Fresenius J Anal Chem (1999) 363 : 655–661

ORIGINAL PAPER

J. W. Einax · A. Aulinger · W. v. Tümpling
A. Prange

# Quantitative description of element concentrations in longitudinal river profiles by multiway PLS models

Received: 13 March 1998 / Revised: 29 October 1998 / Accepted: 17 November 1998

**Abstract** Partial least squares (PLS) models were used to examine the relationships between the distributions of elements in different compartments of a river. These relationships, if existing, enabled predictions to be made of the element concentrations in one compartment by knowing the concentrations in another compartment. The subjects of the study were the element concentrations measured in the water and the sediment of the river Saale as well as in the water and the suspended matter of the river Elbe. Special emphasis was placed on a comparison between two-way and three-way PLS.

## Introduction

Within the framework of a research project concerned with an evaluation of the water quality of the catchment area of the river Elbe and its tributaries, several sampling campaigns on the longitudinal profile of the rivers Saale and Elbe [1–4] have been carried out. The concentrations of certain elements in the sediments, the suspended matter and the dissolved matter were determined and additionally some sum parameters in the river water. One aim of the project was to examine whether a relationship between these compartments of the river water can be shown and if it is possible to deduce the contents of elements in one compartment from known contents of elements in another compartment.

As an example, predictions of element concentrations in the sediments of the river Saale were made from a knowledge of the concentrations in the river water. In the same way, element concentrations were predicted in the suspended matter of the river Elbe from the known concentrations in the water. For this purpose, two- and three-way partial least squares regression models were used.

J. W. Einax
Institute of Inorganic and Analytical Chemistry,
Friedrich Schiller University, D-07743 Jena, Germany

A. Aulinger · W. v. Tümpling · A. Prange
GKSS Forschungszentrum, Max-Planck-Strasse,
D-21494 Geesthacht, Germany

## Theory

It is becoming more and more common in analytical chemistry to calibrate an analytical method by multivariate calibration [5, 6]. A relatively modern method in this field is partial least squares regression (PLSR or PLS). The strength of this method is the ability to eliminate collinearities and noise within the matrix of predictors **X** [7]. This is achieved by extracting partial matrices $\mathbf{X_p}$ from **X**, where $\mathbf{X_p}$ contains – in the optimum case – only such information that is relevant for the description of the matrix of the predictants **Y**. Each partial matrix $\mathbf{X_p}$ is represented by a set of orthogonal vectors **t** and **w**. The vector **t** contains the scores of the objects in $\mathbf{X_p}$ on the latent variable **w**. Likewise two orthogonal vectors **u** and **q** are calculated to reproduce $\mathbf{Y_p}$ as a partial matrix of **Y**. Furthermore, the latent variables are computed under the constraint of maximum covariance between **t** and **u**, in order to optimize the predictive capability of the regression model (Fig. 1, Eq. 1, 2). For a more detailed description of the theory and algorithms see [8–10].

$$x_{ij} = t_i w_j + rx_{ij} \quad (1)$$

$$y_{ij} = u_i q_j + ry_{ij} \quad (2)$$

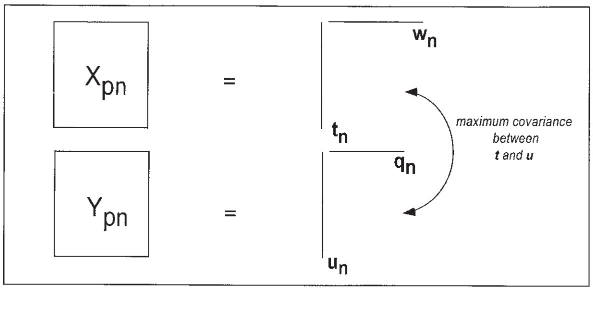
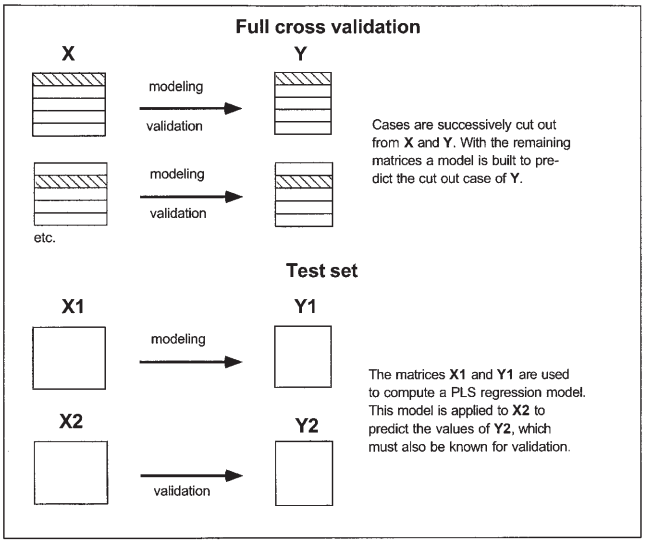


**Fig. 1** Reproduction of the partial matrices $\mathbf{X_p}$ and $\mathbf{Y_p}$ with sets of orthogonal vectors **t**, **w** and **u**, **q**, respectively



Cases are successively cut out from **X** and **Y**. With the remaining matrices a model is built to predict the cut out case of **Y**.

The matrices **X1** and **Y1** are used to compute a PLS regression model. This model is applied to **X2** to predict the values of **Y2**, which must also be known for validation.

**Fig. 2** Validation of the PLS models

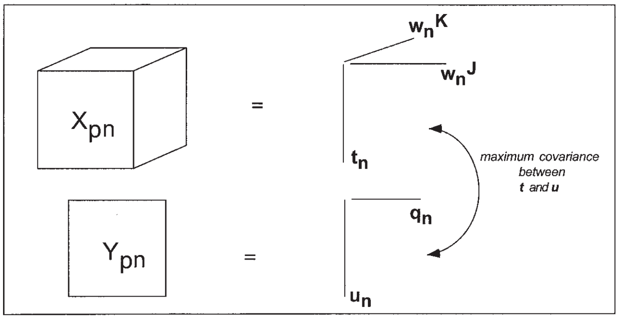


**Fig. 3** Reproduction of a three-dimensional partial matrix $\mathbf{X_p}$ by a set of orthogonal vectors **t**, $\mathbf{w^j}$, $\mathbf{w^k}$

Constraint:

$$\max_{w}\left[\operatorname{cov}(t,u)\middle|\min\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij}-t_iw_i\right)^2\right)\wedge\|w\|=1\right] \quad (3)$$

$x_{ij}$ and $y_{ij}$ are the matrix elements of **X** and **Y**, $t_iw_i$ and $u_iq_i$ are the calculated values of the PLS model and $rx_{ij}$, $ry_{ij}$ are the differences between the model and the real data matrices. Thus, partial matrices $\mathbf{X_p}$ are successively extracted from the initial **X** matrix by minimizing the residuals between the matrix elements in **X** and $\mathbf{X_p}$ by the least squares method. Finally, the partial matrices of **X** are unified into a model matrix $\mathbf{X_{mod}}$, which is used to perform a calibration with **Y**.

In order to decide which model matrices, represented by a certain number of latent variables, are the best for predicting unknown y values, the root mean squared error of prediction (RMSEP) is a useful tool [5]. It contains the differences between the measured and the predicted y values ($y^{meas}$, $y^{pred}$):

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij}^{meas}-y_{ij}^{pred}\right)^2}{I\cdot J}} \quad (4)$$

The model that generates the lowest RMSEP is the best one for the particular calibration problem.

Apparently, for validating the model one needs to obtain a **Y** matrix with known values. For this purpose one can use the same **Y** matrix as for calibration or, if available, a test **Y** matrix (Fig. 2).

As shown in Fig. 3 and Eq. 5 and 6, PLS theory can easily be extended to treat a three-dimensional **X** matrix. The difference is that two loading weights vectors have to be calculated for each latent variable instead of one [10–14]. The idea behind this extension of the PLS algorithm to three or n-way modeling is comparable to some other n-way decomposition methods such as Tucker3 or PARAFAC [15–17].

$$x_{ijk} = t_iw^J{}_jw^K{}_k + rx_{ijk} \quad (5)$$

Constraint:

$$\max_{w^Jw^K}\left[\operatorname{cov}(t,u)\middle|\min\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{ijk}-t_iw_j^Jw_k^K\right)^2\right)\right. \quad (6)$$

$$\left.\wedge\|w^J\|=1,\|w^K\|=1\right]$$

It is, of course, also conceivable to perform a calibration with a three-dimensional **Y** matrix, but this was not investigated in this study, because it is not relevant for this particular practical application.

## Experimental

### Sampling in the river Saale

In the river Saale 17 variables in the sediments and 24 variables in the water at 23 sampling locations along the river were measured [1, 2]. Samplings of the sediments were executed in October 1993, June 1994 and in June 1995, and of the river water monthly between September 1993 and August 1994. The measured element concentrations that have been drawn up as variables in matrices are listed in Table 1.

The water matrix served as matrix of predictors **X** for the prediction of the single sediment matrices one after another; for the two-dimensional models the yearly medians of the variables in **X** were used (Fig. 4).

**Table 1** Variables used for the models of the river Saale

| | |
|---|---|
| Variables in the river water | Redox potential, pH value, conductivity, dissolved oxygen, temperature, DOC, AOX, suspended matter, As, Fe, Mn, Cu, Cr, phosphate, chloride, nitrite, nitrate, sulfate, Mg, Ca, Na, K, Zn, Ni |
| Variables in the sediments | Hg, As, Se, Cd, Zn, Pb, Ni, Cr, Cu, Co, Mn Fe, Mg, Ca, Na, K, TOC |

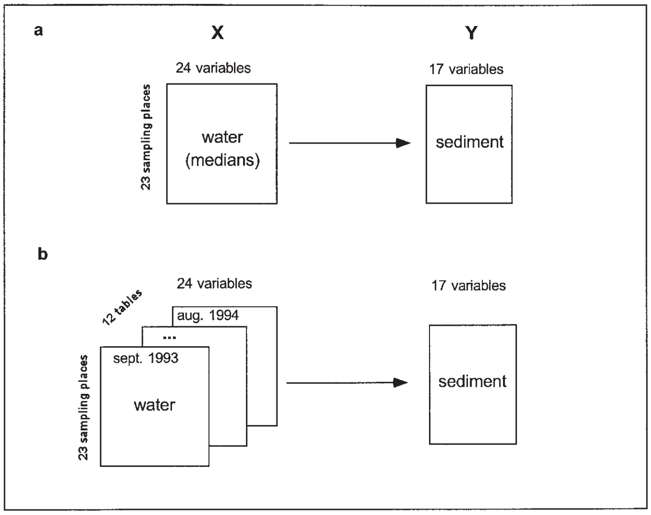


**Fig. 4 a, b** Two (**a**)- and three (**b**)-way-PLS models for the river Saale

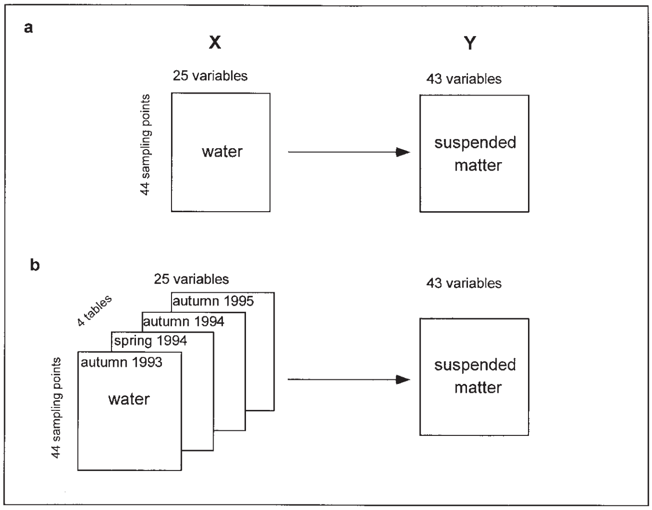


**Fig. 5 a, b** Two (**a**)- and three (**b**)-way-PLS models for the river Elbe

### Sampling in the river Elbe

Whereas for validating the Saale models only full cross validation was possible, as there was only one **X** matrix, the Elbe models could be validated with test sets, because an **X** and a **Y** matrix existed for each campaign [3, 4]. For modeling the river Elbe 43 variables in the suspended matter and 25 variables in the river water at 44 sampling locations along the river Elbe were used. These were measured in four campaigns in autumn 1993, spring 1994, autumn 1994 and autumn 1995 (Table 2, Fig. 5).

Before constructing the PLS models, the matrices had been centered and scaled in such a way that the mean of all variable columns was zero and their standard deviation was one, whereas two versions of calculating the variable means and standard deviations were possible: They could be calculated over sampling locations and sampling dates or over sampling locations only. As the former produced PLS models with a higher prediction ability this method was preferred.

## Results and discussion

### Models of the sediments in the river Saale

Using the yearly medians of the variables in the river water as **X** matrix, the best PLS model resulted from the sediment matrix of June 1994. This is due to the fact that in April 1994 there was high water in the river Saale, which caused the sediment to be washed out and newly formed by the Saale water. The relationship between the water and the sediment – necessary for modeling – was therefore clearer in June 1994 than in September 1993 when the Saale water flowed over old sediment layers [18]. The reason why the prediction of the element concentrations in the Saale sediments of June 1995 becomes poor again is probably related to the time interval between the two sampling events. Even better were the fully cross-validated predictions of the sediment matrix in June 1994 when information about temporal variance of the samples was introduced to the model (Table 3). Hereby, all the monthly measured values were used instead of the yearly medians. This can be done in one of two ways: Either the 12 monthly data tables with the 24 variables measured at 23 sampling locations are strung in a row. This results in a matrix with 23 rows and 288 columns, that can be used to perform a two-way PLS regression (Fig. 1). The other possibility is to stack the 12 data tables together to form a cuboid and to apply a three-way model such as three-way PLS (Fig. 3). However, the prediction ability of the three-way model is not clearly superior to the two-way PLS model with the two-dimensional 23 × 288 matrix. The lowest achievable RMSEP by the three-way model is 0.85 in contrast to 0.86 achievable by applying two-way PLS to the two-dimensional complete matrix.

In Table 3 the differences in predictive quality between the two-way (medians) and the three-way model are quantified as the mean percentile differences between predicted and measured element concentrations *mDiff%* in the sediment of June 1994 (Eq. 7). The elements Cd, Ni, Co, Fe, Mg and K are taken as examples.

**Table 2** Variables used for the models of the river Elbe

| | |
|---|---|
| Variables in the river water | Li, B, Na, Mg, Al, S, K, Ca, Ti, Mn, Fe, Co, Cu, Zn, As, Rb, Sr, Mo, Sb, Ba, content of suspended matter, temperature, dissolved oxygen, pH value, conductivity |
| Variables in the suspended matter | Li, Na, Al, Mg, K, Ca, Sc, Ti, V, Mn, Fe, Co, Ni, Zn, As, Rb, Sr, Y, Nb, Mo, Ag, Sb, Cs, Ba, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, W, Th, U |

**Table 3** Mean percentile differences for certain elements in the sediment matrix of the Saale campaign in June 1994 calculated with the two-way PLS model applied on the yearly medians of the parameters measured in the river water (*mDiff%1*), and with the three-way PLS model (*mDiff%2*)

| | Cd | Ni | Co | Fe | Mg | K |
|---|---|---|---|---|---|---|
| ***mDiff%1*** | 41.41 | 25.53 | 23.44 | 10.79 | 22.31 | 24.10 |
| ***mDiff%2*** | 36.87 | 21.69 | 21.41 | 10.62 | 18.94 | 21.29 |

$$mDiff\% = \frac{\sum_{i=1}^{I} \frac{\left|y_i^{meas} - y_i^{pred}\right|}{y_i^{meas}}}{I} \cdot 100\% \qquad (7)$$

$y_{meas}$, $y_{pred}$ are the measured and the predicted element concentrations; *I* is the number of sampling locations.

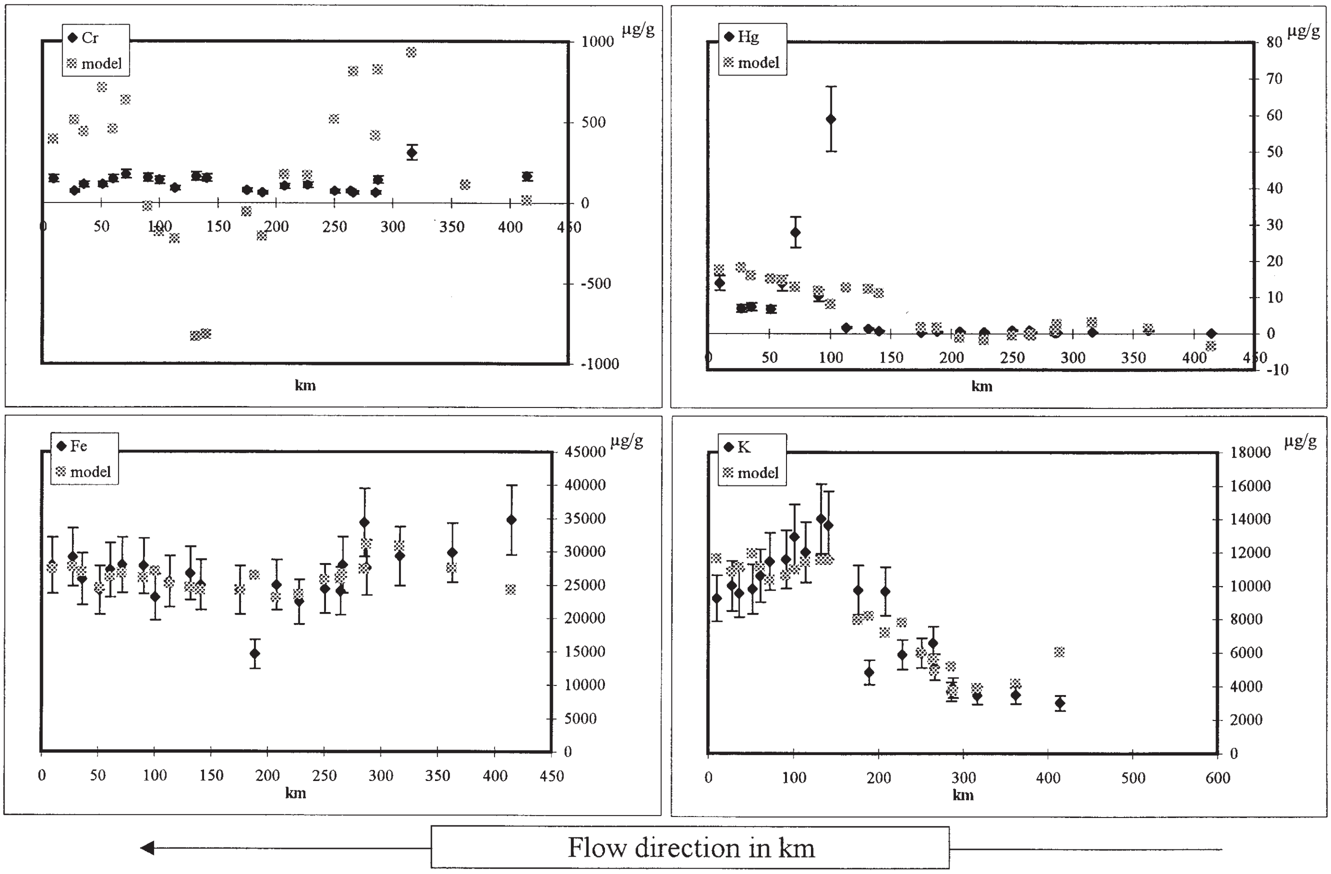
The quality of the prediction is not the same for each variable in **Y**, but for most of the variables it can be seen that the model is able to reproduce the longitudinal curves. The condition for a good model is that the distribution of the variable contents along the river is systematic and not random. This is the reason why the model for e.g. Cr yields very bad results. The predictions for Hg are a little better. However, there is a huge deviation at km 100. In this area the sediment contains much more Hg than in the other part of the river because of the former chlorine alkaline electrolysis in the chemical factory at Buna [1, 2, 18]. This indicates that it is problematic for the model to follow large changes in the sediment body caused by a strong influence of point sources on the condition of the water body. As a result, the extremely high content of Hg at the above sampling location cannot be caught by the model. Examples for a good fit of the modeled data with the real data are represented by Fe and K.

Figure 6 shows the measured and the predicted values of Cr, Hg, Fe and K in the sediment of June 1994 in the longitudinal profile of the river Saale. The predictions were made with the three-dimensional PLS model as described above. The bright markers represent the modeled values and the dark markers the measured ones with an error bar of 15%, which corresponds to a mean total measurement uncertainty.

## Models of the suspended matter in the river Elbe

When the sediment lies in the river bed with the water flowing over it, the suspended matter is transported along with the water body for some distance. One can therefore suppose that a relationship between the suspended matter and the water of a river is more likely than between the sediment and the water. Therefore, the modeling of the Elbe data was expected to produce better results. The two compartments had been separated by filtration immediately after sampling to conserve this relationship as well as possible.

**Fig. 6** Four examples for the prediction ability of the three-way-PLS models for the river Saale



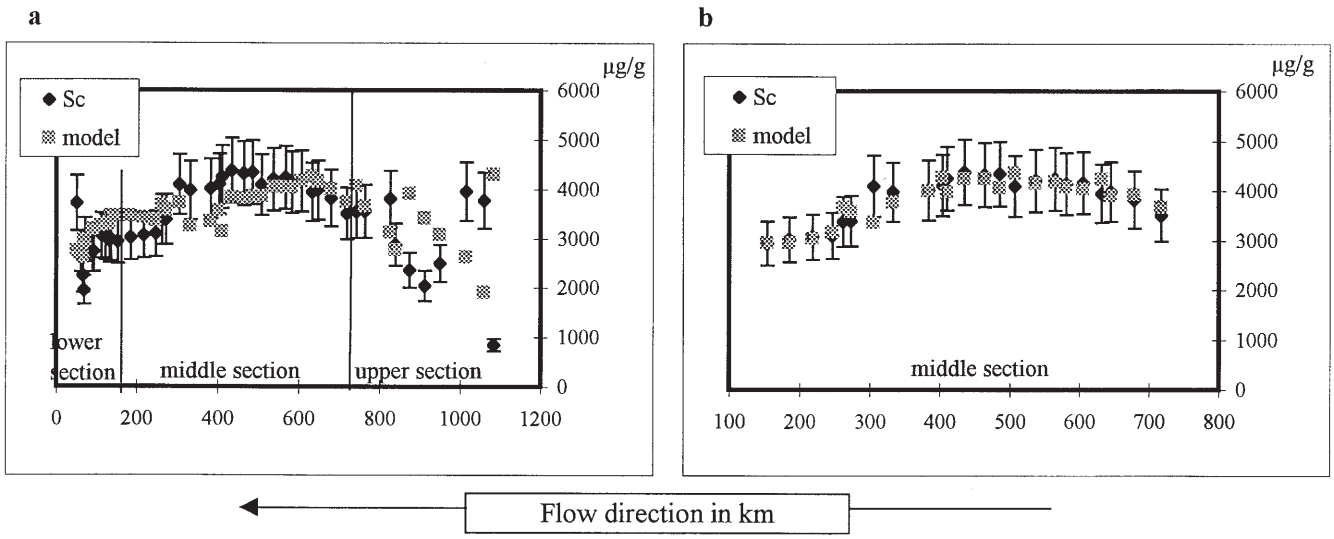


**Fig. 7 a, b** Full cross-validated two-way-PLS models for Sc in the suspended matter of the campaign of autumn 1995. Comparison between the model of the whole river (**a**) and the middle section (**b**)

It is, of course, also true for the modeling of the suspended matter in the river Elbe that it is difficult for the model to grasp large changes in the condition of the water body, as could be seen in the models computed with the data sets of the river Saale. Between the German-Czech border and the weir at Geesthacht the water body is more or less homogenous. In the upper part of the river Elbe the water body is different to that in the middle part because the river is carrying little water and is influenced by a large number of different point sources. In the lower part the influence of the North Sea dominates the water body [3, 4]. Omitting the upper and the lower part and computing a model just for the middle section, which represents the longest part of the river with the most sampling locations, can improve the predictions significantly, as illustrated for Sc in Fig. 7.

So far, the models have only been validated by the full cross-validation method. A better indication of the applicability of the models in practical use is obtained when they can be tested by a test set. For this purpose a regression was executed for the suspended matter and the river water of a certain campaign. The resulting correlation coefficients matrix was applied to the water matrix of another campaign to calculate the accompanying suspended matter matrix and to compare it with the measured values of this campaign.

It was found that the deviations became larger when predicting the suspended matter of an autumn matrix with a model computed with a set of matrices of another autumn campaign and even worse when predicting the same matrix with a model computed with spring matrices. This is the result of differences between the condition of the water body at the time when the samples for the modeling are taken and when the predictions for another campaign are made. The differences in the condition of the river are obviously larger between an autumn and a spring campaign than between two autumn campaigns. So, for example, the deviation between the predicted and the measured values in the suspended matter of autumn 1995 is larger when the regression model to perform these predictions is computed with water and suspended matter measured in spring 1994 than with those measured in autumn 1994 (Table 4).

One can also distinguish between variables whose different behavior in spring and autumn is clearly visible and those whose behavior is similar in spring and in autumn. For the first group of variables the predictions for the suspended matter of autumn 1995 with the model of spring 1994 are very poor (Fig. 8 a). However, the other group of variables can be predicted in the matrix of autumn 1995 with tolerable deviations even with the model of spring 1994 (Fig. 8 b).

Furthermore, computation of a three-way PLS model as described before was performed for the river water measured in autumn 1993, spring 1994 and autumn 1994 as **X** and the suspended matter measured in autumn 1994 as **Y** matrix. The resulting correlation coefficients matrix then was used to predict the element concentrations in the suspended matter of autumn 1995. In this three-way model the variable temporal behavior of the elements was taken into account. This caused a significant improvement in the prediction of the concentrations in the suspended matter, in particular for those elements that could not be modeled by the two-way method since they showed different behavior in autumn and spring (Fig. 8).

**Table 4** Mean percentile differences for certain elements in the suspended matter matrix of the Elbe campaign in autumn 1995 calculated with the PLS models of autumn 1994 (*mDiff%1*), spring 1994 (*mDiff%2*) and with the three-way PLS model (*mDiff%3*)

| | Li | Fe | Sr | Mg | Nd | Zn | Sc |
|---|---|---|---|---|---|---|---|
| *mDiff%1* | 18.9 | 13.1 | 24.4 | 19.3 | 16.4 | 39.0 | 21.7 |
| *mDiff%2* | 35.7 | 32.0 | 31.5 | 29.9 | 36.1 | 40.0 | 41.9 |
| *mDiff%3* | 23.1 | 14.4 | 20.4 | 20.8 | 18.6 | 23.5 | 22.9 |

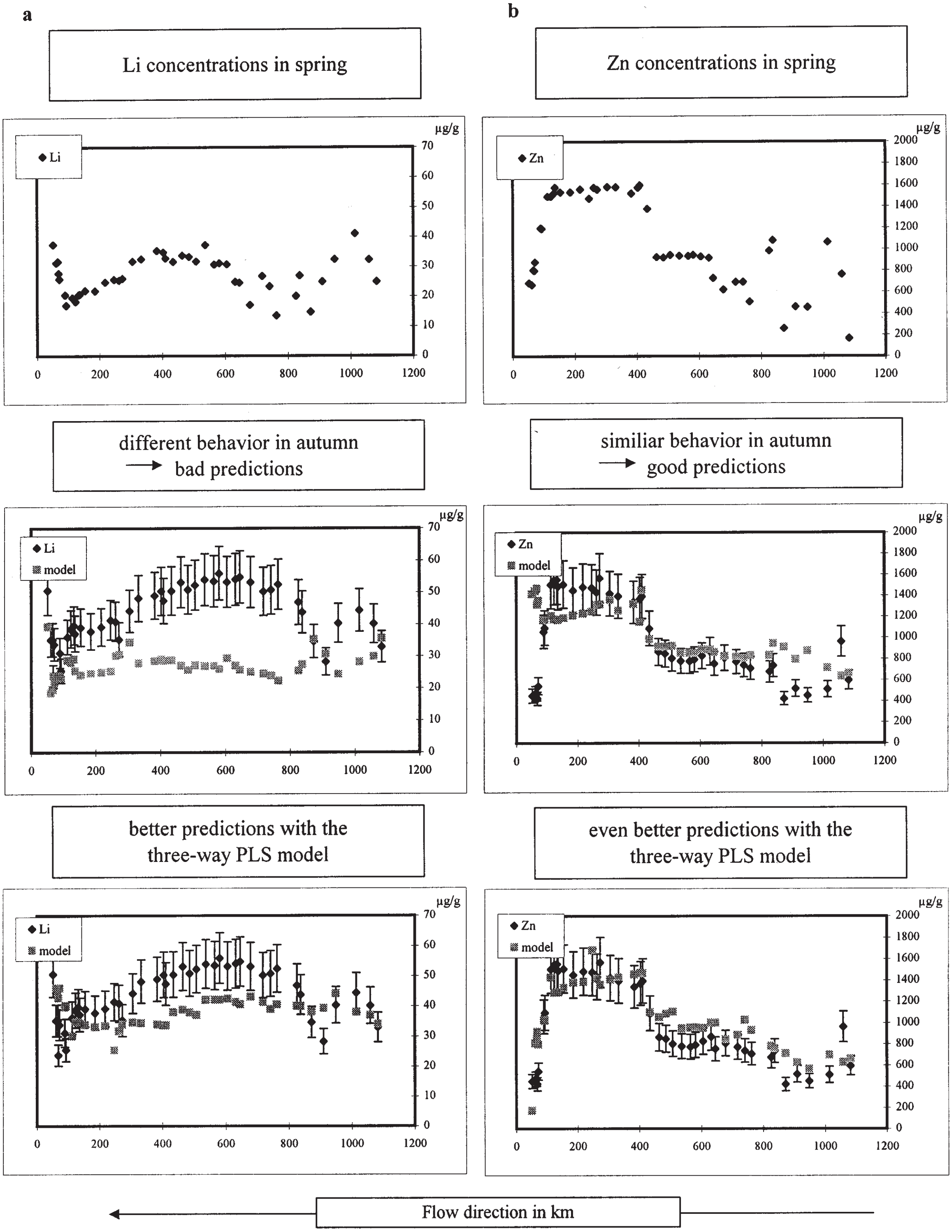


**Fig. 8 a, b** Influence of temporal variability of element concentrations on the predictive ability of a two-way PLS model and its compensation by applying a three-way PLS model, illustrated using the examples of Li (**a**) and Zn (**b**)

## Conclusions

In principle, PLS is suitable for quantitative descriptions of the relations between the different compartments of a river and to predict certain element concentrations along the longitudinal profile. As expected, the modeling of the river Elbe yields more accurate results than the modeling of the river Saale because of the greater correlation between the river water and the suspended matter (Elbe data) than between the river water and the sediment (Saale data). It is not possible to predict values that differ too much from the other values in the data set. Further, modeling becomes problematic if the condition of the river in its longitudinal profile is too variable or if the condition of the river is too different at the time the modeling is performed and at the time when the predictions are made. Especially the latter effect, however, can be tempered by using a three-way PLS model. In any case, for any particular problem, it is necessary to check whether and under what conditions PLS is applicable.

## References

1. Truckenbrodt D, Kampe O, Einax J W (1994) Zur aktuellen Belastungssituation der Saale, Ilm und Unstrut. In: Statusberichte: Die Belastung der Elbenebenflüsse mit Schadstoffen. BMBF; pp 75
2. Truckenbrodt D (1996) Diss A: Untersuchungen zur Belastungssituation in der Saale unter besonderer Berücksichtigung der Schwemetallgehalte der Sedimente. Friedrich-Schiller-Universität, Jena
3. Prange A, v Tümpling jr W, Niedergesäß R, Jantzen E (1995) Wasserwirtschaft Wassertechnik 7 : 22–33
4. Pepelnik R, Prange A, Jantzen E, Krause P, v Tümpling jr W (1997) Fresenius J Anal Chem 359 : 346–351
5. Eriksson L, Hermans J L M (1995) In: Einax J (ed) Chemometrics in environmental chemistry, vol II. Springer, Berlin Heidelberg New York, pp 135
6. Martens H, Naes T (1989) Multivariate calibration. Wiley, Chichester
7. Henrion R, Henrion G (1995) Multivariate Datenanalyse. Springer, Berlin Heidelberg New York
8. Lorber A, Wangen L E (1987) J Chemometrics 1 : 19–31
9. Geladi P (1986) Anal Chim Acta 185 : 1–17
10. Wold S (1987) J Chemometrics 1 : 41–56
11. Bro R (1996) J Chemometrics 10 : 47
12. Xie Y L, Baezabaeza J J, Ramisramos G (1995) Chemom Intell Lab Sys 27 : 211–220
13. Smilde A K (1997) J Chemometrics 11 : 367–377
14. de Jong S (1998) J Chemometrics 12 : 77–81
15. Kroonenberg P M (1992) Statistica Applicata 4 : 619–633
16. Henrion R (1993) J Chemometrics 7 : 477–494
17. Bro R (1997) Chemometrics Intell Lab Sys 38 : 149–171
18. Einax J W, Kampe O, Truckenbrodt D (1998) Fresenius J Anal Chem 361 : 149–154