Polish J. of Environ. Stud. Vol. 16, No. 4 (2007), 613-622

Original Research

Chemometric Treatment of Missing Elements in Air Quality Data Sets

A. Smoliński[1], S. Hławiczka[2]

[1]Central Mining Institute, Department of Energy Saving and Air Protection, Plac Gwarków 1, 40-166 Katowice, Poland
[2]Institute for Ecology of Industrial Areas, ul. 6 Kossutha, 40-844 Katowice, Poland

Received: July 4, 2006
Accepted: February 18, 2007

Abstract

The article reports the results of an exploratory analysis of an air monitoring data set, collected at a monitoring station in the biggest, most congested and most polluted city of the silesian region, Katowice. In order to extract important information on air pollution in this city, the strategy of exploring the data set with missing elements and outliers simultaneously existing in the data was used. The strategy assumed the initial estimation of missing elements based on the application of robust Partial Least Squares (rPLS) and outliers identification based on the so-called robust distance. After outliers identification and replacing them with missing elements, the Expectation-Maximization iterative approach (built into Principal Component Analysis (PCA)) was used for the construction of the final model.

Keywords: air quality, contaminated data, exploratory analysis, missing data, multiple outliers

Introduction

Environmental monitoring data sets often contain missing elements and/or outliers, which may result from insufficient sampling, errors in measurements or faults in data acquisition. Outlying objects do not always mean that there are measurements containing large errors in the data set. According to the definition, any object from the population different from the data majority is considered an unique/outlying object. The existence of outliers and/or incomplete data are a significant obstacle not only for time-series prediction, but also for the calculation of mean values of air pollutant concentrations needed, e.g. for comparison with environmental standards.

Most of the chemometric methods used to explore the knowledge hidden in monitoring data sets work with complete data sets only. Otherwise, missing values have to be filled in. The replacement of missing elements should be performed carefully for any model as using simple statistical methods to solve the problem of missing data in environmental sciences might be misleading. The simplest approach consists in setting the missing elements to zero or to mean values of the measured parameter. However, the aforementioned solutions are not proper for several reasons due to the fact that the mean values can be significantly influenced by outliers, or the true values can be much higher than zero. Moreover, replacing the missing elements by zero or mean values destroys correlation in the data, influencing final interpretation of the studied relationships between objects and variables. A better solution to the problem with missing elements is to estimate the missing values by considering the non-missing, i.e. observed elements only. Such an approach can be presented as minimization of the sum of the squared residuals of the observed elements only, i.e.:

$$\min \sum (W \cdot (X - \hat{X}))_F^2 \quad (1)$$

*Corresponding author; e-mail: smolin@gig.katowice.pl

where matrices X and $\hat{X}$ represent the experimental data set and the reconstructed data set (based on the applied model), respectively. F denotes the complexity of the model whereas W is the matrix, the elements thereof being unity or zero, denoting observed and unobserved elements, respectively. The symbol '•' denotes the element-wise multiplication of the two matrices.

The most popular approach to dealing with missing data relies on expectation maximization (EM) [1-8] or multiple imputation (MI) [9-11] iterative procedure of missing elements estimation. As far as the chemometric methods are concerned, there are many approaches which allow correct identification of outliers, such as the robust Principal Component Analysis (rPCA) presented in this paper [12-15]. The analysis is constructed according to the Croux and Ruiz-Gazen procedure where outliers are identified based on the so-called robust distance. Even though classic chemometric methods deal with the problem of missing elements and outlying elements separately existing in the monitoring data set, it is possible to correct exploratory analysis of data sets with missing elements and outliers simultaneously as proposed by Smoliński et al. [16]. In this article, calculations were performed for a real air quality monitoring data set, which contains measurements of pollutant concentrations performed at a monitoring station in Katowice, Poland.

Theory

In our study, chemometric techniques such as Principal Component Analysis (PCA) [17-21], robust Principal Component Analysis (rPCA) [12-15], Expectation Maximization approach [1-8] and robust Partial Least Squares (rPLS) [22-26] were used to analyze environmental data sets. Their main principles are described below.

Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate procedure which can be applied to reduce data dimensionality with minimal loss of information, to visualize data and to reduce a part of experimental error [17-21]. The experimental data are usually collected into a matrix **X** (m x n), where m and n denote, respectively, the number of samples (objects) and the number of measured parameters (variables). In PCA, matrix **X** is decomposed into two matrices, namely score matrix **S** (m x fn) and loading matrix **D** (n x fn), where fn denotes the number of significant factors. Matrices **S** and **D** are orthogonal, i.e. **S'S=D'D=I**, where **I** is identity matrix. The matrix **E** of dimensions (m x n) represents a residual matrix. The columns of score matrix and rows of loading matrix are called Principal Components (PCs) or eigenvectors. The first PC is a linear combination of original variables that explains the greatest amount of variation of the data; the second PC explains the part of information not explained by the first PC, etc. The sum of the squared elements of each PC is called eigenvalue and represents the portion of the variance which is modeled by the corresponding PC. Therefore, the first principal component is associated with the highest eigenvalue.

Scores vectors (i.e. the columns of matrix **S**) and loading vectors (i.e. the columns of matrix **D**) are used to visualize relationships between the objects and the parameters in matrix **X**, respectively. The relationships between the objects and parameters could be investigated with the use of the score-plots and the loading plots.

Iterative Algorithm for Dealing with Missing Elements

The studied environmental data set contains missing elements (some measurements were not recorded, possibly due to insufficient sample amounts or because the measured value was outside the measuring range of the instrument, or due to instrument malfunction). To construct any model for the data with missing elements, Expectation Maximization (EM) [1-8] approach or Multiple Imputations [9-11] approach can be used. In our study, an iterative algorithm of EM was applied. The main steps of the EM algorithm can be presented as follows:

1. initialization of the missing elements,
2. estimation of model parameters for the actual data set,
3. estimation of the missing elements using the parameters of actual model,
4. return to step 2 until convergence is achieved.

At the first step of the EM procedure, i.e., initialization of missing elements, missing elements are replaced by the expected values, calculated as the means of the corresponding row and column means. The EM algorithm can be built into different computational procedures such as iterative EM/PCA algorithm [2, 3].

Robust PCA and Outliers Detection

The classical PCA is useless when data sets are contaminated by outliers and therefore the correct identification of all outlying objects is of utmost importance. Within the chemometric methods there are numerous approaches enabling outlier identification [27-31]. In order to properly identify outlying objects, a robust PCA model (robPCA) was constructed based on the procedure of Croux and Ruiz-Gazen [12]. The outlying objects were identified using a robust distance [27]. Like in classical PCA, the first step in robust PCA is to center the data using the L1-median [32]. The L1-median, also called spatial median, is defined as a point which minimizes the sum of Euclidian distances to all points in the data. The robPCA is based on the search for the direction in which the projected objects have the largest robust scale. The robust scale estimator, Q_n, is the most effective one [33]. It is defined as the first

quartile of all pair-wise differences between the two data objects. For the univariate data set $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$, it is defined as:

$$Q_n(x) = 2.2219\, c_m\, \{|xi - xj|; i < j\}_{(k)} \quad (2)$$

where $k = \binom{h}{2} \approx \frac{\binom{m}{2}}{4}$, $h = \left[\frac{m}{2}\right] + 1$, and c_m is a small-sample correction factor. The breakdown point of Q_n is 50%.

The main steps of the robPCA algorithm can be presented as follows [12]:

1. Centering data matrix, **X** (m x n), around the L1-median and calculating its rank r≤min(m-1,n); **Xc** = **X**-ones(m,1) L1-median(**X**); **Xnew** = **Xc**;
2. Constructing matrix **A,** containing the normalized rows of matrix **Xnew**; **A**(i,:)=**Xnew**(i,:)./norm(**Xnew**(i,:);
3. Considering all directions described by the data origin and the individual objects of matrix **A** as possible candidates for eigenvectors:
   – projecting all objects on the possible eigenvectors; **Y**=**XnewA**'
   – calculating robust scale of all eigenvectors **Qn**=qn(**Y**),
   – selecting eigenvector with maximal robust scale; i.e., [k j]=max(**Qn**);
4. Constructing the l-th eigenvector with the selected j-th row of **A**; **V**(:,l)=**A**(j,:)';
5. Projecting all objects on the selected eigenvector; **t** = **XcV**;
6. Updating data matrix by its orthogonal complement: **Xnew**(i,:)=(**Xnew**(i,:)'-**V**(:,l)**V**(:,l)'**Xnew**(i,:)')';
7. If the number of eigenvectors, l, is lower than the rank of **Xc,** returning to step 2.

Robust PLS

Partial Least Squares (PLS) is one of the most popular multivariate calibration methods [34-38]. Due to the fact that multidimensional data sets contain inter-correlated measured parameters, it is possible to present many variables as a linear combination of the remaining ones. For instance, a dependent variable **y** (m.1), which is the k-th column of data matrix **X** (m x n), can be presented as a linear combination of the remaining variables $\mathbf{X}_r$ (m x n-1). The PLS model can be written as:

$$\mathbf{y} = \mathbf{X}_r\mathbf{b} + \mathbf{e} \quad (3)$$

where **b** is a vector of n-1 regression coefficients and **e** represents a part of variable **y**, which is not explained by the PLS model.

The studied environmental data set contains missing elements and if they are present in the dependent variable, they should be removed from **y** and $\mathbf{X}_r$. The PLS model is constructed for the remaining data, denoted as $\mathbf{y}_o$ (m_s x 1) and $\mathbf{X}_{ro}$ (m_o x n_o). This model can be used for the prediction of missing elements in **y**. Because environmental data are additionally contaminated by the outliers, it is necessary to use a robust version of the Partial Least Squares. Here the robust PLS approach, rPLS, [proposed in 22, 23]. was used. The rPLS model describes well the data majority, and thus allows us to construct the correct model for $\mathbf{y}_o$ (m_s x 1) and $\mathbf{X}_{ro}$ (m_o x n_o), even if 49% of the data are outlying. The best model is constructed for the so-called clean subset, i.e. the subset of objects without outliers. The constructed model should be characterized not only by good fit ability, but also by a good predictive power. In order to accurately determine the clean subset, genetic algorithm (GA) could be used [39-43], but for the problem discussed, an evolutionary program (EP) [44, 45] appeared more efficient. EP allows replacing the typical operations such as crossovers and mutations with more specific ones. In the EP algorithm, the potential solutions of the investigated problem of finding a clean subset of data are coded in binary chromosomes where the ones denoting presence and zeros represent the absence of the object in model construction.

The EP algorithm convergence is usually achieved after 5-10 iterations (much faster than for GA, which requires hundreds of iterations to gain the convergence). It is important to note that any solution of the problem ought to contain k^* objects. The number of objects must be higher than a maximal number of factors in the PLS model and much more lower than $(1\text{-}p)m_o$, where p is an assumed fraction of data contamination. The k^* objects are used to construct the PLS model, whereas the remaining objects belong to the test set. For the objects from the test set, the residuals were calculated for the model with one, two, etc., factors.

The next step is the calculation of the root mean square error (RMS) for the first w objects from the test set, where $w = m_{obs} - \text{integer}(p \cdot m_{obs}) - k^*$. The calculated RMS are sorted according to the absolute value of their residuals, and the model with minimal value of RMS is chosen as the optimal one. Based on this model, y is predicted for all m_o 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:

$$\text{fitness} = \frac{1}{\text{RMS}} \quad (4)$$

where

$$\text{RMS} = \sqrt{\sum_{i=1}^{m_{obs} - \text{integer}(p \cdot m_{obs})} \frac{(y_{obs} - y_{pred})^2}{m_{obs} - \text{integer}(p \cdot m_{obs})}} \quad (5)$$

Taking into 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 the 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.

The EP algorithm can be summarized in the following steps:

1. randomly selecting initial population of strings;
2. estimating an optimal model for each chromosome and determining its optimal complexity;
3. calculating fitness functions for all chromosomes;
4. reproducing the next generation, using chosen genetic operations;

if convergence is not achieved, return to step 3.

Data

Investigations of the pollutant concentrations were carried out at a monitoring station operated by the Institute for Ecology of Industrial Areas. The station is located on the Institute's premisses, about 5 km westward from the centre of Katowice, a city with a population of 350,000. There are a number of industrial air pollution sources located in the area surrounding the sampling point. Within the range of 10 km from the site are: a non-ferrous metal smelter, two steelworks, a chemical factory and six coal-fired heat and power plants of 3,200 MW capacity. Since 1990, the station has been functioning within Poland's National Air Monitoring Network. In 1999, the station was incorporated into the European Air Monitoring Network, EUROAIRNET, operated by the European Environment Agency. The studied data set presents values of ten different physical and chemical parameters (see Table 1). The air was sampled daily in summer 2003 from the 22nd of June to the 23rd of September. Data are organized in matrix **X** (92 x 10). Each row of matrix **X** represents days when the measurements were taken whereas each column represents measured parameters. Each element of data matrix, x_{ij}, is a mean value of the j-th parameter measured in i-th day. The mean, median and standard deviation of the measured parameters are presented in Table 2. As the measured parameters significantly differ in their ranges, the data set is standardized according to the formula:

$$x_{ij} = \frac{\left(x_{ij} - \bar{x}_j\right)}{s_j} \tag{6}$$

where $\bar{x}_j$ and s_j denote the mean of the j-th column and its standard deviation, respectively.

Results and Discussion

To explore the studied data set and to examine the similarities between the samples, Principal Component Analysis was used. The classical PCA cannot be applied due to the missing elements existing in the studied data. The PCA model can however, be constructed even for a data set with missing elements, based on the EM algorithm incorporated in the PCA technique. Due to the fact that the studied data set contains measurements performed within different magnitude ranges, the PCA model was constructed for the centered and standardized data. The number of significant principal components (PCs) of this data set was determined according to the CV procedure [46]. 91.39% of data variance was described by five PCs.

Table 1. Ten physical and chemical parameters measured in summer 2003 at a monitoring station in Katowice, Poland.

No.	Parameters	Units
1	Wind velocity	m/s
2	Temperature	°C
3	Humidity	%
4	Solar radiation	W/m^2
5	SO_2	µg/m^3
6	PM10	µg/m^3
7	NO	µg/m^3
8	NO_2	µg/m^3
9	CO	mg/m^3
10	O_3	µg/m^3

Table 2. Mean, median and standard deviation of the ten physical and chemical parameters measured in summer 2003 at a monitoring station in Katowice, Poland.

Parameter no.	Mean	Median	Std
1	0.6514	0.3948	0.5500
2	19.3094	18.9740	3.7137
3	68.5752	67.3646	11.5519
4	90.3480	79.3569	41.9433
5	25.3545	23.4688	9.0393
6	29.9290	27.1771	13.3881
7	16.6393	8.3333	19.6788
8	32.3981	30.1250	12.2086
9	0.5190	0.4983	0.1439
10	51.7116	49.0417	17.0018

Score plots and loading plots, which were obtained as a result of this analysis, are presented in Fig. 1.

PC1 reveals a difference between days: the 16th and the 18th of September (samples nos. 86 and 88) and all remaining measuring days. Based on the loading plots, it is possible to conclude that these differences are mainly due to the relatively high values of PM10, NO_2 and CO concentrations (parameters nos. 6, 8 and 9) and relatively low values of the remaining parameters observed for the samples collected on the 16th and the 18th of September. Moreover, the 16th of September (sample no. 86) is characterized by the highest concentration of PM10 and CO (variable nos. 6 and 9). PC2 reflects the difference between days: the 12th and the 4th of September (samples nos. 82 and 74) and all remaining samples, mainly due to the highest humidity (parameters no. 3). PC2 also reveals the uniqueness of the days: the 24th of June and the 27th of July (samples nos. 2 and 35) due to the relatively high value of solar radiation, temperature and concentration of ozone (parameters nos. 4, 2 and 10). PC3 is constructed due to the difference between days: the 30th and the 31st of July (sample nos. 38 and 39) and the 27th of August (sample no. 66). Namely, on the 30th and the 31st of July (samples nos. 38 and 39) relatively high humidity and SO_2 concentration (parameter nos. 3 and 5) were observed. On the 30th of July the highest humidity of summer 2003 was

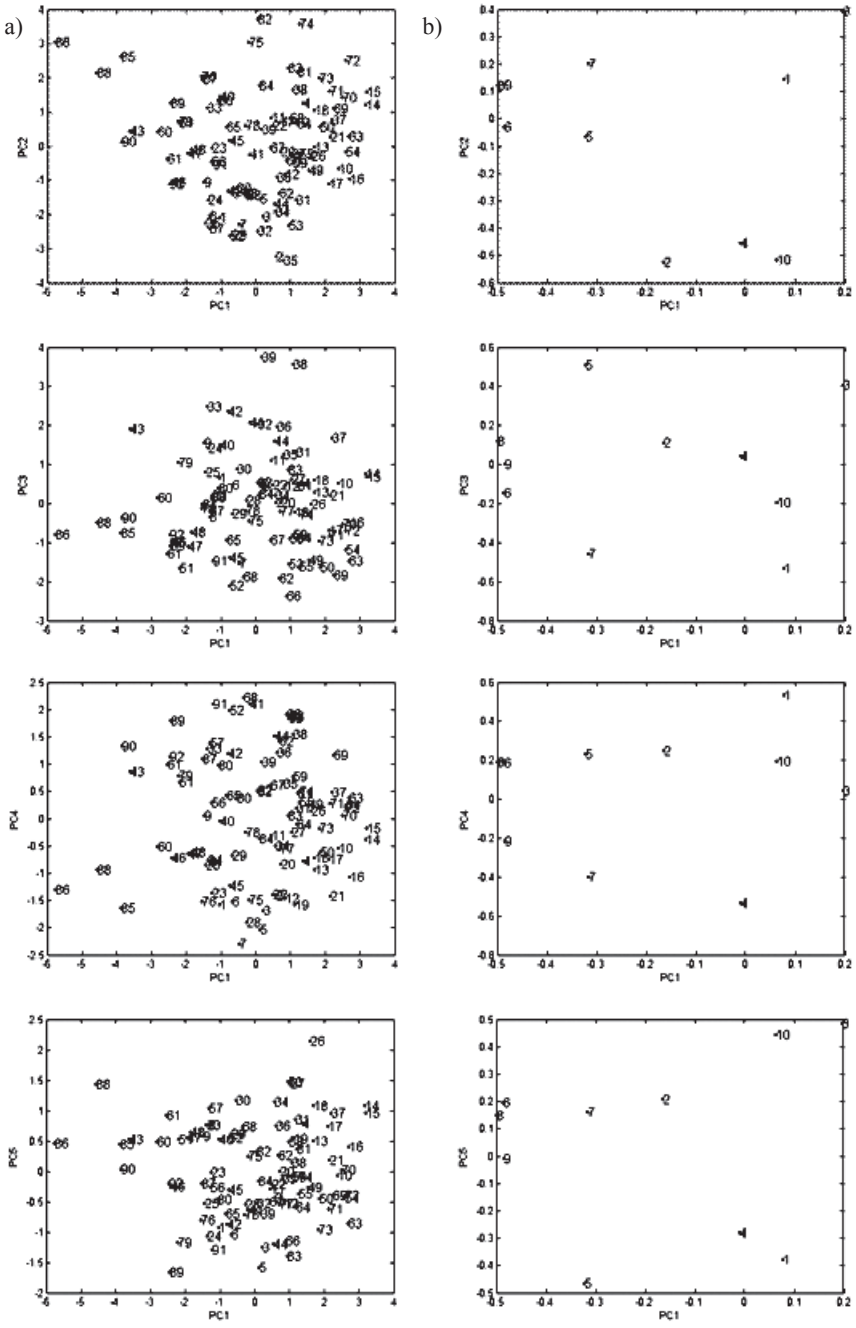

Fig. 1. a) Score plots and b) loading plots as a result of EM/PCA for centered and standardized data **X** (92 x 10).

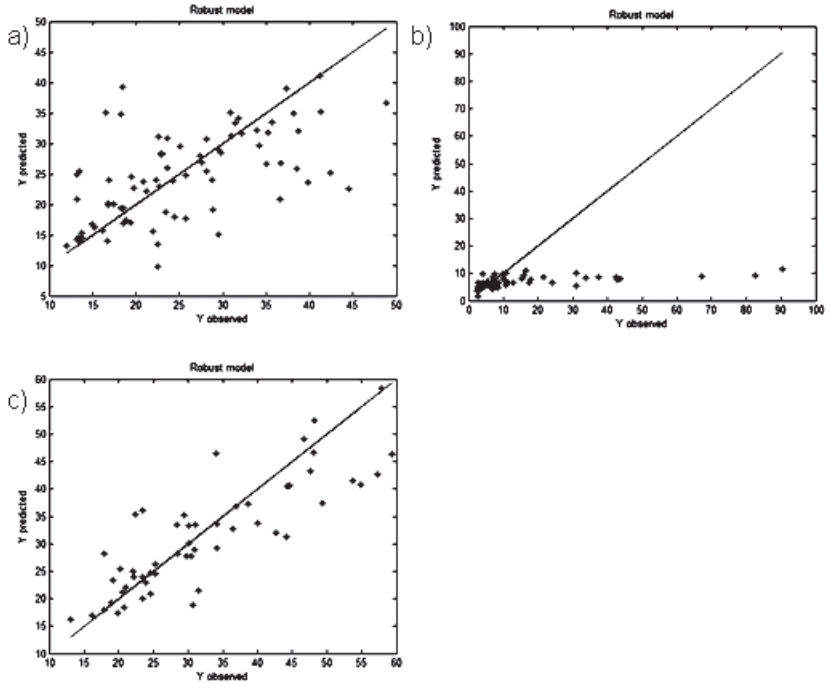

Fig. 2. Robust PLS models of concentration of (a) SO_2 (variable no.5), (b) NO (variable no. 7) and (c) NO_2 (variable no. 8): **y** observed versus **y** predicted on the basis of cross-validation procedure.

observed. Additionally, the 27th of August (sample no. 66) was characterized by the highest value of wind velocity (parameter no. 1). The fourth factor (PC4) reflects the uniqueness of the 29th of June (sample no. 7), whereas PC5 is constructed mainly due to the difference between the 18th of July (sample no. 26) and the 27th of June and the 19th of September (samples nos. 5 and 89). The 29th of June (sample no. 7) is characterized by high solar radiation (parameter no. 4). The 27th of June and the 19th of September (samples nos. 5 and 89) are characterized by relatively high values of solar radiation, wind velocity and SO_2 concentration (parameter nos. 4, 1 and 5), whereas the 18th of July (sample no. 26) is characterized by high concentration of ozone and high humidity (parameters nos. 10 and 3). Loading plots show a negative correlation between temperature and humidity (parameters nos. 2 and 3). Also, a high positive correlation between concentration of NO_2 and CO (parameters nos. 8 and 9), between concentration of SO_2 and PM10 (parameters nos. 5 and 6) and between solar radiation and concentration of ozone (parameters nos. 4 and 10) can be observed.

The conclusions presented above might be inaccurate, taking into account the fact that a reconstructed data matrix suggests the improperness of the model resulting in negative values of missing elements estimates, which in the case of concentration is unacceptable. This fact suggests that the PCA model is strongly influenced by outliers. In order to construct a proper model, it is necessary to correctly identify outlying objects in the data. This may be achieved by the use of a general strategy enabling us to explore contaminated data sets with missing elements proposed by Smoliński et al. [16].

The first step of this strategy is a proper estimation of missing elements, with the purpose of which the three robust PLS models were constructed (see Fig. 2) to predict missing elements of variables 5, 7 and 8 (concentrations of SO_2, NO and NO_2), respectively. In each case all remaining parameters were used to construct these models.

The second step of the above-mentioned strategy is outlier identification based on the robust PCA method (rPCA). It should be kept in mind that the contaminated data set contains measurements performed within different magnitude ranges and that the data were standardized using median and the robust scale [33]. The robust score plots and robust loading plots are presented in Fig. 3.

Outlying objects can be identified using the robust distance [27] (see Fig. 4). Based on the robust distance, which was calculated for five robust scores, eleven unique days (outlying objects) can be identified, namely the 6th – 8th, the 12th and the 21st of August, the 5th, the 6th, the 15th, the 16th, the 18th and the 22nd of September (objects nos. 45-47, 51, 60, 75, 76, 85, 86, 88 and 92).

As the main goal of the data analysis is the formulation of general conclusions about the studied data, all outlying objects, i.e. objects which are different from the data majority, should be identified and eliminated. The uniqueness of these objects should be discussed separately. In an approach presented instead of eliminating entire objects, identified based on robust distances as outliers, only outlying elements of these objects, i.e. parameters nos. 4, 6, 7 and 1 (solar radiation, concentration of PM10 and NO and wind velocity) were treated as missing elements. The last step of the strategy was the construction of the final EM/PCA model for data matrix $\mathbf{X}_2$ containing the initial

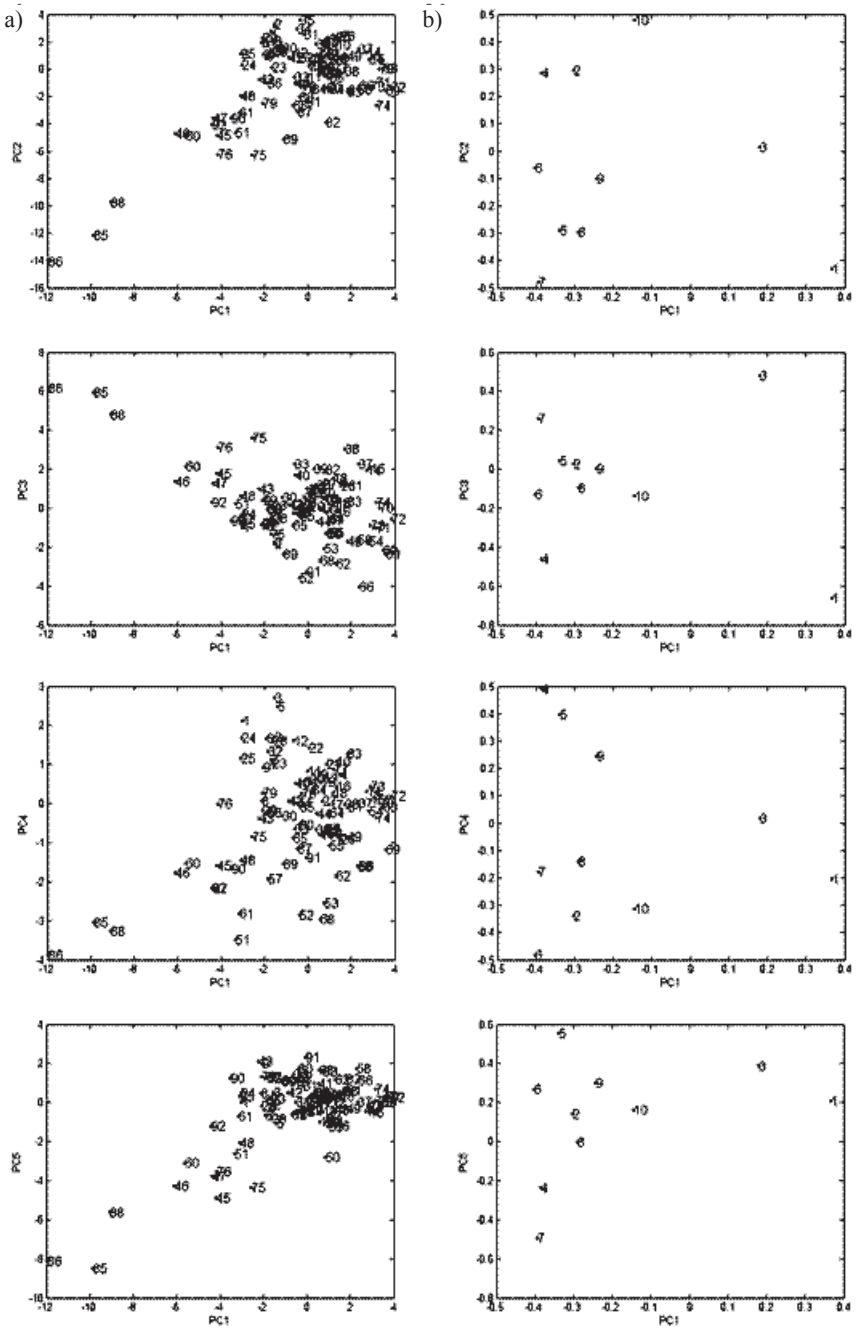

Fig. 3. a) Robust score plots and b) robust loading plots as a result of rPCA for standardized data **X** (92 x 10) with missing elements replaced by values estimated by robust PLS model.

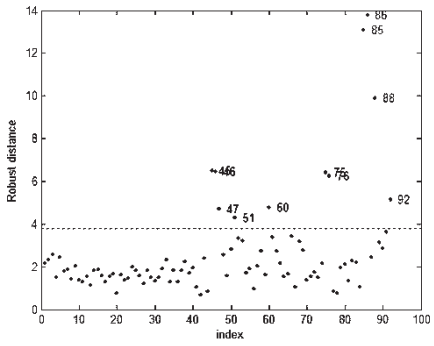

Fig. 4. Robust distances, calculated for five robust scores.

data set with missing elements and all identified outlying elements replaced by missing elements. Five principal components describe 92.62% of total data variance. Score plots and loading plots obtained as a result of EM/PCA of standardized $\mathbf{X}_2$ data set are presented in Fig. 5.

The differences between the results of EM/PCA for $\mathbf{X}_2$ data set (see Fig. 5) and the results of EM/PCA for data $\mathbf{X}_1$ (see Fig. 1) seem to be significant. Based on PC1, it is concluded that the 4th of August, the 16th, the 18th and the 20th of September (objects nos. 43, 86, 88 and 90) differ from the remaining ones due to relatively higher values of NO_2, PM10 and CO concentrations (parameters nos. 8, 6 and 9). Furthermore, the concentration of NO_2 and

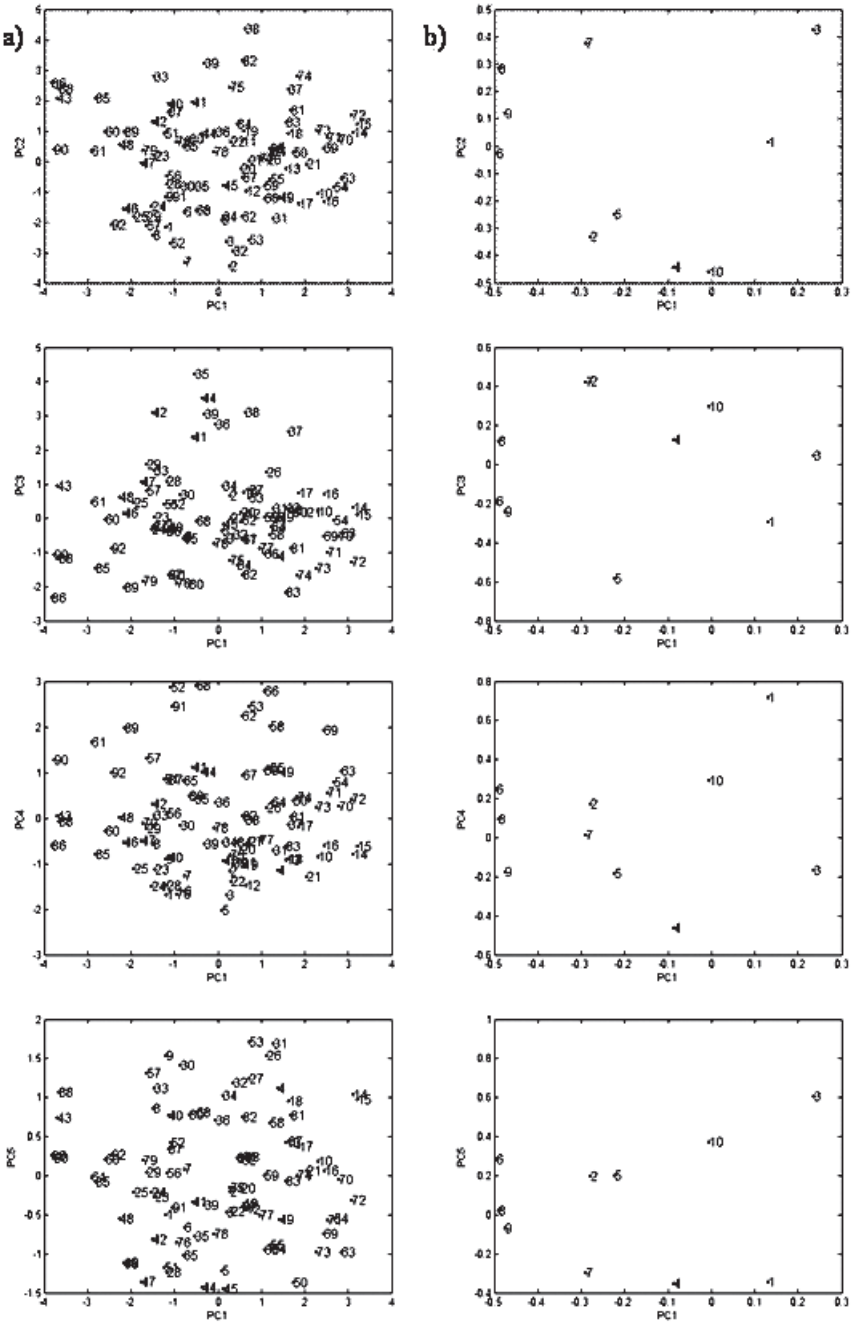

Fig. 5. (a) Score plots and (b) loading plots obtained as a result of EM/PCA of standardized $\mathbf{X}_2$ data set.

CO is the highest during the whole summer. PC2 is constructed mainly because of the difference between the 30th of August (object no. 38) and days: the 24th and the 29th of June (objects nos. 2 and 7), whereas the third principal component (PC3) indicates the uniqueness of the 27th of July (object no. 35). On the 30th of July (object no. 38) the highest humidity during the whole summer (parameter no. 3) is observed. The 24th and the 29th of June (objects nos. 2 and 7) are characterized by a high concentration of ozone and high solar radiation (parameters nos. 10 and 4), whereas the 27th of July (object no. 35) is different from the remaining objects due to high temperature (parameter no. 2). PC4 shows the difference between the 13th, the 27th, the 29th of August (object nos. 52, 66 and 68) and the 27th of June (object no. 5). The days: the 13th, the 27th and the 29th of August (objects nos. 52, 68 and 66) are characterized by high wind velocity (parameter no. 1), whereas the 27th of June (object no. 5) is characterized by relatively high solar radiation (parameter no. 4). PC5 shows the uniqueness of the 5th and the 6th of August (objects nos.53 and 31), mainly due to high concentration of NO (parameter no. 7). Based on loading plots, a positive correlation is identified between temperature and SO_2 concentration (parameters nos. 2 and 5), between solar radiation and ozone concentration (parameters nos. 4 and 10) and high negative correlation between temperature and humidity

(parameters nos. 2 and 3) and between SO_2 concentration and humidity (parameters nos. 5 and 3).

Conclusions

The environmental data are usually very complex and treatment of this type of data requires applying advanced methods of data handling, such as chemometric or environmetric ones. The environmental data sets are often contaminated by outlying elements which occur due to instrument malfunctioning or due to a temporary dramatic change in the environment. As the main goal of data analysis is the generation of general conclusions on the studied data set, these unique objects should be identified and eliminated. Another problem in environmental data analysis is missing elements in the studied data set. The robust statistics and robust methods of data analysis have an advantage over classical ones, especially due to the fact that the data containing outliers can be analyzed, based on the proper estimation for the data majority.

Chemometric methods offer useful tools for solving the problem of missing elements and outliers separately, but they are useless when these problems exist simultaneously in the data set. A strategy enabling the exploration of contaminated data sets with missing elements is widely discussed in this paper. The final EM/PCA model, constructed for the initial data set with correctly identified outliers replaced by missing elements, allows for extracting valuable information about the analyzed phenomenon. It has been shown that the proposed strategy can be successfully applied in environmental studies.

References

1. MCLACHLAN G.J., KRISHNAN T. The EM Algorithm and Extensions.; John Wiley & Sons: New York, **1997**.
2. WALCZAK B. Dealing with missing data. Part 1. Chemom. Intell. Lab. Syst., **58**, 15, **2001**.
3. WALCZAK B. Dealing with missing data. Part 2. Chemom. Intell. Lab. Syst., **58**, 29, **2001**.
4. FISHER R.A. Theory of statistical estimation. Proc. Cambr. Phil. Soc, **22**, 700, **1925**.
5. MUTEKI K., MACGREGOR J.F., Ueda T. Estimation of missing data using latent variable methods with auxiliary information. Chemom. Intell. Lab. Syst., **78**, 41, **2005**.
6. STANIMIROVA I., SIMEONOV V. Modeling of environmental four-way data from air quality control. Chemom. Intell. Lab. Syst., **77**, 115, **2005**.
7. STOICA P., XU L., LI J. A new type of parameter estimation algorithm for missing data problems. Stat. Prob. Letters, **75**, 219, **2005**.
8. WOODWARDA W.A., SAINB S. Testing for outliers from a mixture distribution when some data are missing. Comput. Stat. Data Analysis, **44**, 193, **2003**.
9. RUBIN D.B. Multiple Imputation for Nonresponse in Survey.; John Wiley & Sons: New York, **1987**.
10. HO P., SILVA M.C.M., HOGG T.A. Multiple imputation and maximum likelihood principal component analysis of incomplete multivariate data from a study of the ageing of port. Chemom. Intell. Lab. Syst., **55**, 1, **2001**.
11. HUI D., WAN S., SUA B., KATUL G., MONSONC R., LUO Y. Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. Agricul. and Forest Meteorology, **121**, 93, **2004**.
12. CROUX C., RUIZ-GAZEN A. A fast algorithm for robust principal components based on projection pursuit. (In: A. Prat (Ed.), Compstat: Proceedings in Computational Statists; Heidelberg: Physica-Verlag, pp 211-216, **1996**.
13. HUBERT M., ROUSSEEUW P.J., VERBOVEN S. A fast method for robust principal components with applications to chemometrics. Chemom. Intell. Lab. Syst., **60**, 101, **2002**.
14. STANIMIROVA I., WALCZAK B., MASSART D.L., SIMEONOV V. A comparison between two robust PCA algorithms. Chemom. Intell. Lab. Syst., **71**, 83, **2004**.
15. VANLANDUIT S., CAUBERGHE B., GUILLAUME P., VERBOVEN P., PARLOO E. Reduction of large frequency response function data sets using a robust singular value decomposition. Comp. & Struct., **84**, 808, **2006**.
16. SMOLIŃSKI A., WALCZAK B., EINAX, J.W. Exploratory analysis of data sets with missing elements and outliers. Chemosph., **49**, 233, **2002**.
17. JOLIFFE I.T. Principal Components Analysis.; Springer: New York, **1986**.
18. Wold S. Principal Components Analysis. Chemom. Intell. Lab. Syst., **2**, 37, **1987**.
19. VANDEGINSTE B.G.M., MASSART D.L., BUYDENS L.M.C., DEJONG S., LEWI P.J., SMEYERS-VERBEKE J. Handbook of Chemometrics and Qualimetrics: Part B.; Elsevier: Amsterdam, pp 87-150, **1998**.
20. SINGH C.V., Pattern characteristics of Indian monsoon rainfall using principal component analysis (PCA). Atmosph. Research, **79**, 317, **2006**.
21. KANYA Z., FORGACS E., CSERHATI T., ILLES Z. Reducing Dimensionality in Principal Component Analysis – A Method Comparison. Chromatographia, **63**, 129, **2006**.
22. WALCZAK B. Outlier detection in bilinear calibration. Chemom. Intell. Lab. Syst., **29**, 63, **1995**.
23. WALCZAK B. Outlier detection in multivariate calibration. Chemom. Intell. Lab. Syst., **28**, 259, **1995**.
24. DASZYKOWSKI M., STANIMIROVA I.,, WALCZAK B., DAEYAERT F., DEJONGE M.R., HEERES J., KOYMANSC L.M.H., LEWI P.J., VINKERS H.M., JANSSEN P.A., MASSART D.L. Improving QSAR models for the biological activity of HIV Reverse Transcriptase inhibitors: Aspects of outlier detection and uninformative variable elimination. Talanta, **68**, 54, **2005**.
25. SERNEELSA S., FILZMOSERB P., CROUXC C., VANESPEN P.J., Robust continuum regression. Chemom. Intell. Lab. Syst., **76**, 197, **2005**.

26. VERBOVENA S., HUBERT M. LIBRA: a MATLAB library for robust analysis. Chemom. Intell. Lab. Syst., **75**, 127, **2005**.
27. ROUSSEUW P.J., VANZOMEREN B.C. Unmasking Multivariate Outliers and Leverage Points. J. Amer. Stat. Assoc., **85**, 633, **1990**.
28. FILZMOSERA P., GARRETT R.G., REIMANN C., Multivariate outlier detection in exploration geochemistry. Comp. & Geosciences, **31**, 579, **2005**.
29. CHIANG L.H., PELL R.J., SEASHOLTZ M.B., Exploring process data with the use of robust outlier detection algorithms. J. Process Contr., **13**, 437, **2003**.
30. PIERNA J., JINA L., DASZYKOWSKI M., WAHL F., MASSART D.L. A methodology to detect outliers/inliers in prediction with PLS. Chemom. Intell. Lab. Syst., **68**, 17, **2003**.
31. PIERNA J.A.F., WAHL F., DENOORD O., MASSART D.L. Methods for outlier detection in prediction. Chemom. Intell. Lab. Syst., **63**, 27, **2002**.
32. HUBERT M., ROUSSEEUW P.J., VERBOVEN S. A fast method for robust principal components with applications to chemometrics. Chemom. Intell. Lab. Syst., **60**, 101, **2002**.
33. ROUSSEEUW P.J., CROUX C. Alternatives to the Median Absolute Deviation. J. Amer. Stat. Assoc, **88**, 1273, **1992**.
34. MARTENS H., NAES T. Multivariate Calibration.; John Wiley & Sons: New York, **1989**.
35. WOLD S., MARTENS H., WOLD H. The Multivariate Calibration Problem in Chemistry Solved by the PLS Method, Lecture Notes in Mathematics.; Springer-Verlag: Heidelberg, **1983**.
36. DINC E., USTUNDAG O., Application of Multivariate Calibration Techniques to HPLC Data for Quantitative Analysis of a Binary Mixture of Hydrochlorathiazide and Losartan in Tablets. Chromatographia, **61**, 237, **2005**.
37. HUANG J., BRENNAN D., SATTLER L., ALDERMAN J., LANE B., O'MATHUNA C. A comparison of calibration methods based on calibration data size and robustness. Chemom. Intell. Lab. Syst., **62**, 25, **2002**.
38. GELADI P., Some recent trends in the calibration literature. Chemom. Intell. Lab. Syst., **60**, 211, **2002**.
39. GOLDBERG D.E. Genetic Algorithms in Search Optimization, and Machine Learning.; Addison-Wesley: New York, **1989**.
40. LUCASIUS C.B., KATEMAN, G. Understanding and using genetic algorithms. Part I. Concepts, properties and context. Chemom. Intell. Lab. Syst., **19**, 1, **1993**.
41. LAVINE B.K., DAVIDSON C.E., MOORES A.J. Innovative genetic algorithms for chemoinformatics. Chemom. Intell. Lab. Syst., **60**, 161, **2002**.
42. LAVINE B.K., DAVIDSON C.E., MOORES A.J. Genetic algorithms for spectral pattern recognition. Vibr. Spectr., **28**, 83, **2002**.
43. USTUN B., MELSSEN W.J., OUDENHUIJZEN M., BUYDENS L.M.C. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. Anal. Chim. Acta, **544**, 292, **2005**.
44. MICHALEWICZ Z. Genetic Algorithms + Data Structures = Evolution Programs.; Springer-Verlag: New York, **1992**.
45. JUN Y., XIANDE L., LU H. Evolutionary game algorithm for continuous parameter optimization. Inf. Process. Lett., **91**, 211, **2004**.
46. WOLD S. Cross-validatory estimation of the number of components in factor and principal components models. Technometrics, **20**, 397, **1978**.