JOURNAL OF CHEMOMETRICS
J. Chemometrics **12,** 337–354 (1998)

REGRESSION ON PARAMETERS FROM THREE-WAY DECOMPOSITION

PAUL GELADI,[1]* YU-LONG XIE,[1] ALEXANDR POLISSAR[2] AND PHILIP HOPKE[2]
[1]*Department of Organic Chemistry, Umeå University, S-90187 Umeå, Sweden*
[2]*Department of Chemistry, Clarkson University, Potsdam, NY 13699, U.S.A.*

SUMMARY

This paper presents work on the combination of (a) environmental chemistry, (b) three-way analysis by Parafac constrained to non-negative results and (c) multivariate calibration. For two different environmental examples it is shown how the loading parameters from an independent three-way analysis can be used in a regression against external data. This combination leads to an easier interpretation of the results. The data are from Arctic aerosol studies. The subjectively obtained Parafac loadings are regressed against temperature anomalies. © 1998 John Wiley & Sons, Ltd.

KEY WORDS: regression; three-way analysis; partial least squares; Parafac; positive matrix factorization; receptor modeling

INTRODUCTION

Environmental data from studies involving sampling and analysis of Arctic aerosol may be presented as three-way arrays where the ways are (a) chemical variables, (b) cyclical profile over the year and (c) years. The data follow a receptor model. This is the model for the physical reality from which samples were taken and analyzed. The underlying assumptions are (1) that a source contribution model holds and (2) that the aerosols were produced far away from the Arctic and thoroughly aged and mixed. The source contribution model uses source profiles with concentrations that are positive or zero. The mixing model assumes that the sampled material is a linear combination of contributions from the different sources. The source profiles for long-distance transport are not known, but they can be deduced in part by factor analysis and a subjective interpretation of factors.[1] The mathematical model used for analyzing the data is the three-way factor analysis model. The paper presents a combination of regression analysis and three-way factor analysis applied to some environmental examples. See Figure 1.

Three-way factor analysis

Figure 2 gives the most important three-way factor analysis model used in this paper: the Parafac decomposition. The alternative Tucker decomposition is not used here. These decompositions are described in References 2 and 3 and references cited therein. Positive matrix factorization (PMF)[4–7] allows the calculation of a Parafac model with non-negativity constraints.

* Correspondence to: P. Geladi, Department of Organic Chemistry, Umeå University, S-90187, Sweden. E-mail: paul.geladi.@chem.umu.se

CCC 0886–9383/98/050337–18 $17.50

Received September 1997
Accepted June 1998

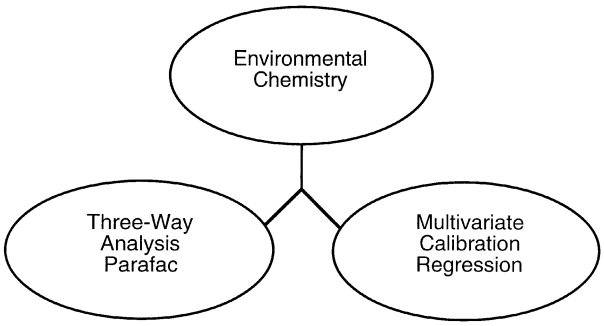

Figure 1. The work presented in this paper is a combination of environmental chemistry, three-way analysis and regression

The Parafac decomposition is

$$\underline{\mathbf{Z}} = \mathbf{a}_1 \otimes \mathbf{b}_1 \otimes \mathbf{c}_1 + \mathbf{a}_2 \otimes \mathbf{b}_2 \otimes \mathbf{c}_2 + \ldots + \underline{\mathbf{G}} \tag{1}$$

where

$\underline{\mathbf{Z}}$ a three-way array ($I \times J \times K$)
$\underline{\mathbf{G}}$ the residual ($I \times J \times K$) sometimes called $\underline{\mathbf{E}}$
$\otimes$ a symbol for tensor product
$\mathbf{a}_1$ the first *a*-loading vector (size $I \times 1$)
$\mathbf{b}_1$ the first *b*-loading vector (size $J \times 1$)
$\mathbf{c}_1$ the first *c*-loading vector (size $K \times 1$).

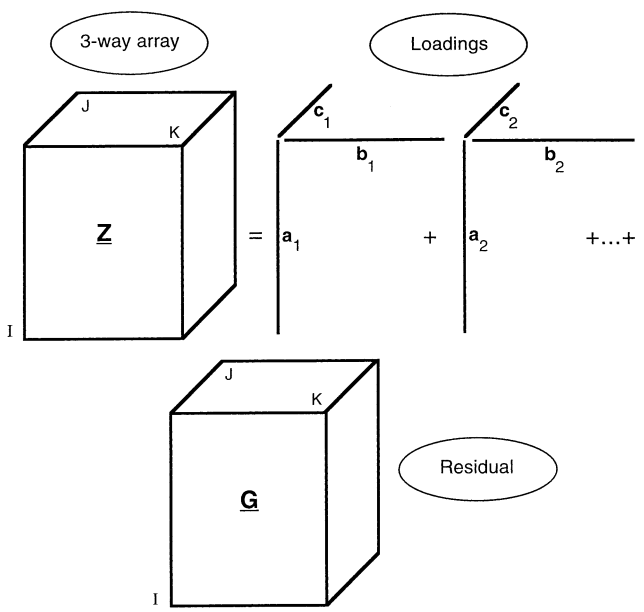

Figure 2. Parafac is a decomposition of a three-way array into sums of tensor products of triplets of loading vectors and a residual. The goal of the decomposition is to have the meaningful information in the loadings and the noise in the residual

The loadings are simply called *A*-loadings, *B*-loadings and *C*-loadings. Usually *R* loading vectors are needed. *R* is called the three-way rank. Three-way rank is very important and difficult to determine.[2,3] The loadings are used to construct loading matrices

$$\mathbf{A} = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \ldots \quad \mathbf{a}_R], \qquad \mathbf{B} = [\mathbf{b}_1 \quad \mathbf{b}_2 \quad \ldots \quad \mathbf{b}_R], \qquad \mathbf{C} = [\mathbf{c}_1 \quad \mathbf{c}_2 \quad \ldots \quad \mathbf{c}_R] \tag{2}$$

For environmental and chemical problems the loadings are preferably non-negative, because negative concentrations and mixture contributions do not make sense. The most general Parafac decomposition does allow negative values. Positive matrix factorization (PMF) forces the loadings in the matrices of (2) to be non-negative. The PMF method has a number of other interesting properties, such as weighting of each element in the array, so that missing data, outliers, etc. can be given high subjective weights. PMF also allows non-linear transformation of the data. It is a well-known fact that environmental data may have a skewed distribution, making a logarithmic transformation necessary.

Three-way factor models possess the properties of slow convergence and non-uniqueness of the results. It is also very difficult to determine the three-way rank of the models, especially in constrained models. Subjective background knowledge and chemical common sense are often used to select a factor solution. In this paper the finally selected factor solutions are tested in a regression model. A functioning regression model gives extra confirmation that the selected factors are meaningful.

Scaling is very important. The three-way array $\underline{\mathbf{Z}}$ can be scaled in many ways. This influences the analysis. Once the scaling of $\underline{\mathbf{Z}}$ is fixed, the resulting decomposition can also be scaled in many ways. Repeating equation (1),

$$\underline{\mathbf{Z}} = [w_{a1}\mathbf{a}_1] \otimes [w_{b1}\mathbf{b}_1] \otimes [w_{c1}\mathbf{c}_1] + [w_{a2}\mathbf{a}_2] \otimes [w_{b2}\mathbf{b}_2] \otimes [w_{c2}\mathbf{c}_2] + \ldots + \underline{\mathbf{G}} \tag{3}$$

where w_{a1}, w_{b1}, etc. are weights.

The weights can be distributed in many ways without affecting the decomposition. An example is

$$w_{a1}w_{b1} = 1/w_{c1}, \quad w_{a2}w_{c2} = 1/w_{b2}, \quad \text{etc.} \tag{4}$$

One has to take these possibilities into account when interpreting the loadings and when using them in further calculations.

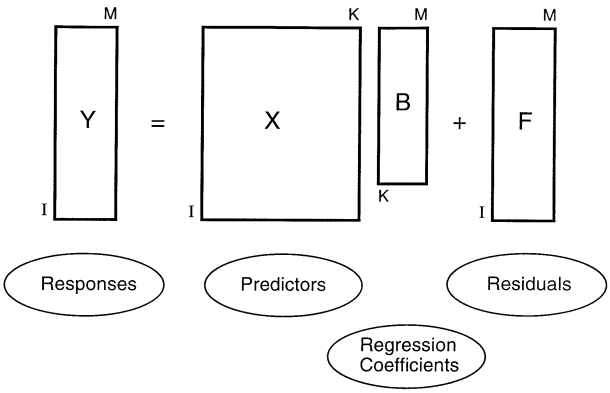

Figure 3. A linear regression model between predictor (**X**) and response (**Y**) variables consists of a matrix **B** of regression coefficients and a residual matrix **F**

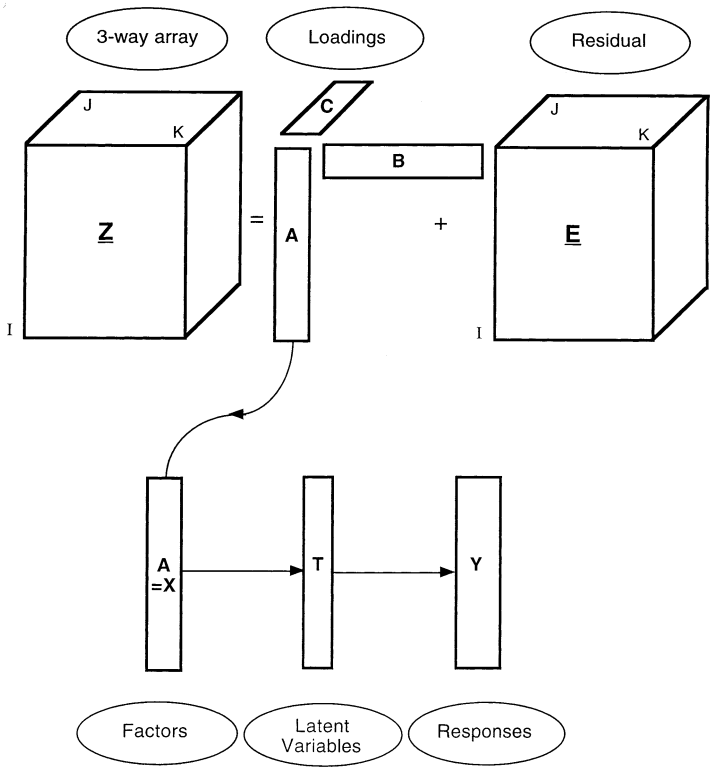

Figure 4. The loadings from a Parafac three-way decomposition can be used as predictor variables in a regression model against external response variables. The residual is **E** or **G**

Regression models

A linear regression model is of the form

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F} \tag{5}$$

where

Y the response data (mean-centered and properly weighted)
X the predictor data (mean-centered and properly weighted)
B the regression coefficients
F the residual.

See also Figure 3.

The differences between the models are based on how **B** is calculated. A very convenient method is partial least squares (PLS) regression. If **Y** has only one variable, PLS1 is used, and if **Y** has many variables, PLS2 that decomposes **Y** is used.[8–12] References to the older PLS literature can be found in Reference 13. PLS is based on latent variables. **X** is decomposed as

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E} \tag{6}$$

and the regression part is between **Y** and **T**:

$$\mathbf{Y} = \mathbf{TQ}^{\mathrm{T}} + \mathbf{F} \tag{7}$$

where

T an orthogonal set of latent variables
P loadings of the decomposition of **X**
Q regression coefficients between **Y** and **T**
E the residual for the decomposition of **X**.

The theory and practical use of these methods can be found in References 8–13. Latent variable decomposition methods have the advantage of being able to separate the noise from the useful structure in **X** (equation (6)). This may also be a disadvantage, since the number of latent variables (also called rank) that is best to use has to be determined. The interesting property of PLS2 is that also the block of *y*-variables is decomposed. This is useful when the *y*-variables are correlated.

A possible test of a regression model is the coefficient of determination, R^2 (assuming that **X** and **Y** are mean-centered):

$$R_x^2 = 1 - \text{trace}(\mathbf{E}^T\mathbf{E})/\text{trace}(\mathbf{X}^T\mathbf{X}) \quad (8)$$

$$R_y^2 = 1 - \text{trace}(\mathbf{F}^T\mathbf{F})/\text{trace}(\mathbf{Y}^T\mathbf{Y}) \quad (9)$$

These quantities can also be calculated for each *x*- or *y*-variable separately. As a rule of thumb, an R_y^2 above 0·8 is considered good and one above 0·9 very good. An R_y^2 below 0·65 means that the regression model does not explain much more than random noise. With a test set, equation (5) becomes

$$\mathbf{Y}_t = \mathbf{X}_t\mathbf{B} + \mathbf{F}_t \quad (10)$$

Also for test sets, a coefficient of determination can be expressed:[14]

$$Q^2 = 1 - \text{trace}(\mathbf{F}_t^T\mathbf{F}_t)/\text{trace}(\mathbf{Y}_t^T\mathbf{Y}_t) \quad (11)$$

where

$\mathbf{F}_t$ the residual for the test set
$\mathbf{Y}_t$ the mean-centered responses for the test set.

A regression model may be used for different purposes. The traditional use is to predict the responses (**Y**) for future measurements of the predictor variables (**X**) when a good set of regression coefficients has been obtained. Regression may also be used to stabilize a model. In this way, regression between the parameters from a model (such as loadings from a three-way decomposition into factor loadings) and some external variable forms a constraint on these parameters. The factor loadings are calculated with the constraint that the regression model should have a high R^2 or Q^2. A third use of regression is for the interpretation of model parameters. In this case the predictor variables in **X** are parameters from a factor analysis model that one wants to study. In the first case the predictor variables **X** are real measured data. In the other two cases the predictor variables **X** are the results (parameters) of a previous analysis. See also Figure 4.

It is important to notice that this paper is about using model parameters from Parafac models as predictor variables. The goal of the studies was *not* to predict temperature anomalies from air pollution data. Temperature anomalies are much easier and cheaper to measure than airborne particulate material and gases.

Regression combined with three-way decomposition

With three-way data arrays of size $I \times J \times K$ one can always run PLS by reorganizing the predictor (X) data to $I \times (J \times K)$ if the responses (Y) block is of size $I \times M$. The interesting thing is to combine the requirements of a PLS model with a three-way decomposition of the data array. A first simple example of real three-way PLS was presented by Wold *et al.*[15] The emphasis was on the regression

for prediction and not on the three-way decomposition. Six 10×10 LC–UV arrays were used as calibration for the concentrations of two aromatic compounds. The test set contained four solutions of the two chemicals. This is three-way PLS2. Based on this work, Ståhle[16,17] presented more theory and a three-way PLS1 example from psychology. The data set contained 4 neurochemicals $\times$ 6 times $\times$ 10 animals. An application for batch processes is given by Nomikos and MacGregor,[18] but this paper is mainly about reorganization to two-way arrays. Nørgaard[19] describes the use of MLR between factor loadings from Parafac and Tucker models. The example is fluorescence and the external variable to be regressed against is the slit width of the spectrometer. A recent systematic study with a fluorescence example is given by Bro.[20]

In this paper, latent variable regression by PLS2 is used between loadings obtained from PMF-Parafac analysis and external variables. See Figure 4. Regression models were built and tested without prior knowledge of the identity of the factors in order to ensure an objective treatment.

When doing the analysis as shown in Figure 4, a number of issues have to be dealt with. The factors from Parafac are not orthogonal, so a new orthogonalization may be useful. A Tucker model would be able to give orthogonal factors, but this type of model is not used here. The factors from Parafac are also not by definition mean-centered. This is certainly the case for PMF with non-negativity constraints. Also, the scaling of the different factors that may be useful for the factor analysis model may not be useful for the regression model. It is better to mean-center the factors and to use scaling by the standard deviation. Orthogonalization is taken care of by the PLS2 algorithm.

THE EXAMPLES

Particulate species from mid-latitudinal industrial sources (Arctic haze) have been observed in the Arctic for many years.[21,22] A number of reviews of the Arctic haze phenomenon have been published.[23–27] The highest concentration of the particulate species has been measured during the winter and spring seasons and the lowest values were measured during the summer.[23,28] It has been shown that such seasonal variations of the aerosol concentration in the Arctic are the result of a combination of a seasonal variability in the long-range transport of air,[29–31] in the atmospheric blocking phenomenon,[32] in pollutant removal processes,[33,34] in the oxidation rate of SO_2[35,36] and in the thickness of surface temperature inversions.[37]

Example 1

Airborne particulate samples have been collected at Alert, Northwest Territories, Canada (latitude 82·3 N, longitude 62·5 W) by the Atmospheric Environment Services (AES) of Canada on a weekly basis since July 1980. Details of the sampling and procedures of chemical analyses can be found in References 34 and 35. Week-long samples were collected on Whatman 41 filters using a high-volume sampler. Major ions were analyzed by ion chromatography (IC) and trace element data were obtained by inductively coupled plasma (ICP) emission spectroscopy and instrumental neutron activation analysis (INAA). Since some measured species have a large number of missing data and values below the detection limit, data for only 24 chemical species measured in samples obtained between September 1980 and August 1991 were used in this analysis (Table 1). Inspection of the time series of the aerosol species concentrations at Alert over the time interval from 1980 to 1991 shows strongly recurring yearly cyclical variations. To examine these cyclical patterns, the time series can be reorganized into a three-way array and this is done by arranging the data as 52 weekly time slices for each of the 24 chemical species over the eleven years for which data are available; see Figure 5. This arrangement forces the week-to-week variations within each year for each factor to be the same from year to year.

In order to convert the data into a three-way array, an additional 40 entries were created in the

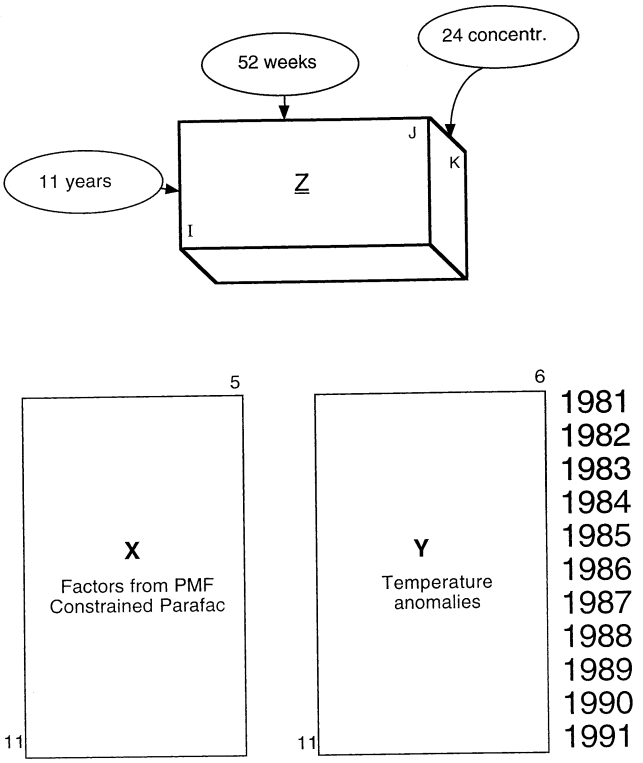

Figure 5. Data of Example 1

appropriate time intervals for all species and their values were assigned as the corresponding geometric mean. The 572 data points of each chemical species were reorganized into a 52×11 matrix, with eleven separate years each containing 52 weeks, giving an $11 \times 52 \times 24$ (years $\times$ weeks $\times$ variables) three-way data array.

A five-factor non-negative solution with rotation was found to fit the data well and was the easiest to interpret physically.[38] The five factors obtained (see Table 1) were airborne soil (Al, Si, Fe, Ti, etc.; F2), sea salt (Na, Cl; F4) and three more factors dominated by sulfate. One sulfate factor has its peak intensity in January and February and includes most of the elements that are normally thought to derive from anthropogenic sources (Pb, Zn, etc.; F3). The seasonal variation of this factor has a

Table 1. Variables and factors for Alert data.[38] Details are described in the text

Variables				Factors
1. Cl	7. NH_4	13. Zn	19. Si	F1. Photogenic
2. Br	8. K	14. Pb	20. As	F2. Soil
3. NO_3	9. MSA[a]	15. Ca	21. La	F3. Anthropogenic
4. SO_4	10. Mn	16. Ti	22. Sb	F4. Sea salt
5. H	11. V	17. I	23. Sm	F5. Biogenic
6. Na	12. Al	18. In	24. Se	

[a] MSA $\equiv$ methane sulfonic acid (CH^3SO_3H).

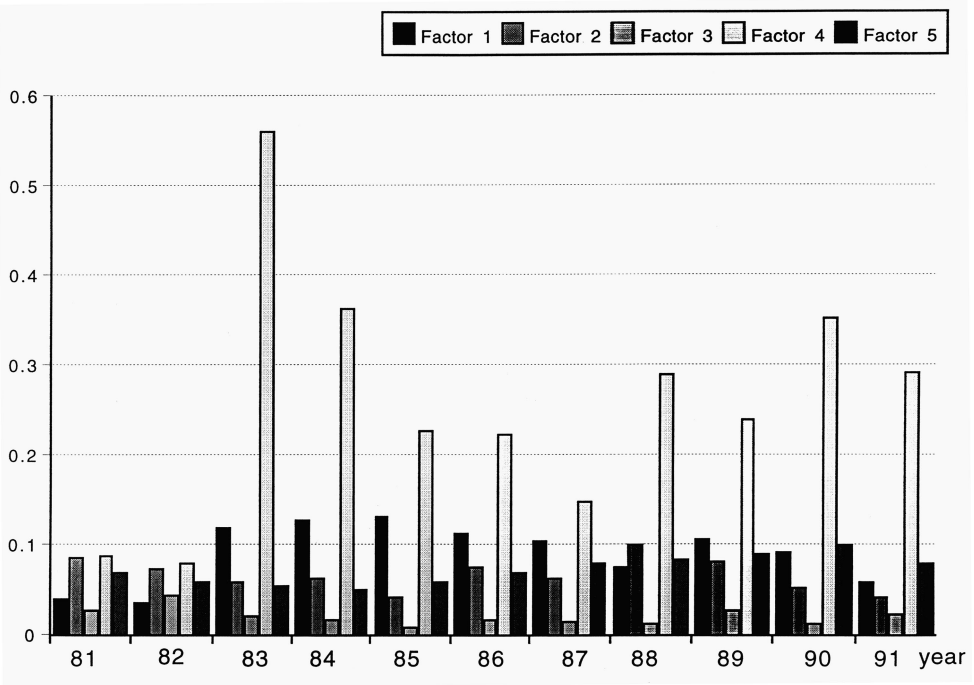

Figure 6. Factors of Example 1

maximum around April, just after polar sunrise. The factors for the years are shown in Figure 6.

In the Arctic atmosphere the air is at sub-freezing temperatures in the dark during the winter. At polar sunrise in the early spring the atmosphere is exposed to sunlight. When the sun rises, photochemical reactions release Br, leading to troposphere O_3 depletion.[34,39] Thus this factor (F1) represents photochemical conversion of SO_2 to acidic sulfate as well as the production of particulate bromine.

The final factor (F5) is characterized by the presence of methane sulfonic acid (MSA), which is the product of the oxidation of sulfur-containing compounds such as dimethyl sulfide (DMS) and dimethyl disulfide (DMDS) emitted by biogenic activity in the surface layer of the ocean. The ratio of MSA to SO_4^{2-} in this factor is $0{\cdot}34 \pm 0{\cdot}04$. Isotopic studies by Li and Barrie[40] have shown that this ratio should be approximately 0·31 for biogenic sulfur aerosol in the Arctic region. Factor F5 therefore represents a biogenic component of the Arctic aerosol. Further corroboration of the origin of this factor can be obtained by examining the temporal variation. There are two peaks in the seasonal variation of the factor: a large one around April/May and another, smaller peak about August. Li *et al.*[41] attributed the spring peak to the sea surface temperature anomalies (SSTAs) in the North Atlantic Ocean west of the coast of continental Europe, and the summer peak to the SSTAs in the ocean region further north in the Atlantic Ocean off the coast of Norway and in the northwestern North Pacific Ocean. Anomalies are the differences in the measured values relative to a long-term average value of that environmental property. Since the sea surface temperature will be related to the atmospheric temperature, the hypothesis can be formed that this factor and possibly others could be correlated to the global temperature, and the relationship between the year-to-year factor scores (*C*-factor) will be examined relative to measures of the temperature described below.

The factors (F1–F5) described above for eleven years become the *X*-variables ($x1$–$x5$) in a PLS

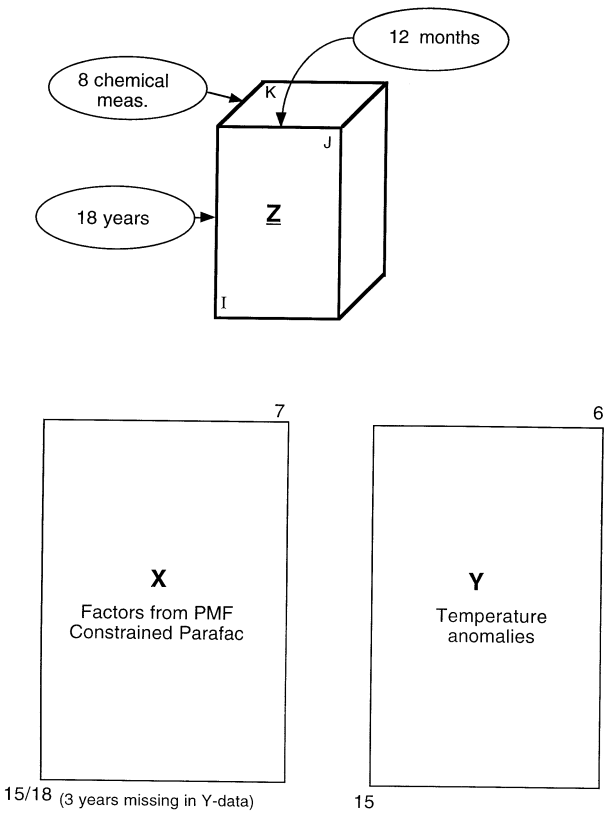

Figure 7. Data of Example 2

model and are tested against global temperature anomaly data for corresponding years, an 11×6 array. Partial least squares regression, PLS2, can handle correlated response variables. The variables and factors are described in Table 1.

Example 2

The Barrow data form an $18 \times 12 \times 8$ (years $\times$ months $\times$ variables) three-way array. A schematic overview of the data structure is shown in Figure 7. The array was analyzed with PMF-type Parafac with non-negativity constraints, leading to seven factors. Parafac models allow many factors, even for small data sets. The factors used were deemed chemically meaningful by subjective interpretation. The variables and factors are described in Table 2.

Surface aerosol data including condensation nuclei counts, black carbon, particle light scattering and aerosol optical depth measurements made at Barrow, Alaska have been reported.[42–45] These parameters are measured in real time at this site. The aerosol scattering and absorption coefficients show a strong annual cycle corresponding with the Arctic haze maximum in the winter and spring and a minimum in the summer and fall.[43,44] On the other hand, annual cycles for the condensation nuclei (CN) data show maxima in March and August.[43] A decreasing long-term trend in the tropospheric aerosol optical depth (AOD) and the surface aerosol scattering coefficient in the March–April period from 1982 to 1991 was observed at Barrow.[45] A similar decrease in the integral atmospheric optical

Table 2. Variables and factors for Barrow data.[49] Explanation is in the text

Variables		Factors	
1. Condensation nuclei	5. O_3	F1. Arctic haze, small particles	F5. $CO + O_3 + CO_2$
2. Black carbon	6. CH_4	F2. Arctic haze, large particles	F6. O_3
3. Scattering at 540 nm	7. CO	F3. Condensation nuclei	F7. $CH_4 + CO_2$
4. Aerosol optical depth	8. CO_2	F4. Volcanic dust	

thickness in the Russian Arctic has been reported by Radionov *et al.*[46] The decrease in the Arctic haze was explained by possible reduction in anthropogenic pollution emissions in Eurasia[45–47] and reduced transport of anthropogenic aerosol to Barrow.[48]

An approach similar to that of Example 1 for examining the time series of these data has been used. Monthly average values have been reorganized into a series of year-to-year slices so that a data set of 8 variables $\times$ 12 months $\times$ 18 years was analyzed by PMF.[49] Two factors (F1 and F2) represent the Arctic haze aerosol. The first Arctic haze factor F1 has seasonal variations with a maximum in January, while factor F2 peaked in March. In these factors the black carbon values are similar, but the winter Arctic haze factor has a higher loading of condensation nucleus concentration, while the spring Arctic haze factor F2 has higher loadings for the scattering coefficient.

The winter factor F1 is primarily related to the finer particle size fraction (high loadings of condensation nucleus), while the spring Arctic haze factor F2 is related to larger-sized particles (high loadings of scattering coefficient). This size increase may be due to the higher rate of photochemical conversion of SO_2 to SO_4^{2-} and the additional particle mass that condenses onto existing particle surfaces. These particles are larger and therefore the spring factor F2 has a lower condensation nucleus concentration loading and higher loadings for scattering coefficient and aerosol optical depth. Thus two major influences on the aerosol concentration are observed: the long-range transport of anthropogenic aerosol giving rise to the winter peak, and the long-range transport plus photochemical oxidation of SO_2 in the spring.

Two factors (F3 and F4) with high loadings for condensation nuclei concentration and aerosol optical depth respectively were identified. Factor F3 with high loadings of condensation nuclei concentration has maxima in March and July–August. These maxima for the condensation nucleus factor could arise from the oxidation of dimethyl sulfide (DMS).[50,51] Part of the DMS may be from the Arctic Ocean.[42] However, it is likely that some material also is transported from more southerly latitudes, similar to what was seen at Alert.

Factor F4 with the highest aerosol optical depth loading has peaks in 1983 and 1992. These peaks are probably related to high concentrations of stratospheric particles after eruptions of Nyamuragira (20 December 1981), Alaid (30 April 1981), El Chichon (4 April 1982) and Pinatubo (14 June 1991). Smaller 1980 and 1986 maxima may be associated with the eruptions of St. Helens (18 May 1980) and Nevada del Ruiz (13 November 1985) respectively. A decrease in integral aerosol optical thickness in the Arctic starting in March 1983 and in March 1992 after the eruptions of El Chichon and Mt. Pinatubo respectively has been reported by Radionov and Marshunova.[52] Maximum in aerosol turbidity was measured in the Arctic in April 1983 and in March 1992.[46,52,53] Since no measurements of aerosol optical depth are made during the polar night, an increase in aerosol optical depth was not observed in the Arctic until a year after an eruption. The time series for the aerosol optical depth factor at Barrow agrees well with the known behavior of volcanic stratospheric aerosol. Thus factor F4 is primarily attributed to the stratospheric part of aerosol optical depth. Three more factors (F5–F7) are mainly related to measured gas concentrations.

The hypothesis is that some factors are correlated with global temperature anomalies. Temperature

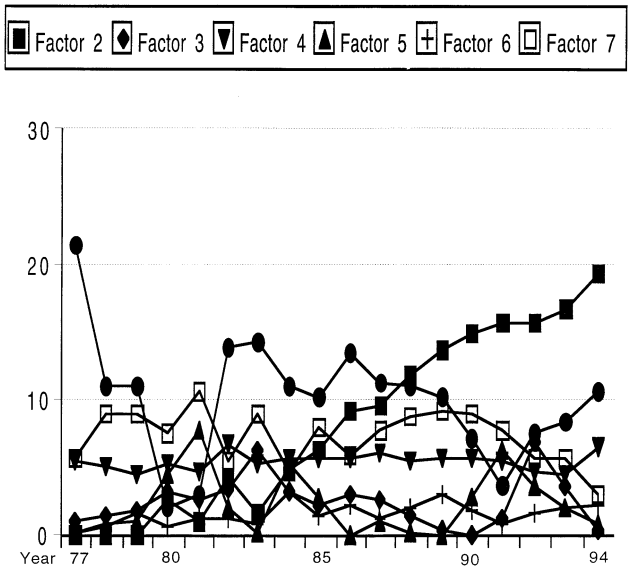

Figure 8. Factors of Example 2. Full circles are factor 1

anomaly data were not available for three of the years, so the Parafac factors were truncated from 18 to 15 years. The factors, 15 years $\times$ 7 factors, are tested in a regression model against temperature anomaly data. Partial least squares regression, PLS2, can handle correlated response variables. The factors are shown in Figure 8. Figures 6 and 8 show the loadings scaled as they came out of the PMF program for the examples.

Temperature anomalies

Temperature anomaly data are available from a number of agencies and research groups. These anomalies are useful in global warming research. They are also useful in air pollution studies where long-range transport of air masses and pollutants is considered.

The difference between the average temperature for a given period, such as a yearly average, and the average for the reference period is termed a temperature anomaly and this is the method by which climatic variations are described in atmospheric science. Global and hemispheric (N and S) annual temperature variations, relative to a 1950–1979 reference period, are available for the years 1854–1991, compiled by Jones[54] and Jones *et al.*[55–57] The anomaly estimates are based on corrected land and marine data. Land data were derived from meteorological data and fixed-position weather ship data that were corrected for non-climatic errors such as station shifts or instrument changes.[55,56] Temperature anomalies are also reported by Angell for the years 1958–1996.[58] They are based on radiosonde measurements for different latitude zones and for different pressure bands in the atmosphere.

The data of Jones, Wigley and Wright of the University of East Anglia for the years 1854–1991[54–57,59,60] form a data matrix of size 138 $\times$ 6. The six variables are northern, southern and global temperature anomalies and the same data repeated with a correction for the El Niño/Southern Oscillation (ENSO) effect. The temperature data are available from the web page http://cdiac.esd.ornl.gov.

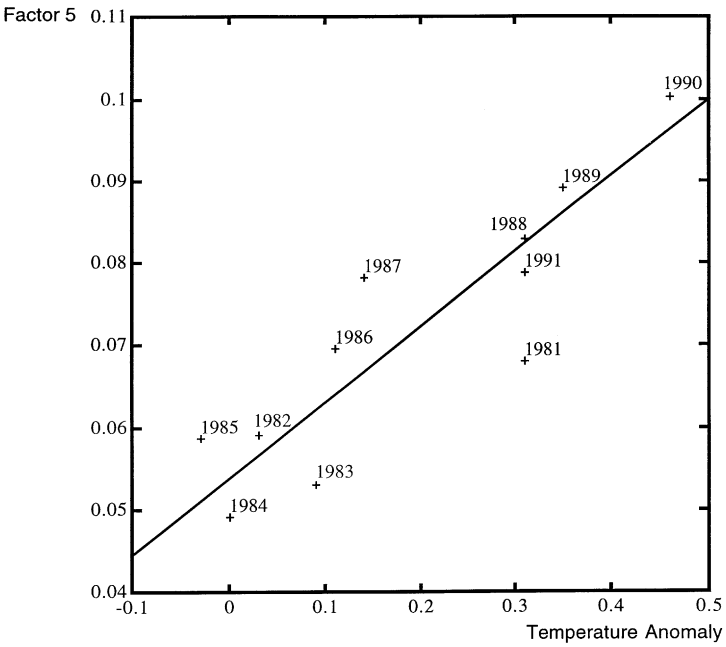

Figure 9. A plot of Parafac factor F5 for Example 1 against temperature anomalies shows a linear relationship. The line is the first principal component

Calculations

The PMF calculations were done with the PMF program of Paatero for DOS.[5–7] All regression calculations were done in Student Matlab for Macintosh.[61–63]

RESULTS AND DISCUSSION

Example 1

Figure 9 shows a plot of factor F5 of the Alert data against the Jones temperature anomalies. The relationship looks rather linear. Such relationships can be spurious, especially when only a small number of objects are available. This is a good reason for testing the complete Jones data with all five PMF factors in a regression model. The data matrices are $\mathbf{X}(11 \times 5)$ and $\mathbf{Y}(11 \times 6)$. $\mathbf{X}$ contains the five factors from PMF and $\mathbf{Y}$ contains the Jones temperature anomalies. PLS2 can handle multiple responses. The PMF factor data have to be mean-centered. The temperature anomaly data are left in their original scaling, but they are mean-centered. The factors in $\mathbf{X}$ have a range of standard deviations of 13 and it was considered wise to scale them to unit standard deviation. With the small number of objects available, cross-validation was considered unnecessary.

Table 3 shows the results obtained for the PLS models for a number of situations: (1) PLS2 of five PMF factors against the six Jones anomalies; (2) PLS2 of six PMF factors against the Jones data; (3) five PMF factors against the Angell temperature anomaly data for the northern hemisphere, a PLS1 model.

After three components, all the PLS models stabilize. Also the two-component models have a high R^2_y. With as few objects and variables available as is the case here, one should be careful to prevent

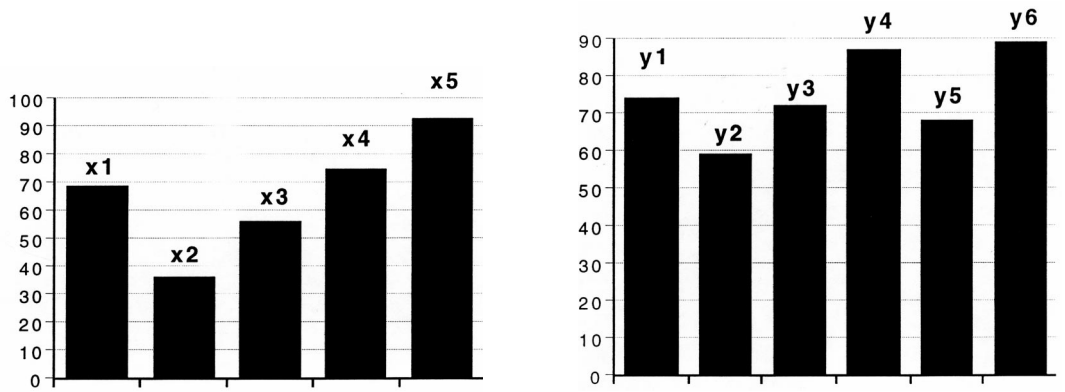

Figure 10. R^2_x and R^2_y for two-component PLS model between five PMF factors ($x1$–$x5$) and six Jones temperature anomalies ($y1$–$y6$)

overfitting. The model of the five PMF factors against the Jones data is best. It is interesting to have a look at what each variable contributes to this model. This is shown in Table 4.

The two-component PLS2 model from Table 4 is also represented as bar-plots in Figure 10. Table 4 and Figure 10 show that factor F5 ($x5$, biogenic) immediately contributes strongly to the PLS model. The anomalies modeled first and best are the ENSO-corrected versions (y-variables 4 and 6). The southern hemisphere anomalies are not well modeled. The Angell data do not give such a good model. Factors $x1$ and $x4$ have a reasonable contribution to the two-component PLS2 model. They represent photogenic processes ($x1$) and sea salt ($x4$). The anthropogenic factor ($x3$) and the soil factor ($x2$) do not seem to work very well in the regression models.

Figure 11 shows a plot of the PLS weights for the **X**-block for components 1 and 2. Also in the plot are the normalized loadings for the **Y**-block. The figure shows that factors F5 ($x5$) and F1 ($x1$) are the

Table 3. R^2_x and R^2_y for PLS2 model of Alert factors and temperature anomalies. Models for five Parafac factors (1), six Parafac factors (2) and five Parafac factors against Angell data (3)

	(1)		(2)		(3)	
Comp.	R^2_x	R^2_y	R^2_x	R^2_y	R^2_x	R^2_y
1	0·24	0·72	0·30	0·47	0·24	0·60
2	0·66	0·79	0·48	0·75	0·58	0·70
3	0·80	0·85	0·79	0·79	0·80	0·73
4	0·88	0·86	0·86	0·83	0·86	0·73
5	1·00	0·87	0·95	0·84	1·00	0·73

Table 4. Individual contributions of each variable to PLS models of rank 1–3. Highest values of R^2_x and R^2_y are in bold. Five PMF factors against Jones data

Rank	$x1$	$x2$	$x3$	$x4$	$x5$	$y1$	$y2$	$y3$	$y4$	$y5$	$y6$
1	13	10	2	1	**93**	71	17	60	**87**	51	**87**
2	69	36	56	75	**93**	74	59	72	87	68	**89**
3	92	39	86	90	**95**	86	60	81	**91**	76	90

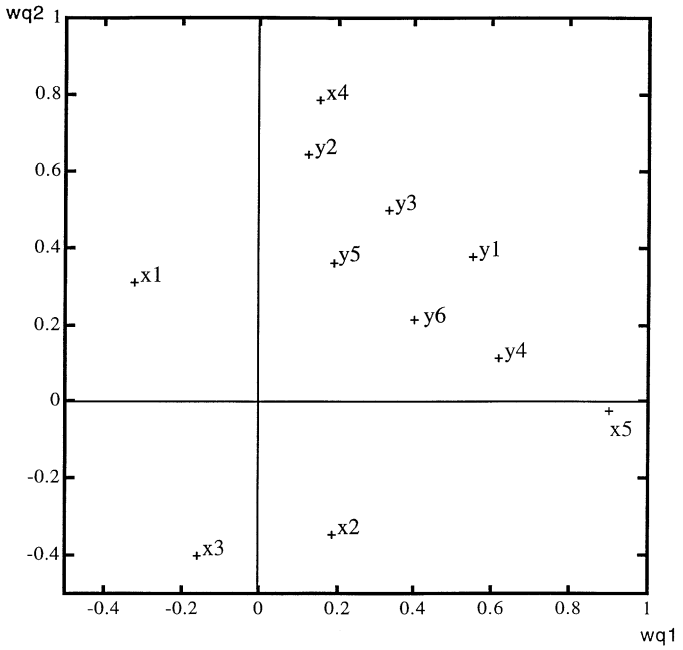

Figure 11. Plot of PLS weights for **X** (*w*) and normalized loadings for **Y** (*q*) for first two PLS dimensions. The plot shows the relationship between *x*- and *y*-variables

most important ones for component 1 that explains *y*4 the best. Factor F4 (*x*4) is the most important one for component 2 that explains the non-ENSO corrected anomalies (*y*1–*y*3) the best. There is also a pattern in the *y*-variables that was not immediately seen in the tables. The global temperature anomalies are the means of the northern and southern hemisphere values. This is seen in the figure.

A useful test is that of prediction ability. The predictor data (*x*1–*x*5) are tested against *y*4, the northern hemisphere anomaly after ENSO correction. The model has an R^2_y of 0·89. Validation was done by making models for ten objects and predicting the object that was left out. Scaling by the standard deviation was used for the **X**-block. In order to avoid overfitting, a two-component model was decided upon. Figure 12 shows the predicted values plotted against the tabulated ones. Taking into account the variability of environmental data, this linear relationship can be considered a good one. The calculated *Q*2 is 0·56. It can be noticed that two years, 1981 and 1982, behave as outliers. With these observations removed, *Q*2 becomes 0·81. The years that behave as outliers (1981 and 1982) are exactly those years when an El Niño effect occurred. This El Niño was the one before the most recent one of 1998.

These findings suggest that two regression models need to be made, one for El Niño years and one for non-El Niño years, but the few available years for the air pollution measurements prevent the construction of such models, since the data sets are too small.

Table 3 also shows the results of PLS regression for a six-factor Parafac model, assumed to be an improvement on the five-factor model. These results are not quite as good as for the five-factor model. The difference is even larger with non-scaled data. This means that the five-factor model is to be preferred over the six-factor model. In this way the regression models can be used for selecting a three-way rank in this example and in many others.

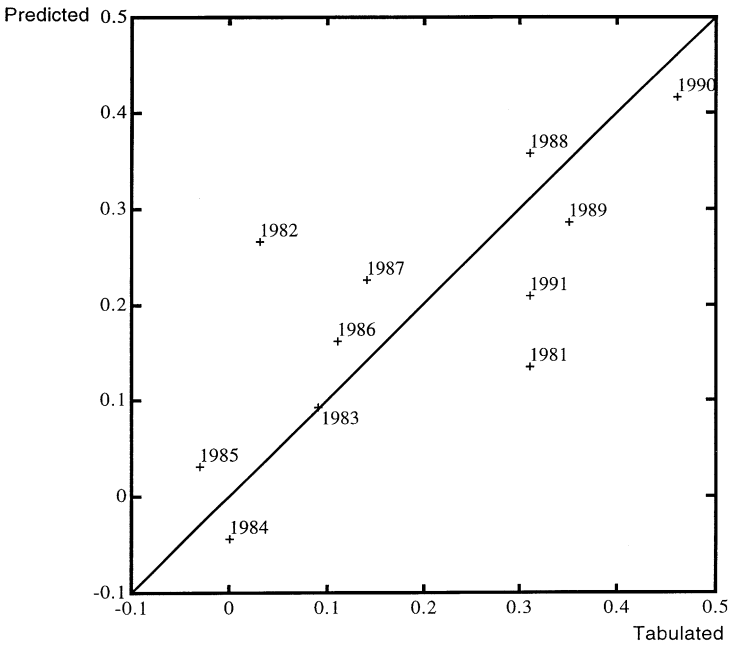

Figure 12. Plot of predicted temperature anomalies against tabulated ones. The years are indicated. A two-component PLS1 regression model is used

Example 2

The Jones temperature data and the Barrow data have corresponding years 1977–1991. For 1992–1994 there were no Jones temperature anomaly data available at the time of writing. The factors from PMF have quite a range of sizes and therefore some different types of scaling may be tested. A first PLS2 model is found in Table 5. This is very similar to what was done for Example 1. Taking into account the measurement noise and the risk of overfitting, it would be wise to select a two- to four-component PLS model. Cross-validation could be used, but it is not very meaningful on small data sets. Again the ENSO-corrected *Y*-variables are better modeled than the non-corrected anomalies. Figure 13 gives an overview of the contributions of the individual variables to a three-component

Table 5. PLS2 models for $\mathbf{X} \equiv$ seven PMF factors from Barrow data and $\mathbf{Y} \equiv$ Jones temperature anomalies. (1) All six *Y*-variables used. (2) Only three ENSO-corrected variables used

	(1)		(2)	
Comp.	R^2_x	R^2_y	R^2_x	R^2_y
1	0·56	0·57	0·56	0·66
2	0·75	0·62	0·74	0·71
3	0·94	0·68	0·94	0·75
4	0·97	0·73	0·96	0·79
5	0·99	0·76	0·99	0·80
6	1·00	0·78	1·00	0·81
7	1·00	0·78	1·00	0·81

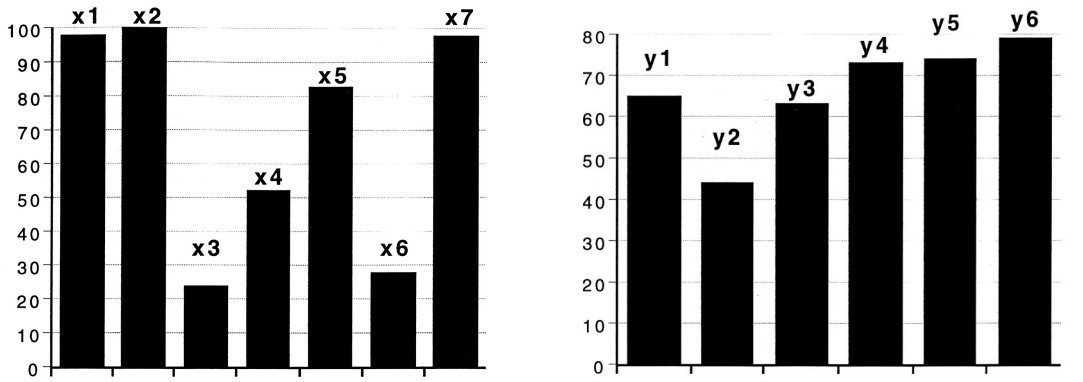

Figure 13. R^2_x and R^2_y for three-component PLS2 model between seven PMF factors of Barrow data ($x1$–$x7$) and six Jones temperature anomalies ($y1$–$y6$)

PLS2 model. Three factors have a high R^2_y: $x1$ and $x2$ (Arctic haze, small and large particles) and $x7$ ($CH_4 + CO_2$). Factor $x5$ ($CO + O_3 + CO_2$) has a medium-size contribution to the model. In principle, one could use the PMF factors for the years 1992–1994 to predict the Jones data for these years. There are, however, no Jones data with which to compare these predictions.

There is a possibility of also testing the local temperatures. Two sets of local temperatures and their standard deviations over the twelve months of the year are available for this purpose and they form four response variables. The PLS2 model of these four variables regressed against factors F1–F7 for 18 years gave an R^2_y of only slightly above 0·5 and therefore it may be assumed that yearly average local temperatures are not important for the extracted PMF factors.

The problem of scaling of the factors also shows up. Different PLS2 models with different interpretations may result from different scalings. Therefore a number of scalings were tested: (1) leaving the scales obtained from PMF as they are; (2) scaling by the inverse standard deviation of each factor; (3) scaling by average estimated errors as obtained from PMF. An overview of all the three-component PLS2 models is generated as follows: an X-variable gets one point if its R^2_x is in the range 0·9–1, and half a point if it is in the range 0·8–0·89. The points are accumulated over the three scaling versions. The results can be found in Figure 14. The second PMF factor $x2$ (Arctic haze, large particles) has a high R^2_x for all the scalings used. PMF factors $x3$ and $x4$ never contribute much to the PLS2 models. These factors represent condensation nuclei ($x3$) and volcanic dust ($x4$).

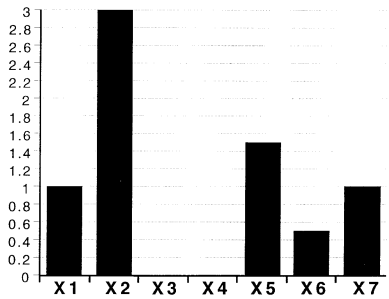

Figure 14. Points scored by accumulating R^2_x results for three PLS2 models with different scalings. PMF factor $x2$ (Arctic haze, large particles) does well for all scalings. PMF factors $x3$ and $x4$ are never much use in the models

CONCLUSIONS

For Example 1, PLS regression between factors from PMF and six temperature anomalies can be used to select PMF factors and anomalies that contribute most to the model. The northern hemisphere temperature anomaly can be modeled by the appropriate loadings (factors) from the Parafac model for Example 1, the Alert data. The two-component PLS model has good predictive properties. Outliers in the prediction are typical El Niño years. The biogenic factor plays an important role in this model. The second most important is the sea salt factor. The soil factor has no role at all. The five-component Parafac model is meaningful. A separate PLS test on a six-component model does not work equally well. PLS models allow the investigator to make sense out of oblique (correlated) Parafac factors and to determine the three-way rank of a Parafac model in a better way. It would of course have been better to have more data (more years). The limited number of years does not allow far-reaching conclusions.

For Example 2, the Barrow data, some of the PMF factors have a strong relation to the global temperature phenomena, while others are unrelated. Arctic haze (large particles) is strongly related to global temperatures, while volcanic ash is not. The ENSO-corrected anomalies also work best for this example. By using the local temperature means and standard deviations, it is shown that local temperature phenomena are not very important in relation to the PMF factors.

The scaling of factors is a very important issue and no good objective criteria are available. It can be assumed that the interpretation of the PLS2 models is dependent on the scaling of the PMF factors. However, some general conclusions can be made by accumulating results from PLS2 models with different scalings of the *X*-variables. The Arctic haze large particles factor contributes very well to all models, while condensation nuclei and volcanic ash do not contribute at all. Other PMF factors have a position in between.

It is an interesting prospect to try multiway regression methods, but already the combination technique where first the PMF-Parafac factors are extracted in a more or less subjective manner and where these factors then are used in a regression model against external data gives much useful information.

REFERENCES

1. P. Hopke (ed.), *Receptor Modeling for Air Quality Management*, Elsevier, Amsterdam (1991).
2. R. Coppi and S. Bolasco (eds), *Multiway Data Analysis*, North-Holland, Amsterdam (1989).
3. H. Law, C. Snyder, J. Hattie and R. McDonald (eds), *Research Methods for Multimode Data Analysis*, Praeger, New York (1984).
4. P. Paatero and U. Tapper, *Environmetrics*, **5**, 111 (1994).
5. P. Hopke, P. Paatero, H. Jia, R. Ross and R. Harshman, in press (1998).
6. P. Paatero, *Chemometrics Intell. Lab. Syst.* **37**, 23 (1997).
7. P. Paatero, *Chemometrics Intell. Lab. Syst.* **38**, 223 (1997).
8. H. Martens and T. Næs, *Multivariate Calibration*, Wiley, Chichester (1989).
9. K. Esbensen, S. Schönkopf and T. Midtgaard, *Multivariate Analysis in Practice*, CAMO, Trondheim (1994).
10. P. Brown, *Measurement, Regression and Calibration*, Clarendon, Oxford (1993).
11. R. Henrion and G. Henrion, *Multivariate Datenanalyse: Methodik und Anwendung in der Chemie und verwandten Gebieten*, Springer, Berlin (1995).
12. A. Höskuldsson, *Prediction Methods in Science and Technology*, Thor, Holte (1996).
13. P. Geladi, *J. Chemometrics*, **2**, 231 (1988).
14. R. Ball, *Appl. Statist.* **12**, 14 (1963).
15. S. Wold, P. Geladi, K. Esbensen and J. Öhman, *J. Chemometrics*, **1**, 41 (1987).
16. L. Ståhle, *Chemometrics Intell. Lab. Syst.* **7**, 95 (1989).
17. L. Ståhle, *J. Pharmaceut. Biomed. Anal.* **9**, 671 (1991).
18. P. Nomikos and J. MacGregor, *Chemometrics Intell. Lab. Syst.* **30**, 97 (1995).
19. L. Nørgaard, *J. Chemometrics*, **10**, 615 (1996).

20. R. Bro, *J. Chemometrics*, **10**, 47 (1996).
21. J. Mitchell, *J. Atmos. Terr. Phys.*, Spec. suppl., p. 195 (1957).
22. K. Rahn, *Atmos. Environ.* **15**, 1345 (1981).
23. L. Barrie, *Atmos. Environ.* **20**, 643 (1986).
24. G. Shaw and M. Khalil, in *The Handbook of Environmental Chemistry*, ed. by O. Hutziger, Vol. 4, Part B, p. 69, Springer, Berlin (1989).
25. I. Sokolik, *Izv. Russ. Acad. Sci., Atmos. Oceanic Phys.* **28**, 509 (1992).
26. A. Vinogradova, *Izv. Russ. Acad. Sci., Atmos. Oceanic Phys.* **29**, 437 (1993).
27. G. Shaw, *Bull. Am. Meteorol. Soc.* **76**, 2403 (1995).
28. K. Rahn, *Atmos. Environ.* **19**, 1987 (1985).
29. W. Raatz and G. Shaw, *J. Climate Appl. Meteorol.* **23**, 1052 (1984).
30. W. Raatz, *Atmos. Environ.* **23**, 2501 (1989).
31. J. Miller, *Atmos. Environ.* **15**, 1401 (1981).
32. T. Iversen and E. Joranger, *Atmos. Environ.* **19**, 2099 (1985).
33. L. Barrie, R. Hoff and S. Daggupaty, *Atmos. Environ.* **15**, 1407 (1981).
34. L. Barrie, M. Olson and K. Oikawa, *Atmos. Environ.* **23**, 2505 (1989).
35. L. Barrie and R. Hoff, *Atmos. Environ.* **18**, 2711 (1984).
36. L. Barrie and M. Barrie, *J. Atmos. Chem.* **11**, 211 (1990).
37. G. Sakunov, A. Timirev and O. Barteneva, *Soviet Meteorology and Hydrology*, transl. by Allerton Press, No. 2, p. 53 (1990).
38. Y.-L. Xie, P. Hopke, P. Paatero, L. Barrie and S. -M. Li, *J. Atmos. Sci.* in press (1998).
39. L. Barrie, J. Bottenheim, R. Schnell, P. Crutzen and R. Rasmussen, *Nature*, **334**, 138 (1988).
40. S. Li and L. Barrie, *J. Geophys. Res.* **98**, 20,613 (1993).
41. S. Li, L. Barrie and A. Sirois, *J. Geophys. Res.* **98**, 20,623 (1993).
42. R. Ferek, P. Hobbs, L. Radke and J. Herring, *J. Geophys. Res.* **100**, 26,093 (1995).
43. B. Bodhaine, *Atmos. Environ.* **23**, 2357 (1989).
44. B. Bodhaine, *J. Geophys. Res.* **100**, 8967 (1995).
45. B. Bodhaine and E. Dutton, *Geophys. Res. Lett.* **20**, 947 (1993).
46. V. Radionov, M. Marshunova, Ye. N. Rusina, K. Ye. Lubo-Lesnichenko and Yu. Ye. Pimanova, *Izv. Russ. Acad. Sci., Atmos. Oceanic Phys.* **30**, 762 (1995).
47. B. Bodhaine and E. Dutton, *Geophys. Res. Lett.* **22**, 741 (1995).
48. D. Jaffe, T. Iversen and G. Shaw, *Geophys. Res. Lett.* **22**, 739 (1995).
49. A. Polissar, P. Hopke, P. Paatero, Y. Kaufman, D. Hall, B. Bodhaine and E. Dutton, submitted (1998).
50. R. Charlson, J. Lovelock, M. Andreae and S. Warren, *Nature*, **326**, 655 (1987).
51. G. Shaw, *Climate Change*, **5**, 297 (1983).
52. V. Radionov and M. Marshunova, *Izv. Russ. Acad. Sci., Atmos. Oceanic Phys.* **29**, 549 (1994).
53. V. Radionov and M. Marshunova, *Atmosphere–Ocean*, **30**, 531 (1992).
54. P. Jones, *Climate Monitor*, **17**, 80 (1988).
55. P. Jones, T. Wigley and P. Wright, *Nature*, **322**, 430 (1986).
56. P. Jones, S. Raper, R. Bradley, H. Diaz, P. Kelly and T. Wigley, *J. Climate Appl. Meteorol.* **25**, 161 (1986).
57. P. Jones, S. Raper and T. Wigley, *J. Climate Appl. Meteorol.* **25**, 1213 (1986).
58. J. Angell, *Mon. Weather Rev.* **114**, 1922 (1986).
59. K. Briffa and P. Jones, *Holocene*, **3**, 82 (1993).
60. P. Jones and K. Briffa, *Holocene*, **2**, 105 (1992).
61. P. Geladi, H. Martens, L. Hadjiiski and P. Hopke, *J. Near Infrared Spectrosc.* **4**, 243 (1996).
62. P. Geladi, *Workbook: Regression for Calibration in Matlab*, Umeå University (1997).
63. The Math Works, *The Student Edition of MATLAB*, Prentice-Hall, Englewood Cliffs, NJ (1992).

