Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Computational Statistics & Data Analysis 46 (2004) 155 – 171

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

# A generalization of STATIS-ACT strategy: DO-ACT for two multiblocks tables

Myrtille Vivien*, Robert Sabatier

*Laboratoire de Physique Moléculaire et Structurale, UMR 5094, Faculté de Pharmacie,*
*15 Av. Charles Flahault, BP 14491, 34093 Montpellier Cedex 5, France*

Received 1 July 2002; received in revised form 27 June 2003

**Abstract**

A new strategy is introduced for analyzing two multiblocks tables: DO-ACT. This method is closely related to the STATIS (or ACT) methodology and the Tucker inter-battery method. The length of two multiblocks are not necessarily the same and the optimal solution obtained is that of a global optimization problem. The advantage of using DO-ACT is that the first step provides a summary of the two multiblocks tables, in the second step two optimal representations (one for each multiblock) of the observations can be plotted and in the third step a global description of each table of each multiblock can be made. An example of DO-ACT performance is illustrated with a real data set. The program implementing the method has been developed using the S-Plus 6.0® (2000) language.

*Keywords:* Multiblocks tables; STATIS (or ACT) method; RV coefficient; Hilbert–Schmidt scalar product; Tucker inter-battery method

## 1. Introduction

Many generalizations of standard linear multivariate analysis like principal component analysis (PCA) or canonical correlation analysis (CCA) have been proposed for study three or more sets of variables, that is to say a multiblock table. In this paper, a multiblock table (or a multiblock) is a set (or group) of matrices measured with the

* Corresponding author. Tel.: +33-4-67-548-088; fax.: +33-4-67-548-649.
*E-mail addresses:* mvivien@pharma.univ-montp1.fr (M. Vivien), sabatier@pharma.univ-montp1.fr (R. Sabatier).

0167-9473/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-9473(03)00146-4

same observations. Most of the extensions of CCA determine the dimension (or rank) of the model step by step, using linear combinations of the variables of each matrix and optimizing a criterion (Horst, 1961; Carroll, 1968; Saporta, 1975; Van de Geer, 1984; Escofier and Pagès, 1984, 1994; Casin, 2001). Only few methods optimize a global criterion and generate in a right way a solution with fixed rank (Gower, 1975; Lavit, 1988; Lavit et al., 1994). Some methodologies have also been developed for multiway contingency tables (Carlier and Kroonenberg, 1996) and for one multiblock contingency table (Bécue-Bertaut and Pagès, 2003) generalizing the correspondence analysis method. In the case of two multiblocks, some methods exist for building regression models. In the chemometric literature, generalizations of the PLS regression method can be found (Bro, 1996; Westerhuis et al., 1998; Smilde et al., 2000). But if the problem is the simultaneous analysis of a pair of multiblock tables, only one method exists: STATICO (Simier et al., 1999). Unfortunately, this method works with two data sets of same dimension, that is to say with data arranged in two three-way arrays. To be able to compare the results of our new methodology, we propose a new version of STATICO which allows the use of two multiblocks (with different number of variables in each matrix).

The purpose of this paper is to introduce a new approach for finding common dimensions inside two multiblocks tables with different length and for describing each one of them. This global approach, referred as DO-ACT (DOuble ACT), is closely related to the STATIS (or ACT) strategy and Tucker inter-battery method (Tucker, 1958) and is made of three successive steps. This multiblock method, can be beneficial to analyze large data sets where the measurements are organized with an important number of tables (or length). The DO-ACT methodology is used to take into account the block structure (two multiblocks) of the data set during the calculations. The Section 2 of this paper gives the general notations; in Section 3 the main features of STATIS and Tucker inter-battery methodologies are briefly presented. Section 4 describes the DO-ACT approach and before concluding, an application on a real data set is made in Section 5.

## 2. Notation

The two multiblocks (or sets of matrices) are referred by $\{X_k\}_k$ $(k = 1,\dots,K)$ and $\{Y_l\}_l$ $(l = 1,\dots,L)$. $K$ and $L$ are the length of the two multiblocks, and it is assumed without loss of generality that $K$ and $L$ are not simultaneously equal to unity. Each $X_k$ and $Y_l$ are $n \times p_k$ and $n \times p_l$ tables, where all the variables are measured on the same $n$ observations. Without loss of generality all the variables, the columns of the matrices, are assumed to be $D$-centered (or/and scaled) with respect of a weighting matrix $D$ of $n \times n$ dimension whose positive diagonal elements sum to 1. In most of the cases $D$ is the uniform weighting: $D = (1/n)Id_n$, where $Id_n$ is the identity matrix. With this notation, the $D$-scalar product between two variables (or vectors) $x$ and $y$ in $\mathbb{R}^n$ is defined by $(x, y)_D = x'Dy$ (where $x'$ is the transposed vector, or matrix, $x$). This scalar product is equal to the covariance between $x$ and $y$ if the two vectors are $D$-centered.

We define $K+L$ statistical triplets $(X_k, Q_{X_k}, D)$ $(k=1,\ldots,K)$ and $(Y_l, Q_{Y_l}, D)$ $(l=1,\ldots,L)$ where $D$ is the metric in $\mathbb{R}^n$ as defined above and $Q_{X_k}$ and $Q_{Y_l}$ are the metrics in the observations spaces $\mathbb{R}^{p_k}$ and $\mathbb{R}^{p_l}$, respectively. $Q_{X_k}$ and $Q_{Y_l}$ are $p_k \times p_k$ and $p_l \times p_l$ symmetrical definite positive matrices. They are used to calculate the scalar products between the observations in $\mathbb{R}^{p_k}$ and $\mathbb{R}^{p_l}$ respectively. The identity matrix ($Id_{p_k}$ or $Id_{p_l}$) is the most widely used metric, but it can be possible to choose another one, see for example (Cailliez and Pagès, 1976; Tenenhaus and Young, 1985). $\{(X_k, Q_{X_k}, D)\}_k$ and $\{(Y_l, Q_{Y_l}, D)\}_l$ are also used to refer to the two multiblocks.

$W_{X_k}D = X_k Q_{X_k} X_k' D$ is an $n \times n$ matrix called operator and denotes the scalar products between observations in $\mathbb{R}^{p_k}$. This matrix is similar to $V_k Q_{X_k} = X_k' D X_k Q_{X_k}$ which is the variance–covariance matrix between the $X_k$ variables if $Q_{X_k} = Id_{p_k}$. To end, we recall that the RV-coefficient (Escoufier, 1973) measures the proximity between two statistical triplets $(X_k, Q_{X_k}, D)$ and $(X_{k'}, Q_{X_{k'}}, D)$ and is defined as

$$RV(W_kD, W_{k'}D) = \frac{tr(W_kDW_{k'}D)}{\sqrt{tr(W_kDW_kD)tr(W_{k'}DW_{k'}D)}}.$$

## 3. The STATIS and Tucker inter-battery methods

In this section, the main purposes of the STATIS and Tucker inter-battery procedures are recalled. Details and proofs can be found in Lavit (1988) and Lavit et al. (1994) for the STATIS methodology and in Tucker (1958) and Chessel and Mercier (1993) for the Tucker inter-battery analysis.

### 3.1. STATIS (or ACT) method

In this part, only one multiblock $\{(X_k, Q_{X_k}, D)\}_{k=1,\ldots,K}$ is considered. For readability, in this section, the subscript $X_k$ in the operator notation will be dropped. The aim of STATIS is to find an operator, called the *compromise*, summarizing the $W_kD$'s $(k=1,\ldots,K)$ the best, in the sense of a criterion. Next, the second goal of STATIS is to analyze this *compromise*, like in PCA, and to plot the observations of the $K$ tables onto the first components.

The *interstructure* step of STATIS begins to calculate the scalar products between the $K$ operators $W_kD$ and next allows to make a graphical representation of the operators (like in multidimensional scaling) in a space of low dimension (two or three). This graphical configuration allows us an overall graphical comparison of the $K$ tables.

The $C$ matrix of dimension $K \times K$ is made of the scalar products between the $W_kD$'s operators. The element on line $k$ and column $k'$ of $C$ is $C_{k,k'} = tr(W_kDW_{k'}D)$. Rather than work with $C$, we can choose to work with the RV-coefficients, which can be seen as the cosine of the angle between the corresponding operators (Escoufier, 1973). Moreover, the user can weight the operators via a diagonal $K \times K$ matrix $\Pi = diag(\pi_k)$, $k=1,\ldots,K$. The diagonalization of the $C\Pi$ matrix leads to $K$ scaled eigenvectors $\{p_r\}_{r=1,\ldots,K}$, belonging to $\mathbb{R}^K$ and associated to the $r$th eigenvalue $\lambda_r$.

Then, the plot given by $(\sqrt{\lambda_r}p_r, \sqrt{\lambda_s}p_s)$ ($r,s$ in $\{1,\dots,K\}$ with $r \neq s$) provides an Euclidean representation of the $K$ operators in the $(r,s)$ plane.

The next step of the STATIS procedure is the compromise. The purpose is to find an $n \times n$ operator $W_cD$. This operator is a "consensus" between the $W_kD$'s operators, according to a criterion. The comprise is required to be a weighted mean of the initial operators: $W_cD = \mu \sum_{k=1}^{K} \pi_k l_k W_k D$, where $\mu$ is a scaling coefficient and $l = (l_1, l_2, \dots, l_K)$ is the weight vector of the $K$ operators. The criterion used to find the $l$ vector is $\|W_cD\|^2 = tr(W_cDW_cD)$, where $\|.\|$ is the Hilbert–Schmidt norm of the $W_cD$ operator (Robert and Escoufier, 1976). The solution of the optimization of $\|W_cD\|^2$ under the scaling constraint for the $l$ vector ($l'l = 1$) is the eigenvector associated with the first eigenvalue of the matrix $C\Pi$. Because it is a positive linear combination of positive semi-definite operators, $W_cD$ is also positive semi-definite (see later for the proof).

The last step, the *intrastructure*, consists in comparing the views of the observations given by the compromise with the view given by the initial operators $W_kD$. Then the PCA of $W_cD$ is performed and all the observations of all the operators are plotted onto the first spaces spanned by the $A$ principal components $\{c_a\}_{a=1,\dots,A}$, elements of $\mathbb{R}^n$ associated with the eigenvalues $\lambda_{c,a}$. The coordinate of the $i$th observation of the $k$th table on the $a$th component is given by the $i$th coordinate of $W_kDc_a/\sqrt{\lambda_{c,a}}$.

### 3.2. Tucker inter-battery method

In this part, only one table $X$ and one table $Y$ (i.e. $L=1$ and $K=1$) are supposed. The purpose of the Tucker inter-battery method is to find two sets of $A$ components: $\{c_{X_a} = XQ_Xa_a\}_{a=1,\dots,A}$ associated with $X$ and $\{c_{Y_a} = YQ_Yb_a\}_{a=1,\dots,A}$ associated with $Y$, which maximize the criterion $cov(XQ_Xa_a, YQ_Yb_a)$ with the constraints $\|a_a\|^2_{Q_X} = a'_aQ_Xa_a = 1$ and $\|b_a\|^2_{Q_Y} = b'_aQ_Yb_a = 1$.

This problem is a compromise between two conflicting objectives: CCA, which maximizes the correlation between $c_X$ and $c_Y$, and PCA of $X$ and $Y$ (with their respective metrics), what is equivalent to diagonalize $W_XD = XQ_XX'D$ and $W_YD = YQ_YY'D$ (Cailliez and Pagès, 1976), and that maximizes variance of the different components.

The successive solutions $c_{X_a}$ and $c_{Y_a}$ of Tucker inter-battery are the eigenvectors of $W_XDW_YD$ and $W_YDW_XD$ associated with the same eigenvalue $\lambda_a^2$. This implies that $\lambda_a$ is also the optimal covariance. Moreover, it is possible to show that the "factors" $a_a$ and $b_a$ are the solution of the diagonalization of the two matrices $X'DYQ_YY'DXQ_X$ and $Y'DXQ_XX'DYQ_Y$. Another difference between Tucker inter-battery, CCA and PCA is that the two sets $\{c_{X_a}\}_{a=1,\dots,A}$ and $\{c_{Y_a}\}_{a=1,\dots,A}$ are not $D$-orthogonal in $\mathbb{R}^n$.

## 4. DO-ACT procedure

In this section the DO-ACT approach is defined and some of its properties are given. We briefly recall notations: two sets of triplets $\{(X_k, Q_{X_k}, D)\}_{k=1,\dots,K}$ and $\{(Y_l, Q_{Y_l}, D)\}_{l=1,\dots,L}$, measured on the same $n$ observations are considered. All the variables (columns of $X$ and $Y$) are $D$-centered.

First, the aim is to study the proximities and the differences between the two sets of $K+L$ triplets and, in a second step, to compare the views of the observations given by the two compromises (one for each multiblock) with those given by the initial tables. Next two compromises are calculated, the closer as possible in the sense of a criterion, and are analyzed. The DO-ACT procedure is organized, as in STATIS, with three successive stages: interstructure, compromise and intrastructure.

### 4.1. Definition

The problem is to find two compromise operators, one per block, $W_XD$ and $W_YD$. Each compromise is a linear combination, with unknown coefficients, of its respective operators, maximizing their scalar product: $tr(W_XDW_YD)$.

Hence we search

$$W_XD = \rho \sum_{k=1}^{K} \pi_k \alpha_k W_{X_k} D \tag{1}$$

and

$$W_YD = \tau \sum_{l=1}^{L} \omega_l \beta_l W_{Y_l} D \tag{2}$$

with $\rho$ and $\tau$ two scaling coefficients, $\Pi = diag(\pi_k)$ and $\Omega = diag(\omega_l)$ two diagonal matrices of a priori operators weights. The two compromises operators $W_XD$ and $W_YD$ are searched through the vectors $\alpha = (\alpha_1, \ldots, \alpha_K)'$ and $\beta = (\beta_1, \ldots, \beta_L)'$. Without loss of generality, it is assumed that the two scaling coefficients are equal to 1. If this is not the case, the operators $W_XD$ and $W_YD$ are substituted for $W_XD/\|W_XD\|$ and $W_YD/\|W_YD\|$.

The compromises are simultaneously searched subject to the constraints

$$\|\alpha\|^2 = \alpha'\alpha = \sum_{k=1}^{K} \alpha_k^2 = 1, \tag{3}$$

$$\|\beta\|^2 = \beta'\beta = \sum_{l=1}^{L} \beta_l^2 = 1. \tag{4}$$

The solutions are found by means of the Lagrange function $L$. Let $\mu$ and $v$ the multipliers associated to the constraints. $L$ is given by

$$L(\alpha, \beta, \mu, v) = tr(W_XDW_YD) + \tfrac{1}{2}\mu(1 - \|\alpha\|^2) + \tfrac{1}{2}v(1 - \|\beta\|^2). \tag{5}$$

### 4.2. Properties

**Proposition 1.** *The solutions of* $\max_{\alpha,\beta}\{tr(W_XDW_YD)\}$ *under constraints* (3) *and* (4) *are given by the two eigen equations*

$$\Omega C' \Pi^2 C \Omega \beta = \mu^2 \beta, \tag{6}$$

$$\Pi C\Omega^2 C'\Pi\alpha = \mu^2\alpha, \tag{7}$$

$$\beta'\beta = 1, \tag{8}$$

$$\alpha'\alpha = 1 \tag{9}$$

*and the maximum of the criterion is equal to $\mu$.*

**Proof.** $tr(W_X D W_Y D) = \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_k \alpha_k \omega_l \beta_l \, tr(W_{X_k} D W_{Y_l} D) = \alpha'\Pi C\Omega\beta = \beta'\Omega C'\Pi\alpha$, where $C$ is the $K \times L$ matrix of the scalar products between the operators: $C_{k,l} = tr(W_{X_k} D W_{Y_l} D)$. Hence, the problem is to find the maximum of a bilinear form under constraints. The associated normal equations are

$$\nabla_\alpha L = \Pi C\Omega\beta - \mu\alpha = 0, \tag{10}$$

$$\nabla_\beta L = \Omega C'\Pi\alpha - v\beta = 0, \tag{11}$$

$$2\nabla_\mu L = 1 - \alpha'\alpha = 0, \tag{12}$$

$$2\nabla_v L = 1 - \beta'\beta = 0. \tag{13}$$

By multiplying (10) on the left by $\alpha'$ and by using constraint (12), $\mu = \alpha'\Pi C\Omega\beta$ is obtained. Similar arguments applied to (11) and (13) give $v=\mu=\alpha'\Pi C\Omega\beta$, the positive (see below) optimal scalar product.

The expression of $\alpha$ given by (10) replaced in (11) gives (7). Similar calculations are made for (6). □

In practice, it is unnecessary to diagonalize two systems (7), (9) or (6), (8). Indeed only one of them is diagonalized and the other coefficients vector is obtained by using Eq. (10) or (11).

If the operator weights (in matrices $\Pi$ and $\Omega$) are equal to 1, the matrices to diagonalize are $C'C$ and $CC'$.

In Eqs. (6) and (7), the elements of the two matrices are non-negative (because the trace of an operators product is positive), and the application of the Perron–Frobenius theorem, see for example Lütkepohl (1996, p. 141), gives that all the entries of the first eigenvector ($\alpha$ and $\beta$) are positive. Hence, the two compromises $W_X D$ and $W_Y D$ are positive semi-definite operators.

The equations in Proposition 1 are similar to those produced by the Tucker inter-battery methodology, but in the usual formula, the covariances ($D$-scalar product) between variables are substituted with the scalar product between operators.

Hence, the definition of the first step of the DO-ACT strategy is

**Definition 2.** Let $\{\alpha_m\}_{m=1,\dots,M}$ and $\{\beta_m\}_{m=1,\dots,M}$ the eigenvectors of (7), (9) and (6), (8) associated with the eigenvalues $\{\mu_m^2\}_{m=1,\dots,M}$. Then, the plot $(\mu_m\alpha_m, \mu_{m'}\alpha_{m'})$ gives an Euclidean representation of the operators $\{W_{X_k}D\}_{k=1,\dots,K}$ onto the plane $(m, m')$. The same definition holds for $(\mu_m\beta_m, \mu_{m'}\beta_{m'})$ and $\{W_{Y_l}D\}_{l=1,\dots,L}$. These two sets of representations are the interstructure of the DO-ACT strategy.

These representations can be interpreted as in the interstructure step of STATIS (Lavit, 1988). The closer all the points are to the first axis, the better the compromise. Moreover, as the goal of DO-ACT is to make the two compromises the closer as possible (that is to say to make the compromise observations scalar products in the $X$-multiblock as close as possible to the compromise observations scalar products in the $Y$-multiblock), the two plots can be compared. If the two compromises are good compromise of their multiblock and if they are close enough, the two plots should nearly be the same.

The next proposition shows that the previous operators are also the solutions of two others problems.

**Proposition 3.** *If* $\|W_X D\|^2 = 1$ *and* $\|W_Y D\|^2 = 1$, *the solutions in Proposition* 1 *are also the solutions of two following optimization problems*:

1. $\max_{\alpha,\beta}\{RV(W_X D, W_Y D)\}$ *and the optimal value is* $\mu$,
2. $\min_{\alpha,\beta}\{\|W_X D - W_Y D\|^2\}$ *and the optimal value is* $2(1-\mu)$.

**Proof.** The proof is straightforward because $RV(W_X D, W_Y D) = tr(W_X D W_Y D)$ and $\|W_X D - W_Y D\|^2 = 2(1 - tr(W_X D W_Y D))$ under the above assumptions. □

The next proposition of this section produces two matrices that generate the two compromise operators $W_X D$ and $W_Y D$.

**Proposition 4.** *Let the two partitioned matrices* $X$ *and* $Y$

$$X = [\sqrt{\rho\pi_1\alpha_1}X_1, \ldots, \sqrt{\rho\pi_K}\alpha_K X_K], \tag{14}$$

$$Y = [\sqrt{\tau\omega_1\beta_1}Y_1, \ldots, \sqrt{\tau\omega_L\beta_L}Y_L]. \tag{15}$$

$W_X D$ *and* $W_Y D$ *are the operators associated with* $X$ *and* $Y$.

The demonstration of this property is without ambiguity.

Proposition 4 shows that the analysis of the unweighted juxtaposition of $X_k$ and $Y_l$, followed by a Tucker inter-battery is not the optimal strategy of the compromise step of DO-ACT. However, the weighting coefficients of matrices can be equal to one, but it is a result of diagonalization (and the structure of the data) and not an a priori choice.

We can now define the two next steps of the DO-ACT methodology:

**Definition 5.** The two compromises of the DO-ACT strategy are $W_X D$ and $W_Y D$. The two intrastructures (one for each of the multiblocks) are the representations of the observations in the Tucker inter-battery analysis of the previous $X$ and $Y$.

These observations representations can be interpreted with the corresponding variables plots given by the inter-battery analysis factors $a_a$ and $b_a$ defined in Section 3.2.

It follows that the plots of the observations viewed by the $K+L$ triplets can be realized by the supplementary representations of the observations of all the triplets onto their respective compromise components. For example, the coordinates of the observations in table $X_k$ on axis $a$ are given by the coordinates of $W_{X_k}DW_YDc_{Y,a}/\sqrt{\lambda_a}$. This is a similar representation to the intrastructure step of STATIS. If all the tables in one multiblock measure the same variables, then the plots for this multiblock are directly comparable. Moreover, in order to explain the observations configurations, correlations circles can be made for each table: they are given by the correlations between the variables in a table and the component of its compromise. Then, the observations plots and the variables plots can be associated and interpreted as in usual PCA.

### 4.3. Some particular cases

The first evident particular case is when the two multiblocks are the same, it is obvious that the STATIS solution is found. In next proposition, the solution when the length of one of the two multiblocks is equal to 1 is explicited.

**Proposition 6.** *Let* $L=1$ (*in this case* $W_YD$ *is the only one operator of this multiblock*) *and for any* $K$, *the* $W_XD$ *solution of Proposition* 1 *is given by*

$$W_XD=\sum_{k=1}^{K}\pi_k\,tr(W_{X_k}DW_YD)W_{X_k}D. \tag{16}$$

The proof is straightforward with Eq. (10).

In Simier et al. (1999), in the case of two multiblocks with the same length ($K=L$) and with the same number of variables at each occasion (that is to say with two three-way data) the authors propose to realize the STATICO approach. Here we present a new version of STATICO in which we use the operators $W_{X_k}D$ and $W_{Y_l}D$ instead of $X_k^tDY_k$ matrices. Hence, this new version of STATICO methodology can work with data set where the number of variables at each occasion in different. The new STATICO is a STATIS methodology applied to the $K$ operators $W_{X_k}DW_{Y_k}D$ (that is to say the operators of the Tucker inter-battery method), which implies that the compromise is $W_cD=\mu\sum_{k=1}^{K}\pi_kl_kW_{X_k}DW_{Y_k}D$. In DO-ACT, one of the two operators products used in the intrastructure step is

$$W_XDW_YD=\rho\tau\sum_{k=1}^{K}\sum_{l=1}^{K}\pi_k\omega_l\alpha_k\beta_lW_{X_k}DW_{Y_l}D. \tag{17}$$

The comparison of these two equalities shows that the DO-ACT compromise works with the $K\times L$ cross products whereas STATICO does not. If $K=L$, choosing one of the two compromises (or methodologies) is dependent on the user and/or the problematical of the study: every tables are linked (really or supposed) with all the tables of the other multiblock, or not.

In the next section, a real data set illustrates the approach. A program implementing the method has been developed using the S-Plus 6.0® (2000) language. In this example, we show the capabilities of this procedure, in particular graphically.

## 5. Application

### 5.1. Data

The data set used in this paper has been collected by Forestier (1994). It can be free obtained from the database of Chessel and Dolédec (1995). This data collection is the result of an experimentation about the reproductibility of *Esolus parallelipipedus* in a fluvial ecosystem. In these data, there are 10 times of sampling (not equally spaced from 1981-10-27 to 1982-12-01), hence it is possible to compare the results of DO-ACT with those of STATICO. Moreover, because the sampling points are not equally distributed in time, the hypothesis of non-influence of two successive measurements, which is assumed in STATICO, is very difficult to carry out. Hence, DO-ACT seems a priori more appropriate than STATICO.

This data set is composed of two multiblocks of same length $K = L = 10$ (time of sampling) measured on $n = 7$ observations (stations). The matrices $X_k$ $(k = 1, \ldots, K)$ contain the measures of seven stages of the evolution of *Esolus parallelipipedus*, that is to say: $p_k = 7$ $(k = 1, \ldots, K)$. *Esolus parallelipipedus* is one of the numerous beetle species. These insects live in sediments of aquatic environments. The second multiblock, $Y_l$ $(l = 1, \ldots, L)$ contains the values of environmental variables. Table 1 is the description of the 11 variables. Notice that a lot of them are null at several occasions: hence the number of variables varies from 9 (matrices 2, 3 and 10) to 11 (matrices 1 and 5–8). The problem is to find the portion of stability of the structure of *Esolus parallelipipedus* in the station typology.

The preprocessing of data, because of different units of variables, consists in centering and scaling in columns, according to the uniform weight (here $D = Id_7/7$). Moreover,

Table 1
Number, name, code and description of the 11 variables of the $Y$ blocks

| No. | Name | Code | Description |
|---|---|---|---|
| 1 | Silt | Silt | Percentage of silt |
| 2 | Sand | Sand | Percentage of sand |
| 3 | Gravel | Grav | Percentage of gravel |
| 4 | Shingle | Shin | Percentage of shingle |
| 5 | Stone | Ston | Percentage of stone |
| 6 | Block | Bloc | Percentage of block |
| 7 | Flag | Flag | Percentage of flag |
| 8 | Depth | Dept | Water depth (in cm) |
| 9 | Speed | Spee | Speed of water (in cm/s) |
| 10 | Periphyton | Peri | Periphyton |
| 11 | Fragments | Frag | Organic fragments (from 0-missing to 2-abundant) |

the calculations in DO-ACT procedure are made with $RV$ option, the two diagonal matrices ($\Omega$ and $\Pi$) as well as $\rho$ and $\tau$ are, respectively, chosen equal to identity and one, and all the metrics for $X$ and $Y$ multiblocks are identity matrices.

### 5.2. Application of DO-ACT

The first results of DO-ACT methodology are the following: the Hilbert–Schmidt norms of the operators of $X$-multiblock vary from 4.238 ($X_9$) to 5.964 ($X_8$) and for $Y$-multiblock from 4.776 ($Y_{10}$) to 7.010 ($Y_8$), the $RV$ coefficients between the two multiblocks are in the interval [0.2978, 0.8210]. Hence, the $K+L=20$ matrices have a comparable variability but the proximities of the different operators are more scattered.

Fig. 1 represents the interstructure in the first principal plane and shows the plot of the operators of the two multiblocks. Like in the STATIS interstructure, the two first axes explain the most important part of the total variability of the operators: 98.74% and 0.86%. We note that the respective positions of the two operators with the same number $k$ (time of sampling) are not necessary in the same place in the plot and vary along the time: for example $X_{10}$ and $Y_{10}$ are relatively close whereas $X_4$ and $Y_4$ are very distant. Hence, this representation confirms us that it is impossible to carry out the non-influence of all the successive measurements. However, it can be noticed that $X_9$ has all its $RV$ coefficients, with the $Y_l$, very high (in [0.6718, 0.8210]) that explains the particular position of this point on the previous graphic. It clearly shows that there is no evolution of the data with time.

The variation on second axis is very small. So, the differences between, for example $Y_6$ and $Y_2$ are only very small. Moreover, as $Y_2$ and $Y_4$ are confounded, we can affirm these two tables are next to be the same. This imply that $\beta_2$ and $\beta_4$ are very close (Table 2).

The $RV$ coefficients between the DO-ACT the compromise $W_X D$ and the $X_k$ operators ($W_{X_k} D$) are in [0.4539, 0.8536]. We have an analogous interval for $Y_l$ matrices and their

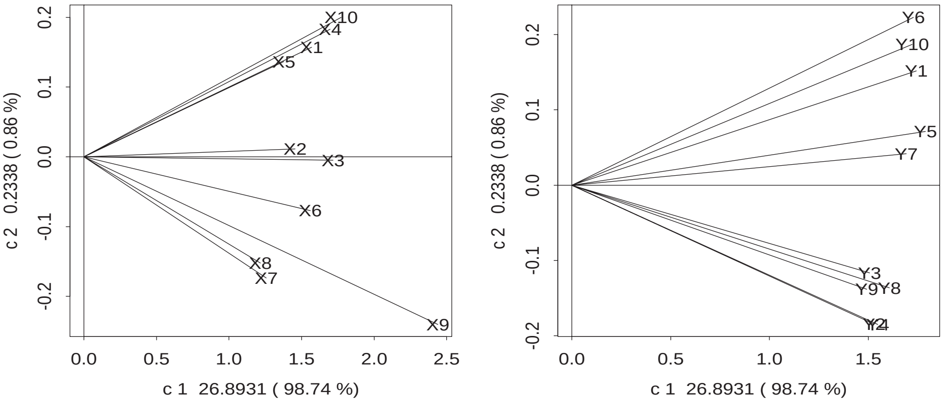

Fig. 1. Inter-structure representation of the two multiblocks in the DO-ACT methodology with the two first principal axes.

Table 2
Coefficients of the $X$ and $Y$ multiblocks compromises in DO-ACT and distances $d$ between the compromise and the operators building the compromise ($d = \|W_XD - W_{X_k}D\|$ or $\|W_YD - W_{Y_k}D\|$)

| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X$ | 0.3028 | 0.2807 | 0.3311 | 0.3274 | 0.2656 | 0.3008 | 0.2421 | 0.2345 | 0.4703 | 0.3419 |
| $d$ | 1.1774 | 0.9559 | 0.9478 | 1.0158 | 0.9060 | 1.3056 | 1.3394 | 1.2668 | 0.9545 | 0.7948 |
| $Y$ | 0.3362 | 0.2952 | 0.2907 | 0.2985 | 0.3452 | 0.3334 | 0.3264 | 0.3101 | 0.2879 | 0.3322 |
| $d$ | 0.8284 | 0.8093 | 0.8498 | 0.8003 | 0.7857 | 0.8633 | 0.7823 | 0.9345 | 0.8213 | 0.8012 |

compromise $W_YD$: [0.9070, 0.9820]. That is to say, the DO-ACT $Y$-multiblock compromise ($W_YD$) is a better summary than the one of $X$-multiblock (in a $RV$ sense). This conclusion corresponds to what can be seen in Fig. 1: in the $Y$-multiblock interstructure representation, all the vectors have near the same length and the angles between them are small enough. In the $X$-multiblock inter-structure the vectors are more differents. In conclusion, the $RV$ coefficient between the two DO-ACT compromise ($W_XD$ and $W_YD$) is equal to 0.8091 and the one between the two STATIS compromises (issued from the two STATIS procedures of each $X$ and $Y$ block) is 0.7549.

The coefficients of the DO-ACT compromises are in Table 2. Note that $X_9$ is the most important for the $X$ compromise whereas it is $Y_5$ in the other multiblock. These coefficients show that all the tables do not influence the compromise in the same manner (in this case all the coefficients would be equal to $1/\sqrt{10} = 0.3162$). Moreover, this table allows to calculate the $\alpha_k\beta_l$ coefficients present in $W_XDW_YD$ (or $W_YDW_XD$) used in the DO-ACT intrastructure step (Eq. (17)). The most important pair of operator is $W_{X_9}D$ and $W_{Y_5}D$ with the value $0.4703 \times 0.3452 = 0.1623$. This value confirms that the STATICO strategy is non appropriate with these data, because this pair of operator does not exist in its compromise. Moreover, Table 2 contains the Hilbert–Schmidt distances between the compromise and the operators which it summarizes. We note that the distances in the $Y$-multiblock are smaller than those for the $X$-multiblock and are very close one with each other. This naturally coincides with the RV values quoted above in the beginning of the section and with the fact that $W_YD$ is a better summary than $W_XD$. We note also that $Y_5$ and $Y_7$ are the closest of $W_YD$ as it could be seen in Fig. 1 on the first axis. To finish this stage we note that the inertia of the two compromises are, respectively, equal to 4.3809 for the $X$ multiblock and 5.5993 for the other. These two values confirm us that the first multiblock is less informative than the second.

The third stage, is the intrastructure step. Fig. 2 represents the positions of the compromise observations (stations) onto the two first components of the $X$ and $Y$ compromises. These two graphics, represent 53.34% of the inertia for the $X$ multiblock and 63.80% for the other. We note that the positions are very similar. That is to say, there is a common typology into both the multiblocks and this structure is rather uni-dimensional.

It appears a first group of very close stations (from 2 to 5) in opposition with stations 6 and 7. Only the station number 1 seems remote from the others, especially

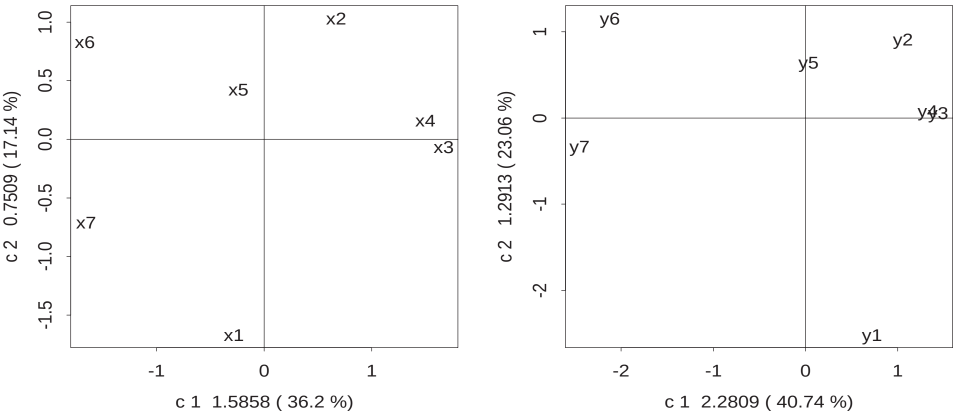

Fig. 2. Compromises observations (stations) into the two first components of the $X$ and $Y$ compromise of the DO-ACT methodology.

due to the second axis. In order to explain these configurations, it is possible to make the variables representations (with the use of the axes $a_a$ and $b_a$ from the inter-battery analysis of the compromises $X$ and $Y$). Some relative and absolute contributions can also be calculated as for PCA, but it gives no more information than what we see on the plots: the stations with the largest contributions (absolute or relative) have the largest coordinates. Fig. 3 shows the correlations between the variables of each matrix $\{X_k\}_{k=1,\dots,K}$ with the two first components of the operators $W_X D$, Fig. 4 is the same representation for $\{Y_l\}_{l=1\dots L}$ and $W_Y D$. So, all the plots are comparable because all the tables measure the same variables. All the variables of the $X$-multiblock are not very well described by the two first components (see the correlations values in Fig. 3), but a kind of gradient of the evolution stages (variables) of the beetles can be noticed with respect to the time of sampling ($X_1$ to $X_{10}$). Stage 7 ($S7$) is always different from the others except in $X_9$ and $S1$ is dissociated itself from the other variables from the first date to the fifth and in the tenth. The stages $S2$ to $S6$ are generally well enough grouped in all the times of sampling except for the sixth and ninth dates of sampling, which seem to be special dates. From the first time ($X_1$) to the third one ($X_3$), the first axis is an opposition between the seventh stage of evolution $S7$. $S1$ is a few remote from the others and it is first characterized by the first axis and next by the second axis. Its position seems to move clockwise (several times) around the origin from the first time of sampling ($X_1$) to the seventh date ($X_7$), and in opposite next. Moreover, stages $S2$ to $S6$ are well enough grouped and their positions in the first principal plane (Fig. 3) move with time: from $X_1$ to $X_5$ and in $X_{10}$, they are on the upper right side of the plane, and in the left upper side in $X_7$ and $X_8$. In $X_6$ and $X_9$, they are scattered on the plot. We note, one more time, that time 9 is very different from the other time of sampling. Here, axis 1 shows an opposition between stages 3, 6, 7 and the other stages of evolution. Fig. 4 shows the correlations of the $Y$-multiblock, environmental variables, with the two first components. This graphic shows that the correlations are larger than in the other multiblock. The global structure of correlations is analogous

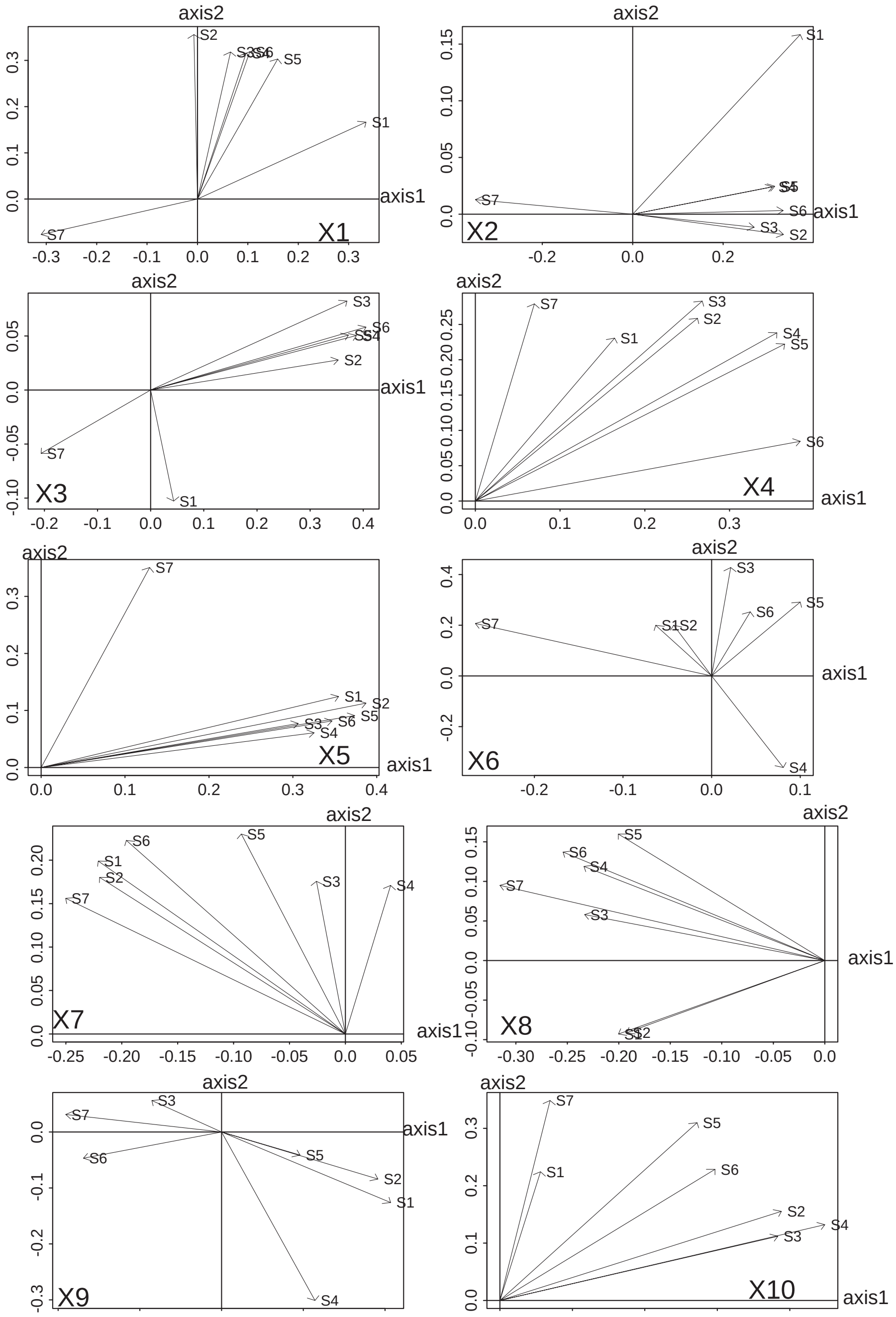

Fig. 3. Correlations of the variables of each $\{X_k\}_{k=1,\dots,10}$ matrices with the two first components (explaining 36.2% and 17.14%) of their compromise in the DO-ACT methodology.

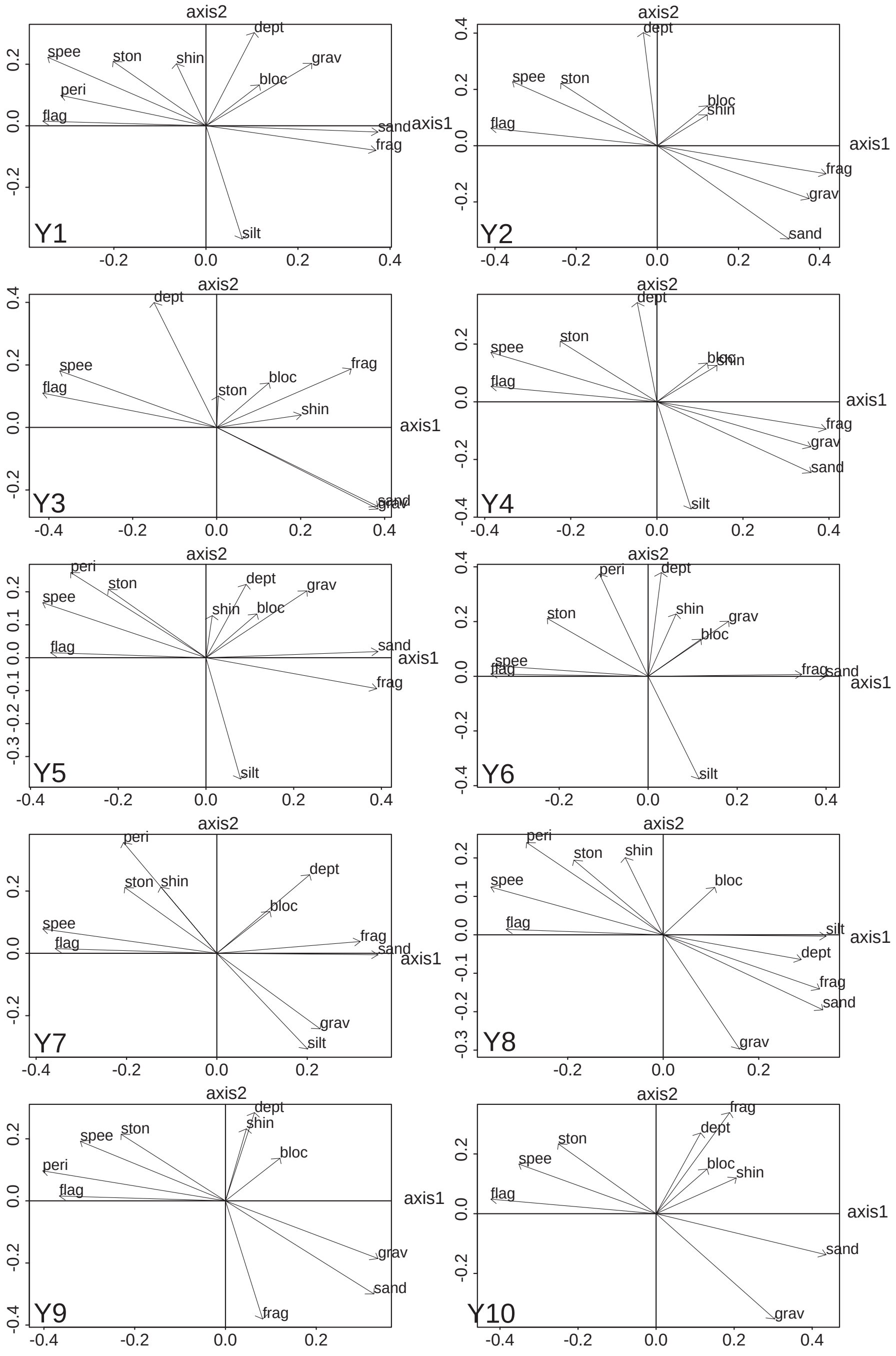

Fig. 4. Correlations of the variables of each $\{Y_l\}_{l=1,\ldots,10}$ matrices with the two first components (explaining 40.74% and 23.06%) of their compromise in the DO-ACT methodology.

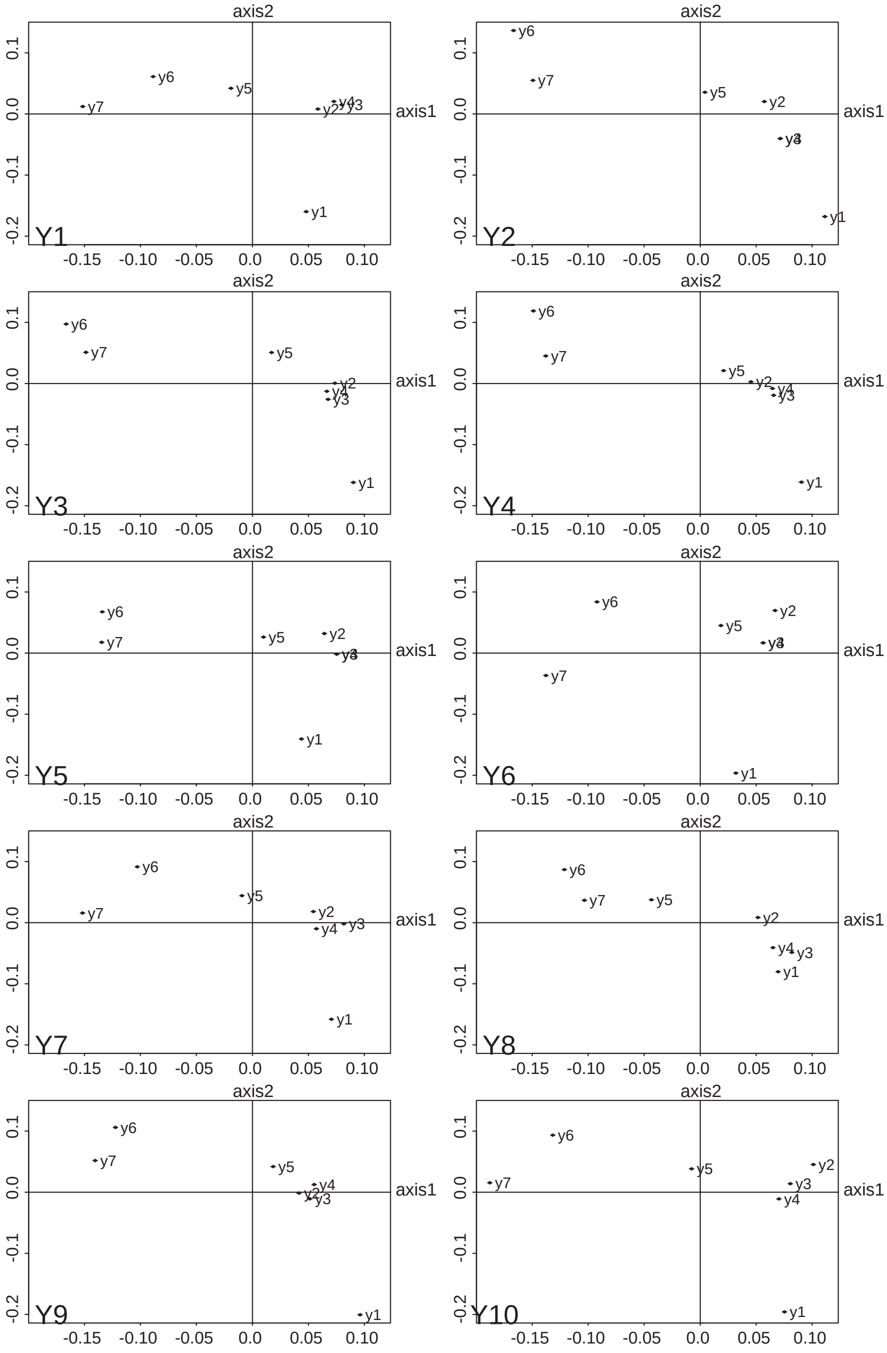

Fig. 5. Intrastructure representations for the observations of each $\{Y_l\}_{l=1,\ldots,10}$ matrices into the two first components (explaining 40.74% and 23.06%) of their compromise in the DO-ACT methodology.

in the 10 matrices and when the silt variable is present (Tables 1, 4–8) it is well described by the second component (except in $Y_8$). This figure confirms that there is a general granulometry gradient (from silt to flag). Now, look at briefly to the position of the observations given by the supplementary projection of each matrix $Y_l$ onto the first intrastructure graph (Fig. 5). These graphs can be associated with Fig. 4, as it is usually done in factor analysis, to explain the stations configuration in each table. Station 7 corresponds to high values of the speed and flag variables, Stations 3 and 4 correspond to low values of fragments and sand variables. We note a great stability of the structure for the most important times of prelevements. The only slightly different graph is the eighth with stations 1, 6 and 7 which are located near the others. In this example we do not produce the other intrastructure plot (for $X$-multiblock), this graph is not more explicit, in an interpretative sense, than the previous one.

So these graphics have shown an evolution of the stages of the beetles with time especially for stages $S1$ and $S7$ and that the granulometry properties of the stations are approximately the same with time except perhaps at time 6.

## 6. Conclusion

In this paper, a new computing linear procedure have been presented. DO-ACT generalizes STATIS and Tucker inter-battery methodologies to take into account data with two multiblocks structures. This methodology uses a natural scalar product between the operators associated with each matrix of the two multiblocks.

The data set we used, with low dimensions ($n=7, K=L=10$), is chosen to illustrate the behavior of the DO-ACT procedure, but this methodology is very efficient with a relatively high observation/variable or length/variable ratio. DO-ACT produces some useful graphics. The most important is the view of the different tables of the two multiblocks: this allows us to do an overall comparison. The last step of DO-ACT, intrastructure step, compares the association between variables and observations and shows broadly the stable structure.

This methodology can be successfully used in most practitioners fields that use many variables (chemometrics, ecology, etc.) with the simple idea of dividing variables into groups (or subsets) with the objective of choosing some of them for the future experimentations.

Some extensions of this methodology can now be set in several ways: data sets with more than two multiblocks, introducing nonlinearity in variables or modeling of one of the two multiblocks in PLS sense.

## Acknowledgements

The authors are grateful to the referees for their suggestions and comments.

## References

Bécue-Bertaut, M., Pagès, J., 2003. A principal axes method for comparing contingency tables: MFACT. Comput. Statist. Data Anal., in press (corrected proof available online since 23 June 2003).

Bro, R., 1996. Multiway calibration. Multilinear PLS. J. Chemometrics 10, 47–61.

Cailliez, F., Pagès, J.P., 1976. Introduction à l'analyse des données. SMASH, rue Duban, Paris, 616p.

Carlier, A., Kroonenberg, P.M., 1996. Biplots and decompositions in two-way and three-way correspondence analysis. Psychometrika 61 (2), 355–373.

Carroll, J.D., 1968. Generalization of canonical correlation analysis to three or more sets of variables. Proceedings of the 76th Convention of the American Psychology Association 3, 227–228.

Casin, Ph., 2001. A generalization of principal component analysis of $K$ sets of variables. Comput. Statist. Data Anal. 35, 417–428.

Chessel, D., Dolédec, S., 1995. ADE software hypercard stacks and quick-basic microsoft program library for the analysis of environmental data, Version 4. Documentation ESA, CNRS 5023, Université Lyon I.

Chessel, D., Mercier, P., 1993. Couplages de triplets statistiques et liaisons espèces-environnement. In: Lebreton, J.D., Asselain, B. (Eds.), Biométrie et Environnement. Masson, Paris, pp. 15–44.

Escofier, B., Pagès, J., 1984. L'analyse factorielle multiple: une méthode de comparaison de groupes de variables. In: Diday, E. (Ed.), Data Analysis and Informatics III. Elsevier, North-Holland, Amsterdam, pp. 41–55.

Escofier, B., Pagès, J., 1994. Multiple factor analysis; afmult package. Comput. Statist. Data Anal. 18, 121–140.

Escoufier, Y., 1973. Le traitement des variables vectorielles. Biometrics 29, 750–760.

Forestier, M.C., 1994. Variabilité spatio-temporelle de distribution d'Esolus parallelipipedus (Muller, 1906) (Coleoptera, Elmidae) à différentes échelles de l'hydrosystème fluvial. Thèse de Doctorat, Université Lyon I.

Gower, J., 1975. Generalized procrustes analysis. Psychometrika 40 (1), 33–51.

Horst, P., 1961. Relation among $m$ sets of variables. Psychometrika 28, 129–149.

Lavit, Ch., 1988. Analyse conjointe de tableaux quantitatifs. Masson, Paris, 252p.

Lavit, Ch., Escoufier, Y., Sabatier, R., Traissac, P., 1994. The ACT (STATIS method). Comput. Statist. Data Anal. 18, 119–997.

Lütkepohl, H., 1996. Handbook of Matrices, Wiley, New York, 304p.

Robert, P., Escoufier, Y., 1976. A unifying tool for linear multivariate statistics methods: the RV coefficient. Appl. Statist. 25 (3), 257–265.

Saporta, G., 1975. Liaison entre plusieurs ensembles de variables et codages de données qualitatives. Thèse de Troisième cycle, Université de Paris VI.

Simier, M., Blanc, L., Pellegrin, P., Nandris, D., 1999. Approche simultanée de $K$ couples de tableaux: application à l'étude des relations pathologie végétale-environnement. Rev. Statist. Appl. XLVII (1), 31–46.

Smilde, A.K., Westerhuis, J.A., Boqué, R., 2000. Multiway multiblock component and covariates regression models. J. Chemometrics 14, 301–331.

S-PLUS (2000). S-PLUS 6.0 for UNIX Programmer's Guide. MathSoft, Seattle, 534p.

Tenenhaus, M., Young, F.W., 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50, 91–119.

Tucker, L.R., 1958. An inter-battery method of factor analysis. Psychometrika 23, 111–136.

Van de Geer, J., 1984. Linear relation among $k$ sets of variables. Psychometrika 49, 79–94.

Westerhuis, J.A., Kourti, K., Macgregor, J.F., 1998. Analysis of multiblock and hierarchical PCA and PLS models. J. Chemometrics 12, 301–321.