JOURNAL OF CHEMOMETRICS
J. Chemometrics 2004; **18**: 173–182
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.855

Quantification of pollution levels by multiway modelling

Geir Rune Flåten*, Bjørn Grung and Olav M. Kvalheim
Department of Chemistry, University of Bergen, Allégaten 41, N-5007 Bergen, Norway

Received 16 January 2004; Accepted 12 March 2004

Environmental surveys are performed regularly in environmental monitoring of possible industrial pollution. The seabed pollution from industry in or near water, e.g. offshore oil production, can be biologically monitored as benthic species, and their diversity can be used as indicators of pollution. The community disturbance index (CDI) was recently proposed as a quantitative measure of pollution calculated from the benthic data. However, the CDI and other measurements are not optimal for quantifying the changes in level of environmental stress over several years. In this work it is suggested that the total environmental stress over several monitoring surveys can be determined by modelling the natural variation for a set of non-polluted sites. The suggested approach uses a multiway method (Tucker3) to describe the natural variation, and the three-way extension of CDI to quantify the pollution. Implicitly, the traditional soft independent modelling of class analogies (SIMCA) classification is extended to its three-way counterpart. The proposed method is tested on the data from four subsequent surveys performed at the North Sea oil field Embla. Copyright © 2004 John Wiley & Sons, Ltd.

KEYWORDS: multiway analysis; multiway modelling; Tucker; benthic communities; environmental monitoring; community disturbance index (CDI); SIMCA; pollution

1. INTRODUCTION

Benthic species and their diversity can be used as indicators of pollution [1], and this is used to monitor environmental stress near possible pollution sources. One example is the oil production fields in Norwegian waters, where the Norwegian Pollution Control Authority (SFT) demands that benthic samples are collected regularly [2]. This results in a data matrix containing the number of each observed species at every sampling site, which can be used to determine the environmental stress for the different sites. However, as the monitoring surveys are performed regularly, a series of data matrices is obtained. These can be put together in a three-way array which contains the number of each species at every sampling site for all surveys. Hence the data array has three directions or modes of information: (i) the sampling site mode, (ii) the species mode and (iii) the time (survey) mode. The question is whether it is possible to extract and quantify the environmental stress for the different sites *and* possible changes over time.

Massart and co-workers [3,4] have suggested an approach for quantifying the level of disturbance in environmental surveys. Their approach is based on soft independent modelling of class analogies (SIMCA) [5]. Massart *et al*. suggested to use the samples collected far from the oil rig as a reference set representing the natural variation in the area, and this reference set is decomposed by principal component analysis (PCA) [6,7], providing a reference model that describes the natural variation. Subsequently, the remaining samples are compared with the reference model, and the degree of similarity, quantified as the community disturbance index (CDI), is used as an estimate for the pollution level. Obviously, the estimate measures the similarity to the chosen reference set, so natural differences and pollution are detected and quantified together. Thus these two sources of variation cannot be separated without additional information, e.g. study of abundance tables and identification of opportunistic species.

There is a natural exchange of species in benthic communities even without any anthropogenic impact. It is assumed that this natural exchange is similar for all non-polluted sites in a survey area. The motivation for the assumption is that, as there is a normal species community structure at an undisturbed site, it is also reasonable to expect that there is a normal change in the community structure, which is likely to be similar in a constrained survey area. Based on this assumption, the two-way approach suggested by Massart *et al*. can be extended to a three-way approach where the multiway reference model describes the natural variation for the chosen reference set over time. Hence an estimate for the

*Correspondence to: G. R. Flåten, Department of Chemistry, University of Bergen, Allégaten 41, N-5007 Bergen, Norway.
E-mail: post@grflaten.net
Contract/grant sponsor: Norwegian Research Council (NFR); Contract/grant number: 128850/410.

environmental stress over several surveys can be obtained. The reference model is calculated by Tucker3, which can be seen as a three-way extension of PCA. PCA decomposes a two-way data matrix into one set of scores and one set of loadings which, for benthic data, describe variation patterns for the sampling sites and species respectively. Analogously, Tucker3 decomposes a three-way array into one set of scores and two sets of loadings which describe variation patterns for the sampling sites, species and time.

It has been shown [8] that multiway analysis can be used for analysing data from a series of monitoring surveys to explore changes in species composition and relative changes in environmental stress for different sites. However, in order to quantify the environmental stress as outlined in the previous paragraph, it is necessary to extend Massart *et al.*'s approach [3,4] to three-way data. To our knowledge, this type of multiway modelling has not been applied before in any field, but several works are using multiway predictions. In general, they discuss the regression problem, i.e. an *N*-way array of observations is modelled and, subsequently, response values for new samples are predicted using this *N*-way model [9–11]. Louwerse *et al.* [12] predict left-out samples from the remaining ones in a three-way array by treating left-out samples as missing values. In an alternative approach, Louwerse *et al.* [12] suggest to successively leave out parts of the three information directions and to use a combination of all possible models to estimate the-left out parts. Neither of these approaches is suitable for our purpose, although the way of extending two-way theory to *N*-way modelling is adopted.

The present paper gives an introduction to the approach for analysing survey data as outlined by Massart and co-workers [3,4]. Secondly, the two-way approach is extended to the three-dimensional case, and certain difficulties concerning three-way (*N*-way) modelling are discussed. Finally, the suggested approach is tested on a series of monitoring surveys performed in four subsequent years at the North Sea oil field Embla.

2. THEORY/METHODS

2.1. Notation and terminology

Italics are used to represent scalars, bold lower-case letters indicate vectors and bold upper-case letters represent matrices (two-dimensional). The three-way data arrays are labelled by underlined bold capitals, e.g. the core array is labelled $\underline{\mathbf{G}}$, and the term 'mode' denotes a direction in the three-way array. Vectors are column vectors, and superscript 't' signifies transposition. Subscripts are used as indices for vectors, matrices or three-way arrays. Subscripts '*a*', '*b*' and '*c*' indicate how the three-way arrays are matricized. Matricization is the term for unfolding, i.e. the two-dimensional representation of the three-dimensional data arrays [13]. '*A*' matricized arrays have the *B* mode objects (or variables) as row vectors in a (two-way) matrix; thus each row vector contains both the *B* mode and the *C* mode observations/measurements for the corresponding object. '*A*', '*B*' and '*C*' correspond to the sampling site mode, the species mode and the survey (year) mode respectively. The symbol '$\otimes$' denotes the Kronecker product.

2.2. Two-way modelling

In the two-dimensional case the data are organized in an $I \times J$ matrix $\mathbf{X}$, where I is the number of samples and J is the number of observed species. The first step is to select the R samples to include in the reference set $\mathbf{X}_{\text{ref}}$. Subsequently the reference set is decomposed by PCA as

$$\mathbf{X}_{\text{ref}} = \mathbf{T}_{\text{ref}}\mathbf{P}_{\text{ref}}^{t} + \mathbf{E} \tag{1}$$

$\mathbf{T}_{\text{ref}}$ is an $R \times L$ matrix containing the L score vectors, and $\mathbf{P}_{\text{ref}}$ is a $J \times L$ matrix containing the L loading vectors. In two-way PCA the number of components L is often decided by cross-validation [14]. The $I \times J$ matrix $\mathbf{E}$ contains the residuals, i.e. the variation in $\mathbf{X}_{\text{ref}}$ not explained by the model.

When the model is established, the samples are fitted to it. The corresponding score values $\mathbf{t}_{\text{fit}}$ are calculated as

$$\mathbf{t}_{\text{fit}}^{\text{t}} = \mathbf{x}_i^{\text{t}}\mathbf{P}_{\text{ref}} \tag{2}$$

where $\mathbf{x}_i^{\text{t}}$ is the $1 \times J$ vector of observations for sample i. The vector $\mathbf{t}_{\text{fit}}^{\text{t}}$ contains a score for each of the L components in the model for sample i.

Fitted score values are calculated for all I samples. Subsequently the deviations from the model (residuals) are calculated as

$$\mathbf{e}_i^{\text{t}} = \mathbf{x}_i^{\text{t}} - \mathbf{t}_{\text{fit}}^{\text{t}}\mathbf{P}_{\text{ref}}^{\text{t}} \tag{3}$$

where the vector $\mathbf{e}_i^{\text{t}}$ represents the $1 \times J$ residual vector for sample i.

The residual vectors are used to calculate the residual standard deviation (RSD):

$$\text{RSD}_i = \sqrt{\frac{\mathbf{e}_i^{\text{t}}\mathbf{e}_i}{J-L}} \tag{4}$$

where RSD_i is the RSD value for sample i. The denominator represents the degrees of freedom. In Equation (5) the calculation of RSD values for the R samples belonging to the reference set is shown for sample r. The expressions are similar, except that the degrees of freedom are modified [5]:

$$\text{RSD}_{\text{ref},r} = \sqrt{\frac{\mathbf{e}_r^{\text{t}}\mathbf{e}_r}{J-L}\frac{R-1}{R-L-1}} \tag{5}$$

After calculating the RSD_i and $\text{RSD}_{\text{ref},r}$ values, it is possible to evaluate whether the tested samples are similar to those of the reference set. This is done by a regular F-test [5]. Firstly the critical F value F_{crit} is calculated at the selected significance level, here $F_{0.9,(J-L),(J-L)(R-L-1)}$. Secondly the F value is combined with the mean RSD value for the reference set, RSD_{ref}:

$$\text{RSD}_{\text{crit}}^2 = \text{RSD}_{\text{ref}}^2 F_{\text{crit}} \tag{6}$$

This gives RSD_{crit}, which can be interpreted as the border of the model. Samples with RSD values larger than RSD_{crit} do not belong to the model, while samples belonging to the model have RSD values smaller than or equal to RSD_{crit}.

Massart [4] interpreted the ratio of the RSD_i value to the RSD_{crit} value as a measurement of the disturbance (pollution) of sample i. He labelled this ratio the community disturbance index (CDI), and the calculation of the CDI value for sample i is

$$\text{CDI}_i = \frac{\text{RSD}_i}{\text{RSD}_{\text{crit}}} \tag{7}$$

2.3. Three-way modelling

In the three-way case the data are organized in a three-dimensional array $\underline{\mathbf{X}}$ with dimensions $I \times J \times K$. The A mode contains the I samples, the B mode contains the J observed species and the C mode contains the K survey years.

Just as in two-way modelling, a reference set $\mathbf{X}_{\text{ref}}$ is selected and a model describing this reference set is calculated. Tucker3 [15,16], which is a three-way generalization of PCA, is used to decompose the reference set. The Tucker3 model for element x_{ijk} in the data array $\underline{\mathbf{X}}$ is

$$x_{ijk} = \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N} a_{il}b_{jm}c_{kn}g_{lmn} + e_{ijk} \tag{8}$$

where a, b and c are the loadings for the element in the respective modes, while L, M and N are the numbers of components in the A, B and C modes respectively. Contrary to two-way PCA, the number of components can differ in the different modes. g_{lmn} is the core element giving the relative importance of the loading combination $a_{il}b_{jm}c_{kn}$. There are no constraints on the loading combinations that are allowed in the general Tucker model, as opposed to PCA where the first score vector corresponds to the first loading vector and so forth. The scalar e_{ijk} is the residual for element ijk.

Analogously to PCA the Tucker loadings a, b and c describe the variation in the A, B and C modes respectively. Hence, by studying e.g. the a loadings, it is possible to learn about patterns and trends in the A mode. Also in most other ways the Tucker model can be interpreted and used as a three-way PCA. The only exception is studies of interaction effects between the three modes, because there are no constraints on loading combinations, as noted above.

The Tucker3 model for the A mode matricized data, i.e. the $I \times J \times K$ data unfolded to $I \times JK$, $\mathbf{X}_a$, implies

$$\mathbf{X}_a = \mathbf{A}\left[(\mathbf{C} \otimes \mathbf{B})\mathbf{G}_a^{\text{t}}\right]^{\text{t}} + \mathbf{E}_a \tag{9}$$

where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ correspond to the loadings in the respective modes and $\mathbf{G}_a$ is the $L \times MN$ A mode matricized core array. The core array $\mathbf{G}$ has the dimensions $L \times M \times N$ and analogously to $\mathbf{X}$ it can be unfolded along each of the three modes to give equivalent representations.

The dimensions of the loading matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ and the matricized core array are $I \times L$, $J \times M$, $K \times N$ and $L \times MN$ respectively. $\mathbf{E}_a$ is the $I \times JK$ A mode matricized residual array.

Introducing $\mathbf{W}_a \equiv (\mathbf{C} \otimes \mathbf{B})\mathbf{G}_a^{\text{t}}$ simplifies the discussion and makes the similarities to the two-way approach more apparent:

$$\mathbf{X}_a = \mathbf{A}\mathbf{W}_a^{\text{t}} + \mathbf{E}_a \tag{10}$$

Assume that the R samples in the reference set $\underline{\mathbf{X}}_{\text{ref}}$ are selected. The reference set is decomposed by Tucker3, giving the loading matrices $\mathbf{A}_{\text{ref}}$, $\mathbf{B}_{\text{ref}}$ and $\mathbf{C}_{\text{ref}}$ and the A matricized core array $\mathbf{G}_{a,\text{ref}}$. Then the model for the A mode matricized reference set $\underline{\mathbf{X}}_{a,\text{ref}}$ can be formulated analogously to Equation (10) by defining $\mathbf{W}_{a,\text{ref}} \equiv (\mathbf{C}_{\text{ref}} \otimes \mathbf{B}_{\text{ref}})\mathbf{G}_{a,\text{ref}}^{\text{t}}$.

Using the simplified three-way expression from Equation (10), the extension of two way SIMCA is straightforward. When the model of the reference set is established, any new sample $\mathbf{x}_{a,i}$ can be fitted to it:

$$\mathbf{a}_{\text{fit},i}^{\text{t}} = \mathbf{x}_{a,i}^{\text{t}}\mathbf{W}_{\text{ref}}(\mathbf{W}_{\text{ref}}^{\text{t}}\mathbf{W}_{\text{ref}})^{-1} \tag{11}$$

The selected definition of $\mathbf{W}$ implies that the core array, $\underline{\mathbf{G}}$ is 'frozen' after the model is established. The fitted samples $\mathbf{a}_{\text{fit}}$ thus have the same core as the R reference samples. This can be motivated by Equation (8): in a two-way model with two components in both the A and B modes the equation reads $x_{ijk} = \sum_{l=1}^{2}\sum_{m=1}^{2} a_{il}b_{jm}g_{lm} + e_{ij}$, which means that the core array $\mathbf{G}$ is a 2×2 matrix. The general Tucker model has no constraints on the core array; however, in the special case which corresponds to the PCA model, $\mathbf{G}$ is a unitary matrix, i.e. a diagonal matrix with ones on the diagonal. In the general Tucker model, as in the PCA model, the core elements give the relations between the different components, but in the Tucker model the core also gives the components' importance, as they usually are normalized, while the size information in a PCA model is usually given by the A mode components.

A central issue in SIMCA modelling is to describe the variation for a group of samples by PCA and to later compare new samples with this model to determine their similarity to the reference set. Applying the Tucker terminology, the model can be formulated as $\mathbf{AGB}^{\text{t}}$. The scores for new samples are found by a projection on $\mathbf{GB}^{\text{t}}$, which equals $\mathbf{B}^{\text{t}}$ as $\mathbf{G}$ is a unitary matrix. However, if the general two-way Tucker model was to be used instead of PCA, the interrelation between the components could be different from the unitary matrix and it would be necessary to project new samples on $\mathbf{GB}^{\text{t}}$. In this case it would make sense to normalize $\mathbf{G}$ and put the size information in $\mathbf{A}$ in order to be able to compare sizes between reference samples and new samples. The other option for using general Tucker in SIMCA would be to group the core matrix with $\mathbf{A}$; hence scores for new samples would be determined by projection on $\mathbf{B}^{\text{t}}$, but the new scores would be the product $\mathbf{AG}$. The size information is kept, but in the general Tucker model the components can be non-orthogonal and, as the components' interrelationships are determined by the core matrix, differences between the reference set and new samples could occur. Clearly this is a non-desirable situation. Thus it makes sense to 'freeze' the core array after calculating the model for the reference set. In this work the core array is also normalized and all size information is put in the A mode components $\mathbf{A}$.

The sample to be fitted in Equation (11) is A mode matricized. The distance to the model for sample i is calculated as

$$\mathbf{e}_{a,i}^{\text{t}} = \mathbf{x}_{a,i}^{\text{t}} - \mathbf{a}_{\text{fit},i}^{\text{t}}\mathbf{W}_{\text{ref}}^{\text{t}} \tag{12}$$

The dimension of $\mathbf{e}_{a,i}^{\text{t}}$ is $1 \times JK$, i.e. a vector with one element for each species each year.

The three-way RSD values for sample i can be calculated as

$$\text{RSD}_i = \sqrt{\frac{\mathbf{e}_{a,i}^{\text{t}}\mathbf{e}_{a,i}}{(J-M)(K-N)}} \tag{13}$$

The correction for degrees of freedom is a direct extension of the correction used in the two-way analysis, Equation (4), i.e.

the degrees of freedom is the number of variables minus the number of components in the model. Just as in the two-way approach, the RSD values for the R samples in the reference set are different with respect to the correction term [5] for degrees of freedom:

$$\mathrm{RSD}_{\mathrm{ref},r} = \sqrt{\frac{\mathbf{e}_{a,r}^{\mathrm{t}}\mathbf{e}_{a,r}}{(J-M)(K-N)}\frac{R-1}{R-L-1}} \tag{14}$$

If the variation in the fitted samples is equal to that in the modelled ones (reference set), their residuals will be similar to the residuals of the reference samples. This can be tested by means of an F-test, and a border $\mathrm{RSD}_{\mathrm{crit}}$ can be established for the model, in analogy with Equation (5):

$$\mathrm{RSD}_{\mathrm{crit}}^2 = \mathrm{RSD}_{\mathrm{ref}}^2 \, F_{\mathrm{crit}} \tag{15}$$

$\mathrm{RSD}_{\mathrm{ref}}$ is the mean of the RSD values for the samples in the reference set and F_{crit} is the critical F value adjusted for degrees of freedom at the selected level of significance, here $F_{0.9,(J-M)(K-N),(R-L-1)(J-M)(K-N)}$.

2.3.1. *Missing values*

In the studied data set there are some missing values due to minor changes in the sampling sites from one year to another. The standard approach to handling missing values is the expectation maximization (EM) algorithm [17], and it is for example used by Andersson and Bro [18] in their N-way toolbox. The EM algorithm is also implemented in the in-house algorithms developed in this project. Obviously, missing values always introduce some extra uncertainty into the analyses. This results in less reliable predictions for new samples, which can be taken into account in the interpretation of the analysis results, as missing values correspond to sites which are not measured in a particular year. Missing values in the reference set, however, affect the model and thus all predictions, so missing values should be avoided in the reference set.

2.3.2. *Plotting procedures*

In a recent work, Kiers [19] suggested plotting procedures for multiway models which take into account that components might not be orthonormal. He also shows how different modes should be visualized in other mode spaces. Basically he suggests routines for finding orthonormal bases and corresponding co-ordinates. Herein his approach to representing combinations of loadings from two modes is utilized. The objective is to visualize the changes in the benthic communities at the different sampling sites during the surveys. This can be achieved by plotting the combination of A mode (sampling sites) and C mode (survey years) loadings in a B mode (species) space. The tracks that arise in the B mode space are called trajectories and give information about the changes in species composition for the different sampling sites over time. It is also possible to visualize the trajectories for both the fitted samples and the reference samples in one figure by using Kiers' routines. Details of the procedure are given in the Appendix.

2.4. Software

All calculations were performed in MATLAB 6.5 (MathWorks Inc., Natick, MA, USA) using a combination of in-house routines and the N-way toolbox [18]. The in-house modelling and prediction routines utilize Gram–Schmidt iterations as proposed by Kroonenberg et al. [20].

3. DATA

The proposed approach is tested on a benthic microfauna data set from the Embla oil field, which is located on the Norwegian continental shelf. The faunal samples are collected and handled according to the prescriptions given by the Norwegian Pollution Control Authority [2,21]. The monitoring surveys from 1990, 1991, 1992 and 1993 are used. Each survey gives a data matrix with the observed species and their abundance at each sampling site. These matrices are collected in a three-way data array. In all but the 1993 survey, 20 sampling sites were monitored and five replicates were taken at each site. All calculations were performed using the replicates, but all results are reported as median values for each site. In 1993 the sites 315/250, 45/250, 135/250, and 225/250* were not monitored. In the three-way array these are labelled as missing values and they are treated as such in the analyses. Over the years the observed species can change owing to natural factors or anthropogenic impact. The total number of species observed for all four surveys is 357, while the highest number observed in a single year is 182. The species not observed in a survey are set to zero for all sampling sites that year in the data array. Additionally, all juvenile species are removed to stabilize the data [22].

4. RESULTS AND DISCUSSION

The first step in the proposed analytical approach is to choose which samples to include in the reference set. Obviously this is a crucial step, as the reference set is the reference for later detection of abnormal variation in the sampling area, i.e. possible pollution. The inclusion of too many samples in the reference set leads to detection of too few abnormal samples, while the inclusion of too few samples leads to detection of too many abnormal samples. The consequences are that possible polluted samples are not detected or that normal samples are classified as polluted respectively. In SIMCA the reference set is usually chosen by using *a priori* information, likewise the approach suggested in this work is a type of supervised classification. In the case of monitoring around an oil field, there is a set of sampling sites located at various distances from the oil rig, which is the suspected pollution source. That the pollution decreases with increasing distance from the source is a reasonable assumption, and the distance from the oil rig is therefore used as the main criterion for choosing the reference set. Additionally, in the regulations given by the Norwegian Pollution Control Authority it is stated that in environmental surveys on the Norwegian continental shelf at least one of the sampling sites along each of the sampling bearings should be unpolluted (Reference [2], p. 30), otherwise the outer sampling sites have to be moved further away from the

*The sampling sites are labelled *aa*/*bb*, where *aa* gives the direction in degrees with respect to the north and *bb* gives the distance in metres to the oil platform along the given direction.

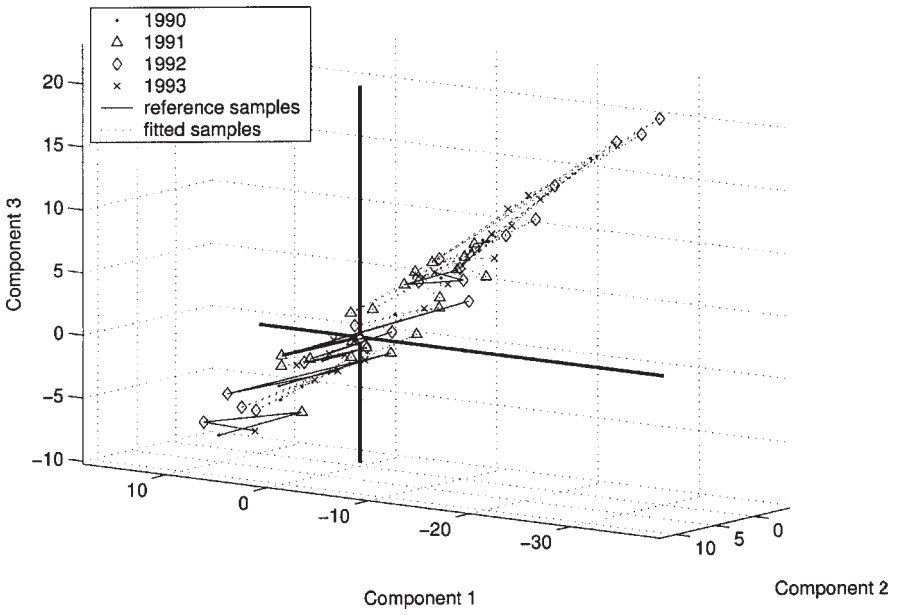

Figure 1. The trajectories for all the samples collected at the Embla field plotted in a species space defined by a (2,3,2) Tucker3 model describing the reference samples. The model explains 49.5% of the variation in the reference set. The lines connect the samples from different years for each site, and the non-reference trajectories are indicated by broken lines.

oil rig. The validity of the reference set is checked by ordination methods, e.g. PCA.

The sampling sites located 1000 m or more away from the oil rig were included in the reference set, and it was validated by PCA ordination, i.e. the sampling sites were verified to be a homogeneous group of samples. The analysis reported later in this paper confirms that the chosen reference set is reasonable.

In Figure 1 the trajectory plot for the Embla data is shown. A (2,3,2) Tucker3 model explaining 57.1% of the variation in the centred reference set was used to produce the plot. In the further analyses a (2,2,2) Tucker3 model is used, but the trajectory plot is more informative for the (2,3,2) model. The trajectory plot depicts, as mentioned earlier, the change in the samples over time in the species space. Hence both differences in species composition and differences in changes in species composition are detectable. The trajectories for the samples not belonging to the reference set are indicated by dotted lines in Figure 1. A first observation is that the reference samples and the non-reference samples are located in different regions of the plot. Obviously this is partly due to the model being built from the reference set, but it also confirms that the chosen reference set is reasonably homogeneous.

Unfortunately the two-dimensional representation of Figure 1 is hard to read, but using a graphical tool with a rotation option gives more information. For instance, the trajectories for the reference samples are generally similar in shape to the one visible in the lower left of the plot. A possible interpretation is that, although the species composition can be slightly different among the reference stations, as indicated by different locations in the species space, there are similarities in the changes over time. In order to investigate this further, it is necessary to study the species abundance tables which were used when building the model. In Table I the 10 most abundant species for sites 180/2000 and 270/1000 in each of the four surveys are shown. Looking at the most abundant species, *Galathowenia oculata*, the figures show that at both stations the highest abundance is found in 1990 and the second highest in 1992, while the abundance is less in 1991 and 1993. *Mysella bidentata* is only ranked among the 10 most abundant species in the 1991–1993 surveys at both sites, while *Phoronis* sp. only makes it to the 10 most abundant species in the years 1990–1992. *Ampharete falcata* shows yet another pattern and is only among the 10 most abundant species in 1991 and 1993 at both sites. Hence, there are some similarities in the changes in species abundance and composition, and, combining the listed and similar observations with knowledge of benthic communities, the nature of the changes over the years can be determined, e.g. it can be determined whether the change over the years is due to normal variation or anthropogenic impact.

A (2,2,2) Tucker3 model explaining 56.9% of the variance in the centred reference set was constructed, and the RSD_i/RSD_{crit} ratios were calculated, Equations (11)–(15). The results are shown in Figure 2. The white bars denote reference samples and each bar is labelled with the distance to the oil rig and the direction with respect to the north.

The ratios reported in Figure 2 comprise the deviation from the model through all the surveys included in the model. Thus they can be interpreted as a measurement of how well the changes in the benthic community at a certain sampling site agree with the overall natural variation, which is described by the model. RSD_i/RSD_{crit} ratios higher than one imply that the corresponding samples are different from the natural variation described by the model, Equation (15). Changes at a certain sampling site that do not correspond to the natural variation, as described by the model, indicate a disturbance (pollution). Natural effects, e.g. landslides, may

Table I. The 10 most abundant species at the 180/2000 and 270/1000 sampling sites for all four surveys. The listed numbers indicate the abundances for the corresponding species

Year	180/2000	270/1000
1990	*Galathowenia oculata* (136)	*Galathowenia oculata* (148)
	Amphiura filiformis (69)	*Amphiura filiformis* (44)
	Phoronis sp. (34)	*Scoloplos armiger* (28)
	Scoloplos armiger (33)	*Goniada maculata* (28)
	Levinsenia gracilis (26)	*Sthenelais limicola* (20)
	Goniada maculata (25)	*Nephtys longosetosa* (14)
	Spiophanes bombyx (22)	*Phoronis* sp. (14)
	Chaetoderma nitidilum (21)	*Levinsenia gracilis* (12)
	Chaetozone setosa (17)	*Eudorellopsis deformis* (11)
	Eudorellopsis deformis (13)	*Harpinia antennaria* (11)
1991	*Mysella bidentata* (91)	*Amphiura filiformis* (78)
	Amphiura filiformis (72)	*Galathowenia oculata* (72)
	Galathowenia oculata (67)	*Mysella bidentata* (41)
	Phoronis sp. (35)	*Phoronis* sp. (31)
	Chaetozone setosa (32)	*Goniada maculata* (24)
	Goniada maculata (28)	*Ampharete falcata* (22)
	Scoloplos armiger (19)	*Scoloplos armiger* (15)
	Chaetoderma nitidilum (16)	*Chaetozone setosa* (15)
	Sthenelais limicola (14)	*Sthenelais limicola* (12)
	Ampharete falcata (13)	*Chaetoderma nitidilum* (11)
1992	*Galathowenia oculata* (107)	*Galathowenia oculata* (157)
	Amphiura filiformis (60)	*Chaetozone setosa* (68)
	Mysella bidentata (50)	*Amphiura filiformis* (50)
	Chaetozone setosa (38)	*Mysella bidentata* (36)
	Harpinia antennaria (27)	*Phoronis* sp. (27)
	Goniada maculata (23)	*Goniada maculata* (26)
	Phoronis sp. (23)	*Harpinia antennaria* (22)
	Chaetoderma nitidilum (18)	*Sthenelais limicola* (15)
	Eudorellopsis deformis (16)	*Tharyx/Caulleriella* sp. (14)
	Nemertini indet. (14)	*Levinsenia gracilis* (13)
1993	*Galathowenia oculata* (75)	*Galathowenia oculata* (90)
	Amphiura filiformis (49)	*Amphiura filiformis* (58)
	Harpinia antennaria (24)	*Mysella bidentata* (40)
	Chaetozone setosa (23)	*Ampharete falcata* (27)
	Mysella bidentata (22)	*Scopelocheirus hopei* (27)
	Goniada maculata (19)	*Chaetozone setosa* (22)
	Sthenelais limicola (18)	*Sthenelais limicola* (21)
	Ampharete falcata (14)	*Goniada maculata* (21)
	Eudorellopsis deformis (14)	*Harpinia antennaria* (19)
	Trichobranchus roseus (13)	*Chaetoderma nitidilum* (13)

occur at certain sites in the sampling area, but usually such effects can be distinguished from pollution effects by considering the list of observed species and their abundance. Inspection of this list is of course mandatory in the analysis of benthic count data.

As can be seen in Figure 2, 270/250 is the sampling site most different from the model, implying that the benthic community at this site differs most from the common features of the reference set as captured by the model. From this it can be interpreted that 270/250 is the overall most disturbed site in the surveys, or that it is the site most influenced by non-natural effects during the monitoring period, e.g. the site with the highest increase in pollution. Whether the deviation from the model is due to pollution can only be confirmed by consulting the species abundance tables, and the 10 most abundant species at 270/250 in the four surveys are listed in Table II. A detailed discussion of the abundance tables is not included here, but a few observations are evident from Table II. Look for instance at the most abundant species, *Chaetozone setosa*, which is known to be an opportunistic species, i.e. it survives in environments with stress levels that would remove other species. The number of observed *Chaetozone setosa* in 1992 is almost six times as high as in 1991, while there is a decline for most of the other species. The *Chaetozone setosa* abundance remains high in the 1993 survey.

Most of the other sites located 250 m away from the oil rig are also clearly different from the model, which indicates that they may be polluted. Generally the sites become more similar to the model the further away from the oil rig they are located. This supports the proposed hypothesis that the pollution is more severe near the oil rig. However, care must be taken, as the model is built using the samples furthest away from the oil rig. If there is a change in the natural environment with increasing distance to the oil rig, such as seabed changes, depth changes or similar changes, the difference reported by the model is just a natural difference and not pollution. Again, pollution can only be confirmed by consulting the species abundance tables or other additional measurements.

The RSD_i/RSD_{crit} ratios depict, as mentioned, the differences between the variation of the non-reference samples and the natural variation over all survey years as described by the model. A natural next question is whether the multiway model also can be used to diagnose the pollution levels for each year. This can be tested by splitting the $1 \times JK$ residual vectors into $1 \times J$ survey-specific vectors. The $1 \times J$ residual vectors are used to calculate RSD values according to Equations (13) and (14), and the survey specific RSD_i/RSD_{crit} ratios are calculated.

The annual RSD_i/RSD_{crit} ratios for the four survey years are shown in Figure 3, where each subfigure is labelled with the corresponding survey year, and the bars corresponding to the reference samples are indicated in white. There are no bars for the last four sites in Figure 3(d), as these sites were not monitored in 1993. Clearly, 1990 is the year with the lowest overall disturbance, Figure 3(a), and for this year only 135/250 is disturbed, i.e. has an RSD_i/RSD_{crit} ratio higher than one. For the other survey years the sites located 250 and 500 m away from the oil rig are disturbed. The exception to this trend is 90/500 in the 1991 survey, which apparently is not disturbed. The 270/250 site, which is the most disturbed site according to the RSD_i/RSD_{crit} ratios for all the years, Figure 2, has low annual RSD_i/RSD_{crit} ratios for both 1990 and 1991. However, in 1992 and 1993 this site is clearly the one which is most different from the model. This strengthens the earlier hypothesis that 270/250 was the sampling site which had the most different changes in species abundance and composition compared with the reference set over the monitoring period. Possibly there is a general trend towards higher RSD_i/RSD_{crit} ratios over the years, and this can indicate that there was an increase in experienced stress on the benthic communities, but, as mentioned earlier such conclusions need to be tested by e.g. consulting the species abundance tables.

The annual RSD_i/RSD_{crit} ratios are not identical to the CDI values proposed by Massart *et al.*, Equation (6), as the multiway model comprises the variation for all the surveys (see previous discussion). CDI values, on the other hand, report deviations from a model describing the natural

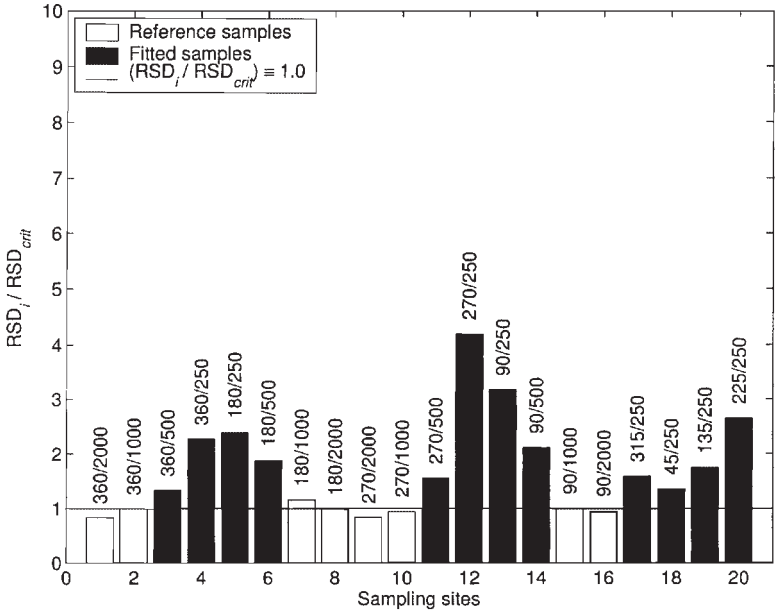

Figure 2. The RSD_i/RSD_{crit} ratios for the stations at the Embla field. Each bar is labelled with the corresponding sampling site.

Table II. The 10 most abundant species at the 270/250 sampling site for all four surveys. The listed numbers indicate the abundances for the corresponding species

1990	1991	1992	1993
Chaetozone setosa (188)	*Chaetozone setosa* (135)	*Chaetozone setosa* (682)	*Chaetozone setosa* (479)
Galathowenia oculata (70)	*Galathowenia oculata* (129)	*Galathowenia oculata* (84)	*Nemertini* indet. (101)
Nemertini indet. (15)	*Phoronis* sp. (68)	*Nemertini* indet. (48)	*Galathowenia oculata* (82)
Levinsenia gracilis (15)	*Mysella bidentata* (55)	*Arctica islandica* (46)	*Tharyx/Caulleriella* sp. (32)
Glycera alba (14)	*Amphiura filiformis* (47)	*Glycera alba* (26)	*Timoclea ovata* (22)
Ampharete falcata (13)	*Goniada maculata* (25)	*Mysella bidentata* (24)	*Ampharete falcata* (21)
Cephalaspidea indet. (12)	*Ampharete falcata* (20)	*Goniada maculata* (16)	*Goniada maculata* (19)
Phoronis sp. (10)	*Levinsenia gracilis* (12)	*Ampharete falcata* (15)	*Thyasira flexuosa* (17)
Chaetoderma nitidilum (8)	*Arctica islandica* (12)	*Phaxas pellucidus* (15)	*Levinsenia gracilis* (15)
Philine scabra (8)	*Nemertini* indet. (8)	*Tharyx/Caulleriella* sp. (14)	*Glycera alba* (11)

variation for a single year. Nevertheless, to compare them with corresponding CDI values is an interesting validation of the annual RSD_i/RSD_{crit} ratios. Hence the CDI values for each survey were calculated according to Equations (1)–(6). The same reference set as used in the three-way modelling was employed and all the models were composed of two PCA components. In Figure 4 the annual RSD_i/RSD_{crit} ratios are plotted versus the CDI values for the four survey years. The symbols denoting the different years are specified in the legend of the figure.

Figure 4 shows that for all surveys except the 1990 survey the correlation between the annual RSD_i/RSD_{crit} ratios and the CDI values is 0.91 or higher. However, it is clear that the correspondence between the two measurements is best for the 1991 survey. The 1992 survey also has a fairly good correspondence, but it seems like the less disturbed samples generally have lower CDI values than RSD_i/RSD_{crit} ratios, while for the most disturbed samples the CDI values are the highest. The results from the 1993 survey are biased towards higher CDI values than RSD_i/RSD_{crit} ratios. The observed deviations for all surveys can be explained by the differences in the two underlying models, as describing the natural variation over several years is likely to incorporate some natural benthic community dynamics which are not captured by a single-year model. A closer look at the 1990 survey demonstrates this point well. In Figure 5 the leverage values for two-way (top) and three-way (bottom) models are shown together with the sum of abundances (middle) for the species observed in the 1990 survey. Leverage values can in this context be regarded as the respective species' importance or weight in the model, and a high value reflects large importance. In the 1990 survey, Figure 5 (middle), *Amphiura filiformis* and *Chaetozone setosa* have similar abundances. However, in the two-way model (top), *Amphiura filiformis* has high importance while *Chaetozone setosa* has lower importance. In the three-way model (bottom) the situation is the opposite: *Amphiura filiformis* has low importance while *Chaetozone setosa* has higher importance. The two-way model also gives a better representation of the characteristics particular to the 1990 species composition, which was expected.

5. CONCLUSION

An approach to multiway modelling of natural variation and quantification of deviations (pollution) from this natural variation is proposed. The multiway approach is deduced

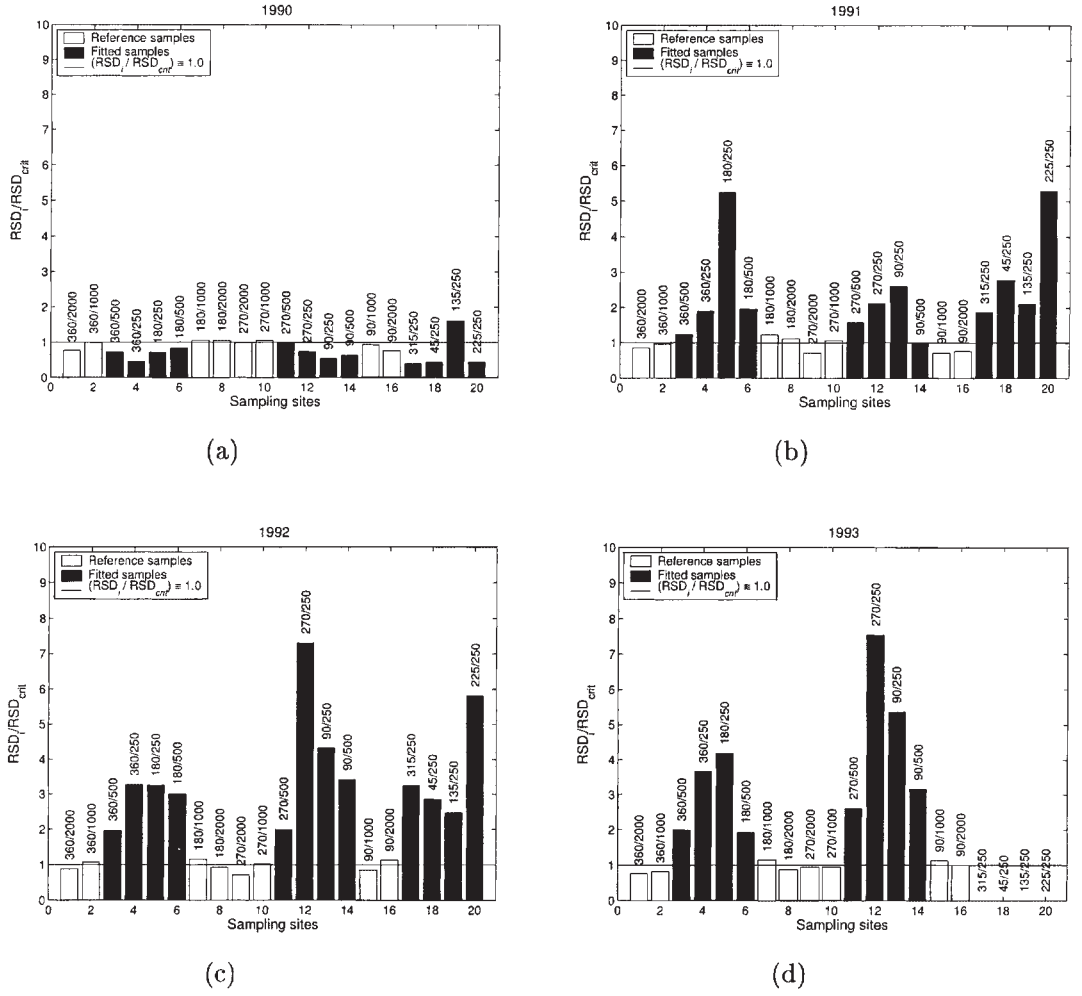

Figure 3. The survey-specific RSD_i/RSD_{crit} ratios for the Embla data. The reference samples are indicated by white bars.

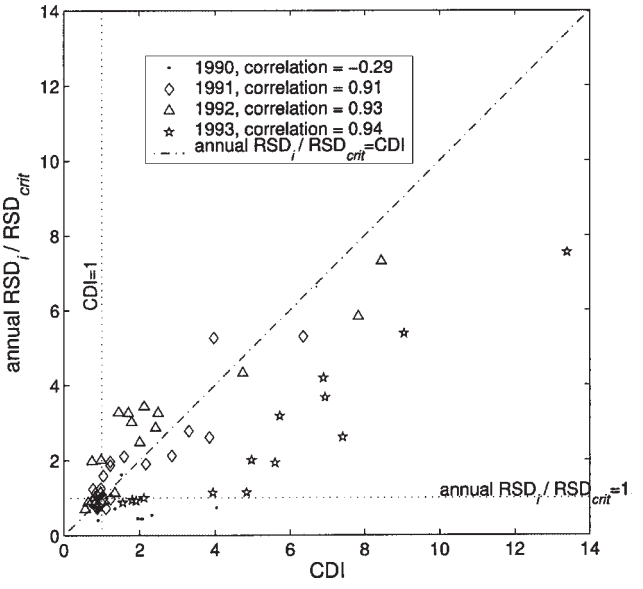

Figure 4. The annual RSD_i/RSD_{crit} ratios plotted against the CDI values for the samples collected at the Embla field. The dotted lines show the border of the models.

from existing two-way theory and tested on a data set consisting of four monitoring surveys performed on an offshore oil field in four subsequent years.

Clearly the proposed method highlights and describes changes that can be confirmed by consulting the species abundance tables. Thus it is a good explorative tool which can be used to spot similarities and differences between sites and the changes in their species composition over time. Additionally, the RSD_i/RSD_{crit} ratios give a quantification of the differences between sampling sites, and, as the RSD_i/RSD_{crit} ratio is the statistic from an *F*-test, a statistical limit for polluted samples is obtained. However, care must be taken, as the RSD_i/RSD_{crit} ratios report differences in general, so it is necessary to confirm that the observed differences are due to pollution by consulting species abundance tables or other additional measurements.

The RSD_i/RSD_{crit} ratios were compared with the CDI values, and generally the correlation between the two measurements was high, more than 0.9. However, there were some deviations, and especially for the 1990 data set the correlation was poor, −0.3. A closer study showed that the deviations were most likely due to different species having

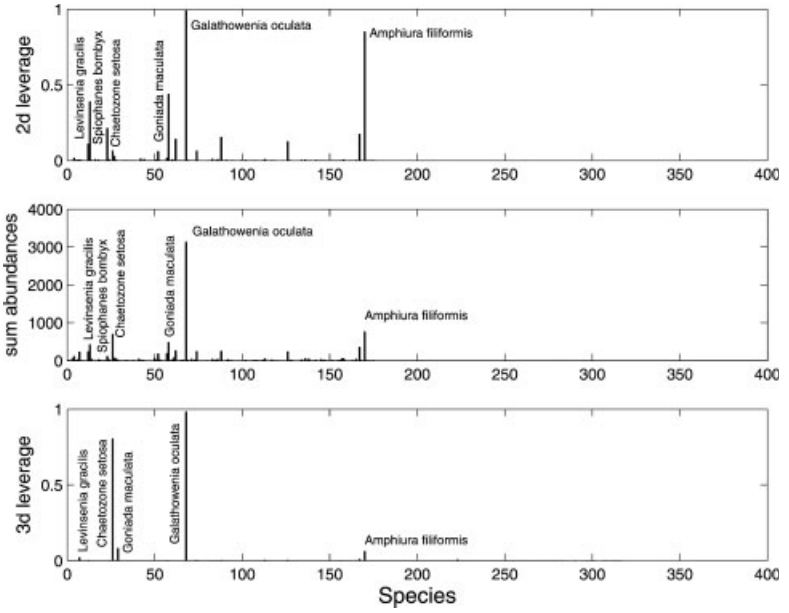

Figure 5. The leverage values for the 2d model used to calculate the 1990 CDI values (top), the total numbers of species summed over all stations in the 1990 survey (middle) and the leverage values for the 3d model used to calculate the 1990 RSD_i/RSD_{crit} ratios (bottom).

different importance in the three-way and two-way models. Hence, if the environmental stress for a single survey is studied, it is advisable to use the CDI values.

Acknowledgments

Norsk Hydro ASA and Phillips Petroleum Company Norway are thanked for giving access to their data from the regular survey and monitor programme of the oil fields on the Norwegian shelf. The Norwegian Research Council (NFR) is thanked for their financial support of the project (grant 128850/410).

APPENDIX. KIERS' PLOTTING PROCEDURE

Kiers [19] recently suggested plotting procedures for multiway models. Below, his procedure for 'trajectory plots', i.e. the combinations of loadings from two modes plotted in the third mode's space, is outlined for the plot of the combination of *A* mode (sampling sites) and *C* mode (survey years) loadings in the *B* mode (species) space.

$\mathbf{W}_b$ is introduced:

$$\mathbf{W}_b \equiv (\mathbf{A} \otimes \mathbf{C})\mathbf{G}_b^{\mathrm{t}} \tag{16}$$

where $\mathbf{A}$ and $\mathbf{C}$ are the *A* and *C* mode loadings, respectively and $\mathbf{G}_b^{\mathrm{t}}$ is the *B* mode matricized core array.

Kiers' procedure requires that an orthogonal basis for $\mathbf{B}$ is found. Subsequently the combinations of *A* and *C* mode loadings, $\mathbf{W}_b$, are plotted utilizing this orthogonal basis.

1. Find $\mathbf{H}$ such that, for $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{H}, \tilde{\mathbf{B}}^{\mathrm{t}}\tilde{\mathbf{B}} = \mathbf{I}$ holds. The transformation matrix $\mathbf{H}$ can be found by Gram–Schmidt orthonormalization, for instance.
2. Plot rows of $\tilde{\mathbf{W}}_b = \mathbf{W}_b(\mathbf{H}^{\mathrm{t}})^{-1}$ to display the *IK* combinations of *A* and *C* mode entities (trajectories).
3. Plot rows of $\tilde{\mathbf{B}}$ to display projections of the *J* original axes onto the subspace.

When 'new' samples are fitted to the existing model, we obtain new values for the sampling site scores $\mathbf{a}_{\mathrm{fit}}^{\mathrm{t}}$. The 'new' $\mathbf{W}_b$ value $\mathbf{W}_{b,\mathrm{fit}}$, is defined as

$$\mathbf{W}_{b,\mathrm{fit}} = (\mathbf{a}_{\mathrm{fit}}^{\mathrm{t}} \otimes \mathbf{C})\mathbf{G}_b^{\mathrm{t}} \tag{17}$$

Apart from exchanging $\mathbf{W}_b$ and $\mathbf{W}_{b,\mathrm{fit}}$, the plotting procedure remains unchanged. Accordingly the same orthogonal basis is used, implying that the trajectories for both the fitted samples and the reference samples can be shown in one figure.

REFERENCES

1. Olsgard F, Gray JS. A comprehensive analysis of the effects of offshore oil and gas exploration and production on the benthic communities of the Norwegian continental shelf. *Marine Ecol. Prog. Ser.* 1995; **122**: 277–306.
2. Nilssen I. Environmental monitoring of petroleum activities on the Norwegian shelf: guidelines. *Tech. Rep. TA:1641/1999*, Norwegian Pollution Control Authority (SFT), Oslo, 1999.
3. Massart B, Kvalheim OM, Libnau FO, Ugland KI, Tjessem K, Bryne K. Projective ordination by SIMCA: a dynamic strategy for cost-efficient environmental monitoring around offshore installations. *Aquat. Sci.* 1996; **58**: 120–138.
4. Massart B. Environmental monitoring and forecasting by means of multivariate methods. *PhD Thesis*, University of Bergen, 1997.
5. Wold S. Pattern recognition by means of disjoint principal components models. *Pattern Recogn.* 1976; **8**: 127–139.
6. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometrics Intell. Lab. Syst.* 1987; **2**: 37–52.

7. Carlson R. 1992. *Design and Optimization in Organic Synthesis, of Data Handling in Science and Technology*, Vol. 8. Elsevier: Amsterdam, 1992.
8. Flåten GR. Dynamic environmental monitoring by means of multivariate modelling. *PhD Thesis*, University of Bergen, 2002.
9. Bro R. Multi-way calibration. Multi-linear PLS. *J. Chemometrics* 1996; **10**(1): 47–62.
10. Smilde AK, Kiers HAL. Multiway covariates regression models. *J. Chemometrics* 1999; **13**: 31–48.
11. Bro R. PARAFAC. Tutorial and applications. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 149–171.
12. Louwerse DJ, Smilde AK, Kiers HAL. Cross-validation of multiway component models. *J. Chemometrics* 1999; **13**: 491–510.
13. Kiers HAL. Towards a standardized notation and terminology in multiway analysis. *J. Chemometrics* 2000; **14**: 105–122.
14. Wold S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 1978; **20**: 397–405.
15. Tucker L. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966; **31**: 279–311.
16. Kroonenberg PM. *Three-mode Principal Component Analysis*. DSWO Press: Leiden, 1983.
17. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B* 1977; **39**: 1–38.
18. Andersson CA, Bro R. The N-way toolbox for MATLAB. *Chemometrics Intell. Lab. Syst.* 2000; **52**: 1–4.
19. Kiers HAL. Some procedures for displaying results from three-way methods. *J. Chemometrics* 2000; **14**: 151–170.
20. Kroonenberg PM, ten Berge JFM, Brouwer P, Kiers H. Gram–Schmidt versus Bauer–Rutishauser in alternating least-squares algorithms for three-mode principal component analysis. *Comput. Statis.* 1989; **5**: 81–87.
21. Paris-Comission. *Guidelines for Monitoring Methods to be Used in the Vicinity of Platforms in the North Sea*. Chameleon: London, 1985.
22. Gray JS, Clarke KR, Warwick RM, Hobbs G. Detection of initial effects of pollution on marine benthos: an example from the Ekofisk and Eldfisk oilfields, North Sea. *Marine Ecol. Prog. Ser.* 1990; **66**: 285–299.

