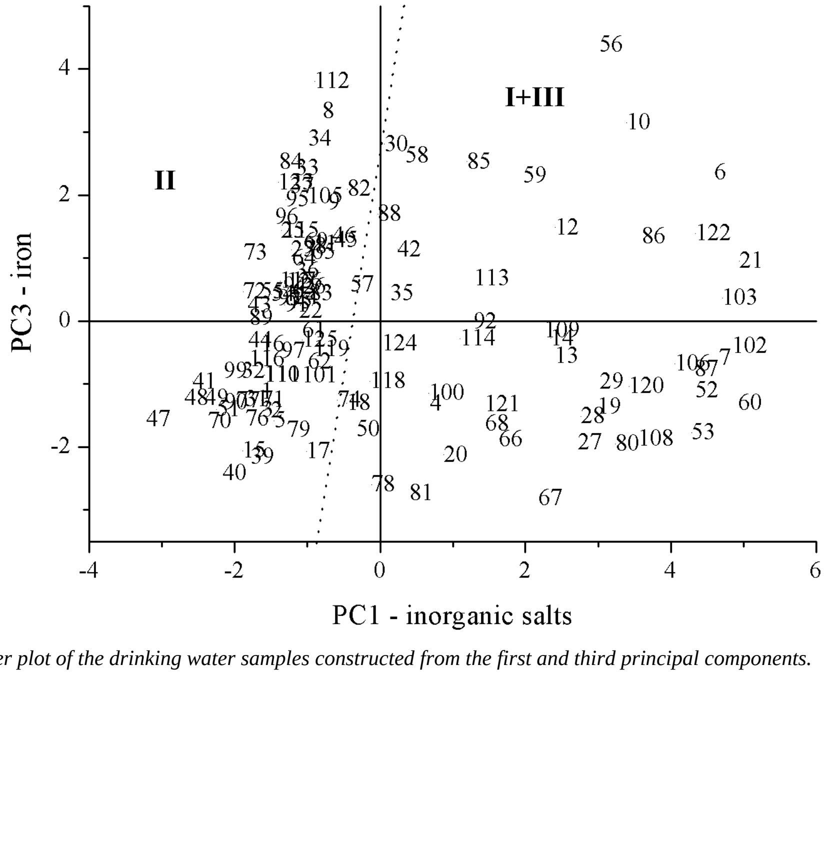

Fig. 2b. Scatter plot of the drinking water samples constructed from the first and third principal components.

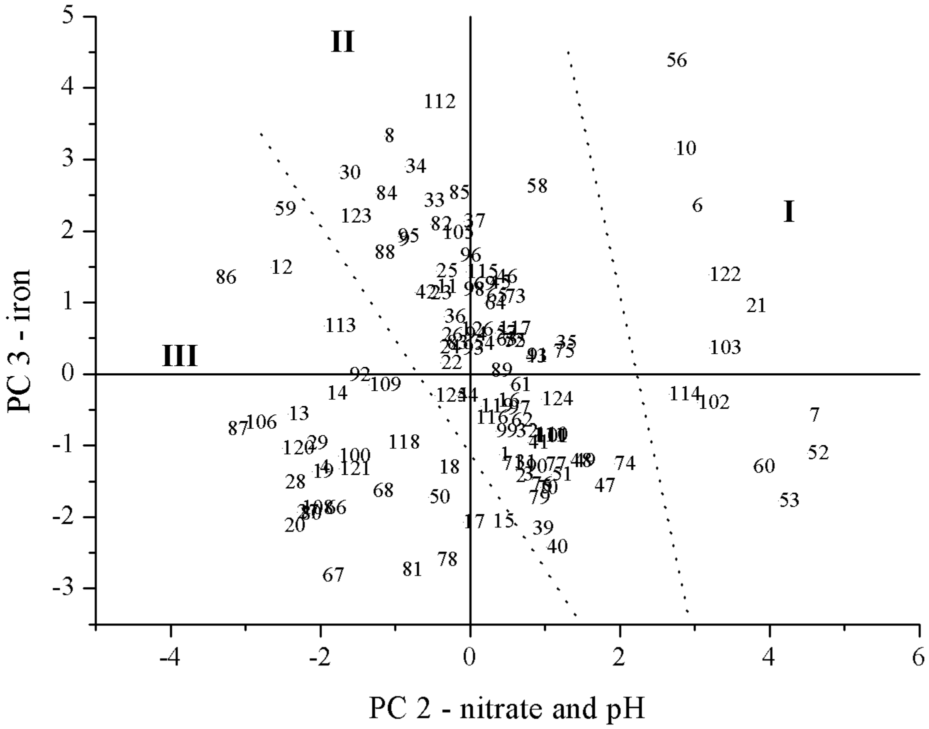

Fig. 2c. Scatter plot of the drinking water samples constructed from the second and third principal components.

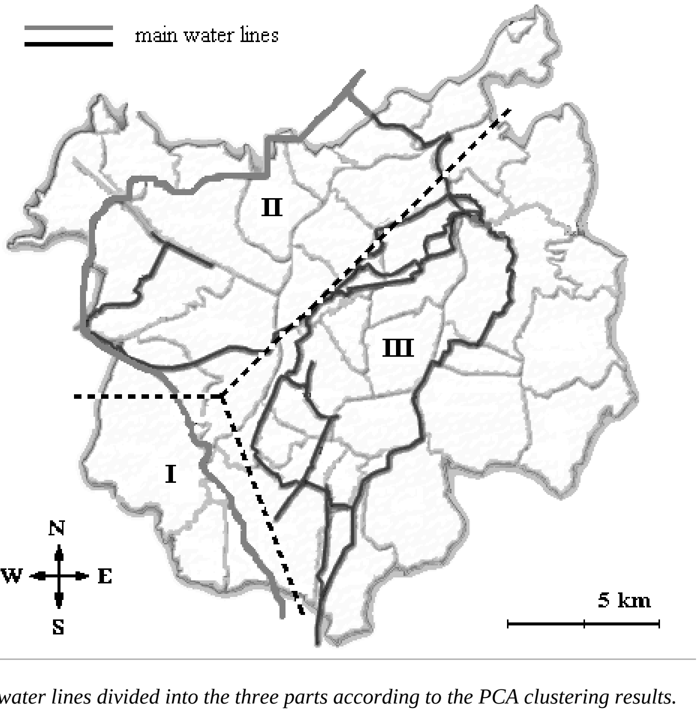

Fig. 3. City map of the main water lines divided into the three parts according to the PCA clustering results.

The PCA scatter plot given in Fig. 2a indicates the three groups of the samples but their 2D projection cannot always show each of them because of their overlapping. Also, it is sometimes hard to decide to which group some samples should be assigned. In order to sort the samples more effectively, the points in Fig. 2a were clustered using the average group and Ward's clustering methods. Unlike average group dendrogram, the Ward's one provided visually the three well organized clusters I to III (Fig. 4), also recognized by an agglomeration distance plot. The basic statistics of each cluster were calculated and summarized in Table 3. It is obvious from these results, that the samples belonging to the cluster I are typical by their higher concentrations of nitrate and lower pH. On the other hand, the samples from cluster III posses the higher amount of inorganic salts. All these findings have been already indicated in Fig. 2a.

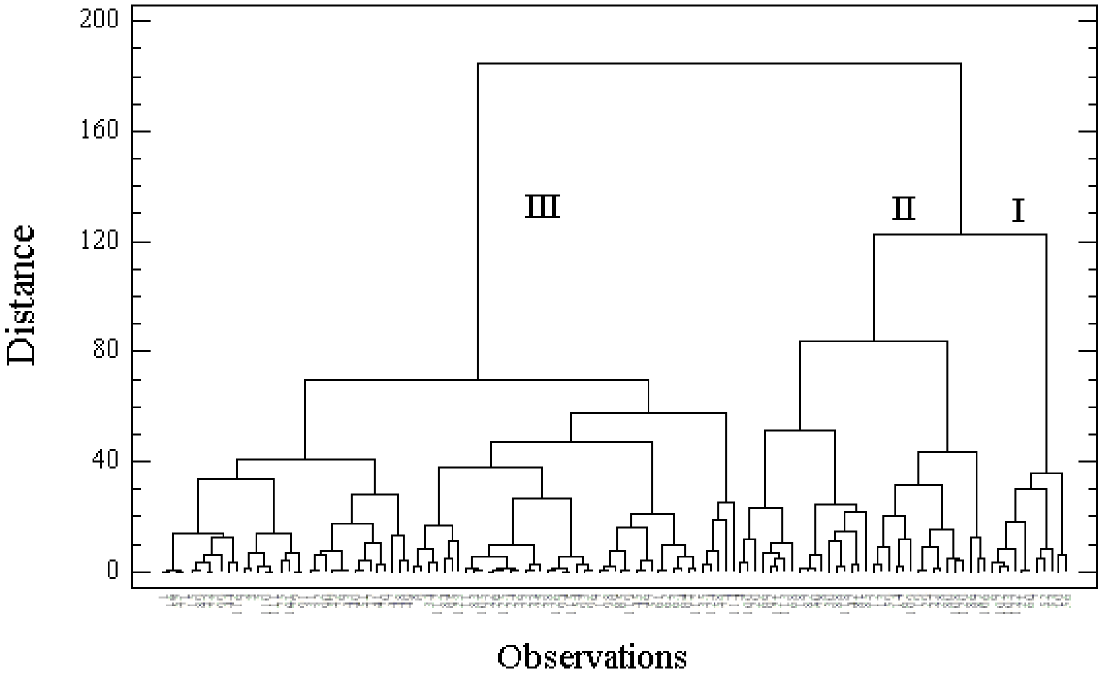

Fig. 4. Ward's dendrogram of the drinking water samples defined by the first and second principal components.

Verification of the PCA results

The results of PCA were compared with those of cluster analysis and the factor analysis applied to the original data set. CA was performed by means of the Ward's method (Fig. 5) because of the same reason given above. The dendrogram manifested almost the same clusters compositions as it was found in Fig. 4. It also confirms that PC_1 and PC_2 contain parameters which are most important for the water quality characterization.

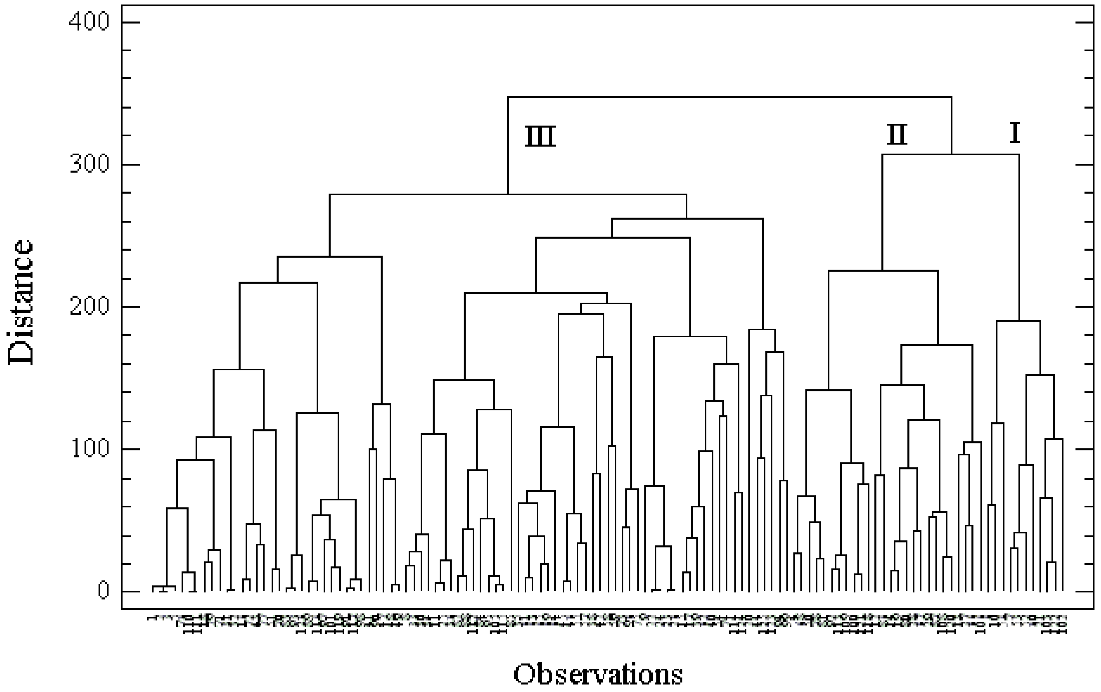

Fig. 5. Ward's dendrogram of the drinking water samples.

The factor analysis, performed for the six factors (indicated by PCA) using the Varimax rotation, revealed nearly the same relationships among the parameters, which was found by PCA, except alkalinity (Tab. 4). As it is obvious from Table 3, for the samples of II and III groups it holds that alkalinity is very close to hardness. It means that hardness is mainly caused by bicarbonate (the so called carbonate hardness). Hardness of the samples I is higher than alkalinity which indicates that hardness is also caused by calcium and magnesium. Both PCA and FA demonstrate that COD-Mn does not play a significant role in the water quality description. It can be explained by the low content of organic compounds in drinking water and also by only an informative character of this parameter because of the low permanganate oxidation efficiency.

Tab. 3. Summary statistics of the revealed clusters.

Parameters	**Cluster I Median** (n=11)	**Cluster I Median st. dev.**	**Cluster II Median** (n=27)	**Cluster II Median st. dev.**	**Cluster III Median** (n=85)	**Cluster III Median st. dev.**
Acidity [mmol/l]	0.6	0.11	0.1	0	0.1	0
Alkalinity [mmol/l]	1.4	0.15	2.5	0.15	1.0	0.05
Ammonia [mg/l]	0.006	0	0.006	0.93	0.006	0
Calcium [mg/l]	73.9	6.71	73.0	4.6	26.2	0.41
COD-Mn [mg/l]	0.3	0.02	0.7	0.08	0.8	0.05
Colour [mg/l]	1.4	2.65	7.3	0.93	9.1	0.46
Conductivity [μS/cm]	54.0	1.31	62.1	4.64	21.3	0.36
Hardness [mmol/l]	2.4	0.12	2.6	0.21	0.9	0.02
Chlorine free [mg/l]	0.08	0.033	0.06	0.015	0.05	0.005
Chlorine total [mg/l]	0.15	0.038	0.15	0.013	0.16	0.013
Iron [mg/l]	0.07	0.066	0.12	0.020	0.13	0.023
Mesophiles [CFU/ml]	0	0.5	1	0.8	0	0.3
Nitrate [mg/l]	32.1	1.61	7.30	0.878	7.66	0.224
Nitrite [mg/l]	0.002	0.0003	0.002	0	0.002	0
pH	6.62	0.045	7.98	0.026	7.81	0.028
Psychrophiles [CFU/ml]	0	1.0	1	0.8	0	0.3
Temperature [°C]	10.5	2.63	12.1	1.12	10.0	0.66
Turbidity [NTU]	0.45	0.163	0.44	0.043	0.47	0.026

Tab. 4. Factor loadings after Varimax rotation.

Parameter	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
Acidity	0.18335	0.90751	-0.16119	-0.03069	-0.02035	-0.03111
Alkalinity	0.95048	-0.23276	0.07378	0.02161	-0.09097	0.01812
Ammonia	-0.07575	-0.08833	-0.01988	-0.12484	0.70434	-0.02772
Calcium	0.93597	0.24434	-0.05355	-0.02790	-0.09865	-0.01408
COD-Mn	-0.36465	-0.51292	-0.18222	-0.14535	0.26795	0.05392
Colour	-0.09838	-0.18588	0.84345	0.13291	-0.02677	0.03160
Conductivity	0.96081	0.17785	-0.02743	0.00859	-0.05480	-0.01141
Hardness	0.96821	0.21454	-0.01172	-0.00182	-0.04323	-0.00507
Chlorine free	0.07805	0.05699	-0.00305	0.95254	0.00104	0.00459
Chlorine total	-0.06443	-0.08361	-0.03887	0.94744	-0.00596	0.03793
Iron	0.07239	0.09400	0.91538	-0.04726	0.16804	-0.06563
Mesophilic b.	-0.04634	-0.04386	-0.01984	0.00299	-0.01522	0.83910
Nitrate	0.09184	0.93936	-0.02983	-0.06970	0.05599	-0.04289
Nitrite	-0.10361	0.01360	-0.00149	-0.04386	0.71719	-0.01283
pH	0.01085	-0.94683	-0.07376	-0.01540	0.01481	-0.00208
Psychrophilic b.	0.03564	-0.02128	0.01708	0.02482	-0.06548	0.83637
Temperature	0.01051	-0.03796	0.04197	-0.38217	-0.67820	0.06889
Turbidity	0.02749	0.06319	0.86172	-0.14681	-0.22534	0.03096

Conclusion

From the PCA findings given above follows that 18 parameters used for the drinking water quality characterization, can be replaced by the 6 principal components explaining about 83 % of the data variance: inorganic salts, nitrate/pH, iron, chlorine, nitrite/ammonia, and bacteria. Regarding the physico-chemical properties and hygienic importance of these parameters, only the six of them can be used for the frequent water quality monitoring: Conductivity, nitrate, iron, free chlorine, nitrite, and mesophiles.

FA mostly confirmed the PCA results and, additionaly, in the case of alkalinity showed relations between hardness and bicarbonate/carbonate concentrations. The first two principal components explaining about 50 % of data contain the key variables of the drinking water supply system: inorganic salts and nitrate/pH.

The PCA scatter plots and dendrograms were used for the samples clustering. Also the combination of scatter plots and cluster analysis was found to be advantages. The revealed clusters gather the drinking water samples according to their origin (surface and ground water).

Multivariate methods were found to be suitable for reducing the water quality parameters and the determination of relationships among them, and also for the samples clustering, as well. These techniques can be helpful for assessors to obtain a global view on the water quality in any urban or other geographical territory when analysing large data sets without *a priori* knowledge about them.

Acknowledgement
This work was supported by the Ministry of Education, Youth and Sport of the Czech Republic (MSM 6198910016).

References

Bartolomeo, A., D., Poletti, L., Sanchini, G., Sebastiani, B., Morozzi, G.: Relationship among parameters of lake polluted sediments by multivariate statistical analysis. *Chemosphere 55 (10), 2004, 1323-1329.*

Cattel, R., D.: The scree test for the number of factors. *Multivariate Behav. Res. 1, 1966, 245-276.*

Ceballos, B., S., O., König, A., Oliveira, J., F.: Dam eutrophication: A simplified technique for a fast diagnosis of environmental degradation. *Water Res. 32 (11), 1998, 3477-3483.*

Everitt, B.: Cluster Analysis (4th edn.). *Hodder Arnold, London, 2001.*

Jolliffe, I., T. Principal component analysis (2nd edn.). *Springer-Verlag, New York, 2002.*

Kaiser, H., F.: The application of electronic computers to factor analysis. *Educ. Psychol. Meas. 20**,** 1960, 141-151.*

Lambarkis, N., Antonakos, A., Panagopoulos, G.: The use of multicomponent statistical analysis in hydrogeological environmental research. W*ater Res. 38 (7), 2004, 1862-1872.*

Malinowski, E., R.: Factor Analysis in Chemistry (2nd edn.). *John Wiley & Sons, New York, 1991.*

Malinowski, E., R., Howery, D., G.: Factor Analysis in Chemistry. *John Wiley & Sons, New York, 1980.*

Praus, P.: Water quality assessment using SVD-based principal component analysis of hydrological data. *Water SA 31 (4), 2005, 1-6.*

Reghunath, R., Murthy, T., R., S., Raghavan,, B., R.. The utility of multivariate statistical techniques in hydrochemical studies: An example from Karnataka, India. *Water Res. 36 (10), 2002, 2437-2442.*

Simeonov, V., Stratis, J., A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis A., Sofoniou M., Kouimtzis T. Assessment of the surface water quality in Northern Greece". *Water Res. 37 (17), 2003, 4119-4124.*

Wunderlin, D., A., Díaz, M., P., Amé, M., V., Pesce, S., F., Hued, A., C., Bistoni M. A. 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), 2001, 2881-2894.*

Zeng, X., Rasmussen, T.,C.: Multivariate statistical characterization of water quality in lake Lanier, Georgia, USA. *J. Environ. Qual. 34 (6), 2005, 1980-1991.*