Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Geochemistry 20 (2005) 2233–2251
Applied Geochemistry
www.elsevier.com/locate/apgeochem

Sequential Factor Analysis as a new approach to multivariate analysis of heterogeneous geochemical datasets: An application to a bulk chemical characterization of fluvial deposits (Rhine–Meuse delta, The Netherlands)

Pieter-Jan van Helvoort [a,*], Peter Filzmoser [b], Pauline F.M. van Gaans [c]

[a] School of Geography and the Environment, University of Oxford, Oxford, United Kingdom
[b] Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria
[c] Department of Physical Geography, Utrecht University, Utrecht, The Netherlands

Received 20 September 2004; accepted 4 August 2005
Editorial handling by Å. Danielsson
Available online 4 November 2005

Abstract

Sequential Factor Analysis (seqFA) is presented here as an enhanced alternative to multivariate factorial techniques including robust and classical Factor Analysis (FA) or Principal Component Analysis (PCA). A geochemical data set of 145 sediment samples from very heterogeneous, mainly riverine, deposits of the Rhine-Meuse delta (The Netherlands) analyzed for 27 bulk parameters was used as a test case. The innovative approach explicitly addresses the priority issues when performing PCA or FA: heterogeneity and overall integrity of the data, the number of factors to be extracted, and which optimum minimal set of key variables to be included in the model. The stepwise decision process is based on quantitative and objectively derived statistical criteria, yet also permitting arguments based on geochemical expertize. The results show that seqFA, preferably in combination with robust methods, yields a highly consistent factor model, and is favorable over classical methods when dealing with heterogeneous data sets. It optimizes rotation of the factors, and allows the extraction of less distinct factors supported by only a few variables, thus uncovering additional geochemical processes and properties that would easily be missed with other approaches. The identification of key variables simplifies the geochemical interpretation of the factors, and greatly facilitates the construction of a geochemical conceptual model. For the case of the fluvial deposits, the conceptual model effectively describes their bulk chemical variation in terms of a limited number of governing processes.

1. Introduction

Factor Analysis (FA) and Principal Component Analysis (PCA) are widely used statistical techniques in environmental geochemistry. These multivariate approaches are used to reduce the large

* Corresponding author. Present address: Zeemanlaan 166, 3572ZH, The Netherlands. Fax: +31 30 2564755.
E-mail addresses: phelv@ouce.ox.ac.uk, helvoort@gmail.com (P.-J. van Helvoort).

0883-2927/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apgeochem.2005.08.009

number of variables that result from extensive laboratory characterization of sediment or soil samples. More importantly, they are applied to identify the main sources of variance within geochemical datasets, and link them to geochemical processes or properties.

In geochemical baseline and exploration studies, FA or PCA has been used to analyze geochemical data for soil or stream sediment samples trying to identify possible imprints of contamination or mineralization over the natural geochemical background composition (Chork and Salminen, 1993; De Vivo et al., 1997; Morsy, 1993; Reimann et al., 2002; Tripathi, 1979). In these types of studies, factor scores have usually been plotted as geochemical maps to identify geochemical anomalies, which are indicative for mineralization or contaminant sources. The geochemical maps have also been combined with other attribute maps like land use, geology, or soil type, to explain the spatial distribution of geochemical anomalies. Factor analysis and PCA have also been applied in sedimentary geochemistry, mainly to identify the effects of provenance and diagenetic processes on the bulk chemistry of unconsolidated material (Hakstege et al., 1993; Huisman and Kiden, 1998; Moura and Kroonenberg, 1990; Tebbens et al., 1999, 2001). Also in many hydrogeochemical studies (Cameron, 1996; Dalton and Upchurch, 1979; Duffy and Brandes, 2001; Evans et al., 1996; Frapporti et al., 1993; Gupta and Subramanian, 1998; Lawrence and Upchurch, 1982; Lee et al., 2001; Meng and Maynard, 2001; Suk and Lee, 1999) multivariate techniques have been used to identify a variety of processes that control water chemistry, including natural mineral dissolution, ground water contamination, salt water intrusion in fresh water aquifers, recharge area, and seasonal variation in surface water composition.

Although FA and PCA are widely applied in geochemistry and hydrochemistry, there are still very few studies that explicitly evaluate the quality of the results and their reproducibility, as these depend on several, often implicit, assumptions and the statistical distribution of the data. Reimann et al. (2002) state some of the most critical issues that should be dealt with when performing FA or PCA, being:

1. What is the role of extreme values (or multivariate outliers, which may not be extreme in any of the individual attribute space directions) on the multivariate results?
2. How many factors should be extracted?
3. Which variables should be included in the factor model?

The first question is relevant considering the reproducibility of the results and the stability of the multivariate model that is adopted. Since in many environmental studies the datasets are large and samples may come from many different locations, these datasets may be subject to a large degree of heterogeneity. This will translate into (groups) of outlying values, which may corrupt the assumption that the dataset meets with some minimum degree of normality, which is needed for a proper application of FA or PCA. It has been proposed in several studies (Reimann and Filzmoser, 2000; Reimann et al., 2002) to apply a robust version of the multivariate statistical methods to overcome this problem. Although there are many different robust methods available and there are still new ones developing, they all use the main principle of selecting subsets of observations that would be most homogeneous and representative for the dataset as a whole. This way, the chance of outlying values distorting the multivariate analysis is minimized.

Answers to the second question, i.e., how many factors should be in the factor model, have been generally formulated as criteria for minimum eigenvalues, explained portion of variance, or scree plots (Cattell, 1966), and more objectively by statistical tests, information criteria, or resampling methods (Basilevski, 1994; Johnson and Wichern, 1998). Answers to the third question are less easily found in the literature, and in most applied studies they are not addressed at all. An exception is the study of Reimann et al. (2002). Although they do not formulate unique answers, they perform an extensive analysis of different subsets of variables. Still, the criteria used to deal with these last two issues appear to remain subjective as they heavily depend on the experience and individual research goals.

In this paper, the main goal is to present sequential Factor Analysis (seqFA) as a new approach to standard FA or PCA. The new approach explicitly addresses the questions 1–3, considering heterogeneity of the input data, the number of factors to be extracted, and the set of variables to be chosen, where it permits both statistical arguments and geochemical expertize. In addition, the approach is developed in robust and non-robust versions and both have been

applied to the same geochemical dataset, which allows comparison. The dataset consists of geochemical data of unconsolidated Late Quaternary deposits as found in the Rhine-Meuse delta plain, The Netherlands. The final result of this case study is, in the terminology of (Meng and Maynard, 2001), a conceptual model for the geochemical properties and processes that govern the chemical composition of these sedimentary deposits. This conceptual model is a clever balance between the available data, geochemical expert knowledge, and statistical arguments.

2. Materials and methods

2.1. The geochemical data set

A geochemical dataset was derived from 145 sediment samples taken from Late Quaternary deposits (Holocene and Pleistocene), mainly of meandering rivers, in the Rhine-Meuse delta, The Netherlands (see Fig. 1). Geochemical characterization of these deposits has been more extensively treated in van Helvoort (2003). As a result of abundant channel avulsing during the Holocene (Stouthamer and Berendsen, 2000), a dense, stacked network of palaeo-channels exists causing a high degree of heterogeneity over short distances (see cross-section in Fig. 1). The sedimentary heterogeneity translates into geochemical heterogeneity because there are close relations between grain size and mineralogy (Huisman and Kiden, 1998; Johnsson, 1993; Moura and Kroonenberg, 1990; Nesbitt and Young, 1996; Passmore and Macklin, 1994; Tebbens et al., 2001). For this reason, the deposits have been grouped into 6 sedimentary facies (Table 1) based on textural and structural properties, using an existing facies classification for fluvial deposits (Miall, 1985, 1996) that has been adapted to this region (Berendsen, 1984; Törnqvist et al., 1994). Each facies has been sampled at various locations and depths along several transects (Fig. 1), covering most of the compositional variation present in these deposits. The sediment samples were dried at 70 °C and mechanically ground (Herzog HSM apparatus). X-ray Fluorescence (XRF) was used for major element (Al_2O_3, CaO, Fe_2O_3, K_2O, MgO, MnO, Na_2O, P_2O_5, SiO_2 and Ti_2O), and trace element (As, Ba, Bi, Cd, Ce, Cr, Cs, Cu, Ga, La, Mo, Nb, Ni, Pb, Rb, S, Sb, Sn, Sr, Th, U, V, Y, Zn and Zr) determinations. Loss on ignition (LOI) was determined at 1150 °C. In addition, the CEC was determined on freeze dried sub samples (unground) using a standard buffered salt method (Hesse, 1971). For a selection of samples, the CEC determination was carried out in quadruplicates and in each batch of 32 samples 3 blanks and 3 ISE standards were analyzed. Soil Organic Matter (SOM) and carbonate contents were determined on ground samples by Thermal Gravimetric Analysis, using a LECO TGA 608 apparatus. Grain size analysis was done by laser-diffraction, after removal of the >2000 μm fraction by sieving, and removal of both organic matter and carbonates using standard methods (van Doesburg, 1996). All samples were analyzed in duplicate using a Coulter LS230 apparatus, which has a detection range between 0.04 and 2000 μm discretized into 116 grain size classes. Quality of all analytical procedures was checked by incorporating random duplicates and international standards.

Prior to statistical analysis, the data were checked for accuracy and measurement artefacts. The observations for Bi, Cd, Ce, La, Mo, Sb, Sn, Th and U were left out of further analysis because over 10% of the observations were below the detection limit. For the other elements, observations under detection were replaced by 2/3 times the detection limit. In addition, log transformation was applied, which for most variables yielded improved normal distributions, or at least improved symmetry. In the non-robust or classical PCA, the removal of obvious outliers was combined with the first step of the seqFA procedure (see Section 2.2).

The authors realize that compositional data is always subject to (some degree of) data closure, because major components analysis usually sum up to 100% (Aitchison, 1981, 1984; Otero et al., 2005). Aitchison (1981) introduced the log-ratio transformation, which diminishes data closure by eliminating the constant-sum and associated curvature in the data set. However, a log-ratio transformation was not applied, because the authors preferred as few data manipulations as possible while focussing on the new approach to factor analysis. Instead of log-ratio transformation, the major component analysis was not normalized on 100%, leading to "open" total analyses varying between 97% and 104%. Also, it was found that curvature in the data was scarce but present, and was reduced after log transformation of all components. In addition, SiO_2 was excluded from factor analysis, being the dominant component in

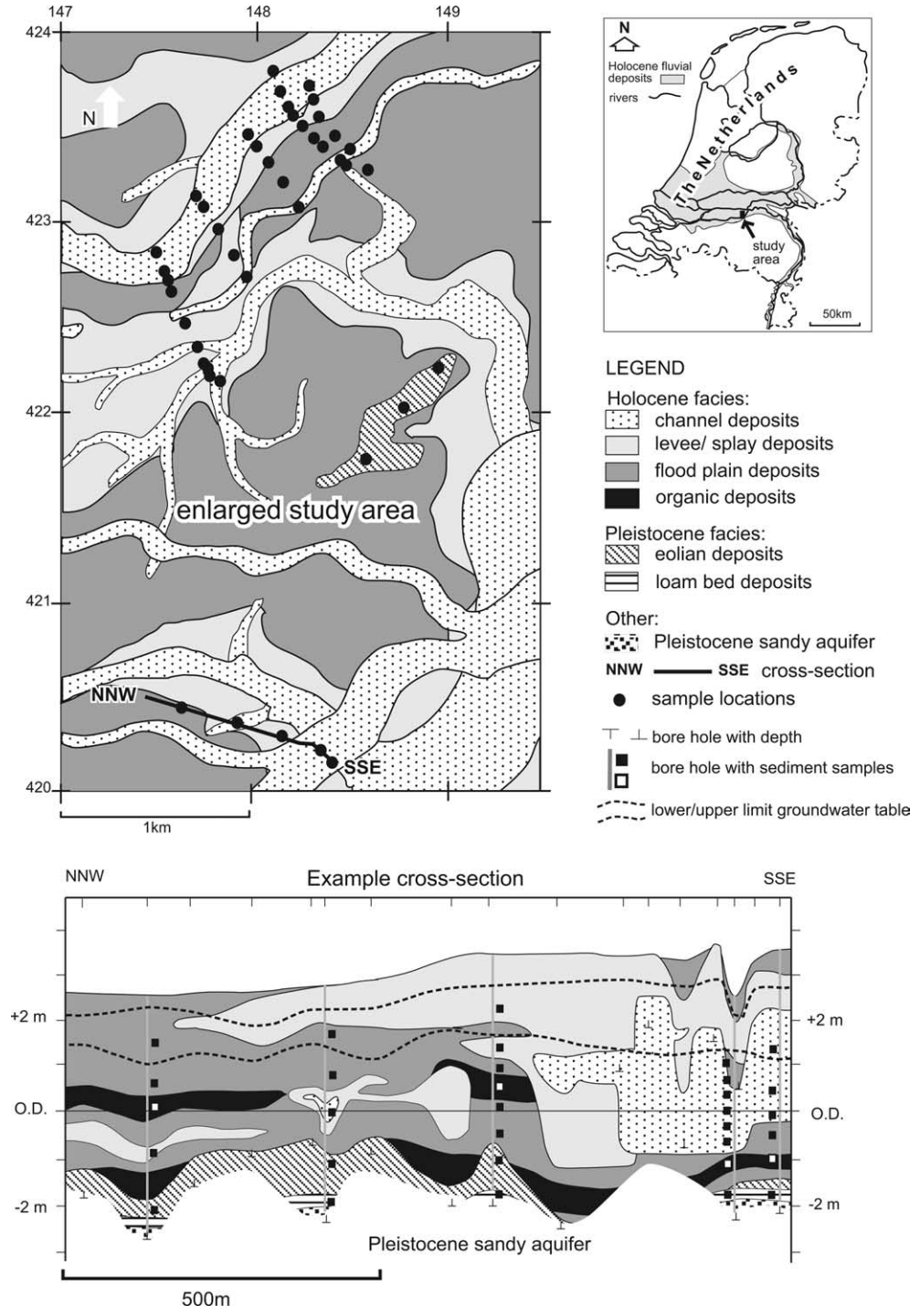

Fig. 1. Location of case study field area, sampling sites, and geological transect.

most samples (up to 95 wt%). The a priori removal of SiO_2 therefore reduced the presence of curvatured relation ships down to a minimum, while the most relevant geochemical relations were preserved. This approach seemed to be satisfying for the goals of this paper.

Table 1
Facies units, facies properties, and sample classification

Facies unit	Lithology	Geometry	Number of samples
Channel deposits	Very fine to coarse sand (105–2000 μm)	5–10 m thick, 50–2000 m wide	43
Natural levee and crevasse-splay deposits	Horizontally laminated sandy–silty clay, small lenses of (very) fine sand (105–210 μm)	Levees: 0.5–1.10 m thick, 50–500 m wide; crevasse-splays: 1–2 m thick, 0.1–5 km wide	23
Flood basin deposits	Massive to very thin laminated clay and humic clay	1–5 m thick, 0.1–10's km wide	30
Organic deposits	Peat	0.1–5 thick, 0.1–10's km wide	20
Eolian dune deposits	Structureless of very fine to fine sand (105–210 μm)	1–10 m thick, 50–2000 m wide	17
Loam bed deposits	Massive sandy–clay to clayey sand; clay in admixture with sand in the fraction 210–300 μm	0.1–1.5 m thick, 1–10's km wide	12
Total			145

2.2. Sequential Factor Analysis approach

2.2.1. General approach

The general procedure of the seqFA approach is summarized in Fig. 2, which shows the 4 consecutive steps. Defining k, m and n as integers, and defining $k = m + n$, these steps can be explained as follows:

1. Define the optimum number of factors (and remove obvious outliers when using non-robust methods). The resulting factor model is called the Complete Factor Model based on k measured variables.
2. Reduce the number of k variables to a set of m key variables, by stripping off n highly correlated variables. The result is the Stripped Factor Model, with only m key variables.
3. Expand the Stripped Factor Model with m key variables back to its full size of k variables by calculating the loadings for the stripped variables. The result is the Expanded Factor Model.
4. Compare the Expanded Factor Model with the Complete Factor Model in terms of explained variance.

Step 1. Optimizing the number of factors extracted and identifying outliers. The optimum number of factors (step 1 in Fig. 2) is found by an iterative process, in which the number of factors (to which any rotation method may be applied), is increased by one at a time, starting from a minimum of two factors. Each time the factor model is extended by a new factor, the distributions of the loadings are examined. When the optimum factor configuration

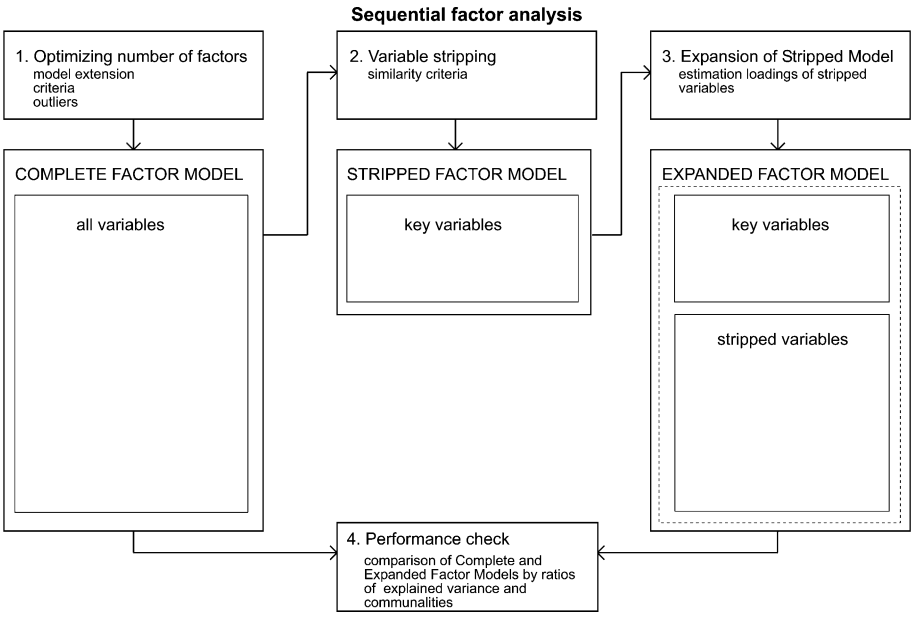

Fig. 2. The general procedure of seqFA, indicating the 4 main steps.

has been reached, the following 3 criteria should apply to the rotated factor loading matrix:

- maximum extension criterion: there should be extracted as many factors as possible until one of the criteria below is not fulfilled;
- minimum loading criterion for factors: each factor should have at least one loading with an absolute value above a threshold value;
- minimum loading criterion for variables: each variable should have at least one loading with an absolute value above a threshold value.

The first criterion is the driving force to extend the numbers of factors in the model to as many as possible. Theoretically, there can be as many factors as variables, but many of them will be meaningless from both statistical and geochemical expert points of view. The other two criteria act as a counter balance against excessive factors, and will be heavier as the threshold is increased. The result of step 1 is the Complete Factor Model based on the full set of k measured variables, and has the optimum number of p factors explaining a portion of variance represented by $S^2_{k\mathrm{C}}$. This step is critical in the whole procedure, because the number of factors determines how many sources of variance will be acknowledged, and how many unique geochemical processes or properties can be isolated from the dataset. In the case of classical PCA, also the removal of obvious outliers is included in this step (see Section 2.2.2).

Step 2. Selecting the key variables by variable stripping. In this step, the number (m) and identity of the principal variables that represent unique sources of variance are identified, giving the factors discerned in the Complete Factor Model their core or 'key' identity. These principal variables are therefore called *key variables*, and have unique loading patterns. The other variables are stripped off using a similarity criterion for their factor loading patterns. This criterion states that if for all factors the difference between the loadings is less than a preset value, one of the two variables can be stripped off from the factor model, because it does not represent a unique source of variance. In other words, the total amount of variance explained by the model does not significantly change when the variable is left out of the model. The main reason to strip off variables is to get rid of collinearity, which generally causes an uneven distribution of variables over the factors. This leads to several imbalances when a rotation method is applied hampering the recognition and interpretation of weaker factors, which could be equally interesting from a geochemical point of view. The model that results from this step is called the Stripped Factor Model, and it only includes a set of m key variables with unique loading patterns (see Fig. 2).

Step 3 and 4. Expansion and performance check of the Stripped Factor Model. In the third step, the Stripped Factor Model is expanded to its original size of k variables by including the stripped variables again, creating the Expanded Factor Model. The expansion is done by calculating the loadings of the stripped variables in the factor space of the Stripped Factor Model. The factor loadings are used to compute the communalities of the stripped variables, which then allow estimating the total portion of explained variance of the Expanded Factor Model ($S^2_{k\mathrm{E}}$) based on k variables. In step 4, the performance of the Expanded Factor Model is tested by comparing the communalities and total portion of explained variance of the Expanded Factor Model with the Complete Factor Model. Step 3 and 4 are elaborated with the following expressions.

The portion of explained variance by the Stripped Factor Model ($S^2_{m\mathrm{S}}$) with p factors and m key variables is given by:

$$S^2_{m\mathrm{S}} = \frac{1}{m}\sum_{i=1}^{m}\sum_{r=1}^{p} a^2_{ir}, \tag{1}$$

where a_{ir} represents the loading of the ith key variable on the rth factor after rotation. The key variable loadings will be summarized in the matrix $\boldsymbol{A}_m$. The term $\sum_{r=1}^{p} a^2_{ir}$ represents the communality h^2_i, which is the fraction of the total variance of key variable j explained by p factors (Davis, 1986). After estimating the factor scores matrix $\boldsymbol{F}_m$ (Davis, 1986), the Stripped Factor Model can be represented as

$$\boldsymbol{X}_m = \boldsymbol{F}_m\boldsymbol{A}^{\mathrm{T}}_m + \boldsymbol{E}_m, \tag{2}$$

where $\boldsymbol{X}_m$ is the standardized data matrix with the m key variables, and $\boldsymbol{E}_m$ an error term. In step 3, using the factor space of the Stripped Factor Model, and supposing that n variables have been stripped, the standardized data matrix of the stripped variables (denoted by $\boldsymbol{X}_n$), can be factorized as

$$\boldsymbol{X}_n = \boldsymbol{F}_m\boldsymbol{C}^{\mathrm{T}}_n + \boldsymbol{E}_n \tag{3}$$

with $\boldsymbol{E}_n$ being an error term, and $\boldsymbol{C}_n$ is the loading matrix referring to the stripped variables. Since the factor model (3) can also be considered as a regression model, the "regression coefficients" $\boldsymbol{C}^{\mathrm{T}}_n$ can be

estimated by multivariate linear regression (Johnson and Wichern, 1998). For a more formal representation on the estimation of the loadings $\boldsymbol{C}_n$ is given in Filzmoser (1997). Now, the portion of explained variance of n stripped variables in the Expanded Factor Model is analogous to (1):

$$S^2_{n\mathrm{E}} = \frac{1}{n}\sum_{j=1}^{n}\sum_{r=1}^{p} c^2_{jr} \tag{4}$$

with c_{jr} being an element of $\boldsymbol{C}_n$, representing the estimated loading of the jth stripped variable on the rth factor. Also, $\sum_{r=1}^{p} c^2_{jr}$ represents the communality h^2_j of the stripped variable j. Combining expressions (1) and (4), the portion of explained variance by the Expanded Factor Model ($S^2_{k\mathrm{E}}$) with k variables can be calculated:

$$S^2_{k\mathrm{E}} = \left(\frac{m}{m+n}\right)S^2_{m\mathrm{S}} + \left(\frac{n}{m+n}\right)S^2_{n\mathrm{E}}. \tag{5}$$

In the evaluation step 4, the overall performance of the Stripped Factor Model can be expressed as the ratio of the explained portions of variance by the Expanded Factor Model ($S^2_{k\mathrm{E}}$) and the Complete Factor Model ($S^2_{k\mathrm{C}}$):

$$\text{Model performance} = \frac{S^2_{k\mathrm{E}}}{S^2_{k\mathrm{C}}}. \tag{6}$$

Accordingly, the performance per variable can be expressed as the ratio of communalities in the Expanded Factor Model and the Complete Factor Model for any variable k:

$$\text{Variable performance} = \frac{h^2_{k\mathrm{E}}}{h^2_{k\mathrm{C}}}. \tag{7}$$

With these two indicators, the effect of stripping off variables on both the overall portion of explained variance and the individual communalities can be assessed.

2.2.2. Computational procedures and outlier replacement

All statistical computations were made in the R environment (version, 1.9.1), a powerful statistical software package which is freely available at http://www.R-project.org. PCA was performed on the correlation matrix with Varimax rotation (Kaiser, 1958). Although any FA method can be used instead in combination with the seqFA procedure, we preferred PCA here because it is the simplest multivariate method and needs no additional assumptions. Hence, the authors used a slightly adapted version of an R-function initially designed for Principal Factor Analysis (PFA), but setting the uniquenesses all to zero. For the robust version, the fast MCD algorithm (Pison et al., 2003; Rousseeuw and van Driessen, 1999) was used to calculate the robust correlation matrix, based on 75% of the data. Thus, a maximum amount of 25% of outliers might be present in the data without affecting the estimation of the correlation matrix, which is considered acceptable (see Pison et al. (2003)).

For the non-robust PCA, the classical sample correlation matrix was used, after some obvious outliers were replaced by median values. First, potential outliers were identified by box plots and Q–Q plots. Second, as a part of the step 1 of seqFA, the factor score distributions of the unrotated PCA model were examined on extreme values, produced by outliers. The outliers responsible for extreme factor scores were eliminated one by one through substitution of median values of the facies to which the cases belonged. It was decided not to leave out the entire case, because the observations for the other variables were not marked as outliers and should not disturb the PCA model. This was repeated, until the unrotated PCA model produced no extreme factor scores anymore. This resulted in replacement of only 6 observations (for Fe_2O_3, MnO, and P_2O_5), occurring in 3 cases, and 2 of them belonging to the organic deposits. The extreme values were associated with dense concentrations of vivianite or Fe/Mn-(hydr)oxides, which had already been spotted during field sampling. Note that the number of replacements was very small compared to the whole data array of 3915 observations (145 cases times 27 variables). The rest of the computational procedure is explained below.

Step 1. In step 1 of seqFA, PCA was repeated while increasing the number of factors one by one, until the loading matrix did not meet with one of the minimum loading criteria. The largest factor configuration that still fulfilled all criteria was marked as the optimum configuration, being the maximum number of factors that should be included in the Complete Factor Model. The minimum loading criteria were set to 0.60 for robust and non-robust PCA, this will be discussed further in the Section 3.1.1.

Step 2. In step 2, variable stripping was applied to the loading matrix belonging to the optimum configuration selected in the previous step. Variable stripping was done as follows:

2.a. The variables were ranked on communality;
2.b. Moving down the list, the loadings were checked on similarity;
2.c. When two variables had similar loadings for all factors, the one with smallest communality was stripped, and the one with the highest was retained. The stripped variable thus was removed from the Complete Factor Model, and the retained one became a key variable;
2.d. After working through the list, the PCA was run again without the stripped variables to generate a new rotated loading matrix, but using the appropriate rows and columns of the initial (robust) correlation matrix estimated for the Complete Factor Model;
2.e. Step 2.a through 2.d were iterated, until no further variables could be stripped off according to the similarity criterion.

The final result is the Stripped Factor Model, with key variables only. The similarity criterion (threshold for loadings) for stripping off was varied to see how this would influence the resulting set of key variables. Depending on the similarity criterion, the number of iterations (step 2.a to 2.d) needed to arrive at the final set of key variables varied. The stripping procedure has been automated by creating a special function in R, and is available from the authors on request.

Step 3 and 4. In step 3, the factor scores produced by the Stripped Factor Model were used to estimate the loadings for the stripped variables by regression (expression 3), creating the Expanded Factor Model. The Expanded Factor Model has a loading matrix of the same dimensions as the Complete Factor Model, which makes a statistical performance check (step 4 of seqFA) of the Expanding Factor Model possible, by using expression 6 (model performance) and expression 7 (variable performance).

3. Results

3.1. The robust and non-robust factor models

3.1.1. Optimizing the number of factors: the Complete Factor Models

The optimum number of factors for the test case is 5 for robust PCA (Table 2). When a 6th factor was added to the robust model, both minimum loading criteria were no longer fulfilled, as Pb had the highest loading of only 0.30 on factor 6. When extending the robust model even further to 7 factors, Na_2O emerged on the last factor (0.42), which also was too low. For the non-robust PCA, the Complete Factor Model could be extended as far as the 9th factor, before the minimum factor loading criterion failed for the 10th factor by 0.26 (for Cu). However, it was decided to develop the non-robust Complete Factor Model also for 5 factors to have exactly the same configuration as for robust PCA. This is necessary for a sound comparison of the final results produced by robust and non-robust PCA.

Table 2
Highest factor loadings on the last extracted factor for robust and non-robust PCA after Varimax rotation

Number of factors	Robust		Non-robust	
1	Cs	0.93	Nb	0.87
2	Carbonate	0.91	Ba	0.77
3	S	0.91	S	0.95
4	Zr	0.90	Carbonate	0.92
5	MnO	0.74	Zr	0.89
6	Pb	0.30	P_2O_5	0.63
7			MnO	0.65
8			V	0.69
9			Pb	0.64
10			Cu	0.26

Fig. 3a and b provide a detailed picture of the loading distributions per variable with the extension of the Complete Factor Model from 2 to 7 factors, both for the robust (left panel) and non-robust (right panel) PCA. Generally, the loadings for both robust and non-robust PCA had bimodality, with typically a single high loading (>0.60) per variable, and several loadings lower than 0.60. However, bimodality was most evident and was maintained with increasing number of factors for the robust loadings. For non-robust PCA, the bimodality tended to weaken with increasing number of factors (Fig. 3b). Adding new factors, the highest factor loadings were weakened, because the Varimax rotation became less efficient in enhancing the loadings. A plausible explanation for the sub optimal rotation results in non-robust PCA is that less obvious outliers start affecting the rotation procedure when the number of factors increases. If the rotation is influenced by these outliers, it will be more difficult to produce enhanced loadings on all factors. Histograms further illustrate that for the robust PCA the minimum loading criterion of 0.60 is optimal, because at this value the divide in the bimodal distribution between high and low loadings is found (see Fig. 4a for the 5 factor configuration, also compare Fig. 3a). For the non-robust PCA

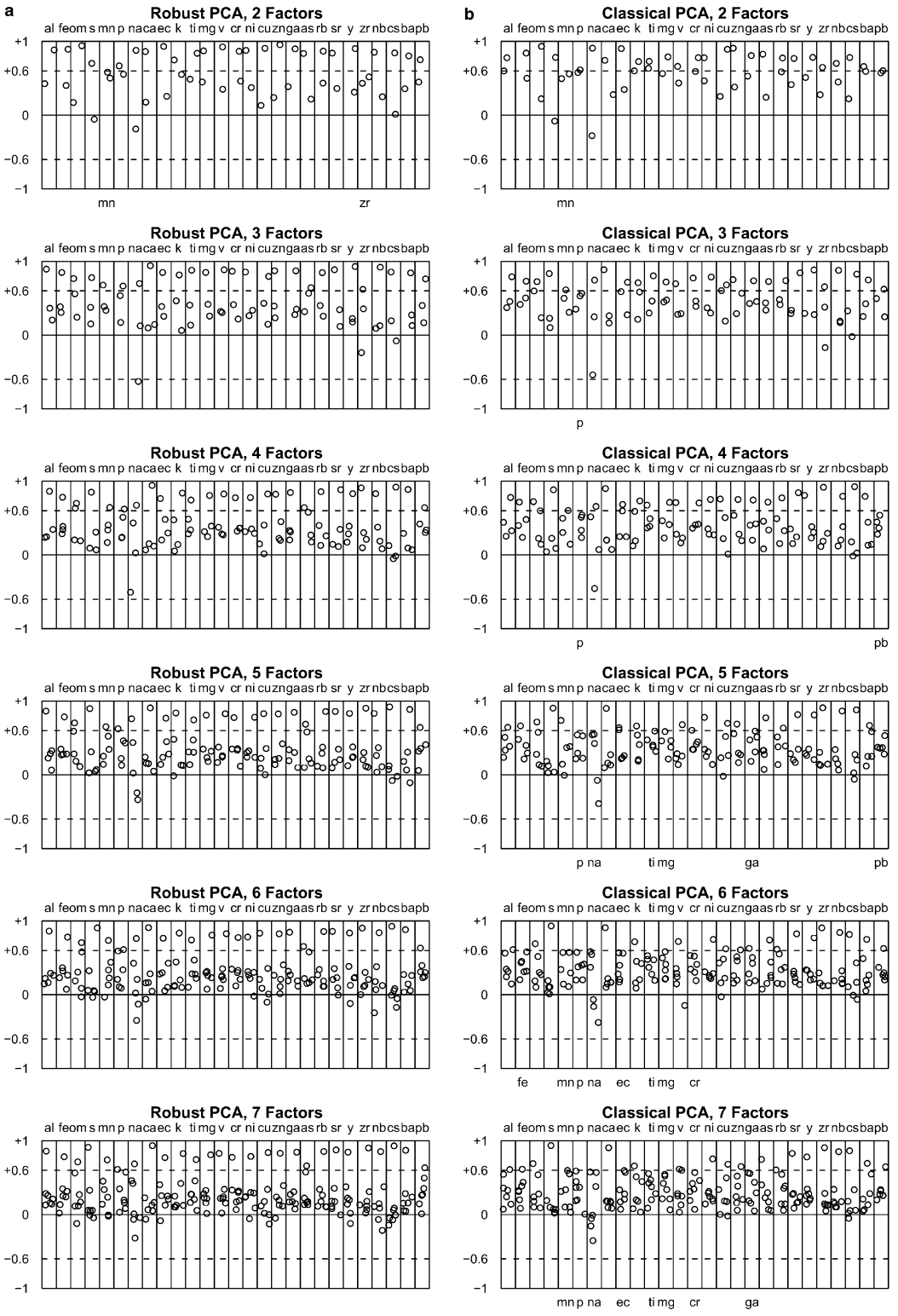

Fig. 3. Diagrams of factor loading distributions per variable (rotated solution) for: (a) robust and (b) non-robust PCA with an increasing number of factors. Although no capitals could be used in the diagrams, the usual chemical symbols are used for trace elements, and variable names of the oxides are abbreviated to single element names. Abbreviations for CEC, SOM, and carbonates, are respectively ec, om and ca. Elements that do not meet with the minimum loading criteria also appear at the bottom end of the diagrams.

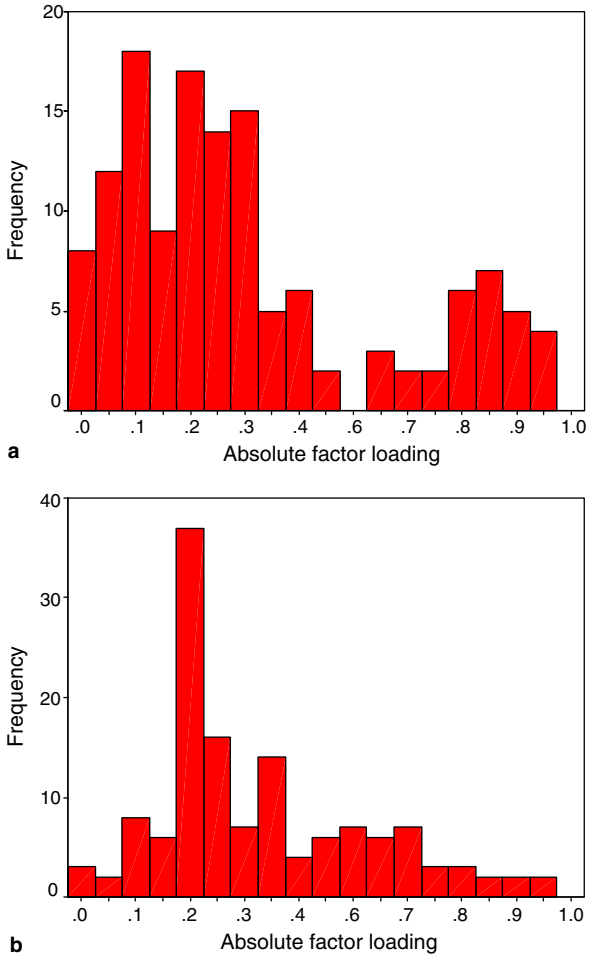

Fig. 4. Bimodality of absolute factor loadings for the: (a) robust and (b) non-robust PCA configuration of 5 factors (rotated solution).

an optimal threshold for minimum loadings cannot actually be defined (Figs. 4b and 3b), hence also the optimum number of factors could not objectively be derived (Table 2). This further justifies the choice of a 5 factor model also for the non-robust case.

Another way to visualize the progressive extension of the Complete Factor Models is presented in Fig. 5a and b. These diagrams show per factor which variable had the highest loading (i.e., the principal variable), depict the relations between consecutive factor configurations, and give a brief geochemical interpretation. The extension of the robust Complete Factor Model shows a very regular pattern, as all new factors originated from the first or the second factor. This is indicated by the dotted connections between the principal variables of the new factors and the factors on which they previously had the highest loadings. In addition, there were very few changes in principal variable of the same factor, and all factors remained in the same order with respect to the relative portion of variance explained. The diagram for non-robust PCA is very different, with several changes of principal variables, and factors that swapped position. This irregular pattern is evidence for instability of the model, as the factors changed identity and rank when including more factors for Varimax rotation. Thus, increasing the number of factors, the whole model changed fundamentally, because the Varimax rotation was very susceptible to a newly included source of variance. From these diagrams it is concluded that robust PCA yields much more stable – i.e., less sensitive to the number of factors chosen – and better reproducible results than non-robust PCA.

3.1.2. Selection of key variables: The Stripped Factor Models

Table 3 lists the key variable sets for the 5 factor configuration resulting from different similarity criteria for the robust and non-robust Stripped Factor Models. As expected, the number of stripped variables always increased with decreasing strictness of the similarity criterion (from 0.10 to 0.30). When similarity was set to 0.40 or more, the 5 factor configuration was not supported by the remaining set of key variables in terms of the minimum loading criteria defined previously, and therefore was considered invalid.

The composition of the key variable set was quite similar for the robust and non-robust PCA, although the stripping usually proceeded more efficiently for the robust PCA (mostly one iteration) versus the non-robust PCA (2 or 3 iterations). With a similarity of 0.30, the common key variables identified are Cs, MnO, S, Na_2O, Zr, Sr, SOM, and Al_2O_3, in order of decreasing communality ratios (see expression 7). The non-robust solution has one more variable (Y) in the Stripped Factor Model. Thus, it is concluded that for the selected factor configuration with 5 factors, the robust and non-robust PCA Stripped Factor Models are highly congruent. The results for the most condensed Stripped Factor Models, i.e., with a similarity stated at 0.30, will be discussed quantitatively in the next section.

3.1.3. The Expanded Factor Models and statistical performance

Tables 4A and 4B show the factor loadings (>0.30) for the Complete Factor Model and the Expanded Factor Model, both for the robust

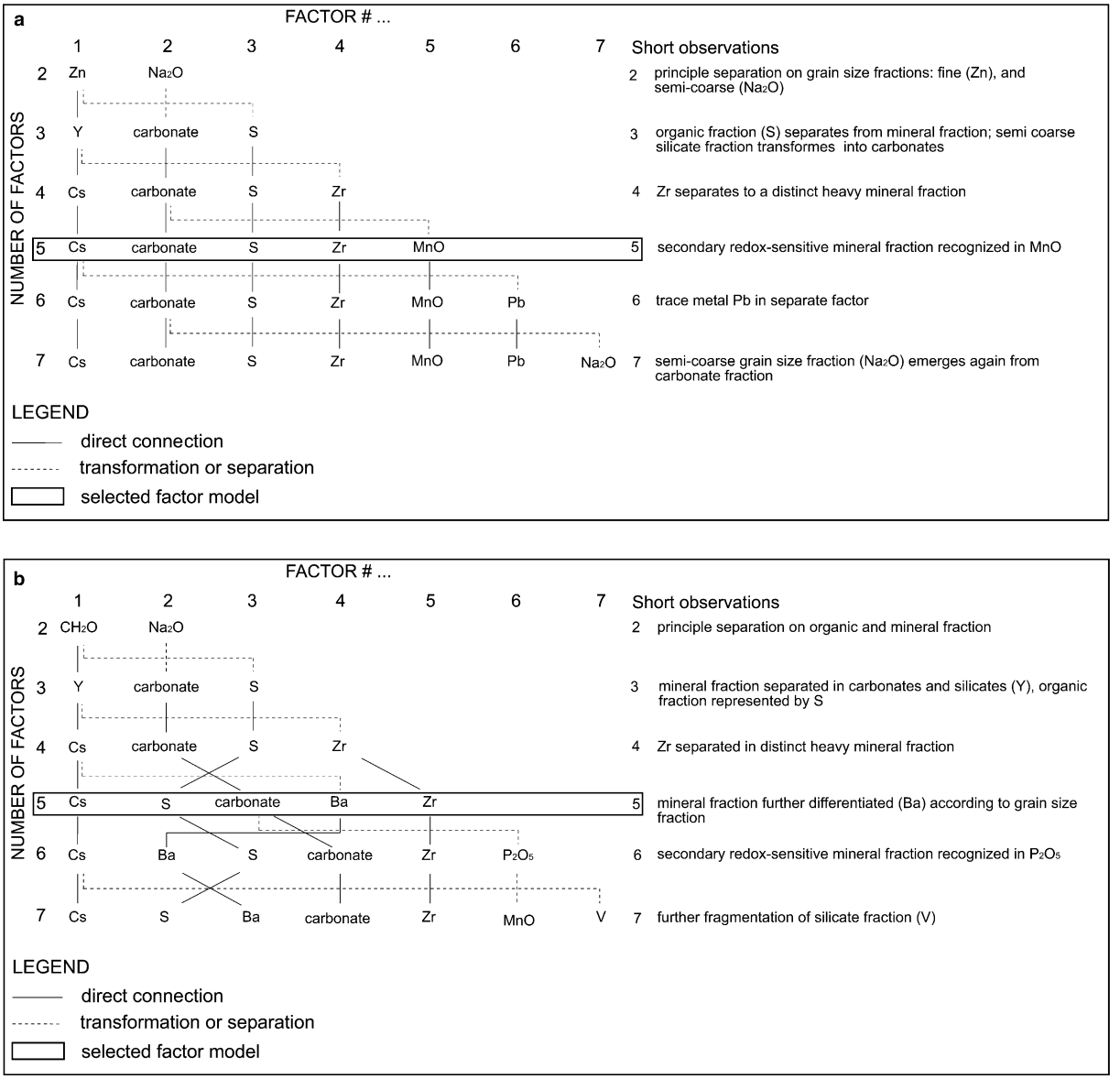

Fig. 5. Diagrams representing the gradual extension of: (a) robust and (b) non-robust Complete Factor Models (after Varimax rotation) and a brief geochemical interpretation (not discussed).

and non-robust solution. The following is observed for *both* the robust and non-robust PCA solutions:

- The ratio of overall explained variance (S^2_E/S^2_C) is close to unity, and shows that the Expanded Factor Model performance is only a fraction less than the Complete Factor Model, which means that the Stripped Factor Model is quite representative in explaining the dominant variance sources of the dataset, including the stripped variables.
- In general, the communality ratios (h^2_{iE}/h^2_{iC}) for the key variables exceed unity (0.96–1.15), whereas for the stripped variables these are all lower than 1. This means that the Stripped Factor Model performs relatively better on the key variables at the cost of the stripped variables, which have been excluded for Varimax rotation. The reason is, obviously, that the Stripped Factor Model was optimized for the key variables.
- The variance explained by the Stripped Factor Model alone (S^2_{mS}) is more evenly distributed over the factors than in the Complete Factor

Table 3
Key variable sets (>0.30) for the 5-factor configuration with different similarity criteria for variable stripping, both for robust and non-robust Stripped Factor Models

All variables	Robust Stripped Factor Models					Non-robust Stripped Factor Models				
	0.10	0.15	0.20	0.25	0.30	0.10	0.15	0.20	0.25	0.30
Al_2O_3	Al_2O_3	Al_2O_3	Al_2O_3	Al_2O_3	Al_2O_3	Al_2O_3	Al_2O_3	Al_2O_3	Al_2O_3	Al_2O_3
As	As	As				As	As			
Ba	Ba	Ba				Ba				
Carbonate	Carbonate	Carbonate				Carbonate				
CEC										
Cr						Cr				
Cs	Cs	Cs	Cs	Cs	Cs	Cs	Cs	Cs	Cs	Cs
Cu	Cu	Cu	Cu			Cu	Cu			
Fe_2O_3	Fe_2O_3					Fe_2O_3				
Ga										
K_2O	K_2O	K_2O				K_2O	K_2O			
MgO										
MnO	MnO	MnO	MnO	MnO	MnO	MnO	MnO	MnO	MnO	MnO
Na_2O	Na_2O	Na_2O	Na_2O	Na_2O	Na_2O	Na_2O	Na_2O	Na_2O	Na_2O	Na_2O
Nb	Nb		Nb			Nb				
Ni										
Pb						Pb	Pb			
P_2O_5	P_2O_5	P_2O_5	P_2O_5	P_2O_5		P_2O_5	P_2O_5	P_2O_5		
Rb	Rb									
S	S	S	S	S	S	S	S	S	S	S
SOM	SOM	SOM	SOM	SOM	SOM	SOM	SOM		SOM	SOM
Sr	Sr	Sr	Sr	Sr	Sr	Sr	Sr	Sr	Sr	Sr
TiO_2	TiO_2	TiO_2	TiO_2			TiO_2	TiO_2	TiO_2		
V	V					V		V		
Y	Y	Y				Y	Y	Y	Y	Y
Zn	Zn	Zn				Zn	Zn	Zn		
Zr	Zr	Zr	Zr	Zr	Zr	Zr	Zr	Zr	Zr	Zr

Model (compare S^2_{mS} with S^2_C for the individual factors in Tables 4A and 4B). The reason is that in the Complete Factor Model most high loadings were found on the first factor (F1), and therefore this factor explains the larger part of the variance. However, in the Stripped Factor Model the highest loadings are more evenly distributed over the factors, because much of the collinearity has been removed by variable stripping.

Comparing results between the robust and non-robust solution, it is evident that the results for the Stripped/Extended Factor Model are far more similar to each other than those for the Complete Factor Model. For the Complete Factor Models, the Varimax rotation obviously leads to distinctly different orientations of the principal axes. The difference is also evident in the communalities of the (key) variables. The Stripped/Expanded Models show highly similar orientations, but for a shuffling of ranks between F3 and F4, and also very similar communalities.

The main implications of these results are 4-fold. First of all, the Stripped Factor Model is the most condensed way to summarize the main sources of variance in a geochemical dataset, without loosing any key information, and without loosing significant explained variance via the Expanded Factor Model. Secondly, the Stripped Factor Model yields higher loadings for the remaining key variables, which facilitates interpretation of the factors. Thirdly, the Stripped Factor Model enhances weaker factors at the cost of stronger ones. This validates the extraction of weaker factors with smaller eigenvalues in the first step of the seqFA. Finally, the Stripped (and Extended) Factor Model is more robust to outliers in minor variables than the Complete Factor Model.

3.2. A conceptual geochemical model for riverine deposits

3.2.1. Interpretation of the factors

The loading matrix of the robust Expanded Factor Model (Table 4A) was used to develop a

Table 4A
Factor loadings (>0.30) for the robust Complete and Expanded Factor Model (variables sorted on communality ratios[b])

	Complete Factor Model						Expanded Factor Model						E/C[b]
	F1	F2	F3	F4	F5	Communality	F1	F2	F3	F4	F5	Communality	
Cs[a]	0.92					0.88	0.96					0.97	1.11
MnO[a]	0.35	0.52			0.65	0.90				0.91		0.98	1.09
S[a]	0.31		0.90			0.91		0.94				0.98	1.07
Na_2O[a]		0.76		0.44		0.94			0.89		0.33	0.97	1.02
Zr[a]	0.30			0.90		0.97					0.91	0.99	1.02
Sr[a]	0.35	0.86				0.94			0.73	0.53		0.95	1.01
SOM[a]	0.70		0.59			0.97	0.60	0.63		0.39		0.97	1.00
Al_2O_3[a]	0.86	0.33				0.99	0.75	0.36	0.31	0.31		0.95	0.96
TiO_2	0.75	0.31	0.31	0.47		0.99	0.62	0.36		0.34	0.50	0.96	0.98
Y	0.83			0.38		0.97	0.74	0.31		0.32	0.44	0.95	0.98
CEC	0.77		0.45			0.95	0.66	0.48		0.36	0.34	0.92	0.97
Fe_2O_3	0.79		0.35			0.97	0.67	0.38		0.47	0.32	0.95	0.97
Ni	0.78		0.32	0.30		0.92	0.67	0.38		0.38	0.34	0.90	0.97
Cr	0.78	0.33	0.35	0.35		0.98	0.64	0.41		0.35	0.39	0.94	0.96
MgO	0.81	0.35	0.32			0.97	0.69	0.39		0.39		0.93	0.96
Nb	0.83			0.37		0.91	0.76				0.42	0.88	0.96
Rb	0.86	0.37				0.98	0.74	0.38	0.32	0.35		0.94	0.96
Zn	0.83		0.42			0.97	0.71	0.47		0.35		0.94	0.96
V	0.83		0.38			0.98	0.71	0.43		0.36	0.31	0.94	0.96
Carbonate		0.91				0.93			0.72	0.57		0.89	0.95
Ga	0.85	0.30	0.34			0.97	0.72	0.41		0.33		0.92	0.95
K_2O	0.84	0.48				0.97	0.72		0.47	0.30		0.92	0.94
Cu	0.83		0.36			0.87	0.72	0.43		0.32		0.81	0.93
Pb	0.64	0.32	0.36	0.41		0.81	0.52	0.38		0.31	0.40	0.75	0.92
As	0.58		0.69			0.89	0.47	0.64		0.35		0.81	0.91
P_2O_5	0.63	0.43			0.46	0.86	0.51			0.58		0.78	0.91
Ba	0.89	0.31				0.93	0.78		0.32			0.83	0.90
							$S^2_{\eta S}$	0.26	0.20	0.19	0.18	0.14	0.97	
							$S^2_{\eta E}$	0.43	0.15	0.08	0.14	0.09	0.89	
S^2_{kC}	0.51	0.15	0.13	0.09	0.05	0.94	S^2_{kE}	0.38	0.16	0.11	0.15	0.11	0.92	0.97

[a] Key variables.
[b] Ratio of communality in Expanded Factor Model (E) to Complete Factor Model (C).

geochemical model for the Late Quarternary deposits in the Rhine-Meuse delta by interpreting each factor carefully. The factor scores were plotted in Fig. 6 to illustrate the geochemical differences between the facies. The non-robust Expanded Factor Model is not discussed in a separate section, because of its similarity to the robust model.

Factor 1. Variation in clay content. The first factor represents the variation of the finest grain size fraction in the riverine deposits. The factor scores of F1 reflect the textural difference between facies very well, placing them in order of increasing clay content. As a result of (hydrodynamic) sorting processes, clay content increases from eolian dune, channel, loam bed, crevasse-levee, organic to flood plain deposits (see Table 1). Note that in the organic deposits the clastic matrix has been diluted by SOM, leading to lower clay contents than in the flood plain deposits. Factor 1 has the highest loadings for the key variables Cs (0.96) and Al_2O_3 (0.75), and almost all other trace elements that have been stripped off (Table 4A). Cesium is highly adsorptive to clay mineral surfaces (Gier and Johns, 2000; Shahwan and Erten, 2001), whereas Al_2O_3 is the most important building block of clay minerals. However, contrary to other regional studies using factor analysis to describe geochemical variation in sedimentary deposits (Huisman and Kiden, 1998; Moura and Kroonenberg, 1990; Tebbens et al., 2001), Al_2O_3 was not found to be the principal variable describing clay mineral content. The explanation is that Al_2O_3 does not occur uniquely in clay minerals, but also in other silicates that occur in larger grain size categories (see F3).

Factor 2. Variation in reduced sulfur and SOM contents. Factor 2 is interpreted as the variation in

Table 4B
Factor loadings (>0.30) for the non-robust complete and Expanded Factor Model (variables sorted on communality ratios[b])

	Complete Factor Model							Expanded Factor Model						E/C[b]
	F1	F2	F3	F4	F5	Communality		F1	F2	F3	F4	F5	Communality	
MnO[a]	0.37	0.38	0.74			0.85				0.90			0.98	1.15
S[a]		0.90				0.87			0.94				0.98	1.13
Cs[a]	0.88					0.90		0.96					0.97	1.08
Na_2O[a]			0.43	0.54	0.55	0.94					0.90	0.34	0.97	1.04
Sr[a]			0.82	0.35		0.94		0.31		0.60	0.63		0.95	1.01
Zr[a]	0.31				0.91	0.98					0.32	0.90	0.99	1.01
SOM[a]	0.58	0.72				0.96		0.57	0.65	0.42			0.96	1.00
Y[a]	0.81				0.35	0.98		0.77		0.34		0.42	0.97	0.99
Al_2O_3[a]	0.64	0.39	0.34	0.51		0.99		0.72	0.36	0.35	0.35		0.95	0.96
Fe_2O_3	0.66	0.48	0.40			0.96		0.67	0.39	0.49			0.95	0.98
TiO_2	0.59	0.40	0.32	0.38	0.47	0.99		0.61	0.36	0.36	0.31	0.48	0.97	0.98
CEC	0.64	0.61				0.95		0.65	0.52	0.38			0.92	0.97
P_2O_5	0.53		0.54			0.76		0.50		0.58			0.73	0.97
Zn	0.69	0.56				0.97		0.70	0.47	0.39			0.94	0.97
Cr	0.60	0.45	0.34	0.42	0.35	0.98		0.64	0.40	0.38	0.31	0.36	0.94	0.96
Rb	0.63	0.38	0.39	0.51		0.98		0.71	0.37	0.38	0.36		0.94	0.96
Carbonate			0.90			0.94				0.64	0.63		0.89	0.95
Cu	0.70	0.52				0.88		0.71	0.44	0.36			0.84	0.95
MgO	0.59	0.46	0.38	0.46		0.96		0.66	0.41	0.40	0.31		0.91	0.95
Ni	0.78	0.31	0.35			0.91		0.71		0.40			0.85	0.94
V	0.68					0.68		0.64					0.64	0.94
Ga	0.59	0.47	0.31	0.51		0.97		0.69	0.43	0.34			0.90	0.93
K_2O	0.54		0.41	0.66		0.97		0.67		0.33	0.53		0.91	0.93
Nb	0.87				0.34	0.94		0.76		0.31		0.41	0.87	0.92
Ba	0.59			0.67		0.93		0.72			0.40		0.82	0.88
As	0.33	0.77		0.33		0.90		0.43	0.66	0.38			0.78	0.87
Pb	0.37	0.37		0.53	0.36	0.77		0.48	0.31	0.31	0.35	0.31	0.64	0.83
							S^2_{mS}	0.29	0.19	0.19	0.16	0.14	0.97	
							$S^2_{\eta E}$	0.40	0.14	0.16	0.09	0.07	0.86	
$S^2_{\kappa C}$	0.34	0.19	0.16	0.14	0.09	0.92	$S^2_{\kappa E}$	0.36	0.16	0.17	0.12	0.09	0.89	0.97

[a] Key variables.
[b] Ratio of communality in Expanded Factor Model (E) to Complete Factor Model (C).

reduced S and SOM. As reduced S precipitates in sulfide compounds like pyrite (FeS_2), it is a very good indicator for low redox environments. Sulfides commonly occur with SOM (Berner, 1971), because the mineralization of SOM results in low redox potentials leading to SO_4 reduction and subsequent precipitation of metal sulfides. This is well demonstrated in Fig. 6, where the organic facies rich in SOM has clearly the highest factor scores for F2. The covariation of As (Table 4A) suggests that this element has been incorporated into sulfides, which is not uncommon in Dutch subsoil sediments (Huisman, 1998).

The smaller loadings of Al_2O_3 and most other (trace) elements on this factor suggest also a grain size effect, because low redox potentials are well preserved in finely grained deposits with low permeability. This effect is shown by the high median factor score on F2 for the flood plain deposits (Fig. 6). Low redox potentials are also found in the loam bed deposits, which have been buried since the Holocene ground water rise (Berendsen, 1998; Berendsen and Stouthamer, 2001), and now occur at depth. The channels and crevasse-levee deposits, with their high permeability forming the main aquifers, are the most oxic facies.

Factor 3. Variation of albitic feldspar and carbonate contents. Factor 3 has high loadings for the key variables Na_2O (0.89) and Sr (0.73). Sodic plagioclase (albite, $NaAlSi_3O_8$) has been suggested to be the most important source of Na_2O in Holocene and Pleistocene sandy sediments in The Netherlands (Huisman and Kiden, 1998). However, the weak covariation of K_2O (0.47) suggests that albite occurs in a mixture with K-feldspar [$KAlSi_3O_8$], or that Na_2O and K_2O jointly occur in albitic K-feldspar

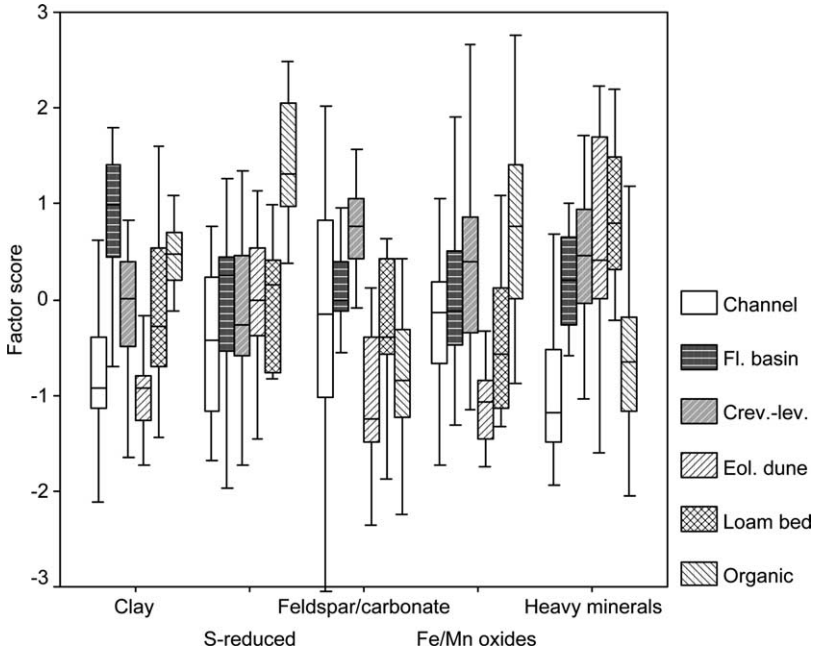

Fig. 6. Boxplots showing the factor score distributions per factor and per facies. Boxes represent interquartiles, with median indicated by the horizontal bar. Whiskers show maximum and minimum scores that fall within 1.5 times the interquartile range of the box measured from the upper and lower quartiles, respectively.

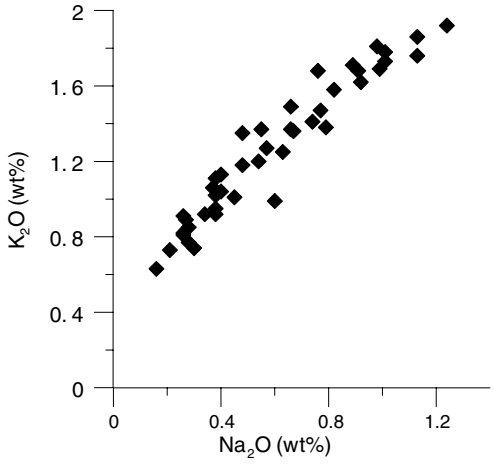

Fig. 7. K_2O vs. Na_2O (weight percentages) for the sandy eolian dune and channel facies.

[$(Na,K)AlSi_3O_8$]. Fig. 7 shows a steady increase of K with Na in a ratio of about 3:2 (weight percentages) in the sandy facies, suggesting that the albitic K-feldspar or the mineralogical mixture has a constant composition. According to Fig. 8, these silicates are enriched in the 20–150 μm grain size fraction, and should be abundant in the channel, eolian dune and crevasse-levee facies. The weak loading for Al_2O_3 on this factor confirms that the variation of Al_2O_3 is not solely dictated by clay mineral content.

The high loading for Sr translates into the variation of carbonate content, because Sr is a common substitute for Ca in carbonates and aragonites (Kinsman and Holland, 1969). This is clear from the identical loading patterns for Sr and carbonate in the Expanded Factor Model. In these deposits, carbonate has mainly been identified as being present as detrital fragments of biogenic origin, which have been concentrated in silty facies (crevasse-levee deposits) along with Na-bearing silicates (see Fig. 6) because of their similar weight. For this reason, carbonate content covaries with Na_2O content, and loads on the same factor.

As F3 has a mixed geochemical significance, the factor score distributions in Fig. 6 should be interpreted with care. On geochemical grounds, it might be better to consider the principal variables separately, i.e., the Na_2O and Sr/carbonate contents. Though Table 4A and Fig. 8 suggest high carbonate contents occurring with intermediate grain sizes, for the eolian dune facies this relation is not present because all carbonate was leached out during the Early Holocene when the eolian dunes were uncovered. The covariation of carbonate and feldspar contents in the crevasse-levee facies tends to dominate this

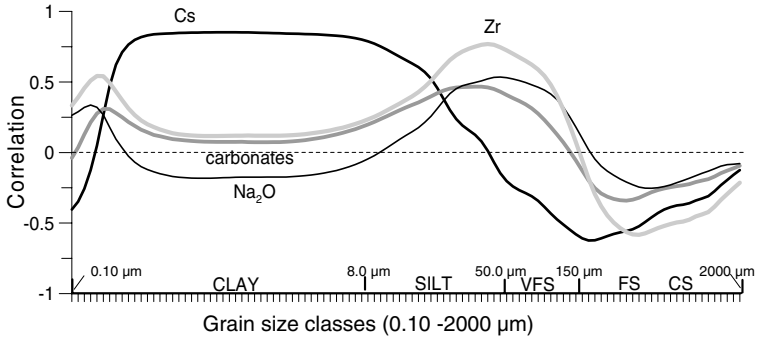

Fig. 8. Correlation of selected key variables to laser particle data (0.10–2000 µm).

relationship, but it should be assessed for the other facies individually.

Factor 4. Variation in Mn and P mineral contents. This factor has the highest loading for key variable MnO (0.91), and weaker loadings for P_2O_5, Fe_2O_3, carbonate, and Sr. The factor is interpreted as the variation in secondarily formed Mn and P-bearing minerals like Mn-(hydr)oxides, vivianite [$Fe_3(PO_4)_2 \cdot 8H_2O$], and apatite [$Ca_5(PO_4)_3(OH)$]. In aquifer sediments, Mn (hydr)oxides commonly occur with Fe(III)-(hydr)oxides under (sub)oxic conditions (Appelo and Postma, 1994; Heron and Christensen, 1994; Larsen and Postma, 1997), while vivianite has been reported as a sink for P and Fe(II) in anaerobic, clayey abandoned channel deposits of the Meuse river (Tebbens et al., 1999). In addition, apatite forms under alkaline conditions in the presence of Ca, which is often the case in the shallow subsoil under arable land as a result of excessive Ca and P-fertilizer application (Sposito, 1989). Hence, MnO, P_2O_5 and Fe_2O_3 are related via secondary mineral phases and therefore occur in the same factor. Fig. 6 shows that the organic and crevasse-levee deposits have been enriched in Mn/P bearing minerals. This is complementary to field observations of vivianite occurrence in reduced organic deposits, and Mn/Fe (hydr)oxides in suboxic crevasse-levee deposits occurring close to the ground water table. The covariation of Sr/carbonate with MnO largely follows grain size (see Fig. 8), because the silty crevasse-levee facies is enriched in both Fe/Mn (hydr)oxides and carbonates. The eolian dune deposits are depleted of Mn/P minerals.

Factor 5. Variation in heavy mineral content. This factor has the highest loading for Zr, and weaker ones for TiO_2, Nb, and Y. It is interpreted as reflecting the variation in heavy mineral content, including zircon (Zr), rutile (TiO_2), and associated trace elements Nb and Y, which are known to be least mobile during weathering processes (Humphris and Thompson, 1982; Thompson, 1973). Fig. 8 suggest that Zr is enriched in the silt fraction, and Fig. 6 confirms that heavy minerals have indeed been enriched in the silty crevasse-levee and loam bed facies. In addition, heavy minerals have been enriched in the sandy eolian dune facies, which are not silty. The explanation is that the sandy eolian dunes have been depleted of most other components by leaching and eolian sorting processes, leaving a relative concentration of heavy minerals.

3.2.2. A conceptual geochemical model

The robust Expanded Factor Model can be translated into a conceptual geochemical model, describing the chemical variation in the sedimentary deposits of the Rhine-Meuse delta plain in terms of independent physical and chemical processes. This approach is very similar to Meng and Maynard (2001), who formulated a conceptual model describing the dominant processes governing water chemistry in a Brazilian aquifer. Tables 5A and 5B list the 4 processes derived from the Expanded Factor Model, being: depositional sorting, peat formation, redox processes, and dissolution/precipitation of secondary minerals. These processes have been grouped as either syn-depositional or post-depositional processes.

Table 5A shows that F1 and F5 represent only one process, whereas the other factors incorporate several processes. For this reason, F1 and F5 are labelled as ''pure'' factors, and F2, F3, and F4 as ''mixed'' factors. The pure factors were easiest to interpret but for the mixed factors, additional geochemical expertize or field observations are needed to assess their meaning properly. Likewise, variables can be classified as ''pure'' or ''mixed'',

Table 5A
Representation of governing processes by the individual factors of the robust Expanded Factor Model

Process	F1	F2	F3	F4	F5
Syn-depositional					
Depositional sorting (clay minerals, feldspar, heavy minerals, carbonate fragments)	×	×	×		×
Peat formation (SOM accumulation)		×			
Post-depositional					
Redox (formation of sulphides, Mn/Fe oxides, vivianite)		×		×	
Precipitation/dissolution (carbonate leaching, apatite precipitation)			×	×	

Table 5B
Representation of governing processes by the key variables in the robust Expanded Factor Model

Process	Cs	Al_2O_3	SOM	S	Na_2O	Sr	MnO	Zr
Syn-depositional								
Depositional sorting	×	×	×	×	×	×		×
Peat formation			×	×				
Post-depositional								
Redox				×			×	
Precipitation/dissolution						×	×	

dependent on whether they are associated with a single or several processes. Table 5B shows that the key variables Cs, Al_2O_3, Na_2O, and Zr are ''pure'', and associated with only one process (syn-depositional grain sorting). The distribution of other constituents has been affected by several processes, and thus are of the ''mixed'' type. Note that Tables 5A and 5B are complementary because a factor can only be ''pure'' when the key variable with the highest loading is ''pure'' as well.

It is concluded that the robust Expanded Model is more than just 5 separated sources of variance extracted from a geochemical dataset, but also accurately describes many geochemical and mineralogical properties of the sedimentary deposits in the Rhine-Meuse delta. These properties can be linked to 4 governing independent physical and geochemical processes, leading to a conceptual model that helps understanding of the geochemical variation of these deposits in detail.

4. Discussion and conclusions

Over the years, there has been extensive discussion about how FA or PCA should be applied (Garrett, 1993; Reimann and Filzmoser, 2000; Reimann et al., 2002). Important issues invariably have been the effect of outliers, and the number of factors and variables that should be included in the multivariate solution. The novel sequential approach presented in this paper clarifies these issues substantially.

Using a heterogeneous geochemical dataset, the authors took the opportunity to assess the effect of multivariate outliers on PCA by comparing the robust and non-robust solutions. Using the robust estimate of the correlation matrix as input for PCA, it was observed from the gradual extension of the Complete Factor Model that the rotated factor solutions are very consistent whereas the non-robust solutions are not. Also, the change from the Complete Factor Model towards the Stripped/Extended Factor Model seems to be more moderate for the robust solution. Therefore, it is concluded that a robust approach is superior to a non-robust approach, and the authors recommend always applying robust PCA or FA to geochemical datasets, minimizing the effect of multivariate outliers on the rotated factor solutions.

The number of factors that should be extracted has been identified in an objective manner. A minimum loading criterion has been identified from the rotated loading distributions produced by robust PCA. The loading distributions showed a persistent bimodality, indicating that Varimax rotation works optimally for the robust case. The minimum loading criterion should therefore be set at the lower boundary of the high end distribution, and used for determining objectively the optimum number of factors that should be extracted. This is an important achievement of seqFA, because so far, there were no adequate objective guides for factor extraction. However, the researcher may decide to deviate from

the optimum, but in doing so should realize that the factor model tends to become over or underspecified with too many or too few factors respectively.

The issue of variable extraction has been addressed in a systematical approach. In this study, similarity criteria and communality sorting were used to identify key variables. The results are therefore objective and statistically optimized, highly condensed, and easy to interpret. However, the results could also be generated in a more flexible way. For instance, if researchers are interested in trace element chemistry, they could manually preset for each factor the trace element with the highest communality as the key variable. Although they would not find the statistical optimum, they would be developing the factor model towards trace element chemistry, because the factor rotation is manipulated (optimized) towards the preset keys. This makes the variable stripping procedure flexible, as the results can always be checked by comparing the explained variance and communalities of the Complete Factor Model and the Expanded Factor Model.

In general, it is concluded that seqFA is a very useful approach to explore heterogeneous geochemical datasets multivariately. Using robust statistics, seqFA leads to a balanced set of factors and variables, because the results are produced in several steps. Within each step, the researcher obtains new information concerning the multivariate structures in the dataset, the stability of the factor model, and the developing identities of the extracted factors. This opens the way to a broad range of applications of robust PCA or FA, including multivariate outlier detection and stability analysis, variance source identification, and variable clustering. In addition, the identification of hidden geochemical processes and properties is improved relative to traditional approaches. This applied study demonstrates that with seqFA, the authors were able to derive a consistent geochemical conceptual model to explain the compositional variability of the heterogeneous Late Quaternary deposits in the Rhine-Meuse delta (The Netherlands).

References

Aitchison, J., 1981. A new approach to null correlations of proportions. Math. Geol. 13, 175–189.

Aitchison, J., 1984. Reducing the dimensionality of compositional data sets. Math. Geol. 16, 617–634.

Appelo, C.A.J., Postma, D., 1994. Geochemistry, Groundwater and Pollution. Balkema, Rotterdam.

Basilevski, A., 1994. Statistical Factor Analysis and Related Methods. Theory and Applications. Wiley, New York.

Berendsen, H.J.A., 1984. Problems of lithostratigraphic classification of Holocene deposits in the perimarine area of the Netherlands. Geol. Mijnbouw 63, 351–354.

Berendsen, H.J.A., 1998. Birds-Eye view of the Rhine-Meuse delta (The Netherlands). J. Coastal Res. 14, 740–752.

Berendsen, H.J.A., Stouthamer, E., 2001. Late Weichselian and Holocene palaeogeography of the Rhine-Meuse delta, The Netherlands. Palaeogeog. Palaeoclim. Palaeoecol. 161, 311–335.

Berner, R.A., 1971. Principles of Chemical Sedimentology. McGraw-Hill, New York.

Cameron, E.M., 1996. Hydrochemistry of the Fraser River, British Columbia: seasonal variation in major and minor components. J. Hydrol. 182, 209–215.

Cattell, R.B., 1966. The scree test for the number of factors. Multivar. Behav. Res. 1, 245–276.

Chork, C.Y., Salminen, R., 1993. Interpreting exploration geochemical data from Outokumpu, Finland: MVE-robust factor analysis. J. Geochem. Explor. 48, 1–20.

Dalton, M., Upchurch, S., 1979. Interpretation of hydrochemical facies by factor analysis. Ground Water 16, 228–233.

Davis, J.C., 1986. Statistics and Data Analysis in Geology. Wiley, New York.

De Vivo, B., Boni, M., Marcello, A., Di Bonito, M., Russo, A., 1997. Baseline geochemical mapping of Sardinia (Italy). J. Geochem. Explor. 60, 77–90.

Duffy, C.J., Brandes, D., 2001. Dimension reduction and source identification for multispecies groundwater contamination. J. Contam. Hydrol. 48, 151–165.

Evans, C.D., Davies, T.D., Wigington Jr., P.J., Tranter, M., Kretser, W.A., 1996. Use of factor analysis to investigate processes controlling the chemical composition of four streams in the Adirondack Mountains, New York. J. Hydrol. 185, 297–316.

Filzmoser, P., 1997. Finding structures of interest in a large dataset using factor analysis. Austrian J. Statistics 26, 27–34.

Frapporti, G., Vriend, S.P., van Gaans, P.F.M., 1993. Hydrochemistry of the shallow Dutch groundwater; interpretation of the national ground water monitoring network. Water Resour. Res. 17, 2993–3004.

Garrett, R.G., 1993. Another cry from the heart. Explore 81, 9–14.

Gier, S., Johns, W.D., 2000. Heavy metal-adsorption on micas and clay minerals studied by X-ray photoelectron spectroscopy. Appl. Clay Sci. 16, 289–299.

Gupta, L.P., Subramanian, V., 1998. Geochemical factors controlling the chemical nature of water and sediments in the Gomti River, India. Environ. Geol. 36, 102–108.

Hakstege, A.L., Kroonenberg, S.B., van Wijk, H., 1993. Geochemistry of Holocene clays of the Rhine and Meuse rivers in the centra-eastern Netherlands. Geol. Mijnbouw 71, 301–315.

Heron, G., Christensen, T.H., 1994. The role of aquifer sediment in controlling redox conditions in polluted groundwater. In: Dracos, T., Stauffer, F. (Eds.), Transport and Reactive Processes in Aquifers. Balkema, Rotterdam, pp. 73–77.

Hesse, P.R., 1971. Cation and anion exchange properties. In: A Textbook of Soil Chemical Analyses. John Murray, London, pp. 88–105.

Huisman, D.J., 1998. Geochemical Characterization of Subsurface Sediments in The Netherlands. Netherlands Institute of Applied Geoscience, TNO.

Huisman, D.J., Kiden, P., 1998. A geochemical record of Late Cenozoic sedimentation history in the soutern Netherlands. Geol. Mijnbouw 76, 277–292.

Humphris, S.E., Thompson, G., 1982. A geochemical study of rocks from the Walvis Ridge, South Atlantic. Chem. Geol. 36, 253–274.

Johnson, R., Wichern, D., 1998. Applied Multivariate Statistical Analysis. Prentice-Hall, London.

Johnsson, M.J., 1993. The system controlling the composition of clastic sediments. In: Johnsson, M.J., Basu, A. (Eds.), Geological Society of America Special Paper 284, vol. 284. GSA, pp. 1–19.

Kaiser, H.F., 1958. The Varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 187–200.

Kinsman, D.J.J., Holland, H.D., 1969. The co-precipitation of cations with $CaCO_3$ – IV. The co-precipitation of Sr^{2+} with aragonite between 16 and 96 °C. Geochim. Cosmochim. Acta 33, 1–17.

Larsen, F., Postma, D., 1997. Nickel mobilization in a groundwater well field: Release by pyrite oxidation and desorption from manganese oxides. Environ. Sci. Technol. 31, 2589–2595.

Lawrence, F.W., Upchurch, S.B., 1982. Identification of recharge areas using geochemical factor analysis. Ground Water 20, 680–687.

Lee, J.Y., Cheon, J.Y., Lee, K.K., Lee, S.Y., Lee, M.H., 2001. Statistical evaluation of geochemical parameter distribution in a ground water system contaminated with petroleum hydrocarbons. J. Eviron. Qual. 30, 1548–1563.

Meng, S.X., Maynard, J.B., 2001. Use of statistical analysis to formulate conceptual models of geochemical behavior: water chemical data from the Botucatu aquifer in Sao Paulo state, Brazil. J. Hydrol. 250, 78–97.

Miall, A.D., 1985. Architectural-elements analysis: a new method of facies analysis applied to fluvial deposits. Earth-Sci. Rev. 22, 261–308.

Miall, A.D., 1996. The Geology of Fluvial Deposits. Springer, Heidelberg.

Morsy, M.A., 1993. An example of application of factor analysis on geochemical stream sediment survey in Umm Khariga area, Eastern Desert, Egypt. Math. Geol. 25, 833–850.

Moura, M.L., Kroonenberg, S.B., 1990. Geochemistry of Quarternary fluvial and eolian sediment in the southeastern Netherlands. Geol. Mijnbouw 69, 359–373.

Nesbitt, W., Young, G.M., 1996. Petrogenesis of sediments in the absence of chemical weathering: effects of abrasion and sorting on bulk composition and mineralogy. Sedimentology 43, 341–358.

Otero, N., Tolosana-Delgado, R., Soler, A., Pawlowsky-Glahn, V., Canals, A., 2005. Relative vs. absolute statisical analysis of compositions: A comparative study of surface waters of a Mediterranean river. Water Res. 39, 1404–1414.

Passmore, D.G., Macklin, M.G., 1994. Provenance of fine-grained alluvium and late Holocene land-use change in the Tyne basin, northern England. Geomorphology 9, 127–142.

Pison, G., Rousseeuw, P.J., Filzmoser, P., Croux, C., 2003. Robust factor analysis. J. Multivar. Anal. 84, 145–172.

Reimann, C., Filzmoser, P., 2000. Normal and lognormal data distribution in geochemistry: dead of a myth. Consequences of geochemical and environmental data. Environ. Geol. 39, 1001–1014.

Reimann, C., Filzmoser, P., Garrett, R.G., 2002. Factor analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem. 17, 185–206.

Rousseeuw, P.J., van Driessen, A., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Shahwan, T., Erten, H.N., 2001. Thermodynamic parameters of Cs+ sorption on natural clays. J. Radioanal. Nucl. Chem. 253, 115–120.

Sposito, G., 1989. The Chemistry of Soils. Oxford University Press, New York/Oxford.

Stouthamer, E., Berendsen, H.J.A., 2000. Factors controlling the Holocene avulsion history of the Rhine-Meuse delta (The Netherlands). J. Sed. Res. 70, 1051–1064.

Suk, H., Lee, K.K., 1999. Characterization of a ground water hydrochemical system through multivariate analysis: clustering into ground water zones. Ground Water 37, 358–366.

Tebbens, L., Veldkamp, A., Kroonenberg, S.B., 1999. Natural compositional variation of the river Meuse (Maas) suspended load: a 13 Ka geochemical record from the upper Kreftenheye and Betuwe Fromations in Northern Limburg. Geol. Mijnbouw 79, 391–409.

Tebbens, L., Veldkamp, A., Kroonenberg, S.B., 2001. The impact of climate change on the bulk and clay geochemistry of fluvial residual channel infillings: The Late Weichselian and Early Holocene River Meuse sediments (The Netherlands). J. Quatern. Sci. 13, 345–356.

Thompson, G., 1973. A geochemical study of the low-temperature interaction of seawater and oceanic igneous rocks. EOR (Trans. Am. Geophys. Union) 54, 1015.

Törnqvist, T.E., Weerts, H.T.J., Berendsen, H.J.A., 1994. Definition of two new members in the upper Kreftenheye and Twente Formations (Quaternary, the Netherlands): a final solution to persistent confusion? Geol. Mijnbouw 72, 251–264.

Tripathi, V.S., 1979. Factor analysis in geochemical exploration. J. Geochem. Explor. 11, 263–275.

van Doesburg, J.D.J., 1996. Particle-size analysis and mineralogical analysis. In: Buurman, P., van Lagen, B., Velthorst, E.J. (Eds.), Manual for Soil and Water Analysis. Backhuys Publishers, Leiden, pp. 251–278.

van Helvoort, P.J., 2003. Complex Confining Layers. A physical and geochemical characterization of heterogeneous unconsolidated fluvial deposits using a facies-based approach. Netherlands Geographical Studies, Utrecht University.