PERGAMON

Chemosphere 49 (2002) 233–245

CHEMOSPHERE

www.elsevier.com/locate/chemosphere

Exploratory analysis of data sets with missing elements and outliers

A. Smoliński [a], B. Walczak [a,*], J.W. Einax [b]

[a] *Institute of Chemistry, Silesian University, 9 Szkolna Street, 40-006 Katowice, Poland*
[b] *Institute of Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Lessingstrasse 8, D-07743 Jena, Germany*

Received 27 February 2002; received in revised form 10 June 2002; accepted 17 June 2002

Abstract

The main goal of the presented paper was to develop a general strategy allowing exploration of contaminated data sets with missing elements, based on application of robust PLS for initial estimation of missing elements. Using robust distance, the outlying elements were identified. After their identification and replacing by missing elements, the expectation-maximization algorithm (which can be built in into different computational procedures, such as principal component analysis and its generalisation to the *N*-way data—the TUCKER3 model) was used for construction of the final model.

Keywords: Exploratory analysis; Missing data; Multiple outliers

1. Introduction

Experimental data sets can simultaneously contain missing elements and outliers. Missing elements can result from not recorded measurements (possibly due to high analysis costs or because they were outside the scope of the measuring instrument). Outliers can be caused by the instrument's malfunctioning or because they do not make part of the population data majority.

The existing methods allow treating these problems separately, but none of the established approaches allows proper estimation of missing elements when a data set contains outliers, or identify outliers in data sets with missing elements.

Problems associated with exploratory analysis of data sets with missing elements and outliers are demonstrated on the real monitoring data set, which contains measurements of pollutants' concentration performed in different sampling sites over a certain period of time.

The main goal of our study was to elaborate the general strategy, which might allow exploration of contamined data sets with missing elements. The proposed approach is based on application of robust PLS (Walczak, 1995a,b) for the initial estimation of missing elements. Once the missing elements are properly estimated, the techniques of outliers identification can further be applied. After identification of outliers (or the outlying elements), a final exploratory model can be constructed, using the Expectation-Maximization (EM) iterative approach (McLaachlan and Krishnan, 1997) to deal with missing data.

2. Theory

2.1. Methods of exploratory analysis

The most popular technique of exploratory analysis of multivariate data sets is principal component analysis

* Corresponding author.
E-mail address: beata@tc3.ich.us.edu.pl (B. Walczak).

0045-6535/02/$ - see front matter
PII: S0045-6535(02)00326-0

(PCA) (e.g. Joliffe, 1986; Wold, 1987; Vandeginste et al., 1998). It allows reduction of data dimensionality, its visualization, and interpretation of the objects and variables relationships. In this approach, matrix, $\mathbf{X}$ ($m \times n$), is decomposed into the two matrices, $\mathbf{S}$ ($m \times fn$) and $\mathbf{D}$ ($n \times fn$). Symbols m and n denote, respectively, the number of objects and variables, $\mathbf{S}$ represents the scores matrix, $\mathbf{D}$ represents the loading matrix, $\mathbf{E}$ is the residuals matrix and fn denotes number of significant factors. Scores and loadings matrices are orthogonal, i.e., $\mathbf{S}'\mathbf{S} = \mathbf{D}'\mathbf{D} = \mathbf{I}$. Columns of matrix $\mathbf{S}$ are called principal components (PCs), or eigenvectors. Each PC is constructed as a linear combination of original variables with the weights maximizing description of data variance (i.e., $\mathbf{S} = \mathbf{XD}$). Sum of squared elements of each eigenvector (PC) is called eigenvalue and represents a portion of variance, modeled by a corresponding PC. The first PC describes the highest part of the data variance, so that the associated eigenvalue has the highest value as well. Sum of eigenvalues defines the total data variance.

Scores vectors (i.e. columns of matrix $\mathbf{S}$) and loading vectors (i.e. columns of matrix $\mathbf{D}$) are used to visualize relationships between objects and parameters in a considered matrix $\mathbf{X}$. For example, score plot PC1–PC2 and loading plot PC1–PC2 represent, respectively, relationships between objects and parameters.

Generalization of PCA for the data sets with the N-way structure is the TUCKER3 model (Tucker, 1966; De Ligny et al., 1984; Van der Kloot and Kroonenberg, 1985; Geladi, 1989). In this approach, the three-way array, $\underline{\mathbf{X}}$, is decomposed into three loading matrices $\mathbf{A}$ ($I \times D$), $\mathbf{B}$ ($J \times E$), and $\mathbf{C}$ ($K \times F$), and the core array $\underline{\mathbf{G}}$ ($D \times E \times F$), where D, E and F denote number of factors in each individual mode.

Square elements of the core array explain the importance of combinations of the corresponding factors in variance description. Their sum gives variance of the data described by the TUCKER3 model. Mathematically, the TUCKER3 decomposition of matrix $\mathbf{X}^{(I\times JK)}$ can be written as:

$$\mathbf{X}^{(I\times JK)} = \mathbf{A}^{(I\times D)}\mathbf{G}^{(D\times EF)}\left(\mathbf{C}^{(K\times F)} \otimes \mathbf{B}^{(J\times E)}\right)' + \mathbf{E}^{(I\times JK)}$$

where '$\otimes$' denotes the Kronecker product of matrices, and $\underline{\mathbf{E}}$ denotes the three-way array, containing the model's residuals.

Number of factors in each mode is not necessarily the same. Loading vectors, i.e. columns of matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are used to visualize the relationships between parameters in the individual modes. For example, loading plots A1–A2, B1–B2 and C1–C2, represent relationships between the parameters in the first, the second and the third mode, respectively.

2.2. Exploratory analysis of data sets with missing elements

The PCA or TUCKERs models can be constructed even for the data sets with missing elements, using such approaches as, e.g., EM (McLaachlan and Krishnan, 1997) or multiple imputations (Rubin, 1997). In our study, the EM approach was used. EM is an iterative procedure, which can be incorporated in any modeling technique.

The PCA/EM algorithm can be summarized as:

1. fill in missing elements with their initial estimates,
2. perform singular value decomposition of the complete data set,
3. reconstruct $\mathbf{X}$ with the predefined number of factors,
4. replace missing elements with the predicted values and return to step 2 until convergence is attained.

To start the procedure, missing elements are replaced by, e.g., the expected values, calculated as the means of the corresponding row's and column's means. Good performance is also observed, when the initial estimates of missing elements are evaluated, based on the matching procedure (Little and Rubin, 1987). The main idea of this approach is, as follows: the missing elements for an ith object are replaced by the corresponding values, observed for the most similar object from the studied data set. Obviously, similarity is evaluated, based on the subset of the observed variables only. When the PCA model is constructed for the centered or standardized data, updating of the data mean and standard deviation ought to be incorporated in the iterative procedure.

When the number of significant factors is unknown, a cross-validation (CV) procedure is incorporated into the algorithm. The more general algorithm has the following form:

1. fill in missing elements with initial estimates
   for $fn = 1$: A (maximal number of factors (PCs)),
2. perform singular value decomposition,
3. reconstruct $\mathbf{X}$ with fn factors,
4. replace missing elements with the values predicted with the fn factor and return to step 2 until convergence is attained
   perform CV for a model with the fn factors to compute RMSCV and
5. select an optimal number of factors.

If the N-way data contain missing elements, then the EM procedure can be incorporated in the ALS procedure. The ALS/EM approach for the TUCKER3 model has the following form:

1. fill in the missing elements with their initial estimates,

2. initialize **B** and **C**,
3. $[\mathbf{A}] = \mathrm{svd}(\mathbf{X}^{(I\times JK)}(\mathbf{C}\otimes\mathbf{B}), L)$,
4. $[\mathbf{B}] = \mathrm{svd}(\mathbf{X}^{(J\times IK)}(\mathbf{C}\otimes\mathbf{A}), M)$,
5. $[\mathbf{C}] = \mathrm{svd}(\mathbf{X}^{(K\times IJ)}(\mathbf{B}\otimes\mathbf{A}), N)$,
6. $\mathbf{Z} = \mathbf{A}'\mathbf{X}(\mathbf{CB})$,
7. $\mathbf{X}^{\text{predicted}} = \mathbf{AZ}(\mathbf{C}'\otimes\mathbf{B}')$,
8. replace missing elements with their predicted values,
9. return to step 3 until convergence of the model is attained,
10. return to step 3 until convergence of missing elements is attained.

Fit of the final model is expressed by percent of explained variance for the observed elements only (because the estimated missing elements fit the model perfectly well).

2.3. Outliers detection

There are many approaches, aiming at identification of outliers (Rousseeuw and Leroy, 1987). In our study, the robust PCA (robPCA) model was constructed according to the procedure proposed by Croux and Ruiz-Gazen (1996), and outliers were identified, based on the robust distance (Rousseuw and Van Zomeren, 1990).

The main idea of the robPCA method is to search for the direction in which the projected objects have the largest robust scale. As proved by Croux and Ruiz-Gazen, the robust scale estimator, Q_m, proposed by Rousseeuw and Croux (1992) is the most efficient one.

The Q_m estimator of robust scale is defined as the first quartile of all pairwise differences between the two data objects. More formally, for the univariate data set $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$, it is defined as:

$$Q_m(x) = 2.2219 c_m \{|x_i - x_j|;\ i < j\}_{(k)}$$

where

$$k = \binom{h}{2} \approx \binom{m}{2}\Big/4, \quad h = \left[\frac{m}{2}\right] + 1$$

and c_m is a small-sample correction factor. The breakdown point of Q_m is 50%.

The algorithm of robPCA, in the version proposed by Croux and Ruiz-Gazen (1996), consists of the following steps:

1. Center data matrix, $\mathbf{X}$ $(m \times n)$, around the L1-median and calculate its rank r

   $(r \leqslant \min(m-1, n)$;

   $\mathbf{Xc} = \mathbf{X} - \mathrm{ones}(m,1)\,\text{L1-median}(\mathbf{X})$; $\mathbf{Xnew} = \mathbf{Xc}$;
2. construct matrix **A**, containing the normalized rows of matrix **Xnew**;

   $\mathbf{A}(i,:) = \mathbf{Xnew}(i,:)/\mathrm{norm}(\mathbf{Xnew}(i,:))$;
3. consider all directions, described by the data origin and the individual objects of matrix **A** as possible candidates for eigenvectors:
   - project all objects on the possible eigenvectors; $\mathbf{Y} = \mathbf{XnewA}'$,
   - calculate robust scale of all eigenvectors $\mathbf{Qm} = qn(\mathbf{Y})$,
   - select eigenvector with maximal robust scale; i.e., $[kj] = \max(\mathbf{Qm})$;
4. construct lth eigenvector with the selected jth row of **A**; $\mathbf{V}(:,l) = \mathbf{A}(j,:)'$;
5. project all objects on the selected eigenvector; $\mathbf{t} = \mathbf{XcV}$;
6. update data matrix by its orthogonal complement:

   $$\mathbf{Xnew}(i,:) = (\mathbf{Xnew}(i,:)' - \mathbf{V}(:,l)\mathbf{V}(:,l)'\mathbf{Xnew}(i,:)')';$$
7. if the number of eigenvectors, l, is lower than the rank of **Xc**, return to step 2.

The L1-median mentioned above (i.e. in step 1), also known as spatial median, is a highly robust estimator of location (Hubert et al., 2002).

2.4. Robust PLS model for initial estimation of missing elements

In multidimensional data sets, measured parameters usually are intercorrelated to certain extent and any variable can be presented as linear combination of the remaining ones. For instance, the jth column of the data matrix $\mathbf{X}$ $(m \times n)$ can be considered as a dependent variable (denoted as **y**) and modeled by the remaining variables (denoted as **Xr**) with the observed elements. If the dependent variable contains missing elements, these elements ought to be removed from both **y** and **Xr**, and the remaining data will further be denoted as $\mathbf{y}_{\text{obs}}$ $(m_{\text{obs}} \times 1)$ and $\mathbf{Xr}_{\text{obs}}$ $(m_{\text{obs}} \times n_{\text{obs}})$. The PLS model constructed for $\mathbf{y}_{\text{obs}}$ and $\mathbf{Xr}_{\text{obs}}$ can be used for prediction of missing elements in $\mathbf{y}_{\text{obs}}$. However, if the data set contains outliers, it is necessary to construct the robust PLS model for $\mathbf{y}_{\text{obs}}$ and $\mathbf{Xr}_{\text{obs}}$. The robust model ought to well describe data majority, i.e., a subset of objects containing at least 51% of all objects in the initial data set (Walczak, 1995a,b). The subset of objects, for which the best model is observed is called the 'clean subset', i.e., it is considered as a subset without outliers. For all those familiar with data modeling it is obvious, that the good model is a model with the simultaneous good fit and good predictive ability.

The clean subset could be determined using genetic algorithm (GA) (Goldberg, 1989; Lucasius and Kateman, 1993), but for the problem discussed, more efficient proves evolutionary program (EP) (Michalewicz, 1992), which allows replacement of the typical genetic

operations, i.e. cross-over and mutations, with certain more specific ones.

Potential solutions of the investigated problem of finding a 'clean subset of data' are coded in binary chromosomes. Each chromosome is a binary vector $(1 \times m_{obs})$ with ones denoting presence and zeros denoting absence of the objects in model construction.

The main steps of EP are very similar to the typical steps of GA, and namely:

1. randomly select initial population of strings,
2. estimate an optimal model for each chromosome and determine its optimal complexity,
3. calculate fitness functions for all chromosomes,
4. reproduce the next generation, using chosen genetic operations,
5. if convergence is not achieved, return to step 3.

Any solution of the problem at hand ought to contain k^* objects, where: $k_{min} < k^* < k_{max}$ and k_{min} denotes a maximal number of factors in the PLS model, $k_{max} \ll (1-p)m_{obs}$, and p is an assumed fraction of data contamination. For instance, if the data set contains 100 objects and an assumed fraction of contamination equals 0.1 (i.e. 10% objects can be outliers), then k_{max} ought to be much lower than 90 and k_{min} ought to be equal 10 (if we assume, that the PLS model does not require more than 10 factors).

These conditions are equivalent to k^* 1's in any chromosome, considered as a potential solution to our problem. The k^* number of the selected objects are used for construction of a bilinear regression model. Then the residuals for the others objects, which belong to the test set (i.e. the objects with 0's in the considered chromosome) are calculated for the model with one, two, three etc. factors. Then the root mean square error (RMS) is calculated for the first w objects from the test set (where $w = m_{obs} - \text{integer}(p \cdot m_{obs}) - k^*$, which are sorted according to the absolute value of their residuals). Model with the minimal value of RMS is considered as the optimal one. This model is used for prediction of **y** for all m_{obs} objects. Squared residuals, i.e. squared differences between the observed and the predicted **y** values, are sorted and the set of k_{max} objects with the lowest residuals is used for reproduction. The sum of the k_{max} squared residuals is used to calculate the fitness function:

$$\text{fitness} = \frac{1}{\text{RMS}}$$

where

$$\text{RMS} = \sqrt{\sum_{i=1}^{m_{obs}-\text{integer}(p \cdot m_{obs})} \frac{(y_{obs} - y_{pred})^2}{m_{obs} - \text{integer}(p \cdot m_{obs})}}$$

Taking into the account the residuals of the k_{max} objects, i.e. more data than the number of objects used for model construction, we can estimate both the model fit and its predictive ability.

Fitness function for any chromosome containing k^* 1's is calculated, based on the k_{max} objects and these k_{max} objects are used in reproduction step. Chromosomes representing children are constructed by randomly selecting the k^* objects from the set containing k_{max} objects, which are used to evaluate the parent chromosome.

Usually the algorithm converges to a good solution after some 5–10 iterations only (contrary to GA, which requires hundreds of iterations for a good exploration of a whole space of possible solutions). This fast convergence of EP is associated with the fact, that using small subsets of objects (i.e. only k^* objects) to construct the model, we enhance a probability that these subsets will not contain any outliers. A similar idea is explored in the FastMCD approach (Rousseeuw and Van Driessen, 1999).

3. Data

The monitoring data concerns wet precipitation sampling in the Austrian part of the European Air Chemistry Network (Simeonov et al., 1999). The five sampling places ((1) Reutte, (2) Kufstein, (3) Innervillgaten, (4) Haunsberg and (5) Werfenweng) are located in the two Austrian regions, and namely three of them (Reutte, Kufstein, and Innervillgaten) are placed in the Austrian country of Tyrol and the other two (Haunsberg and Werfenweng) in the country of Salzburg. The time period of sampling embraced the years from 1984 to 1993. From the geographical point of view, these samples might well be considered as originating from the North Alpine rim (Reutte, Kufstein, and Haunsberg, collected at the altitudes between 520 and 930 m a.s.l.) and from the inner Alpine area (Innervillgaten and Werfenweng, collected at the altitudes between 940 and 1730 m a.s.l.).

All samples were analysed by use of ion chromatography for the following ions: (1) hydrogen, (2) ammonium, (3) sodium, (4) potassium, (5) calcium, (6) magnesium, (7) chloride, (8) nitrate and (9) sulphate. Initially, this data was organized in the matrix form with the dimensionality (50×9), as presented in Table 1.

4. Results and discussion

The data of mean annual wet deposition loads of the major ions (hydrogen, ammonium, sodium, potassium, calcium, magnesium, chloride, nitrate and sulphate) of

Table 1
The studied data set (Simeonov et al., 1999)

No.	Site[a]	Year	Parameters (kg/ha)								
			1 H^+	2 NH_4^+	3 Na^+	4 K^{2+}	5 Ca^{2+}	6 Mg^{2+}	7 Cl^-	8 NO_3^-	9 SO_4^{2-}
1	1	1984	0.307	5.095	1.014	0.655	**10.369**	**1.159**	2.779	3.766	7.199
2		1985	0.291	5.886	1.874	1.376	**10.369**	**1.159**	4.772	4.126	6.770
3		1986	0.324	4.315	1.356	1.182	**10.369**	**1.159**	4.407	3.872	6.310
4		1987	0.349	5.130	1.363	0.717	**10.369**	**1.159**	3.490	4.149	5.581
5		1988	0.274	5.532	3.629	1.912	9.827	0.825	8.148	5.819	7.606
6		1989	0.124	6.465	7.472	2.976	12.306	1.296	15.173	5.016	8.832
7		1990	0.076	6.197	4.494	1.201	14.395	1.329	14.952	4.573	7.066
8		1991	0.102	6.925	1.897	0.689	8.398	1.094	6.147	4.782	6.184
9		1992	0.083	6.930	1.703	0.977	9.010	1.234	6.665	3.648	5.136
10		1993	0.308	8.765	2.067	1.118	8.278	1.177	5.681	4.545	6.330
11	2	1984	0.571	10.366	1.011	2.415	**4.191**	**0.554**	3.377	7.084	12.770
12		1985	0.506	7.446	2.039	1.214	**4.191**	**0.554**	3.382	5.794	8.566
13		1986	0.373	6.006	1.429	0.723	**4.191**	**0.554**	3.418	4.128	6.420
14		1987	0.476	7.352	1.500	0.618	**4.191**	**0.554**	3.340	4.186	5.670
15		1988	0.511	7.213	1.522	1.175	5.160	0.625	4.975	7.587	9.597
16		1989	0.434	6.832	1.399	0.994	5.139	0.565	5.135	6.107	9.309
17		1990	0.324	7.577	1.681	1.001	4.410	0.569	4.924	5.361	7.830
18		1991	0.338	7.009	1.252	0.474	4.093	0.594	4.204	6.618	9.133
19		1992	0.168	6.109	0.723	0.309	3.087	0.474	2.206	3.966	5.320
20		1993	0.248	10.814	1.648	0.640	3.257	0.499	2.756	5.602	6.905
21	3	1984	0.196	3.409	0.859	0.572	**2.789**	**0.374**	2.932	3.078	6.220
22		1985	0.331	3.714	0.867	0.527	**2.789**	**0.374**	1.587	1.916	7.474
23		1986	0.156	3.635	0.689	0.311	**2.789**	**0.374**	1.611	1.887	4.448
24		1987	0.136	3.316	0.457	0.448	2.845	0.350	2.300	1.947	4.278
25		1988	0.094	2.757	1.004	0.696	3.720	0.447	1.352	1.514	3.644
26		1989	0.125	3.415	0.831	0.393	2.384	0.283	2.191	1.936	3.864
27		1990	0.062	3.520	0.599	0.279	2.314	0.372	1.792	1.960	3.192
28		1991	0.048	2.631	0.409	0.387	2.098	0.353	1.678	1.376	2.490
29		1992	0.039	2.918	0.653	0.348	3.328	0.436	2.569	1.524	2.874
30		1993	0.021	2.225	0.445	0.264	2.834	0.377	1.745	1.373	2.358
31	4	1984	0.250	10.206	2.192	1.552	9.144	0.842	6.652	5.684	11.061
32		1985	0.236	7.709	1.905	1.071	17.215	1.199	11.472	5.807	11.290
33		1986	0.222	7.183	1.506	1.104	12.815	0.889	9.593	5.356	9.424
34		1987	0.279	8.790	2.015	1.133	8.871	0.853	8.476	6.238	9.393
35		1988	0.256	9.328	2.096	1.819	10.499	1.002	10.064	7.598	10.113
36		1989	0.340	9.000	1.613	0.645	9.440	0.800	5.703	6.983	11.160
37		1990	0.238	8.754	5.524	2.260	7.630	0.724	7.239	6.274	8.957
38		1991	0.138	12.158	3.551	1.358	4.644	0.642	8.041	6.083	8.494
39		1992	0.116	8.777	2.737	1.095	5.054	0.573	6.729	4.796	5.913
40		1993	0.183	7.388	7.114	1.224	9.888	0.627	13.530	3.813	7.384
41	5	1984	0.212	5.064	6.060	1.950	4.789	0.885	3.087	3.516	5.907
42		1985	0.110	4.112	2.637	1.130	8.771	2.268	8.012	3.498	8.309
43		1986	0.122	2.993	2.797	0.862	7.986	1.120	2.882	2.689	5.019
44		1987	0.154	3.327	6.900	1.359	7.208	1.752	5.679	3.626	5.526
45		1988	0.083	3.125	2.971	1.407	9.520	1.696	10.537	3.831	5.424
46		1989	0.184	4.594	3.254	1.676	9.132	2.084	3.474	4.129	7.784
47		1990	0.127	4.946	1.933	1.381	4.251	1.142	4.660	3.500	5.012
48		1991	0.085	5.115	2.588	1.991	3.467	4.770	3.098	3.328	4.781
49		1992	0.054	4.868	2.517	2.517	6.956	1.161	2.949	3.109	3.823
50		1993	0.018	4.428	1.882	0.836	43.919	3.777	4.290	3.606	4.469

[a] Sampling sites: 1—Reutte, 2—Kufstein, 3—Innervillgaten, 4—Haunsberg, 5—Werfenweng.

the samples collected in the five sites (Reutte, Kufstein, Haunsberg, Innervillgaten, and Werfenweng) during ten years (i.e. from 1984 to 1993) were taken from (Simeonov et al., 1999). In this data set (denoted as **X**), twenty

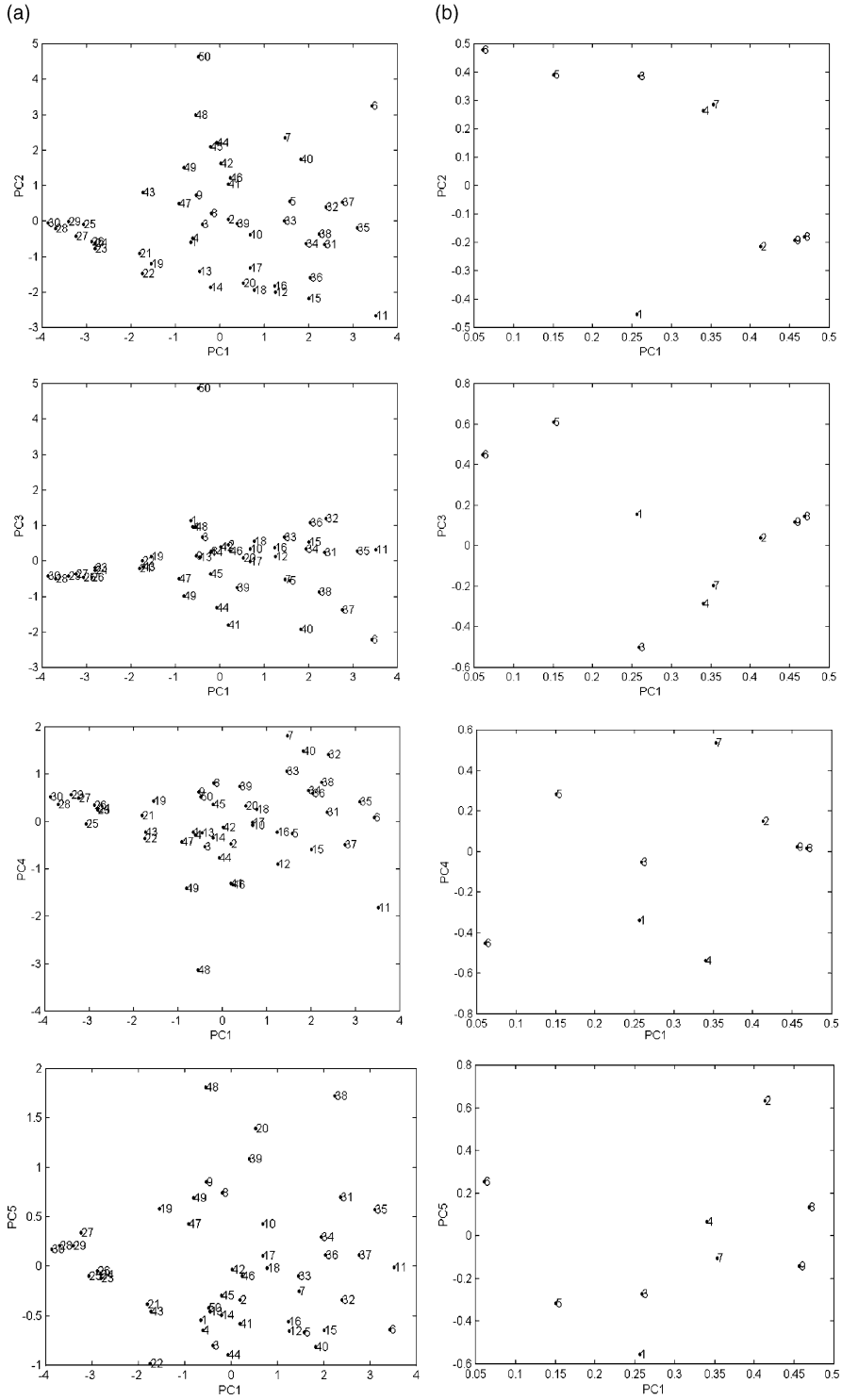

Fig. 1. (a) Score plots and (b) loading plots as a result of PCA for centered and standardized data set **X**.

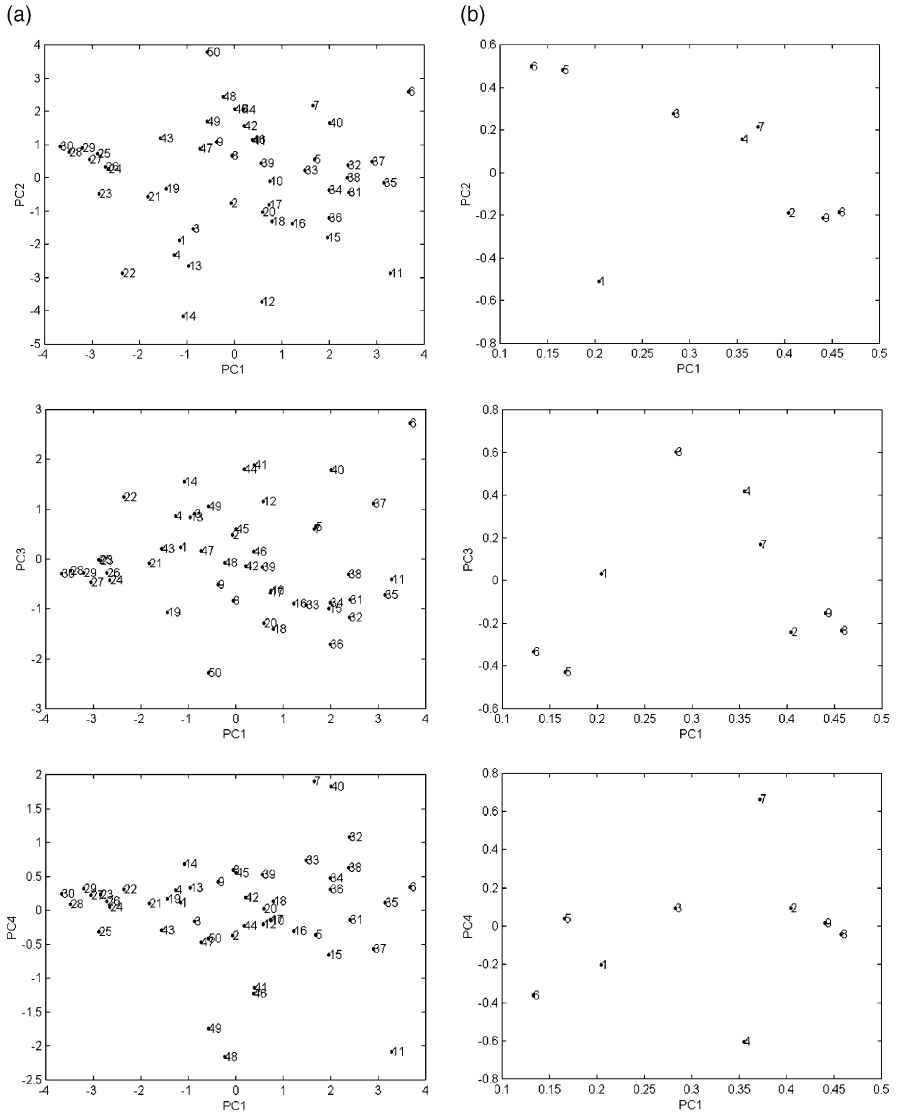

Fig. 2. (a) Score plots and (b) loading plots of EM-PCA for centered and standardized **X1** data set.

Table 2
Estimates of missing elements

Indices of objects	Original data		EM/PCA		Robust PLS	
	Variable 5	Variable 6	Variable 5	Variable 6	Variable 5	Variable 6
1	10.369	1.159	−10.374	−1.234	5.514	0.514
2	10.369	1.159	−3.623	−0.090	5.833	0.646
3	10.369	1.159	−11.348	−1.111	5.947	0.610
4	10.369	1.159	−15.946	−1.961	4.479	0.549
11	4.191	0.554	−5.156	0.370	7.871	0.605
12	4.191	0.554	−21.939	−2.506	5.856	0.586
13	4.191	0.554	−17.114	−2.153	4.752	0.552
14	4.191	0.554	−28.966	−3.791	3.442	0.541
21	2.789	0.374	−2.776	−0.245	5.580	0.512
22	2.789	0.374	−22.973	−2.762	5.944	0.444
23	2.789	0.374	−4.528	−0.480	3.419	0.414

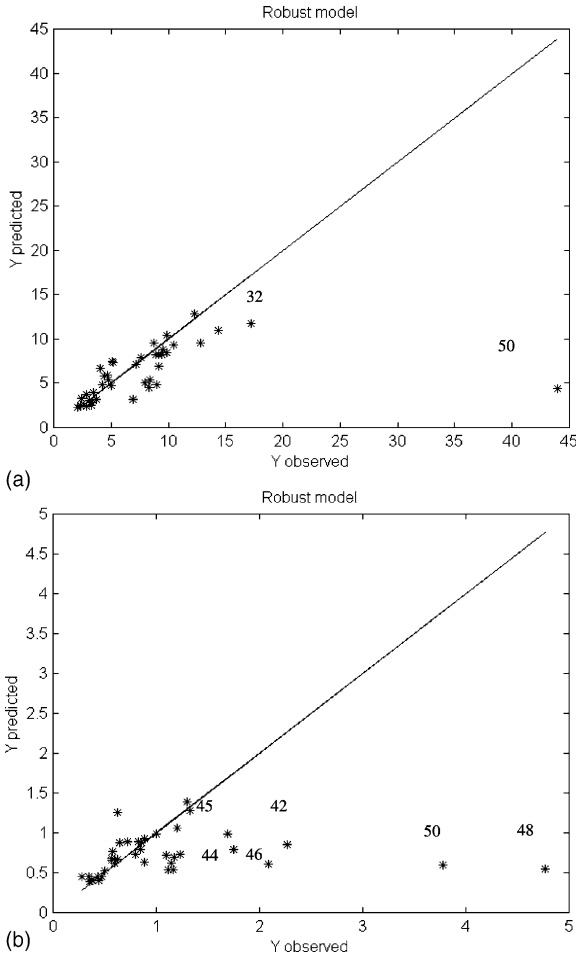

Fig. 3. Robust PLS model of (a) variable 5 and (b) variable 6; assumed fraction of contamination $p = 0.3$.

two missing elements were substituted by the mean values calculated for each sampling site and in this form the data set is presented in Table 1.

Let us have a look at the results of exploratory analysis of this data set, and namely of the PCA. As this data set contains measurements performed within the different magnitude ranges, a comparison of the pollution's profiles for the different sampling sites was performed for the centered and standardized data. For determination of the number of significant components (PCs) of this data set, the percent of modeled variance (Wold, 1987), residual variance (Cattell, 1966), Malinowski's *F*-test (1988) and the RMS error of cross-validation (RMSCV) (Wold, 1978) were used. The PCA model with five significant PCs describes 91.9% of data variance. Score plots and loading plots, which were obtained as a result of this analysis, are presented in Fig. 1.

PC1 reflects the difference between samples from Innervillgaten (samples nos. 21–30, sampling site no. 3) and all the remaining samples. PC2 is constructed mainly due to the difference between object no. 50 (sampling site no. 5, the year 1993) and objects from Kufstein (objects nos. 11–20, sampling site no. 2). PC3 reveals difference between object 50 (sampling site no. 5, the year 1993) and the remaining objects from the same sampling site, whereas PC4 reveals uniqueness of object no. 48 (sampling site no. 5, the year 1991). Based on the loading plots, we can conclude that the objects from Innervillgaten (samplig site no. 3) differ from the remaining objects mainly due to the very low concentration of the NO_3^- and SO_4^{2-} ions (parameters 8 and 9). Difference observed between object 50 (sampling site no. 5, the year 1993) and the objects from Kufstein (sampling site no. 2) is mainly due to the different concentrations of the Ca^{2+} and Mg^{2+} ions (i.e. to parameters 5 and 6), i.e. object 50 characterizes by a very high value of the concentration of these ions. The fourth factor distinguishes object no. 48 with the high values of Mg^{2+} and K^+ (i.e. of parameters 6 and 4) from the remaining ones.

Loading plots reveal high correlation between parameters 7 and 4 (i.e. between the chloride and the potassium ions) and among 2, 9, and 8 (i.e. among the ammonium, sulphate and nitrate ions).

The above conclusions can, however, be inaccurate because substitution of missing elements by the groups means can lead to distortion of inner relationships of the data. Certain solution to this problem is offered by such approaches as, e.g., the EM algorithm. Let us have a look at the results of EM-PCA. This data set, in which the group of means is replaced by missing elements, is denoted as **X1**. According to the CV procedure (Wold, 1978), there are four significant factors explaining 90.7% of the data variance.

Distribution of objects and variables in the score and loading plots is presented in Fig. 2. Nothing is really wrong with this figure, but a closer look at the reconstructed data matrix suggests that the constructed model is rather improper, because the estimates of missing elements are negative (see Table 2). This fact suggests that missing elements were estimated to compensate the influence of the data outliers.

In order to conclude about objects' similarity and the relationships among the variables, outliers ought to be identified and eliminated before construction of the model. To properly identify the outlying objects in a data set with missing elements, we ought to start with proper estimation of missing elements. Taking into the account the presence of outliers, estimation of missing elements was performed with aid of the robust PLS models. One model was constructed in order to predict missing elements of the variable Ca^{2+} (no. 5), and the second one to predict missing elements in the variable

Mg^{2+} (no. 6), in each case using all the remaining variables. These models are presented in Fig. 3.

Estimates of missing elements based on robust PLS models are listed in Table 2. They differ to a large extent from the values used in the original data and, contrary to the estimates of EM-PCA, they all are positive.

Once the missing elements are properly estimated, any approach to outlier identification can be applied. In our study, the robust PCA method was used to the standardized data. Standardization was performed, using the data median and the robust scale (Rousseeuw and Croux, 1992). Projections of objects and variables on the robust PCs are presented in Fig. 4.

Using the robust distance, proposed by Rousseuw and Van Zomeren (1990), two outlying objects can be identified, namely objects 48 and 50 (i.e. Werfenweng, 1991 and 1993, respectively; see Fig. 5). Object no. 50 is far away from the data majority, whereas object no. 48 is outlying to a smaller extent.

It ought to the stressed that identification of outlying objects not necessarily means that these outliers result from erroneous measurements. According to definition,

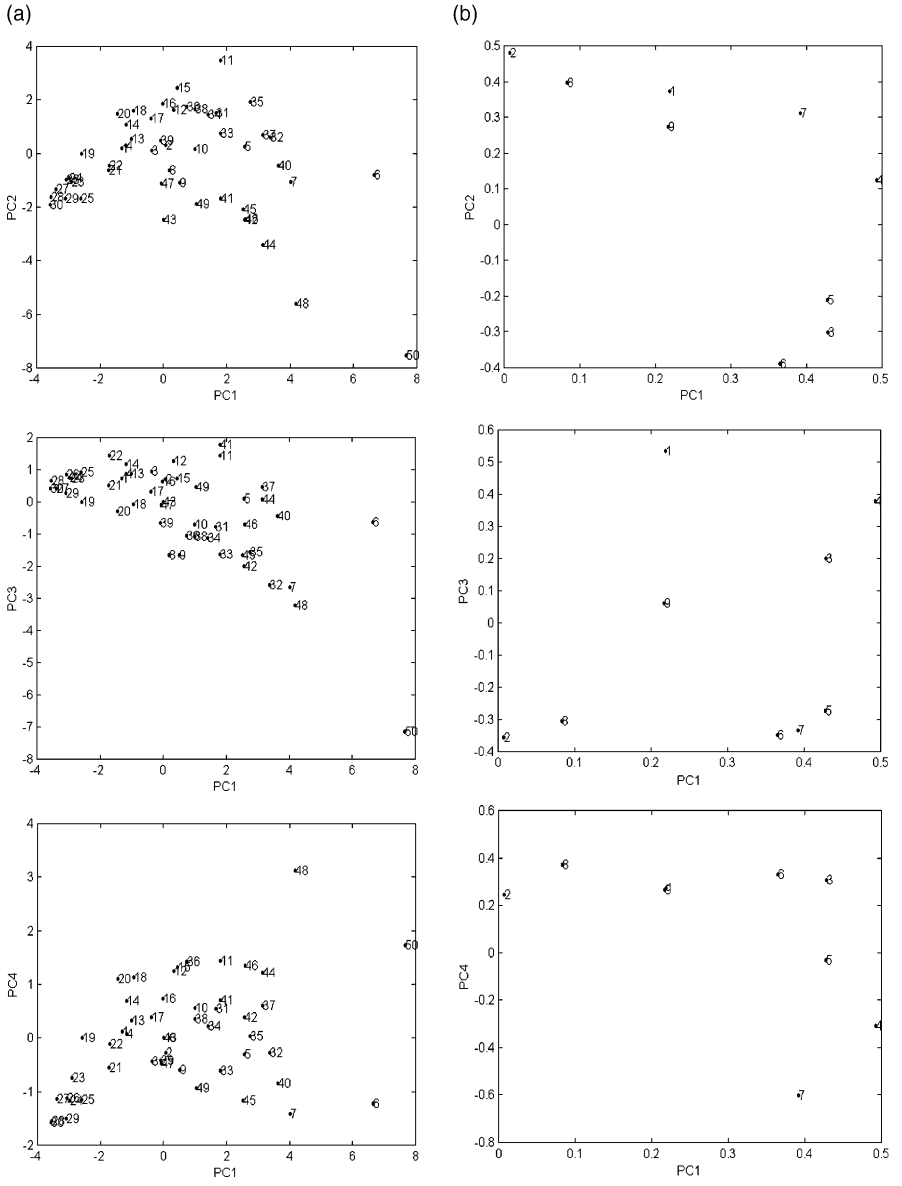

Fig. 4. (a) Score plots and (b) loading plots of robust PCA for standardized data set with missing elements replaced by values estimated by robust PLS model.

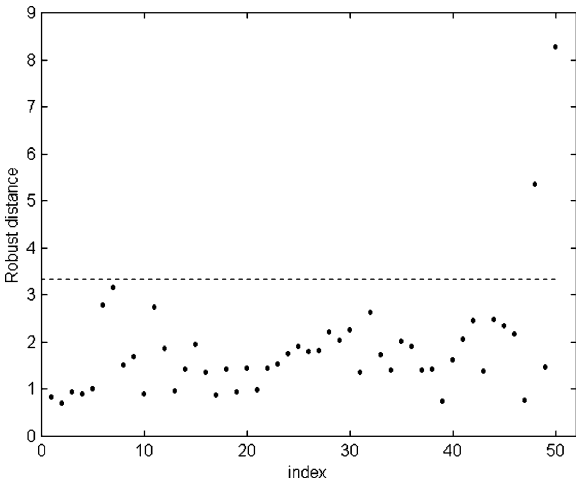

Fig. 5. Robust distances, calculated for four robust scores.

any object from the population different from that of the data majority can be considered as outlying or unique. If we are interested in formulation of general conclusions about the studied data, this type of objects ought to be eliminated and their uniqueness discussed separately.

In our study, instead of eliminating the two objects (i.e. nos. 48 and 50), a proposal was made to treat the outlying elements of these objects (i.e., (48, 5), (48, 6), (50, 5) and (50, 6)) as the missing elements.

As **X2** let us denote the matrix containing missing elements in the variable Ca^{2+} (no. 5) and in the variable Mg^{2+} (no. 6) for objects nos. 1–4 (Reutte, 1984–87), for objects nos. 11–14 (Kufstein, 1984–87), for objects nos. 21–23 (Innervillgraten, 1984–86), for object no. 48 (Werfenweng, 1991) and for object no. 50 (Werfenweng, 1993).

The results of the EM-PCA model for this data set are presented in Fig. 6. Five significant PCs describe 94.7% of data variance.

The results of EM-PCA for matrix **X2** (Fig. 6) differ from the results of EM-PCA for the data set **X1** (Fig. 2) and also from the results of PCA performed for the original data set **X** (Fig. 1). The main difference is observed in the relative location of objects nos. 48 and 50 in the score plots and in the correlation structure of measured variables. In the loading plots of **X** there are two subgroups of correlated variables, whereas in the data set **X2**, three subgroups of correlated parameters are revealed:

PCA(**X**)

Cluster 1: NH_4^+, NO_3^-, SO_4^{2-} (parameters nos. 2, 8 and 9),
Cluster 2: K^+, Cl^- (parameters nos. 4 and 7),
Not correlated: H^+, Na^+, Ca^{2+}, Mg^{2+} (parameters nos. 1, 3, 5 and 6).

EM-PCA(**X2**)

Cluster 1: NH_4^+, NO_3^-, SO_4^{2-} (parameters nos. 2, 8 and 9),
Cluster 2: K^+, Ca^{2+}, Cl^- (parameters nos. 4, 5 and 7),
Cluster 3: Na^+, Mg^{2+} (parameters nos. 3 and 6),
Not correlated: H^+ (parameter no. 1).

The PCA loading plots are easy to interpret, whereas it is difficult to trace any changes of the ions concentration in the function of time and to interpret the main differences among the five studied regions, because on the score plots these two modes (i.e. regions and time) are mixed. This difficulty can be omitted by the Tucker3 approach. The investigated data set can be organized in the form of a 3-way array. Matrices **X** and **X2**, presented as arrays of dimensionality [9 10 5], will further be denoted as $\underline{\mathbf{X}}$ and $\underline{\mathbf{X2}}$. The Tucker3 models with complexity [2 2 2], constructed for the data standardized in the first mode, describe 90.7% and 92.86 % of data variance for $\underline{\mathbf{X}}$ and $\underline{\mathbf{X2}}$, respectively. The loading plots, which present the measured parameters, the sampling places and the sampling time, are given in Fig. 7.

The core matrices, **G1** and **G2**, in an unfolded form are presented below:

$$\mathbf{G1} = \left[\begin{array}{cc|cc} 42.1682 & -0.4864 & -0.0954 & 2.1109 \\ -0.0731 & 2.3692 & 8.7062 & 2.4002 \end{array}\right]$$

$$\mathbf{G2} = \left[\begin{array}{cc|cc} 44.9850 & -0.1599 & -0.0152 & 3.2307 \\ -0.0469 & 2.0468 & 9.1971 & 0.7977 \end{array}\right]$$

The two parts of matrices **G1** and **G2** refer to the region mode, i.e. the first part is associated with the first component of the region mode, and the second (right) part is associated with the second component of this mode. Rows refer to the components associated with the time mode. Columns within each part of **G** are associated with the factors of the parameters' mode.

The squared elements of $\underline{\mathbf{G}}$ estimate the importance of interactions among factors. Based on them, we can conclude that the most important are elements (1 1 1) and (2 1 2) of the two core matrices, **G1** and **G2**, i.e. the first factors of parameters, time and regions modes and the second factor of parameter mode, the first factor of time mode, and the second factor describing regions (2 1 2).

Taking into the account the sign of a core element and the signs of all loadings on the factors associated with this core element, for the data $\underline{\mathbf{X}}$ we can conclude that within the scrutinized period of time the regions 4, 2 and 1 have the relatively higher values of all the parameters than the remaining regions 3 and 5, but this difference is the highest for the parameters 8, 2, and 9 and the smallest for the parameters 5 and 6. It means that the interaction among the first factors of all modes reveals the difference between the northern Alpine rim (sampling sites 1, 2, 4) and the inner Alpine rim (sam-

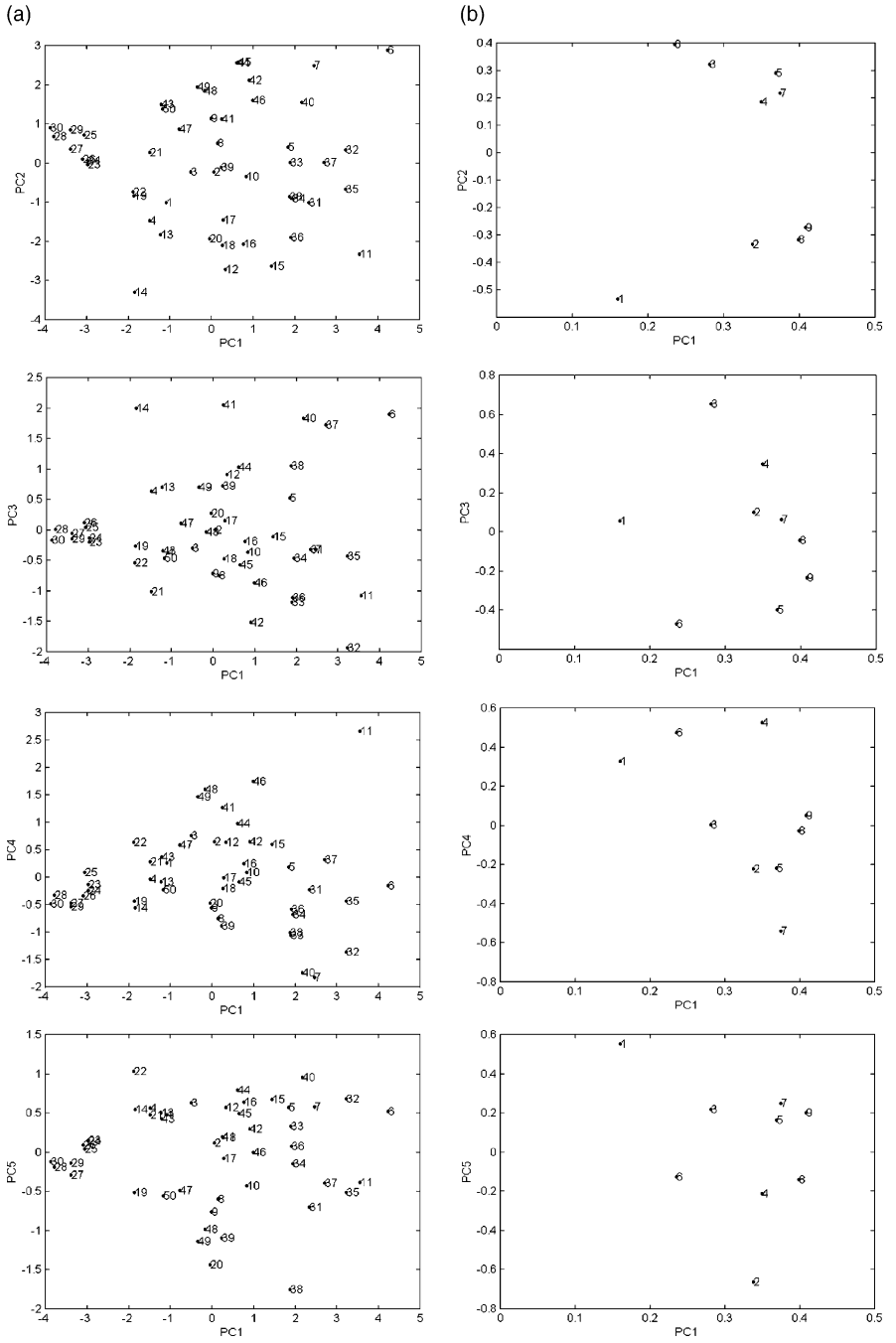

Fig. 6. (a) Score plots and (b) loading plots as a result of EM-PCA of standardized **X2** data set.

pling sites 3 and 5), mainly due to the differences between concentration of three of 'antropogenic' (2, 9, 8) and 'crustal' (5 and 6) parameters. Although all the parameters studied have a positive loading on factor 1, they are ordered along this factor in the following order: crustal (Ca^{2+} (5), Mg^{2+} (6)), 'mixed salt' (K^{+} (4), Na^{+} (3), Cl^{-} (7)) with one of the 'anthropogenic' (H^{+} (1)) and the rest of antropogenic (NH_4^{+} (2), SO_4^{2-} (9) and NO_3^{-} (8)) parameters.

Interaction of factors (2 1 2) reveals that within the studied period of time, region 2, and to a very small extent, regions 3 and 4 also, have the relatively

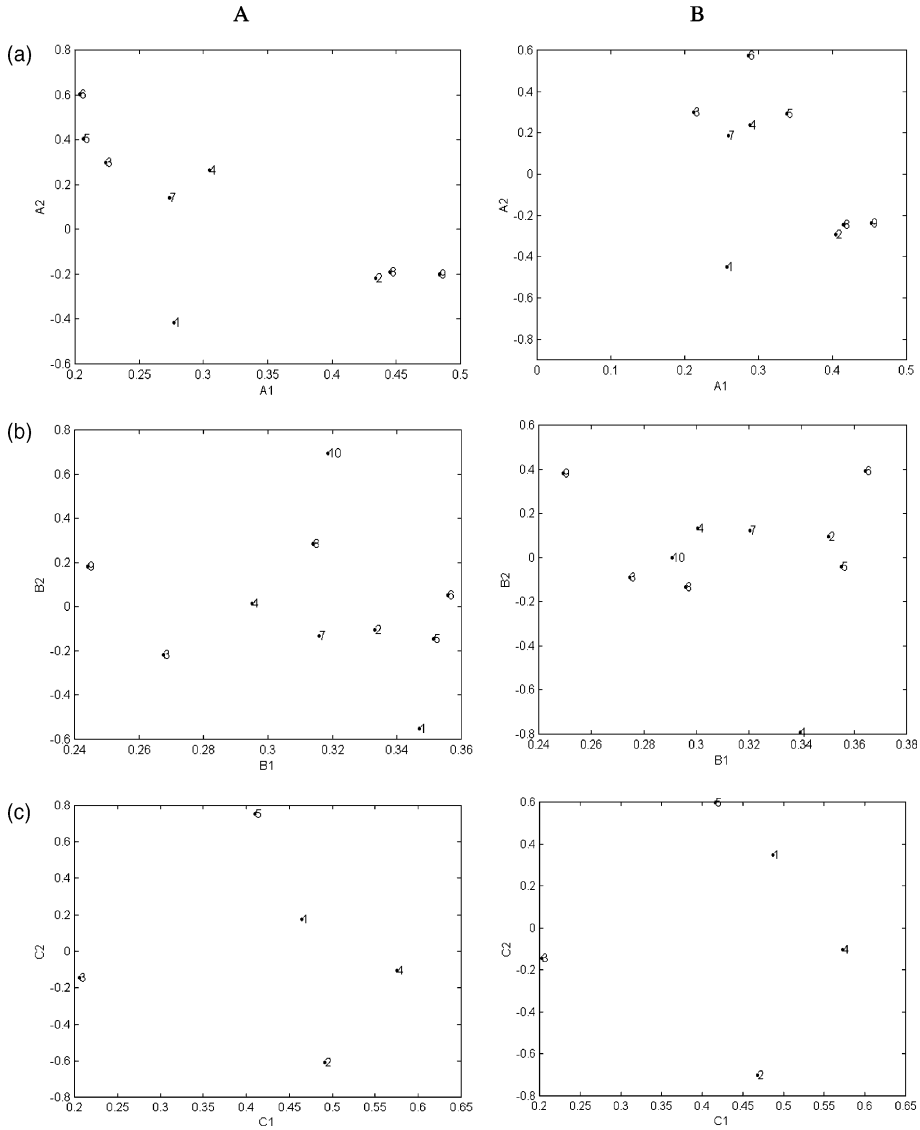

Fig. 7. Loading plots of TUCKER3 and EM-TUCKER3 model of (A) **X** and (B) **X2** data arrays standardized in the first mode; loading plots (a) parameters, (b) sampling time and (c) sampling places.

lower values of parameters 6, 5, 4, 3, and 7 than the remaining regions, and the high values of parameter 1, whereas regions 5 and 1 have the relatively lower values of parameter 1 and, to the lower extent, of parameters 9, 2 and 8, and the high values of parameters 6, 4, 5, 3 and 7.

According to the Tucker3 model of **X2**, the main difference between the northern Alpine rim (sampling sites 1, 2 and 4) and the inner Alpine rim (sampling sites 3 and 5) is not caused by the concentration of crustal and antropogenic parameters, but by the differences in concentration of mixed salts and antropogenic parameters. Parameters 5 and 6 (crustal parameters) changed their position on the loading plot and now along the first factor the studied parameters are ordered as: mixed salts with one of the antropogenic (H^+ (1)), crustal and the rest of antropogenic parameters.

Similar as in the case of **X**, the second factor additionally differentiates all sites within the northern Alpine rim and the inner Alpine regions, and the main difference is observed between the sites 5 and 2, due to the differences in concentration of parameter 1 and of the remaining parameters from the group of antropogenic and crustal parameters. Now an overall difference between sites 5 and 1 is even lower, than this for the **X** data. Some differences are also observed in the time

mode, and namely the different position of variables 8 and 10 along the first factor is observed.

5. Conclusions

Correct estimation of missing elements, when the data set contains the outliers plays an important role in an exploratory data analysis. In this paper the general strategy was shown, allowing exploration of contaminated data sets with missing elements, based on application of robust PLS for an initial estimation of missing elements. After identification of outlying elements and their replacement by missing values, it was possible to construct the EM-PCA and EM-TUCKER3 models, to explore the data set studied. Final conclusions based on these models differ to certain extent from the conclusions based on the models constructed for the data set with missing elements, substituted by the groups' means and containing the outliers.

References

Cattell, R.B., 1966. The Scree test for the number of factors. Multivariate Behavioral Research 1, 245–276.

Croux, C., Ruiz-Gazen, A., 1996. A fast algorithm for Roubust principal components based on projection pursuit. In: COMPSTAT: Proceedings in Computational Statistics. Physica-Verlag, Heidelberg, pp. 211–217.

De Ligny, C.L., Spanjer, M., van Houwelingen, J.C., Weesie, H.M., 1984. Three-mode factor analysis of data on retention in normal-phase high-performance liquid chromatography. Journal of Chromatography 301, 311–323.

Geladi, P., 1989. Analysis of multi-way multi-mode data. Chemometrics and Intelligent Laboratory Systems 7, 11–30.

Goldberg, D.E., 1989. Genetic Algorithms in Search Optimization, and Machine Learning. Addison-Wesley, New York.

Hubert, M., Rousseeuw, P.J., Verboven, S., 2002. A fast method for robust principal components with applications to chemometrics. Chemometrics and Intelligent Laboratory Systems 60, 101–111.

Joliffe, I.T., 1986. Principal Components Analysis. Springer, New York.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. John Wiley & Sons, New York.

Lucasius, C.B., Kateman, G., 1993. Understanding and using genetic algorithms. Part I. Concepts, properties and context. Chemometrics and Intelligent Laboratory Systems 19, 1–33.

Malinowski, E.R., 1988. Statistical *F*-tests for abstract factor analysis and target testing. Journal of Chemometrics 3, 49–60.

McLaachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. John Wiley & Sons, New York.

Michalewicz, Z., 1992. Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag, New York.

Rousseeuw, P.J., Croux, C., 1992. Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association 88, 1273–1283.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. John Wiley & Sons, New York.

Rousseeuw, P.J., Van Driessen, K., 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics 41, 212–223.

Rousseuw, P.J., Van Zomeren, B.C., 1990. Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association 85, 633–639.

Rubin, D.B., 1997. Multiple Imputation for Nonresponse in Survey. John Wiley & Sons, New York.

Simeonov, V., Puxbaum, H., Tsakovski, S., Sarbu, C., Kalina, M., 1999. Classification and receptor modeling of wet precipitation data from central Austria (1984–1993). Environmetrics 10, 137–152.

Tucker, L.R., 1966. Some mathematical notes on three-mode factor analysis. Psychometrika 31, 279–311.

Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J., 1998. Handbook of Chemometrics and Qualimetrics: Part B. Elsevier, Amsterdam, p. 87–150.

Van der Kloot, W.A., Kroonenberg, P.M., 1985. External analysis with three-mode principal component models. Psychometrika 50, 479–494.

Walczak, B., 1995a. Outlier detection in multivariate calibration. Chemometrics and Intelligent Laboratory Systems 28, 259–272.

Walczak, B., 1995b. Outlier detection in bilinear calibration. Chemometrics and Intelligent Laboratory Systems 29, 63–73.

Wold, S., 1978. Cross-validatory estimation of the number of components in factor and principal components models. Technometrics 20, 397–406.

Wold, S., 1987. Principal Components Analysis. Chemometrics and Intelligent Laboratory Systems 2, 37–52.