Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 73 (2004) 219–233

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

STATIS, a three-way method for data analysis. Application to environmental data

I. Stanimirova[a], B. Walczak[a,b], D.L. Massart[a,*], V. Simeonov[c], C.A. Saby[d], E. Di Crescenzo[d]

[a] ChemoAC, Pharmaceutican Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium
[b] on leave from Silesian University, 9 Szkolna Street, 40-006 Katowice, Poland
[c] Faculty of Chemistry, University of Sofia 'St. Kliment Okhridski', J. Bourchier Blvd.1, 1126 Sofia, Bulgaria
[d] Total (CReS), Chemin du Canal, B.P.22, 69360 Solaize, France

Received 3 October 2003; received in revised form 16 March 2004; accepted 17 March 2004
Available online 1 June 2004

Abstract

The present paper deals with the data exploration of three-way environmental data with the use of "Structuration des Tableaux A Trois Indices de la Statistique" (STATIS). The performance of the method is compared with Tucker3 and PARAFAC2, two more commonly used methods in chemometric *N*-way data analysis. The features of STATIS are demonstrated on real data sets. Due to its robust properties, lack of special requirements for data preprocessing and ability to deal with sets of two-way tables (matrices) that do not have the same dimension for columns or rows, STATIS appears as a very attractive three-way exploratory tool.

Keywords: Exploratory analysis; Three-way data; STATIS; PARAFAC2; Tucker3

1. Introduction

Environmental data sets can be multidimensional and have a complex structure. Usually, they are collected as sets (tables) of objects and variables obtained under different experimental circumstances or for various sampling periods, etc. Putting all tables together results in data with three-way structure [1]. An example for such data is when in samples collected at different sampling sites, the concentrations of several chemical components are measured during certain period of time (sites × parameters × time). There are many tools helping to explore and interpret three- or higher way structure of the data. The most popular ones in chemometrics are PARAFAC [2,3], PARAFAC2 [4,5] and Tucker3 [6,7]. A software tool, called CUBATCH, for applying these methods was recently presented [8].

The aim of this paper is to present a method, called STATIS [9–11], which can also be applied for exploratory analysis of three-way data sets, and to compare its performance with *N*-way methods for the analysis of environmental data. The abbreviation STATIS stands for "Structuration des Tableaux A Trois Indices de la Statistique", which could be translated in English as "structuring three-way data sets in statistics".

2. Theory

2.1. STATIS

STATIS is an exploratory tool for three-way data analysis. Its main idea is to compare different data tables (matrices) obtained under various experimental conditions, but containing the same number of rows and/or columns [12]. By analogy to *N*-way methods, the three-way data set is denoted by $\underline{\mathbf{X}}$ with dimensions *I*, *J* and *K*, corresponding to the number of rows, columns and tables, respectively [1]. Thus, an element of $\underline{\mathbf{X}}$ is x_{ijk}, where $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

Each direction is called a mode and the number of levels in the mode is called dimension (see Fig. 1a). A table $\mathbf{X}_k$ is a slice of $\underline{\mathbf{X}}$ of dimension $I \times J_k$ (see Fig. 1b) obtained by fixing the index in the third mode. It can also be named a frontal slab or layer. The tables $\mathbf{X}_i$ and $\mathbf{X}_j$ are called horizontal and vertical slabs, and can be obtained by fixing

[*] Corresponding author. Tel.: +32-2-477-4734; fax: +32-2-477-4735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.03.005

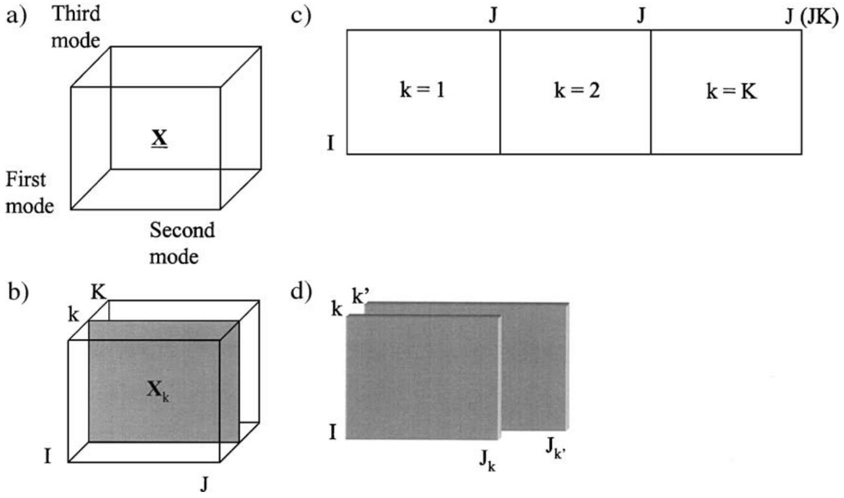


Fig. 1. (a) A three-way data array; (b) definition of k-th frontal slab; (c) unfolding of $\underline{\mathbf{X}}$ ($I \times J \times K$) to $\mathbf{X}$ ($I \times JK$); (d) the case when k-th and k' -th frontal slabs have different column dimension ($J_k \neq J_{k'}$).

the first or the second mode index. This notation, usual for N-way methods, is introduced here in STATIS, in order to make the description of the method comparable to that used in N-way analysis.

To illustrate the main steps of the method [12], let us first consider that different $\mathbf{X}_k$ tables, for example sites × para-parameters ($I \times J_k$) are put next to each other, which results in a two-way matrix of dimension $I \times JK$ or sites × (parameters × time) (see Fig. 1c). Such a matrix could be analyzed by principal component analysis (PCA). This is known as unfolding PCA. PCA performed on the variance–covariance matrix of this composite table gives information about similarity between sites based on parameters measured during different sampling times. The variance–covariance matrix for the unfolded table is obtained by summing the individual, $\mathbf{W}_k$, variance–covariance matrices:

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{W}_k \tag{1}$$

However, the knowledge how similar the individual data tables are cannot be obtained and atypical (sites × para-parameters) tables then have too large an influence in the final result. In STATIS, the variance of each table is weighted according to the similarities among the tables, which gives a three-way character to the method and makes results more robust and more interpretable.

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{a}_k \mathbf{W}_k, \tag{2}$$

where $\mathbf{a}_k$ is a vector of weights. As a consequence, if the weights are equal to 1, results from unfolded PCA and STATIS will be the same. In fact, STATIS can be considered as an unfolded PCA with a special weighting of the variance of individual tables. This weighting confers its three-way character to the method.

How the weighting is performed will be shown considering the aforementioned example. First, the tables (sites × parameters) for different sampling time are compared. For each data table, the variance–covariance matrix, $\mathbf{W}_k$ ($I \times I$), reflecting the similarities between I objects within this data table, is computed:

$$\mathbf{W}_k = \mathbf{X}_k \mathbf{Q}_k \mathbf{X}_k^T, \tag{3}$$

where $\mathbf{X}_k$ is a matrix of dimension $I \times J_k$ and $\mathbf{X}_k^T$ is its transpose matrix. When all $\mathbf{X}_k$ matrices of $\underline{\mathbf{X}}$ have the same number of variables, $\mathbf{Q}_k$ is usually the identity matrix of dimension $J \times J$, whereas if there are differences in number of variables in individual $\mathbf{X}_k$ matrices, they can be, if necessary, compensated by using $\mathbf{Q}_k$ ($J_k \times J_k$) with the diagonal elements equal to $1/J_k$.

The similarities between two variance–covariance matrices $\mathbf{W}_k$ and $\mathbf{W}_{k'}$ for data tables k and k' , computed according to Eq. (3), can be calculated as follows:

$$\langle \mathbf{W}_k, \mathbf{W}_{k'} \rangle = \mathrm{trace}(\mathbf{W}_k \mathbf{D} \mathbf{W}_{k'} \mathbf{D}), \tag{4}$$

where $\mathbf{D}$ is a matrix of dimension $I \times I$, the diagonal elements of which are equal to $1/I$.

The most commonly used measure of closeness between the variance–covariance matrices is the so-called RV coefficient, which has been introduced by Robert and Escoufier [13]. For the k-th and the k' -th data tables, it is defined in the following way:

$$\mathrm{RV}(\mathbf{W}_k, \mathbf{W}_{k'}) = \frac{\langle \mathbf{W}_k, \mathbf{W}_{k'} \rangle}{\sqrt{\langle \mathbf{W}_k, \mathbf{W}_k \rangle \langle \mathbf{W}_{k'}, \mathbf{W}_{k'} \rangle}} \tag{5}$$

Since $\langle \mathbf{W}_k, \mathbf{W}_k \rangle$ and $\langle \mathbf{W}_{k'}, \mathbf{W}_k' \rangle$ are equal to the sum of the squared diagonal elements of $\mathbf{W}_k$ and $\mathbf{W}_{k'}$, respectively, they are the norm of the k-th and k' -th variance–covariance matrices.

The RV coefficients are non-negative and scaled between 0 and 1, and organized into a square matrix ($K \times K$). The closer RV (k, k') is to 1, the more similar the two variance–covariance matrices k and k' are. The similarities among tables can be visualized in the space of principal components, after performing PCA on the RV matrix, which is called interstructure analysis. The first eigenvector obtained after PCA of the non-centered RV matrix is a global size variable representing the "agreement between tables". Its elements are normalized in such a way that their sum is equal to 1 and used as weights ($\mathbf{a}_k$) in order to define what in STATIS is called a "compromise" among K tables as a weighted sum of the variance–covariance matrices $\mathbf{W}_k$, yielding Eq. (2).

As it was mentioned above, the STATIS compromise, $\mathbf{W}$, differs from the usual way of defining the compromise as a mean of the covariance matrices as in two-way PCA of unfolded matrix. Including weights proportional to the agreement between tables makes the STATIS compromise more robust. The weight of the outlier will be closer to zero with respect to the other weights. PCA of the compromise matrix $\mathbf{W}$ ($I \times I$) gives information about the similarity of objects in the first mode. Their distribution can be visualized in the space spanned by the principal components and the representation is called compromise plot. Additionally, it is possible to project individual covariance matrices, $\mathbf{W}_k$, on the compromise plot. Such a plot displays the location of each object on the compromise plot as a weighted center of K individual locations of this object. From the scores of the compromise matrix, $\mathbf{W}$, the loadings of individual variance–covariance matrices, $\mathbf{W}_k$, can be obtained and this is important for visualizing the "hidden" modes. The coordinates (loadings) of the objects from the k-th table for f principal components on the compromise plot are given by the following equation:

$$\mathbf{C}_k = \mathbf{W}_k\mathbf{L}\mathbf{E}, \tag{6}$$

where $\mathbf{E}$ is a diagonal matrix ($f \times f$), the diagonal elements of which are the inverse of the square roots of the compromise eigenvalues. Matrix $\mathbf{L}$ contains the scores of the PCA of the compromise and has dimension $I \times f$.

STATIS offers another possibility, which makes this approach quite attractive for exploratory analysis. Because the variance–covariance matrices are calculated as the first step, in the case, when $\mathbf{X}_k$ and $\mathbf{X}_{k'}$ have the same dimension in the first mode ($I_k = I_{k'}$), but they differ in the second mode dimension ($J_k \neq J_{k'}$) (see Fig. 1d), the compromise for I objects can be obtained using the same algorithm. For example, when the concentrations of chemical components (parameters) are measured during different sampling periods at each sampling site (parameters × time × sites), the compromise for the parameters can be obtained.

If all tables have the same dimension for columns and rows, i.e. when the structure of the data is perfect, then three different data arrangements are possible, leading to three strategies of data analysis with STATIS. In this way, a compromise can be obtained for each of the modes by a different rearrangement of $\underline{\mathbf{X}}$. When, all data tables do not have the same dimension for rows or columns, i.e. the structure of the data is imperfect; the compromise is obtained on the mode, for which all tables have the same dimension. For the aforementioned example, it means that compromise can be obtained on the parameters and the sampling sites.

The algorithm of STATIS can be summarized as follows:

1. Calculate the K variance–covariance matrices as

$$\mathbf{W}_k = \mathbf{X}_k\mathbf{Q}_k\mathbf{X}_k^T$$

2. Calculate the matrix of RV coefficients ($K \times K$)

$$\mathrm{RV}(\mathbf{W}_k, \mathbf{W}_{k'}) = \frac{\langle \mathbf{W}_k, \mathbf{W}_{k'} \rangle}{\sqrt{\langle \mathbf{W}_k, \mathbf{W}_k \rangle \langle \mathbf{W}_{k'}, \mathbf{W}_{k'} \rangle}}$$

3. Perform PCA of the RV matrix [**S**,**V**,**S**] = SVD(**RV**), SVD is the singular value decomposition version of PCA [14]

$$\mathbf{a}_k = \mathbf{s}_1 / \sum_{k=1}^{K} \boldsymbol{S}_{\boldsymbol{1k}},$$

where $\mathbf{s}_1$ ($K \times 1$) is the first column vector of the matrix $\mathbf{S}$

4. Calculate the compromise among tables as:

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{a}_k\mathbf{W}_k$$

5. Perform PCA of $\mathbf{W}$

$$[\mathbf{L}, \mathbf{V}, \mathbf{L}] = \mathrm{SVD}(\mathbf{W})$$

6. Display the compromise score plot

2.2. Tucker3

Tucker3 is a method for data decomposition [6,7,15], considered as a generalization of two-way principal component analysis to N-way arrays. The original N-way data array, $\underline{\mathbf{X}}$, of dimension $I \times J \times K$ is decomposed into three matrices $\mathbf{A}(I \times S)$, $\mathbf{B}(J \times M)$ and $\mathbf{C}(K \times N)$, the elements of which are called loadings, where S, M and N are the number of factors extracted on the first, second and third mode, respectively. The interactions between different modes are explained by the core array $\underline{\mathbf{G}}(S \times M \times N)$, arranged as $S \times MN$. The unfolded original data to $\mathbf{X}(I \times JK)$ can be reconstructed by Tucker3 model as:

$$\mathbf{X} = \mathbf{A}\mathbf{G}(\mathbf{B} \otimes \mathbf{C})^T, \tag{7}$$

where $\otimes$ is the Kronecker product. For example, for two matrices $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}$ is of dimension $I \times J$, it is defined as:

$$\mathbf{X} \otimes \mathbf{Y} = \begin{bmatrix} x_{11}\mathbf{Y} & \cdots & x_{1J}\mathbf{Y} \\ \vdots & & \vdots \\ x_{I1}\mathbf{Y} & \cdots & x_{IJ}\mathbf{Y} \end{bmatrix}$$

The data decomposition is done by means of alternating least square algorithms [15,16]. The main steps of Tucker3 algorithm are described in Ref. [15].

Tucker3 allows easy visualization of the distribution of the objects in the factor space in all modes, but the interpretation is quite complicated. The difficulty comes from the fact that one has to take into account the elements of Tucker3 core array, $\underline{\mathbf{G}}$, which gives information about important interactions between modes.

2.3. PARAFAC2

In some cases, the slabs (tables) constituting $\underline{\mathbf{X}}$ do not have the same numbers of rows or columns. The N-way method, which can deal with that problem, is PARAFAC2. The objective of the method is to model new $\underline{\mathbf{Y}}$ data containing the covariance matrices of the set of two-way $\mathbf{X}_k$ matrices of $\underline{\mathbf{X}}$. If $\mathbf{X}_k$ matrices of $\underline{\mathbf{X}}$ are arranged as frontal slabs and have different columns dimension, the new $\underline{\mathbf{Y}}$ has dimension $I \times I \times K$. After unfolding of $\underline{\mathbf{Y}}$ as $\mathbf{Y}^{(K \times II)}$, the PARAFAC2 model can be written as follows:

$$\mathbf{Y}^{(K \times II)} = (\mathbf{C}|\otimes|\mathbf{C}^T)^T diag(\text{vec}\mathbf{H})(\mathbf{A} \otimes \mathbf{A})^T, \qquad (8)$$

where $|\otimes|$ is the Khatry-Rao product [15]. Matrix $\mathbf{A}(I \times F)$ contains the first mode loadings and matrix $\mathbf{C}(K \times F)$ holds the third mode loadings, respectively, where F is the number of factors extracted. $\mathbf{H}$ is the cross-product matrix of $\mathbf{B}$ ($\mathbf{H} = \mathbf{B}^T\mathbf{B}$), where $\mathbf{B}$ holds the second mode loadings.

3. Data preprocessing

Data preprocessing takes an important place in data analysis [17]. Several types of data pretreatment are known, the most usual ones in two-way data analysis are centering and scaling. Centering removes the differences in the size of rows and/or columns. In row centering the corresponding row mean is subtracted from each element of the data matrix. Column centering is done by subtracting from each data element the corresponding column mean. Centering can be done sequentially on rows and columns, which is known as double centering. The order of centering does not affect the final result. Scaling can also be done on rows and columns of the data matrix. Centering and scaling can be combined. This is the case for autoscaling, which is applied among the others when the variables are in different units. Its aim is to give the variables the same importance, by making the standard deviation of each variable equal to 1. These preprocessing methods can be extended to N-way arrays. The preprocessing of N-way data requires more caution, and several rules on how to do this can be found in the literature [17]. Single data centering is done across one mode and can be followed, if necessary, by sequential centering in other modes. Scaling in one mode will not change the data structure, whereas scaling to unit standard deviation in two modes is not possible. Combinations of centering and scaling for a given mode can be performed in the same way as for the two-way data. They are not problematic, when scaling within one mode is combined with centering across other modes [17]. Several scalings can be performed sequentially, in both preprocessing of two-way and preprocessing of N-way data, but will generally need iterations and may not converge. If centering is applied when this is not necessary, i.e. when there is no offset to be corrected [17], it can contribute additional artificial variation, which will destroy the data structure and will lead to spurious results obtained by both two-way and N-way methods.

In STATIS, depending on the problem at hand, preprocessing can be done for each table separately or on unfolded data; similar to the way this would be done for the unfolded two-way data.

4. Data

The data set consists of the annual mean concentrations of nine chemical components (H^+, NH_4^+, Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, NO_3^- and SO_4^{2-}) monitored during 12 years at six sampling sites (Reutte, Kufstein, Innervillgraten, Sonnblick, Nasswald and Lobau), 15 years at Haunsberg and Werfenweng, 10 years at Litschau and Lunz, 9 years at the Nassfeld site [18]. The data do not have perfect trilinear structure, since for one mode (years) the dimensions of the data are not the same for all data tables. In order to obtain also data with perfect trilinear structure, the time dimension of the original data was set to be 8 years for each sampling site, but as we will show later, it is also possible to work with data with an imperfect trilinear structure and to use all the data.

5. Results and discussion

First, STATIS is performed on the non-preprocessed data set with perfect trilinear structure. Each data table $\mathbf{X}_k$ ($I \times J_k$) contains different sampling sites ($I = 11$), characterized by chemical components ($J = 9$) measured in a certain year k. The K tables are arranged as frontal slabs in $\underline{\mathbf{X}}$. PCA of the RV matrix reflecting the similarity between tables is presented in Fig. 2a.

Tables 1 (year 1990), 2 (year 1991), 3 (year 1992) and 6 (year 1995) are the most different from the mean covariance

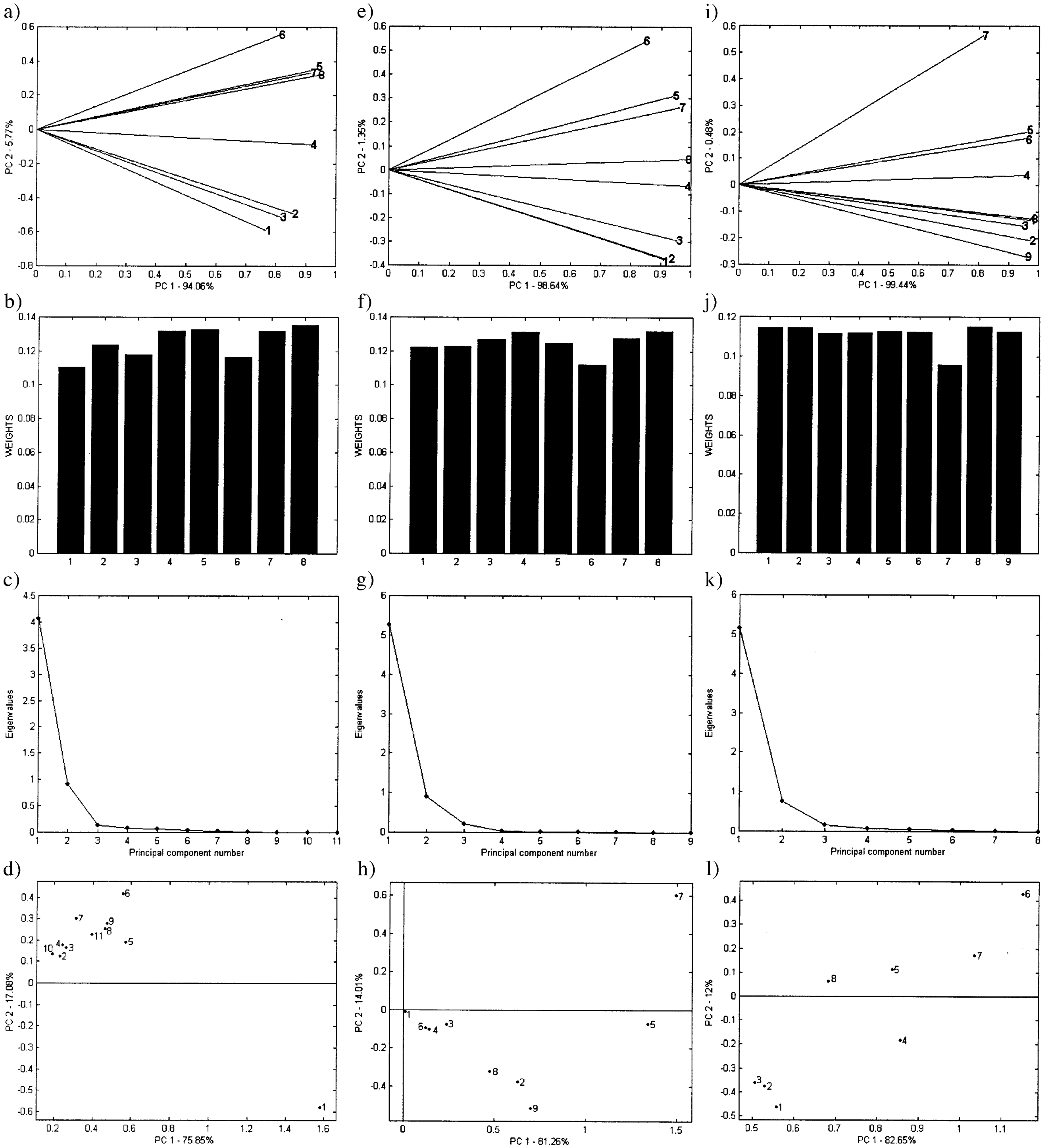


Fig. 2. Results of STATIS for the non-preprocessed data set with perfect trilinear structure: (a) PCA of the RV matrix: PC1–PC2 plot for 8 years; (b) bar plot of 8 weights for 11 sites; (c) eigenvalues scree plot of PCA of the site compromise matrix; (d) PC1–PC2 plot of the compromise of 11 sites; (e) PCA of the RV matrix: PC1–PC2 plot for 8 years; (f) bar plot of 8 weights for 9 variables; (g) eigenvalues scree plot of PCA of the variable compromise matrix; (h) PC1–PC2 plot of the compromise of 9 variables; (i) PCA of the RV matrix: PC1–PC2 plot for 9 variables; (j) bar plot of 9 weights for 8 years; (k) eigenvalues scree plot of PCA of the years compromise matrix; (l) PC1–PC2 plot of the compromise of 8 years.

(see Fig. 2a) and the weights for these tables are therefore smaller with comparison to the others (see Fig. 2b).

The weights define the compromise for the different sites. Two principal components, after PCA of the compromise, explain 92.9% of the variance (see Fig. 2c). All sites, except Haunsberg (1), form a single group in the left upper corner in the space spanned by the first two PCs (see Fig. 2d). These are Innervillgraten (2), Reute (3), Kufstein (4), Litschau (5), Lobau (6), Lunz (7), Nassfeld (8), Nasswald (9), Sonnblick (10) and Werfenweng (11), which show similar ion concentration patterns for all chemical components measured during the period of sampling. The Haunsberg site is different from the compact group.

The same compromise pattern reflecting the similarities between sites can be obtained when the tables are arranged as vertical slabs. In this case, the tables are compared with respect to the measured properties. This means that each slab (table) contains information about the ion concentrations of one chemical component monitored at 11 sampling sites during 8 years and $\underline{\mathbf{X}}$ is of dimension 11 × 8 × 9.

The compromise of variables is obtained after reorganizing $\underline{\mathbf{X}}$ in such a way that variables are considered the first mode ($I=9$), sampling sites the second ($J=11$) and years of sampling the third ($K=8$). Fig. 2e presents the similarities among tables. The sixth table (year 1995) agrees the least with the compromise, which is reflected by its smaller weight (see Fig. 2f). Two principal components explain 95.3% of the variance, after PCA of the compromise (see Fig. 2g). There are two groups of variables determining the sample composition at different sites (see Fig. 2h). The first contains H^+, Na^+, K^+, Mg^{2+} ion concentrations (marked as 1, 3, 4 and 6, respectively) and the second comprises NH_4^+, NO_3^- and SO_4^{2-} (marked as 2, 8 and 9). The samples content mainly differs with respect to the Ca^{2+} (5) and Cl^- (7) ion concentrations.

If $\underline{\mathbf{X}}$ is reshaped as $I=8$, $J=11$ and $K=9$ or $I=8$, $J=9$ and $K=11$, then a compromise plot reflecting the similarity between the years can be obtained. Results for the first combination, where tables are compared according to the variables (see Fig. 2i) are presented. The compromise is defined by the weights with respect to the variables (see Fig. 2j). Two PCs, of the PCA of the compromise matrix, explain 94.7% of variance (see Fig. 2k). A group of objects nos. 1 (1990), 2 (1991) and 3 (1992) can be found in the left lower corner on the compromise plot. The others 4 (1993), 5 (1994), 6 (1995), 7 (1996) and 8 (1997) differ from each other and from the compact group (see Fig. 2l).

Because the weights are not very different (due to not very large differences between covariance matrices), the STATIS results are comparable to those obtained by two-way PCA of unfolded data.

Before interpreting these results, we will first compare the performance of STATIS with N-way methods and the effect of pretreatment of the data. Tucker3 was applied to the same non-preprocessed data with perfect trilinear structure. The Tucker3 method was chosen, because similarly to PCA, in Tucker3 the resulting loading matrices are orthogonal. The results are presented in Fig. 3.

The variance explained for each combination of model complexity, starting from [1 1 1] to [5 5 5], is calculated (see Fig. 3a). The decomposition model with two factors in each mode, [2 2 2], explains 91.6% of the total data variance. Further increase of the model complexity does not change much the explained data variance. The pattern observed for the three modes, sites, variables and time resembles the pattern of those modes obtained by the STATIS method (see Figs. 2d,h,l and 3b,c,d). When the results of Tucker3 have to be interpreted further, the core array $\underline{\mathbf{G}}$ (see Fig. 3e), the elements of which reflect the interactions between the modes, is taken into account.

The results were not fully interpreted for the non-preprocessed data, because the measured properties (variables) have different units, which leads to a large difference in the variables range. Only the comparison of the performance of both methods is presented. Scaling, to the unit standard deviation within the mode containing the variables, gives them the same importance.

The results of STATIS for preprocessed data are shown in Fig. 4. The patterns on the compromise plots differ from those observed for non-preprocessed data (see Figs. 2 and 4). The compromise plot of the sites is shown for two principal components explaining 87.9% of the variance (see Fig. 4a). The same number of principal components is selected to visualize the distributions of the variables and years on the compromise plots, and the variance explained is 89.7% and 93.2%, respectively (see Fig. 4b and c). Three groups of sites can be distinguished along PC1 on the site compromise plot (see Fig. 4d). The first group contains the Innervillgraten (2), Reutte (3) and Sonnblick (10) sites; the second one can be split into two subgroups, Kufstein (4), Werfenweng (11) and Lunz (7), Nassfeld (8) and Nasswald (9); and the third group contains Haunsberg (1), Litschau (5) and Lobau (6). The second PC separates the Haunsberg (1) samples from the rest. On the variable compromise (see Fig. 4e), PC1 is a factor reflecting the NH_4^+ (2), NO_3^- (8) and SO_4^{2-} (9) ion concentrations of the samples. The second latent factor contrasts the H^+ (1) ion concentration with the K^+ (4), Ca^{2+} (5), Mg^{2+} (6) and Cl^- (7) ion concentrations. The compromise plot for years reveals a diffuse structure. A segregation of 1990 (1)–1993 (4) and 1995 (6)–1996 (7) periods can be observed along PC2. Putting all the information together, it can be concluded that the samples can be ranked according to their NH_4^+ (2), NO_3^- (8) and SO_4^{2-} (9) content for the whole sampling period. The Innervillgraten (2), Reutte (3) and Sonnblick (10) samples have low, the Kufstein (4), Lunz (7), Nassfeld (8), Nasswald (9) and Werfenweng (11) samples intermediate and Litschau (5), Lobau (6), and Haunsberg (1) high NH_4^+ (2), NO_3^- (8) and SO_4^{2-} (9) content. Samples from Innervillgraten (2), Reutte (3) and Sonnblick (10) have also a somewhat lower acidity. The Haunsberg (1) samples can be distinguished from the others by their high Ca^{2+} (5) and Cl^- (7) content in 1995 and 1996.

It is not necessary to obtain the compromise for the three modes to interpret the results. Another way of representation can be used [10]. The location of each point on the compromise plot is a weighted center of K individual locations of this point. This can be visualized by drawing convex hulls through individual locations projected on the compromise plot. The smaller dispersion of the individual

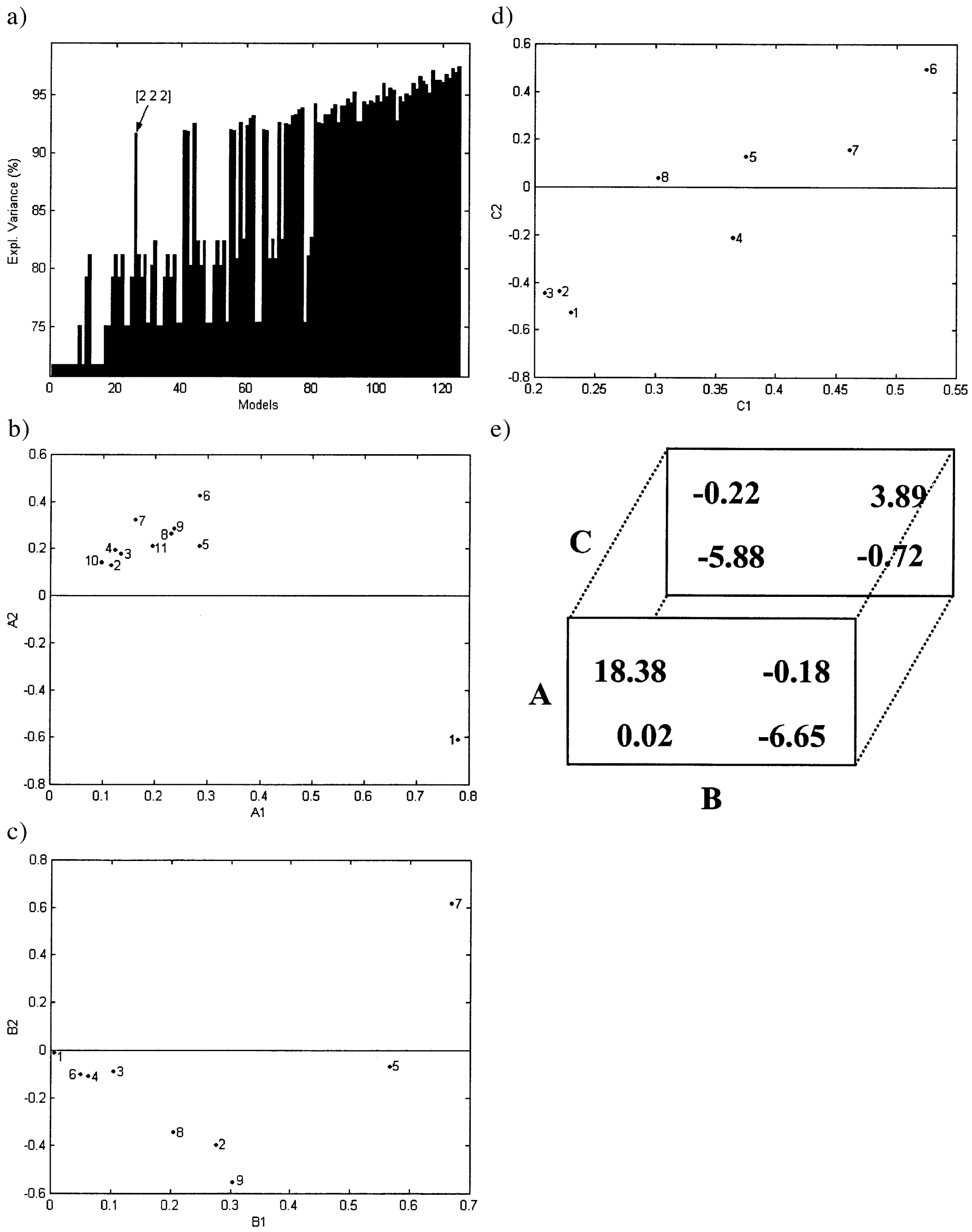
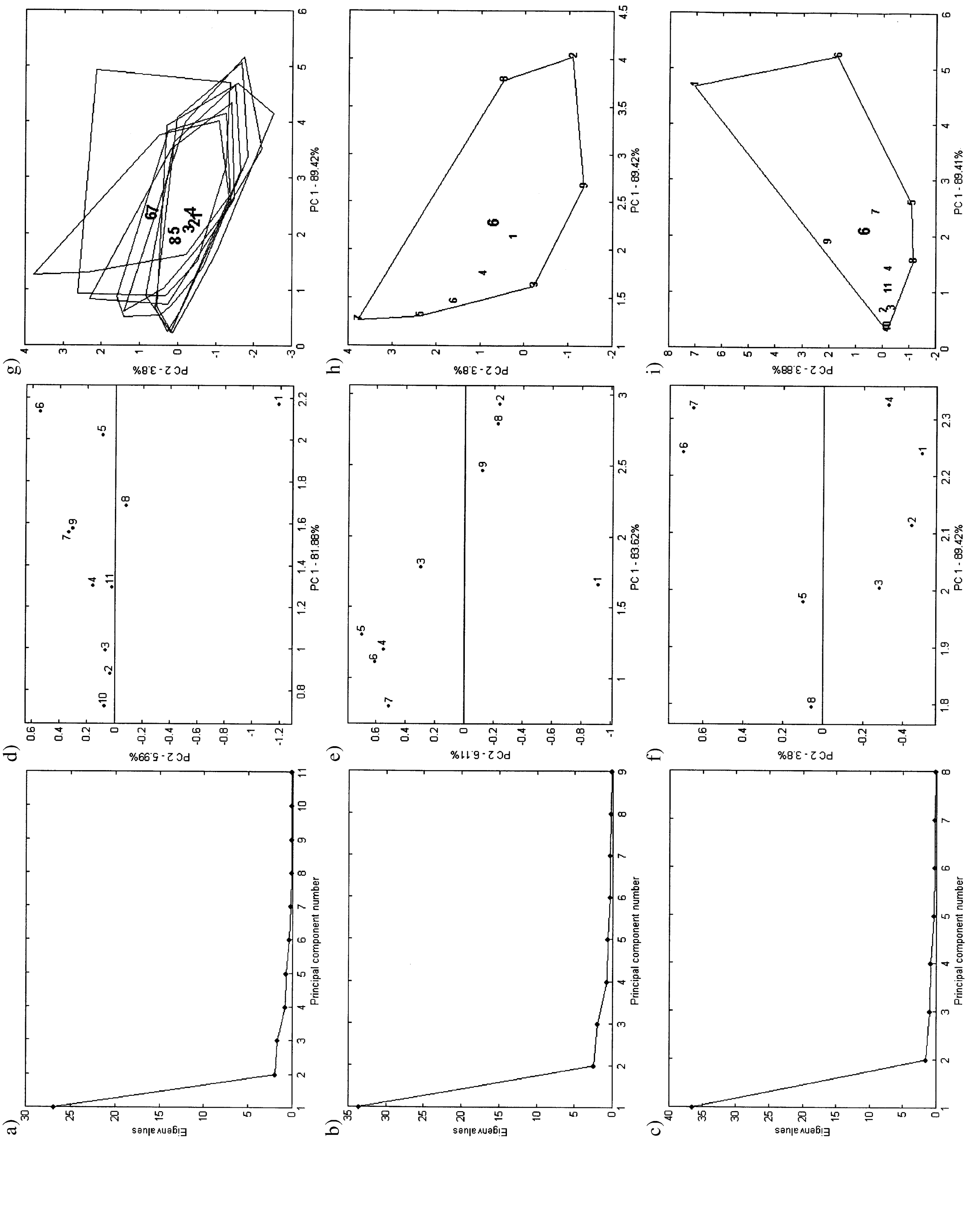


Fig. 3. Results of Tucker3 for the non-preprocessed data: (a) explained variance by models of different complexities; (b) projection of sites on the plane defined by A1 and A2; (c) projection of variables on the plane defined by B1 and B2; (d) projection of years on the plane defined by C1 and C2; (e) core array $\underline{\mathbf{G}}$(2 2 2).

locations around its compromise the better agreement among tables in the compromise. If the convex hulls do not overlap, it can be concluded that the compromise points are different. For example, almost all convex hulls formed by the individual locations (different chemical components, 1–9) on the year compromise plot (see Fig. 4g) overlap, except the convex hulls for 1995 (6) and 1996 (7) years. Drawing all eight (for each year) convex hulls makes the plot unreadable in details. For this reason, to illustrate how such plot can be interpreted the convex hull for the 1995 year (6) is drawn separately in Fig. 4h. In this way, the information about the concentration profile for this year can be obtained. To be able to observe the relation with the sampling sites for the same year, the convex hull passing through the different sampling sites is constructed. This convex hull is presented in Fig. 4i.

Along PC1, sites nos. 2, 3, 4, 5, 7, 8, 9, 10 and 11 are situated near the compromise point (6), whereas sites 1 and 6 are far away in Fig. 4i. Site 1 is also segregated from the others along PC2. In Fig. 4h, parameters 2 and 8 are far away from the compromise point along PC1, whereas parameters 5 and 7 are different from the others along PC2. Combining the information of Fig. 4i and h, it can be concluded that all samples collected during 1995 (year


a)
Eigenvalues
Principal component number
b)
Eigenvalues
Principal component number
c)
Eigenvalues
Principal component number
d)
PC 2 - 5.99%
PC 1 - 81.88%
e)
PC 2 - 6.11%
PC 1 - 83.62%
f)
PC 2 - 3.8%
PC 1 - 89.42%
g)
PC 2 - 3.8%
PC 1 - 89.42%
h)
PC 2 - 3.8%
PC 1 - 89.42%
i)
PC 2 - 3.88%
PC 1 - 89.41%

6), except those from sites Haunsberg (1) and Lobau (6), have low ion concentrations of NH_4^+ (parameter 2) and NO_3^- (parameter 8). The samples from Haunsberg (1) are richer in Ca^{2+} (5) and Cl^- (7).

Results for the same preprocessed data set were obtained by Tucker3. The chosen model complexity is [3 3 2], i.e. three factors in the first and the second modes, and two factors in the third mode. The total variance explained is 89.7% (see Fig. 5a). Because of the rotational freedom of the Tucker3 model, the selected model is the one with a straightforward interpretable core matrix.

The patterns observed for sites (see Figs. 4d and 5b), variables (see Figs. 4e and 5d) and years (see Figs. 4f and 5f) are slightly different from those obtained by STATIS. For example, year 1995 (6) can be segregated from year 1996 (7) along the second factor of the third mode C2 in Fig. 5f, whereas this is not possible along PC2 in Fig. 4f. The reason is that the variance explained in each of the separate STATIS analyses is different from the total variance explained by the complex Tucker3 model.

To interpret the relationships between the elements in different modes the core array $\underline{\mathbf{G}}$ is needed (see Fig. 5g). First, the more important elements from the core array are selected. They are (1,1,1), (3,3,1), (2,2,1) and (3,2,1). Thus, the first three loading vectors in the first two modes and one loading vector in the third mode should be considered in interpretation. How to interpret the results will be demonstrated with an example. The elements along the first factor of the first, second and third mode have sign '+', the core element (1,1,1) has sign '+' and their product is '+'. Three groups of sites can again be observed along the first factor in the first mode A1 (see Fig. 5b). The same groups were observed along PC1 on the STATIS site compromise plot for the standardized data (see Fig. 4d). The first factor of the second mode B1 (see Fig. 5d) reflects mainly NH_4^+ (2), NO_3^- (8) and SO_4^{2-} (9) ion concentration of the samples. A diffuse distribution of the objects can be seen along the first factor in the third mode C1 (see Fig. 5f). It can be concluded that the samples from all sites are ranked according to their increasing content of NH_4^+ (2), NO_3^- (8) and SO_4^{2-} (9) during the whole sampling period. This was also the first conclusion made from the STATIS results. In the same manner the interpretation for the other important elements is made. For all sites, except Haunsberg (1), Lobau (6) and Werfenweng (11) a substantial decrease of Na^+ (3), K^+ (4), Ca^{2+} (5), Mg^{2+} (6), Cl^- (7) concentrations combined with increase of acidity of the samples during 1995 (6) and 1996 (7) is observed. Samples from Haunsberg (1), can be recognized from the others by their high Ca^{2+} and Cl^- ion concentrations during 1995 (6) and 1996 (7) years as well as Lobau (6) samples by their very high proportion of K^+ (4) and Mg^{2+} (6) during 1995 (6).

Another possibility is to perform STATIS on autoscaled data. Autoscaling combines centering and scaling to unit standard deviation (standardization). It removes differences in variables range and gives them the same importance in the data analysis. This type of pretreatment gives results that are more interpretable from a chemical point of view, but because we want to compare with Tucker3 and PARAFAC2, where autoscaling is giving, for our data sets, much more complex models, which are difficult to interpret, we insist here more on comparing results obtained on raw data or data after standardization. The results of autoscaled data obtained by STATIS for the first three PCs in each mode are presented in Fig. 6.

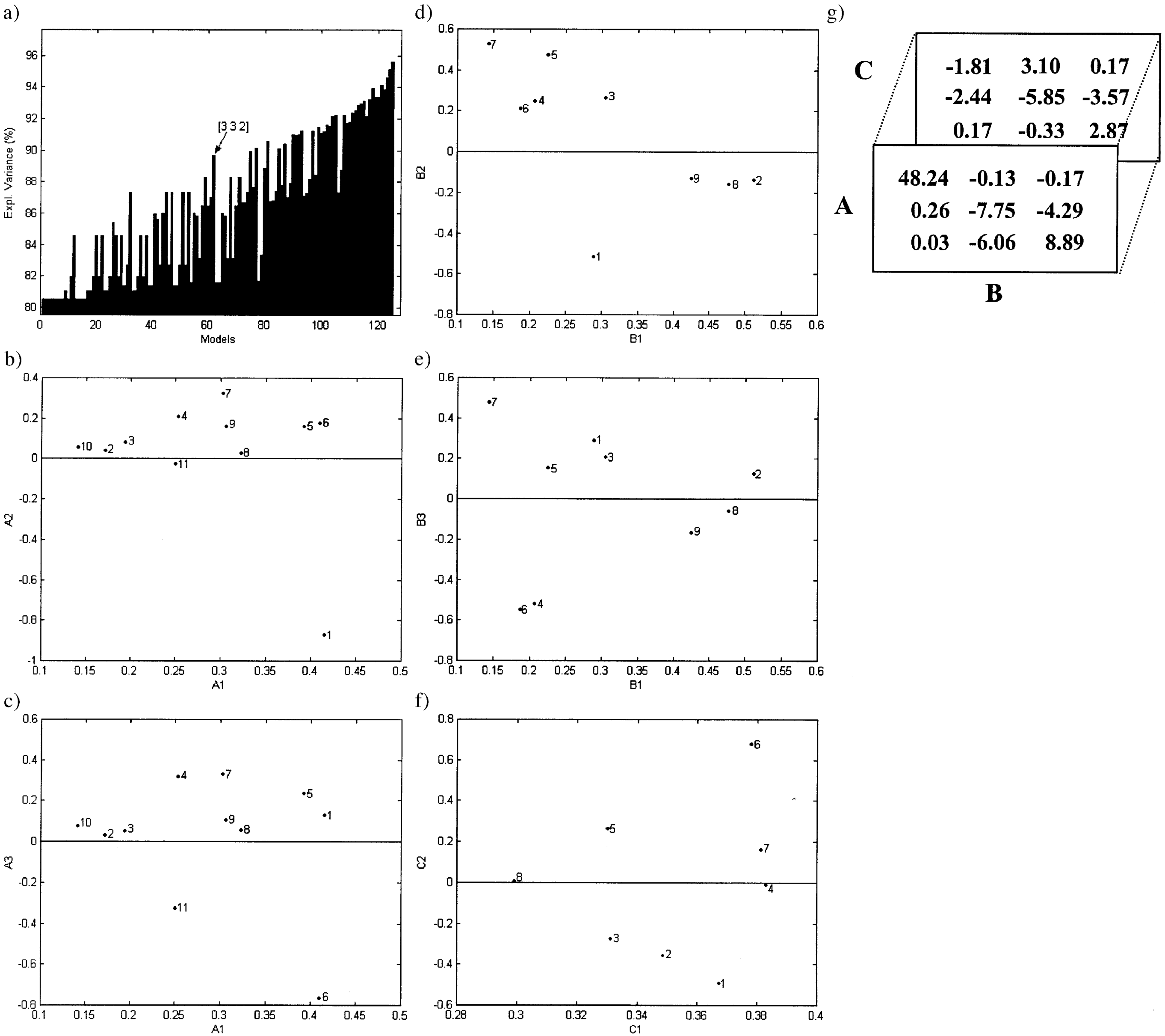
The first latent factor on the variable compromise (see Fig. 6a) plot reflects now the total ionic content (except H^+) of the samples, while the first PC was explaining more the NH_4^+, NO_3^-, SO_4^{2-} content of the samples when the data were only standardized. The second PC describes the acidity of the samples and the third latent factor is a "mixed salt" factor contrasting Cl^- (7) on one hand and K^+ (4) and Mg^{2+} (6) ion concentrations on the other (see Fig. 6b). Along PC1, the sites on the site compromise (see Fig. 6c) are ranked according to their total ionic content (except H^+). Sonnblick (10), followed by Innervillgraten (2) and Reutte (3) has the lowest ionic content. Intermediate values are found for Kufstein (4), Lunz (7), Nasswald (9) and Werfenweng (11), Nassfeld (8) has the higher ionic content and the highest contents are found for Litschau (5), Lobau (6) and Haunsberg (1). Haunsberg samples can be distinguished from the others mainly because of their higher proportion of Cl^- (7) during 1995 (6), and Lobau samples because of their high K^+ and Mg^{2+} ion concentrations during the same year, 1995 (see Fig. 6e and f).

5.1. Applying STATIS to data with a non-perfect structure

STATIS can be used also when the data set does not have the same dimension for columns and/or rows (imperfect trilinear structure). In the case of the studied data, the properties of the samples collected at each site are measured during different sampling periods. The performance of the STATIS method on such data will give information about variables and sites distribution on the compromise plots. In order to obtain the variable compromise, $\underline{\mathbf{X}}$ is arranged as $I=9$ (chemical components), $J_k \neq J_{k'}$ (different sampling period of each site) and $K=11$ (sampling sites). The results of standardized data are presented in Fig. 7.

Two principal components explain 89.8% of the compromise matrix variance. The object distribution on the

Fig. 4. Results of STATIS for the standardized data set with perfect trilinear structure: (a) eigenvalues scree plot of PCA of the site compromise matrix; (b) eigenvalues scree plot, of PCA of the variable compromise matrix; (c) eigenvalues scree plot of PCA of the year compromise matrix; (d) PC1–PC2 plot of the compromise of 11 sites; (e) PC1–PC2 plot of the compromise of 9 variables; (f) PC1–PC2 plot of the compromise of 8 years; (g) convex hulls of 9 variables on the year compromise plot; (h) convex hull of 9 variables drawn on the compromise for year 1995 (6); (i) convex hull of 11 sites drawn on the compromise for year 1995 (6).



-1.81 3.10 0.17
-2.44 -5.85 -3.57
0.17 -0.33 2.87

48.24 -0.13 -0.17
0.26 -7.75 -4.29
0.03 -6.06 8.89

Fig. 5. Results of Tucker3 for the standardized data: (a) explained variance by models of different complexities; (b) projection of sites on the plane defined by A1 and A2; (c) projection of sites on the plane defined by A1 and A3; (d) projection of variables on the plane defined by B1 and B2; (e) projection of variables on the plane defined by B1 and B3; (f) projection of years on the plane defined by C1 and C2; (g) core array $\underline{\mathbf{G}}$ (3 3 2).

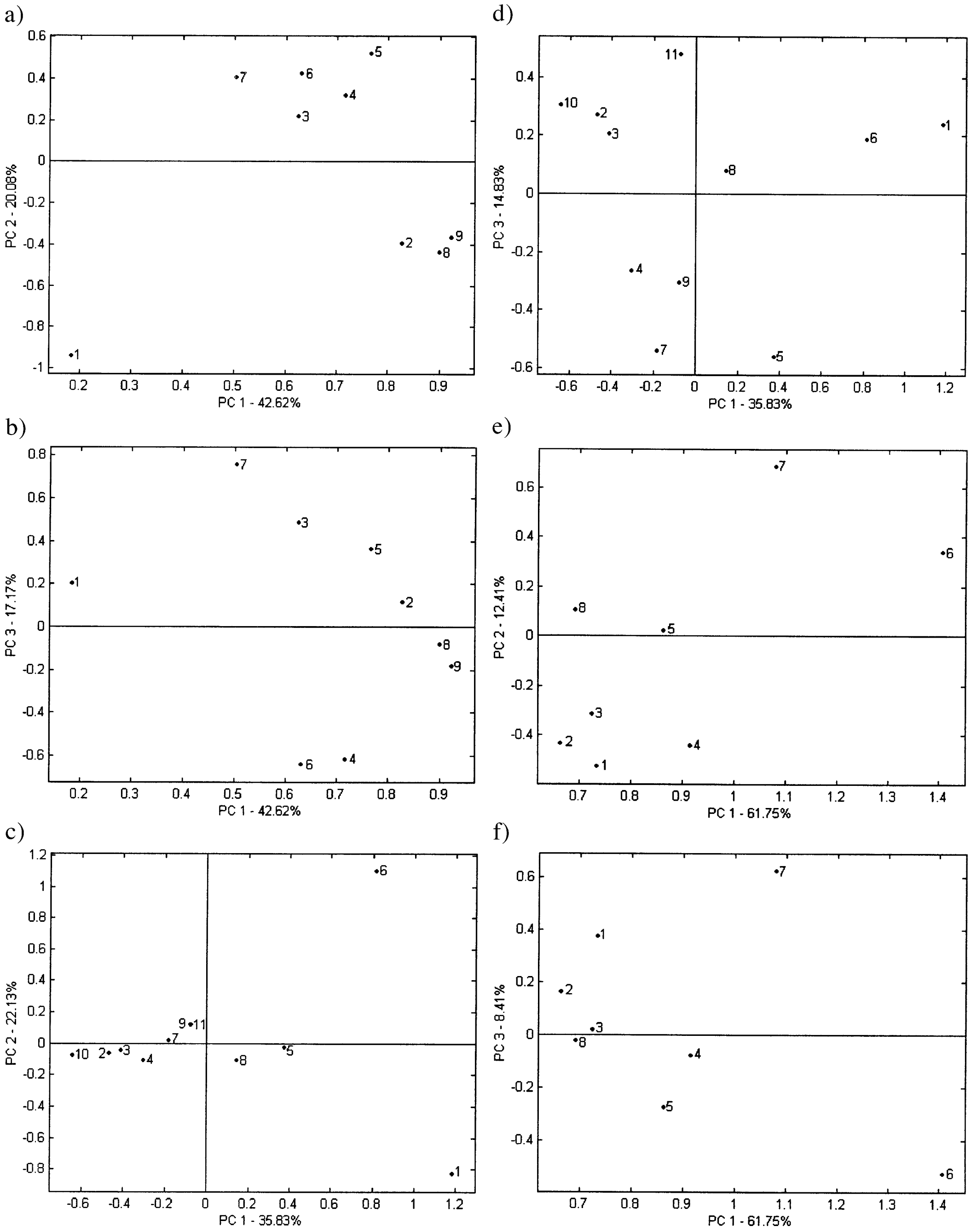


Fig. 6. Results of STATIS for the autoscaled data set with perfect trilinear structure: (a) PC1–PC2 plot of the compromise of 9 variables; (b) PC1–PC3 plot of the compromise of 9 variables; (c) PC1–PC2 plot of the compromise of 11 sites; (d) PC1–PC3 plot of the compromise of 11 sites; (e) PC1–PC2 plot of the compromise of 8 years; (f) PC1–PC3 plot of the compromise of 8 years.

variable compromise plot (see Fig. 7a) is almost the same as in the case of preprocessed data with perfect trilinear structure (compare Figs. 4e and 7a). On the projection PC1–PC2, there are again three groups of parameters.

Due to the different sampling period of each site, $J_k \neq J_{k'}$, the variance–covariance matrix for each individual site in $\underline{\mathbf{X}}$ is calculated, which leads to new data $\underline{\mathbf{Y}}$ of dimension $I \times I \times K$. $\underline{\mathbf{Y}}$ of dimension $K \times I \times I$ is used as input data in STATIS to obtain the compromise for sites.

The site compromise plot, constructed for two principal components, explaining 91.9% of the variance, is given in Fig. 7b. Three groups of sites can again be distinguished along PC1. Haunsberg (1) and Lobau (6) sites can be clearly segregated from the others sites along PC2.

The PARAFAC2 method, able to deal with imperfect trilinear structures, was also applied to the studied data set. The results are presented for the variables and sites in Fig. 7c and d, respectively. The projection of variables on the

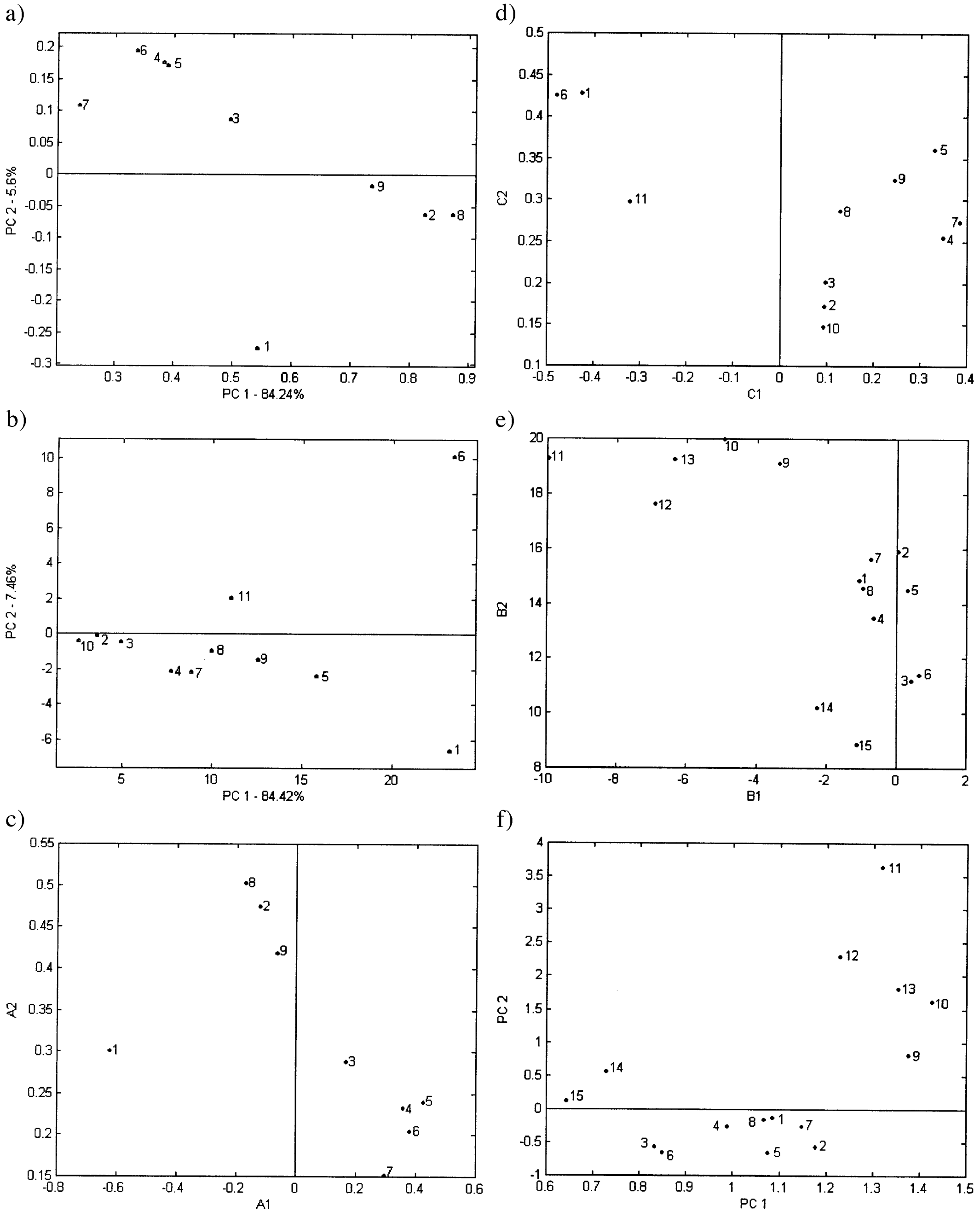


Fig. 7. PC1–PC2 plot of the compromise of: (a) 9 variables and (b) 11 sites, after applying STATIS to standardized data with a non-perfect structure; projection of: (c) variables on the plane defined by A1 and A2, and (d) sites on the plane defined by C1 and C2, after applying PARAFAC2 to standardized data with a non-perfect structure; (e) projection of years on the plane defined by B1 and B2 for the Haunsberg site (1), after applying PARAFAC2; (f) projection of years on the plane defined by PC1 and PC2 for the Haunsberg site (1), after applying STATIS.

plane defined by the first two factors accounts for 89.3% of the total variance. The pattern of variables is similar to that obtained with STATIS for the variable compromise (compare Fig. 7a and c). However, the sequence of factors obtained in both methods is reversed. The first factor of PARAFAC2 corresponds to PC2 of STATIS. The same situation is observed for the site distribution (see Fig. 7d). Again the first PARAFAC2 factor corresponds to the second STATIS principal component. Moreover, the pattern is different. The objects 1 (Haunsberg) and 6 (Lobau) can be distinguished along PC2 on the STATIS site compromise (see Fig. 7b), whereas they come close together on the projection of sites in the plane defined by two PARAFAC2 factors (see Fig. 7d). Additionally, information about the concentration profile during the whole sampling period of each site can be obtained. For instance, the concentration profile for Haunsberg (1) explains its outlying character (see Fig. 7e). During

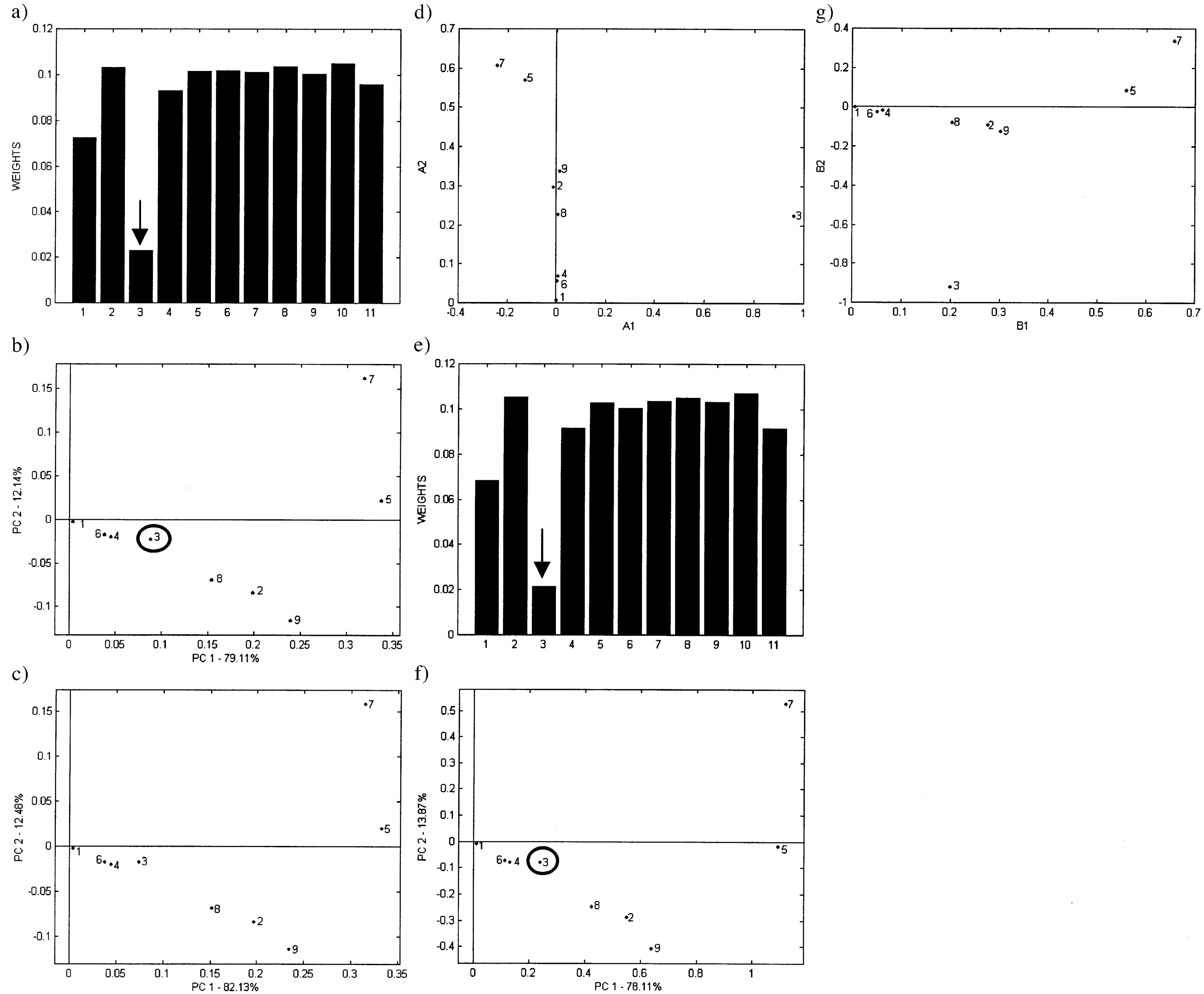


Fig. 8. (a) Bar plot of 11 weights for 9 variables for contaminated data with imperfect structure; (b) PC1–PC2 plot of the compromise of 9 variables for contaminated data with imperfect structure; (c) PC1–PC2 plot of the compromise of 9 variables for non-contaminated data with imperfect structure; (d) projection of variables on the plane defined by A1 and A2, after applying PARAFAC2 to contaminated data; (e) bar plot of 11 weights for 9 variables for contaminated data with perfect trilinear structure; (f) PC1–PC2 plot of the compromise of 9 variables for contaminated data with perfect structure; (g) projection of variables on the plane defined by B1 and B2, after applying Tucker3 to contaminated data.

1993–1997 (objects nos. 9–13), it differs from the profile in other sampling years. The reason is the high concentrations of Ca^{2+} (5) and Cl^{-} (7).

The same information about the concentration profile during the whole sampling period of each site can be achieved by the STATIS method. From the scores of PCA ($\boldsymbol{L}$) of the variable compromise, set of loadings ($\boldsymbol{P}_k$) for each of the K individual tables, X_k, constituting $\underline{\mathbf{X}}$ for f principal components can be obtained using the following equation:

$$\mathbf{P}_k = \mathbf{X}_k^T\mathbf{L}\mathbf{E}, \tag{9}$$

where **E** and **L** have the same dimensionality as **E** and **L** in Eq. (6).

Fig. 7f shows the concentration profile for Haunsberg during 1985–1999 (objects nos. 1–15). The pattern observed resembles the pattern obtained by PARAFAC2 (see Fig. 7f and e). However, the sequence of factors is again reversed.

Some differences between the results of STATIS and the Tucker3 as well as PARAFAC2 methods can be expected due to the different objectives of the methods. Tucker3 and PARAFAC2 are decomposition models, which fit the original data as well as possible, whereas STATIS reveals object distributions on the compromise plot.

5.2. Robust properties of STATIS

The robust properties of STATIS are demonstrated on data sets (with imperfect and perfect trilinear structure), where deliberately a few outlying objects were introduced. For the complete data, i.e. the data with imperfect trilinear structure, the concentration of Na^{+} (3) was increased 25 times for the first year (1988), 9 times for the second year (1989) and 16 times for the third year (1990) for the Reutte site (table 3). The results for STATIS and PARAFAC2 are given in Fig. 8a–d.

As expected the third site (Reutte) has now the smallest weight (see Fig. 8a). The consequence of this is that the variable compromise plot remains unchanged (see Fig. 8b). For comparison, the variable compromise plot of the original data is presented in Fig. 8c. PARAFAC2 is however very sensitive to the presence of outliers for the same contaminated data set. Fig. 8d shows that the variable pattern observed is influenced to a high extent by the high concentrations of Na^{+} (3) for Reutte site.

For the same site Reutte (3) the concentration value of Na^{+} was increased 25 times for the first, the second and the third year in the data set with perfect trilinear structure. As a result of the STATIS comparison between tables, the third table (site) has the smallest weight (see Fig. 8e). The variable compromise plot constructed for two PCs explaining 91.9% of variance remains unchanged (see Figs. 2h and 8f).

For comparison, Tucker3 was applied to the same contaminated data set. The complexity of the model is [3 2 3] explaining almost the same amount of variance as in STATIS, 91.7%. The variable distribution is far more highly influenced by the outlying object 3 (Na^{+}) (see Fig. 8g).

6. Conclusions

STATIS is a three-way method for exploratory data analysis. It is best understood starting from an unfolded two-way table. For an $I \times J \times K$ data set this is obtained by juxtaposition of K ($I \times J$) two-way tables. To analyze the resulting table by PCA, the variance–covariance matrix is used. This is normally obtained by summing the K variance–covariance matrices of the K individual tables constituting the unfolded table. STATIS first weights the variance–covariance matrices of each table according to the similarities between them. The tables least similar to the mean are given the lowest weight. In this way, the three-way character of the method is obtained.

To demonstrate the method, we have compared it with PARAFAC2 and Tucker3, which are considered the standard methods in chemometrics for the analysis of three-way data. We wanted to see if they give similar results when applied to a chemical data set. Since such results depend on the pretreatment of the data, we considered pretreatments that are feasible also with PARAFAC2 and/or Tucker3. It was found that STATIS and Tucker3 lead to the same results for non-preprocessed data. When the data are standardized, the Tucker3 model gives similar results but requires for our data set a higher complexity to explain the same amount of variance as in STATIS.

An advantage of STATIS, shared among N-way methods only by PARAFAC2, is that it can deal also with imperfect trilinear data structure, i.e. data for which one or more rows or columns are missing in the data cube. Some small, but unimportant, differences are observed in the results for standardized data with both methods. STATIS has several other features that make it a useful tool for exploratory analysis. The first such feature is that there are no special requirements in STATIS on how to preprocess the data. All pretreatments that would be acceptable for the unfolded 2-way table can be applied.

Another appealing feature of STATIS as an exploratory tool is its robust properties. The results are not affected by the presence of large outliers. A third feature, that should be outlined, is its very good visualization properties. The compromise plots with convex hulls drawn through individual objects give an impression about the similarities and dissimilarities among them and help to identify the variables responsible for the dissimilarities, thereby making the interpretation of the STATIS results easier.

Additionally, contrary to N-way methods, the STATIS algorithm is very computer time efficient since it is non-iterative.

STATIS is also subject to some limitations. It cannot be generalized to more than three-way data and it is only an exploratory tool since it cannot be used in a modeling context as can be done with other *N*-way models.

References

[1] A. Rizzi, M. Vichi, Representation, synthesis, variability and data preprocessing of three-way data set, Computational Statistics & Data Analysis 19 (1995) 203–222.

[2] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149–171.

[3] N. Faber, R. Bro, P. Hopke, Recent developments in CONDECOMP/PARAFAC algorithms: a critical review, Chemometrics and Intelligent Laboratory Systems 65 (2003) 119–137.

[4] H. Kiers, J. Ten Berge, R. Bro, PARAFAC2—PART I. A direct fitting algorithm for the PARAFAC2 model, Journal of Chemometrics 13 (1999) 275–294.

[5] R. Bro, C. Andersson, H. Kiers, PARAFAC2—PART II: Modeling chromatographic data with retention time shifts, Journal of Chemometrics 13 (1999) 295–309.

[6] R. Henrion, N-way principal component analysis. Theory, algorithms and applications, Chemometrics and Intelligent Laboratory Systems 25 (1994) 1–23.

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

[8] S. Gourvénec, G. Tomasi, C. Durville, E. Di Crescenzo, C.A. Saby, D.L. Massart, R. Bro, G. Oppenheim, CuBatch, a MATLAB interface for *N*-dimensional data analysis, Chemometrics and Intelligent Laboratory Systems, in press.

[9] Ch. Lavit, Y. Escoufier, R. Sabatier, P. Traissac, Computational Statistics & Data Analysis 18 (1994) 97–119.

[10] P. Schlich, Defining and validating assessor compromises about product distances and attribute correlations, in: T. Naes, E. Risvik (Eds.), Multivariate Analysis of Data in Sensory Science, Elsevier, Amsterdam, The Netherlands, 1996, pp. 259–305.

[11] A. Carlier, Ch. Lavit, M. Pages, M. Pernin, J. Turlot, A comparative review of methods, which handle a set of indexed data tables, Multiway Data Analysis, 1989, pp. 85–101.

[12] Ch. Lavit, Analyse conjointe de tableaux quantitatifs, Masson, Paris, 1988.

[13] P. Robert, Y. Escoufier, A unifying tool for linear multivariate statistical methods. The RV-coefficient, Applied Statistics 25 (1976) 257–265.

[14] B.M.G. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998.

[15] R. Bro, Multi-way analysis in food industry. Models, algorithms and applications, thesis, 1998.

[16] P. Kroonenberg, J. de Leeuw, Principal component analysis of three-mode data by means of alternating least squares algorithms, Psychometrika 45 (1980) 69–97.

[17] R. Bro, A. Smilde, Centering and scaling in component analysis, Journal of Chemometrics 17 (2003) 16–33.

[18] F. Kalina, H. Puxbaum, A high density network for wet only precipitation chemistry sampling in Austria, Quarterly Journal of the Hungarian Meteorological Service 100 (1996) 159–170.